News-oriented Stock Index Prediction with Natural Language Processing and Machine Learning

Abstract

In this paper, by using Latent Dirichlet Allocation (LDA) and long short-term memory (LSTM), I developed a straightforward but effective methodology to extract valuable information from news scripts and predict stock index (SSE composite index). The accuracy of the final model predicting upward or downward trends of stock index is 73.3%. Furthermore, by historical simulation, in an environment where short selling is not allowed the annualized return rate is 11.65%, while the annualized return rate can reach 31.54% if short selling is permitted. To achieve such results, I adopted a special official news source Xinwen Lianbo given its special properties and relations to the domestic market.

1.Introduction

Stock market prediction is a classic problem in the intersection of finance and computer science, which if efficient can be used as a tool to maximize profits from the trend of stocks and intermediate risks.

Traditional trials to predict the trend of stocks depend on technical analysis and fundamental analysis. In technical analysis, analysts try to find patterns in stock price, trade volume, etc., in order to predict the stock trend. Meanwhile, fundamental analysis is performed through the study of the financial state of the company, the markets and macro-economy condition changes. However, the effectiveness of the two kinds of analysis can be of doubt. The philosophies behind the two types of analysis are challenged by efficient-market hypothesis. The "weak-form" efficient-market hypothesis thinks that all the past trading data have been reflected in the current stock price and therefore the study of past trading data like past stock price or volume cannot bring any new understanding of the stock so as the profits. "Semi-strong-form" efficient-market hypothiss, on the other hand, holds that all the public data, including financial statements, the market information and macro-economic information, are reflected in current stock price, so the study of such factors are not helpful. By far, the evidence supporting technical analysis remains mixed and inconclusive, while fundamental analysis is generally accepted and widely utilized by both academia and industry professionals.

Recent years, with the rapid growth of the Internet and the increasingly advanced machine learning techniques, some new data sources emerge, and more and more people become interested in these data sources so as to gain an informative advantage. These new alternative data are data that originate outside of the standard repertoire of market data but are considered useful for predicting stock prices, detecting different risk exposures and discovering new price movement indicators. For instance, Adam Atkins et al. used news data to extract helpful information and studied their prediction effects on stock volatility and stock close price.

Unfortunately, unlike technical analysis or traditional fundamental analysis that enjoys a relatively mature set of methodologies, in terms of how to take advantage of those alternative data, there is not a common set of recognized tools to help us and even if some research has developed some resulting methodologies, these methodologies are quickly replaced by daily advancing new technologies. Sentiment analysis of textual data for example, old ways of extracting sentimental information include Dictionary-Based Approach whose dictionary is derived by manual annotation. But nowadays, we have deep learning or even LLM (large language model) to do this work which by velocity or outcome are far better.

In this paper, following some research in this field, I have developed a straightforward yet effective methodology to extract useful information from news data and predict the stock index (SSE composite index). To be specific, starting from November 2022, I gathered approximately one year of data for both Xinwen Lianbo news scripts and SSE index returns, resulting in data from 268 valid trading days. For these 268 days, I applied LDA (Latent Dirichlet Allocation) topic modeling to the news scripts to divide each day's news script into several topics and obtained the proportions these topics occupied. The time series formed by these proportions are the 'information' we extract. After this division, these topic proportions together with historical SSE index returns were fed into the LSTM (long short-term memory) model to predict SSE index returns. Because we are trying to extract information from the news and use this information to predict stock index returns, the choice of news source largely determines whether we have satisfying results or not. If the news reveals little about the market or even contains rumors, however good the methodology is, the results will not be satisfying. Furthermore, to match the properties SSE holds, the news source should also cover the whole of domestic affairs instead of focusing on one particular field and we hope the news source is authentic and of as high quality as possible. For the above reasons, I chose Xinwen Lianbo as the news source in this research. More detailed reasons for news source choosing will be described in Part 3.

In the final part of this paper, to validate the effectiveness of the methodology, various methods were applied. First, I compared the model (LSTM) based on news-extracted information to the exact same model but based on only past historical price data to evaluate how well the information was extracted from the news by LDA. Second, a simple investment simulation was constructed according to the prediction from the model whose results showed that a simple investment strategy developed from this methodology could achieve a very decent annualized return. Third, several mathematical formulas were constructed and combined with charts to further illustrate the effectiveness.

