Zeng Zhe MC364969 Section 2

Finding Patterns in News Reports and Stock Index

Introduction:

Text data have always been an important data source for human information, and news reports are one of the most important types in text data. Recent years, as the traditional approaches in financial analysis, technical analysis and fundamental analysis, are reaching their limitations, many are searching for new data sources to strengthen their understandings in the financial market. These kinds of new data sources are called alternative data. For example, the electricity usage of the factory of Tesla is considered to be a perfect instance for alternative data, as the electricity usage can be an indicator for the factory production. Text data is also considered as alternative data and with new advanced techniques such as machine learning, it is