To sum up the contributions of this paper:

1. Using a special news source, a new straightforward methodology based on news text analysis is developed to make stock index prediction.
2. Several validation methods are used to show the effectiveness of such methodology.

The remaining part of this paper is organized as follows. Part 2 reviews the related and recent literature. Part 3 describes the empirical analysis including data source choosing, nature language processing model, machine learning model and their results. Part 4 provides the evaluation for the methodology. Part 5 is the conclusion and some perspectives on future work.

2.literature review

Stock prediction is a challenging task due to the highly volatile and non-linear nature of financial markets but also an attractive topic since its central position in stock investment. In general, traditional approaches to predict stocks can be categorized into two primary approaches: technical and fundamental analysis, according to the various types of information they mainly relied on. Technical analysis focuses on historical price movements and trading volumes to identify patterns and trends that can predict future stock behavior. In contrast, fundamental analysis assesses a stock's intrinsic value by examining financial reports, macroeconomic indicators, and other economic factors. While these methods have laid the groundwork for stock prediction, recent advancements have introduced significant shifts in both research methodologies and the types of data employed.

There are two notable trends in the evolution of stock prediction research:

1. Methodological Shift: Researchers are moving away from traditional time series statistical models towards more advanced techniques, including machine learning, deep learning, and complex multimodal neural networks. These modern approaches can capture intricate patterns and non-linear relationships within the data, potentially improving prediction accuracy.
2. Data Evolution: The scope of data used for stock prediction has expanded beyond basic market data like historical prices and trading volumes. It now encompasses fundamental data such as financial statements and macroeconomic indicators, as well as non-traditional data sources like social media sentiment and news texts. Incorporating new alternative data types allows for a more comprehensive analysis and can enhance the robustness of predictive models.

In Methodological Shift, many models are used to achieve better results. For simple ANN, Moghaddam et al. (2016) investigated the use of artificial neural networks (ANNs) for NASDAQ index prediction, showing that integrating historical prices and day-of-week effects can yield better performance compared to traditional statistical models. Their work demonstrates the value of ANN in modeling complex, non-linear relationships in financial data.

However, simple ANN is not always enough to the increasingly complicated financial markets, so much more deep learning architectures are developed. For so many models, there are some summary papers. Jiang (2021) reviewed the application of deep learning techniques in predicting stock market prices, categorizing various neural network architectures such as Feedforward Neural Networks (FNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). The paper emphasizes the superior performance of deep learning models over traditional linear methods and highlights the recent advances in model types, data sources, and model reproducibility. The survey also illustrates that stock market prediction has benefited from a variety of deep learning architectures. RNNs, particularly Long Short-Term Memory (LSTM) networks, are commonly used due to their ability to handle time-series data, capturing temporal dependencies effectively. CNNs have also been applied to extract spatial patterns from financial data presented as time-series images or even to model market sentiment derived from news and social media. Furthermore Hu et al. (2021) conducted a survey summarizing various deep learning methods applied to stock market and foreign exchange (Forex) predictions. Their survey highlights the rise of hybrid models that combine various deep learning approaches, such as LSTM combined with Deep Neural Networks (DNN) or Reinforcement Learning. Reinforcement

Learning, in particular, has demonstrated notable potential by optimizing decision-making strategies through continual learning from past trades and market reactions. The authors also point out that the usage of advanced architectures like Hybrid Attention Networks (HAN), Convolutional Neural Networks (CNN), and models designed for specific purposes (e.g., self-paced learning mechanism and Wavenet) reflects the trend toward more sophisticated models capable of learning deeper representations of financial data. This enables these models to capture complex patterns and dependencies that may exist between different market indicators, thus improving prediction accuracy.

With the development in methodologies, particularly in nature language processing technologies, it becomes possible to take advantage of alternative data including news scripts, social media posts, and so on, especially as traditional methods based on historical price and volume often lack the ability to capture complex market behavior.

As the Internet grows rapidly, social media platforms become a new hot to be used as an alternative data source. For instance, Zhang et al. (2011) analyzed tweets to predict major stock indices, such as Dow Jones, NASDAQ, and S&P 500. The study found significant negative correlations between emotional tweets (like "fear" or "worry") and these indices, indicating that heightened public concern is associated with downward stock movement. They concluded that social media could be a reliable source of predictive information for short-term market movements, as public emotions directly influence trading behavior.

Also, several studies have shown the effectiveness of financial news in predicting stock market behavior. For example, Hu et al. (2018) proposed a deep learning framework for stock trend prediction that addresses the challenges of chaotic and varying quality in online news content. They designed a Hybrid Attention Network (HAN) that considers sequential dependencies between news items and the diverse influence of different news. This model integrates a self-paced learning mechanism to enhance the learning process by initially focusing on more informative data and progressively tackling challenging news samples. Their findings indicated that the hybrid approach could significantly improve prediction accuracy and suggested that a simple trading strategy based on this model could achieve better annualized returns compared to traditional baselines. Besides García-Méndez et al. (2023) focused on the automatic detection of relevant information, predictions, and forecasts in financial news using LDA for topic modeling. Their system analyzes financial news to extract meaningful insights, using temporal analysis to distinguish speculative statements and future-oriented predictions. By combining multi-paragraph topic segmentation with co-reference resolution, their approach aims to assist investors in filtering relevant information and separating forecasts from less pertinent content. The emphasis on temporality at the discursive level helps identify relevant predictions, providing investors with a tool to analyze investment strategies efficiently.

In summary, the integration of sophisticated machine learning and deep learning techniques, along with the utilization of a broader range of data sources, marks the current trajectory of research in this field.

Although there has been much work on how to tackle news data to make stock prediction, the methodologies in the recent research are relatively complicated and computationally expensive. In this paper, a straightforward methodology to extract useful information from news text will be developed to predict stocks return.

3.empirical analysis

In this section, an empirical text regression analysis that follows the procedures in the introduction will be conducted to predict the index daily return, and the details of the methodology will be described including data source choosing, nature language processing model, machine learning model and their results.

3.1data set choosing and its explanations

The data used in this paper can be divided into two parts: First, the news scripts from which useful information is extracted and second, the stock market index daily returns to be predicted. Here are the detailed introductions of the news source used and the stock market index to be predicted:

1. Xinwen Lianbo as the news scripts source: Xinwen Lianbo is a daily news programme produced by China Central Television (CCTV), a state broadcaster. It is shown simultaneously by all local TV stations in mainland China, making it one of the world's most-watched programmes. It has been broadcast since 1 January 1978. Also, this program is used as a medium for the state to announce government announcements and meetings, commentaries on major economic and policy issues, and the activities of national leaders, so the program reflects official positions of the Chinese government on a wide range of matters.
2. SSE Composite Index as the dependent variable: The SSE Composite Index (Shanghai Stock Exchange Composite Index) also known as SSE Index is a stock market index of all stocks (A shares and B shares) that are traded at the Shanghai Stock Exchange. The SSE Composite Index encompasses all listed companies on the Shanghai Stock Exchange weighted by their market capitalization, which includes large-cap, mid-cap, and small-cap stocks across various sectors. This broad inclusion makes it a comprehensive barometer of the Chinese stock market.

To further illustrate the reasons why I chose Xinwen Lianbo as the news scripts source and the value behind it, I have to give an example. Before the COVID-19 full opening, it could be witnessed that the frequency of the topic of COVID-19 was going through a high peak, and skeptically this movement can be seen as a preparation in public opinion. So, the value of choosing this particular news source can be summarized in three aspects. First, Xinwen Lianbo is an important official comprehensive broadcaster that enjoys great significance domestically and therefore fits the need to predict SSE index both in content coverage and quality. Second, as it reflects official positions, it may have leading effects in public behaviors including the financial market. Third, just as the example I give, it may have hidden knowledge or motives of the government. With all the three points, I think Xinwen Lianbo is a good choice to be used as news source to predict SSE Composite Index.

3.2machine learning models

In this paper, there are mainly two models used. They are the first LDA (Latent Dirichlet

allocation) topic model and the second long short-term memory (LSTM) model. LDA are used to divide news script of each day and get the topic proportions of each day. Together there are 268 days of proportions, and they will form a time series for each topic whose length is 268. By dividing news scripts into topics and getting their proportions, we can find structures in the news scripts and these structures lie in those time series. The descriptions of the two machine learning models are as follows.

1.LDA (Latent Dirichlet allocation): Latent Dirichlet allocation (LDA) is an unsupervised generative probabilistic model for modeling a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Latent Dirichlet allocation (LDA), first introduced by Blei, Ng and Jordan in 2003, is one of the most popular methods in topic modeling. LDA represents topics by word probabilities. The words with the highest probabilities in each topic usually give a good idea of what the topic is. And these word probabilities can be obtained from LDA. LDA assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well.

Given a corpus D consisting of M documents each of length N_d, LDA models D according to the following generative process:

(a) Choose a multinomial distribution ϕ_t for topic t ($t \in \{1,\ldots,T\}$) from a Dirichlet distribution with parameter β.

(b) Choose a multinomial distribution θ_d for document d ($d \in \{1,\ldots,M\}$) from a Dirichlet distribution with parameter α.

(c) For a word w_n($n \in \{1,\ldots,N_d\}$) in document d,

I. Select a topic z_n from θ_d.

II. Select a word w_n from ϕ_{zn}.

In the above generative process, words in documents are only observed variables while others are latent variables (ϕ and θ) and hyper parameters (α and β). The probability of observed data D is computed and obtained of a corpus as follows:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d$$

From the generative process whereby the documents are created, the various distributions (the set of topics, their associated word probabilities, the topic of each word, and the particular topic mixture of each document) can be learned through statistical inference. There are many strategies to estimate LDA parameters such as Gibbs sampling, Variational Bayes inference and Expectation Maximization.

2.Long Short-Term Memory model: Long Short-Term Memory (LSTM) is a specialized type of recurrent neural network (RNN) architecture designed to effectively capture and retain long-term dependencies in sequential data. Introduced by Hochreiter and Schmidhuber in 1997, LSTMs address the vanishing gradient problem that often hampers traditional RNNs, enabling them to learn from extended sequences without losing critical information over time. The core of an LSTM consists of a cell state and three key gates: the input gate, forget gate, and output gate. These gates regulate the flow of information, deciding which data to retain, update, or discard at

each step, thereby maintaining a stable and flexible memory mechanism. This capability makes LSTMs particularly powerful for a wide range of applications, including natural language processing, speech recognition, time series forecasting, and any task that involves understanding and predicting patterns over time. By effectively managing and leveraging long-term dependencies, LSTMs have become a foundational tool in deep learning, enabling more accurate and robust models in complex sequential tasks.

3.3model results

LDA model will generate the according words for each topic, but it needs a given number as the number of topics. There are many statistical methods to choose the number of topics, but in this project, by trials I think checking how well the words for each topic form a topic will give the best result. By checking the topics and the according words manually, 10 as the number of topics gives the best result. The main result of the LDA model is as follows.

Table 1 Topic Proportion Sheet for News Scripts

Date	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Return
2022/11/4	0.42	0.05	0.0	0.0	0.08	0.23	0.0	0.23	0.0	0.0	2.43%
2022/11/7	0.0	0.0	0.0	0.0	0.12	0.46	0.13	0.29	0.0	0.0	0.23%
2022/11/8	0.01	0.04	0.0	0.03	0.0	0.51	0.0	0.26	0.1	0.05	-0.43%
...	...	...	...	...	...	...	...	...	...	...	...
2023/12/7	0.43	0.0	0.0	0.02	0.48	0.07	0.0	0.0	0.0	0.0	-0.09%
2023/12/8	0.0	0.23	0.09	0.0	0.24	0.28	0.03	0.04	0.0	0.09	0.11%

In Table 1, every trading day's news script is divided into multiple topics represented by their proportions and 'Return' stands for the SSE index return for that day. In case you are interested in what each topic really is, the words of the highest frequency for each topic are presented as Table 2. Please notice that the LDA will not summarize what the topic is automatically but only give the words of highest possibility for each topic, so what each topic really stands for should be summarized according to these words afterwards. For example, for the ten topics, I summarized them as international affairs, party affairs, transportation etc. From the ten topics, we can see a roughly comprehensive coverage over every aspect of the country which matches SSE index as a general financial market index.

Table 2 Feature Words for Each Topic

TOPIC	WORDS
1.INTERNATIONAL AFFAIRS	会见, 倡议, 领导人, 峰会, 论坛, 战略伙伴, 人文, 共识, 互利, 对话, 两国人民, 原则, 开幕式, 协作, 高度, 国事访问, 经贸, 高峰论坛, 合 作伙伴, 主义, 伙伴, 气候变化, 朋友, 共同利益, 伙伴关系
2.PARTY AFFAIRS	理论, 干部, 文章, 讲话, 监督, 宣传, 统一, 马克思主义, 法治, 意识, 革命, 全党, 决策, 斗争, 党和国家, 决定性, 开局, 中华, 单位, 领悟, 党组, 培训, 党校, 运用, 实际
3.TRANSPORTATION	旅客, 消费, 人们, 游客, 春运, 客流, 体验, 节目, 招待会, 金牌, 冰 雪, 交通, 比赛, 公路, 市民, 大使馆, 民众, 兔年, 景区, 小时, 祖国, 女子, 水路, 技能, 冠军
4.AEROSPACE	航天员, 神舟, 载人, 空间站, 航天, 飞船, 乘组, 耕地, 天舟, 座谈会, 货运, 劳动, 太空, 组合体, 实验, 事故, 人口, 盐碱地, 京津冀, 空间, 火箭, 协同, 发射场, 文章, 调研
5.FINANCE	金砖, 金融, 座谈会, 责任, 经济社会, 力度, 协同, 格局, 倡议, 环境 保护, 监管, 决策, 质量, 负责同志, 调研, 进出口, 主体, 合理, 意见, 大局, 民生, 动力, 生物, 试验区, 改革开放
6.ECONOMY	消费, 地带, 医院, 月份, 互联网, 物流, 博会, 社区, 基层, 外贸, 转 型, 网络, 进口, 制造业, 博览会, 总体, 数字化, 冰雪, 出口, 动能, 进 出口, 汽车, 智能, 武装, 供给
7.DISASTER RELIEF	应急, 防汛, 调研, 大学生, 高校, 核污染, 运动会, 面积, 秋粮, 大运 会, 粮食, 高温, 救援, 部队, 全力, 毕业生, 大火, 设施, 民众, 公园, 运动员, 天气, 代表团, 火炬, 救灾
8.COVID-19	疫情, 防控, 新冠, 病例, 通报, 社区, 医疗, 疫苗, 肺炎, 患者, 宣讲 团, 病毒, 精准, 调整, 本土, 负责人, 药品, 医院, 基层, 方案, 人群, 检测, 重症, 成员, 老年人
9.OTHER AFFAIRS	管道, 事件, 救援, 粮食, 地震, 制造业, 集团, 农村, 事故, 调研, 民 众, 集群, 人才, 小麦, 质量, 银行, 武器, 基地, 天然气, 贷款, 布局, 有限公司, 农田, 抗议, 公司
10.LEGAL AFFAIRS	草案, 委员, 宪法, 主席团, 代表团, 委员长, 人大代表, 妇女, 党和国 家, 建议, 残疾人, 部长, 秘书长, 决议, 法律, 全体会议, 汇报, 议案, 伟业, 常务主席, 人选, 名单, 任命, 审查, 主任委员

In Table 1, there are ten topic time series, and one index returns time series whose length are all 268. Next, the topic time series and returns time series were used as features and divided into the training data set accounting for 80% of all the data samples and the test data set accounting for the rest 20%. Then, a two-layer LSTM model was run on the training data set to model the returns. To predict today's return, the LSTM model didn't use today's news, since today's news was a conclusion not prediction for today's events but used past historical news information and past historical returns as the input. Meanwhile, the features put in the LSTM were normalized.

The result of LSTM model on the test data set is as Figure 1. From Figure 1, if we take into consideration that we are predicting a market index which is usually thought to be the systematic risk impossible to reduce, then the overall fitting effect of the LSTM model can be seen as good or at least acceptable.

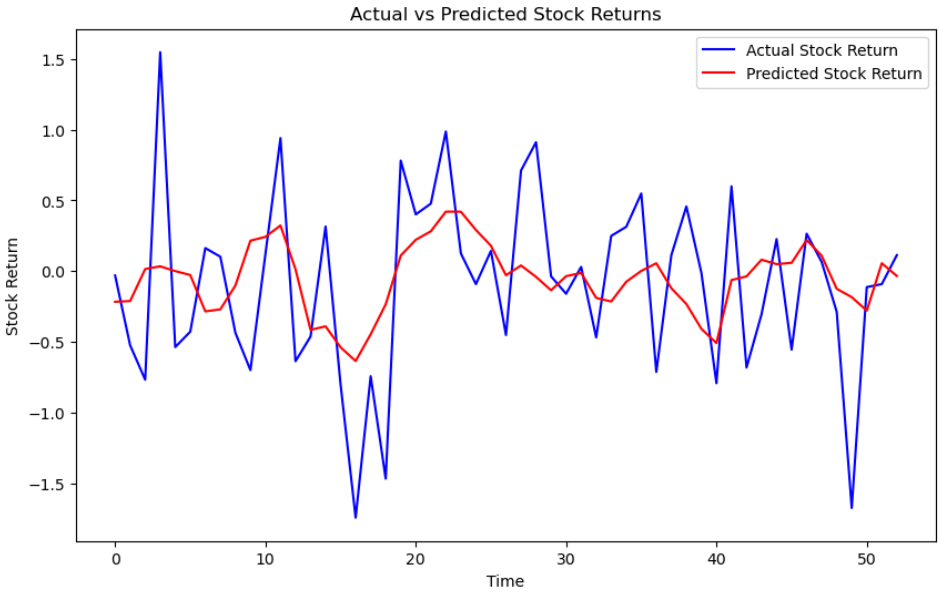

Figure 1 Actual vs Predicted Stock Returns

4.evaluation

In this section, several methods will be adopted to validate the effectiveness of the methodology.

4.1return rates as feature only

To demonstrate whether the methodology extracted information from the news scripts efficiently, one good way is to remove the topic time series from the LSTM model and examine the difference. So, I constructed another LSTM model with historical return rates as the only feature. The result on the test data set is as Figure 2. It can be seen clearly that the model with return rates as its only feature hardly captures any trends of stocks compared with the one with news topic information. So, it is safe to say the methodology developed in this paper can effectively extract useful information from news and significantly increase the model's capability to capture stock trends.

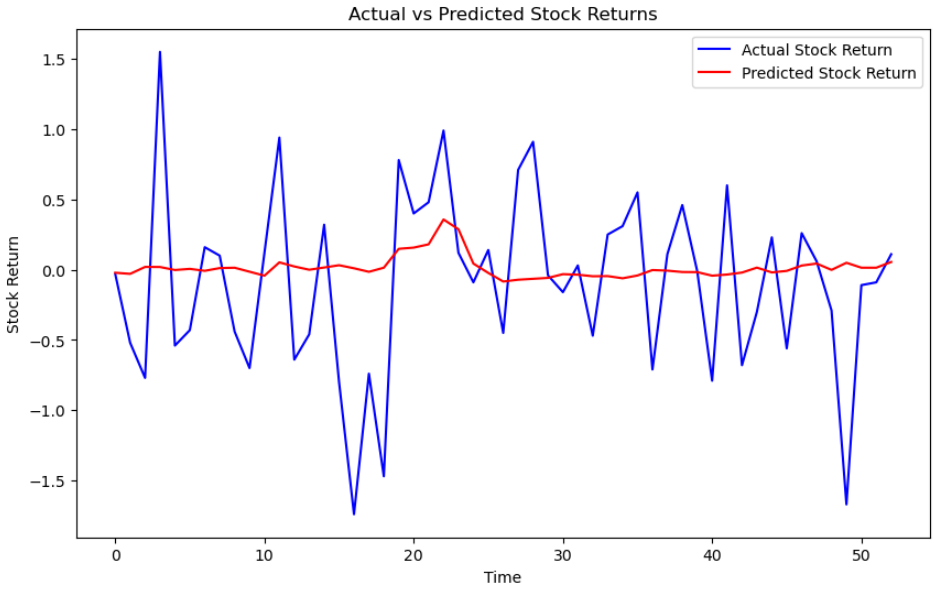

Figure 2 LSTM with historical returns as feature only

4.2mathematical explanations

Let's assume that the return rate R_t and news information N_t (N_t can be a vector including historical data) are both the reflection of the latent economic fundamentals a_t with a small stochastic error term of white noise. This assumption can be expressed by the following mathematical formulas.

$$①R_t = f(a_t) + u_t$$
$$②N_t = g(a_t) + v_t$$

From ① and ②:

$$a_t = g^{-1}(N_t - v_t)$$
$$R_t = f \cdot g^{-1}(N_t - v_t) + u_t$$

Let $h(x) = f \cdot g^{-1}$, since v_t is very small and based on Taylor's theorem, we have:

$$R_t = h(N_t - v_t) + u_t$$
$$③R_t = h(N_t) - h'(N_t) \cdot v_t + u_t$$

LSTM model as a nonlinear regression technique uses N_t (news information) as the regressor to fit R_t which is expressed as ③. The result of LSTM should be close to $E(R_t|N_t)$ and can be expressed as ④:

$$④\widehat{R_t} = E(R_t|N_t) = h(N_t) - h'(N_t) \cdot \widehat{v_t}$$

Based on law of total variance, we have $var(R_t) = E(var(R_t|N_t)) + var(E(R_t|N_t))$. From this formula, we can deduce:

$$⑤var(\widehat{R_t}) = var(E(R_t|N_t)) = var(R_t) - E(h'(N_t)^2) \cdot E(var(v_t|N_t)) - var(u_t)$$

⑤ can explain the phenomenon that the predicted return volatility is smaller than the actual volatility and give the specific volatility gap formula under assumptions ① and ②.

Now we have a prediction formula ④ for LSTM and formula ① for the original return rates. Figure 1 can be seen as a visual demonstration for ① and ④. From these formulas, we can see

the return rates of ① and ④ all consist of two parts, one deterministic part plus a small stochastic error and according to ⑤, the volatility of ① is bigger than ④. The deterministic parts of the two formulas are time series which if the model is correct should be similar, so smoothing techniques will be applied to mitigate the stochastic parts and see how well the model fits. The result is as Figure 3. Compared to Figure 1, because the stochastic parts are mitigated, Figure 3 is a more effective way to visualize the fitness of the model. From Figure 3, it can be considered that the fitness of the model is satisfying.

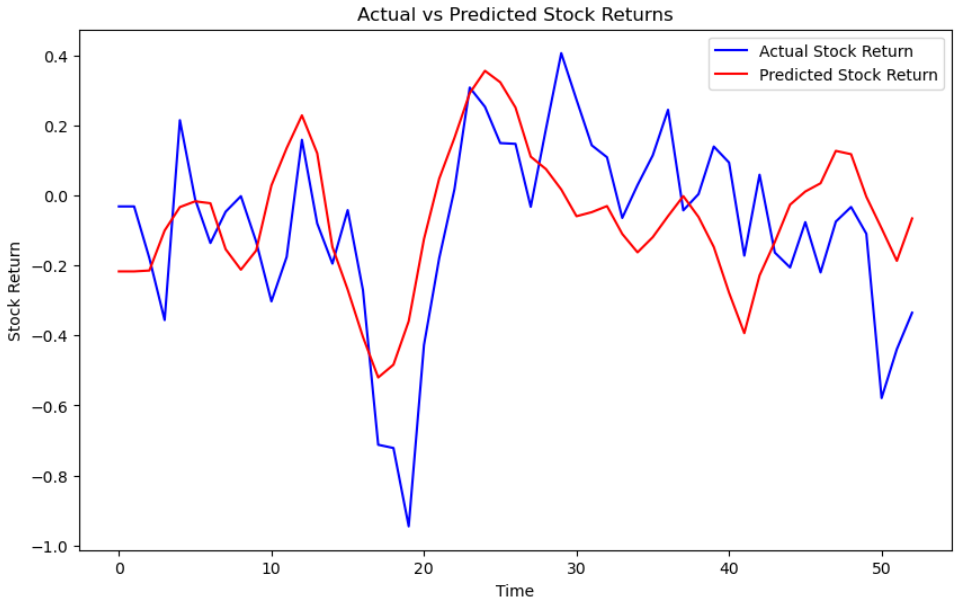

Figure 3 Smoothed Returns

4.3confusion matrix and investment simulation

To examine the financial effectiveness of the methodology, investment simulation is inevitable. So, in this part, a simple investment strategy will be constructed and simulated on real market data. Just as we have topic COVID-19 in our data set which may not exist in other periods of time, the patterns the model captures can be unstable if the model try to predict a relatively far future, so only the former month of the test data set will be used in this simulation.

Before the simulation, let's look at the confusion matrix, so we can have a general idea about how well the model can predict financially. The confusion matrix is as Table 3. 'Positive' represents the upward trend of the index and 'negative' represents the downward. The accuracy rate of the actual and predicted returns being in the same direction is 73.3%. Compared to similar research, it's a very good result.

Table 3 Confusion Matrix

Total 30	Predicted Positive	Predicted Negative
Actual Positive	10	4
Actual Negative	4	12

Setting the initial capital as 1, the strategy is to go long the index when the model predicts a positive gap between the return rate and the transaction cost, and if allowed to go short the index when the model predicts a downward trend of the index whose profits can cover the transaction cost. The trading frequency is on a daily basis, and it assumes that transaction cost is 2‰. The cumulative profit curves are illustrated in Figure 4. In Figure 4, we can see steady positive trends in the profits in both situations. Given there are 240 trading days in a year, the annualized return rates of the strategy are 11.65% and 31.54% respectively. Also compared to similar research, given the simplicity of the methodology, these return rates are rewarding.

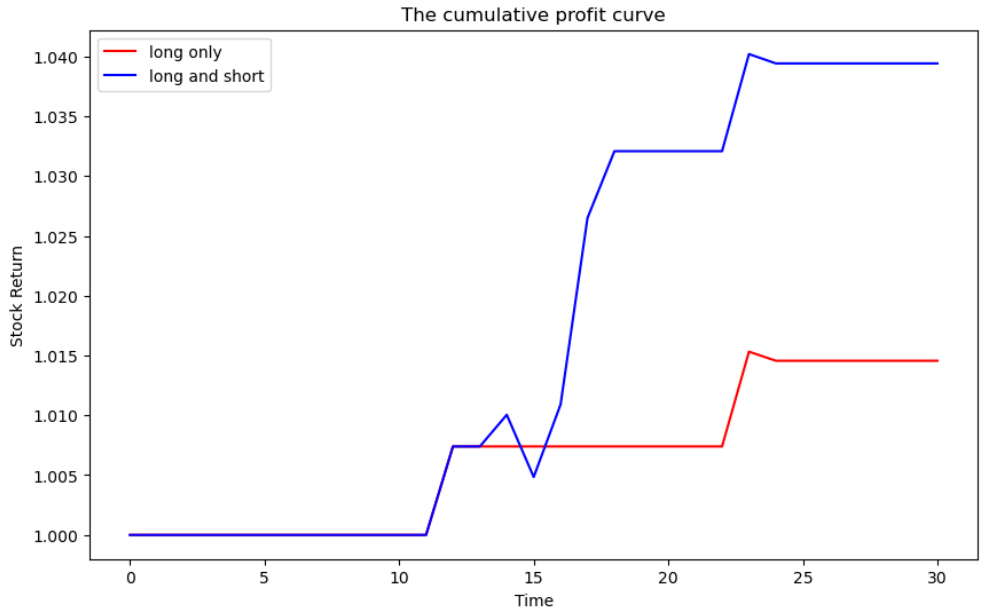

Figure 4 The Cumulative Profit Curves

5.conclusion

Using a very special news source, I developed a simple but effective methodology to predict market index (SSE composite index) whose performance is not inferior to related research models that adopt sophisticated deep learning network and use a vast amount of textual data.

Of course, there are limitations. First, the news source is unique to some extent, so whether the methodology flow in this paper can be generalized onto other sources of news is a question to answer. Second, even though many evaluation methods are applied to validate the performance, the pattern itself caught by the methodology can be unstable through time and therefore to make the methodology more trustworthy, similar experiments should be conducted in different periods of time separately to test its applicability capturing different patterns.

[1] Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, *184*, 115537.

[2] Park, C.-H., & Irwin, S. H. (2007). What do we know about the profitability of technical analysis? *Journal of Economic Surveys, 21,* 786–826.

[3] Atkins, A., Niranjan, M., & Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science, 4(2)*, 120–137.

[4] Hardeniya, T., & Borikar, D.A. (2016). Dictionary Based Approach to Sentiment Analysis - A Review. *International Journal of Advanced Engineering, Management and Science, 2.*

[5] Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T.-Y. (2018). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)* (pp. 261–269). Association for Computing Machinery.

[6] Hedayati Moghaddam, A., Hedayati Moghaddam, M., & Esfandyari, M. (2016). Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, *21*(41), 89–93

[7] Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, *184*, 115537.

[8] Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, *4*(1), 9.

[9] Hansen, K. B., & Borch, C. (2022). Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. *Big Data & Society*, *9*(1).

[10] Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through Twitter: "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, *26*, 55–62.

[11] García-Méndez, S., de Arriba-Pérez, F., Barros-Vila, A., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with Latent Dirichlet Allocation. *Applied Intelligence*, *53*(16), 19610–19628.

[12] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

[13] Jelodar, H., Wang, Y., Yuan, C., et al. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*, 15169–15211.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.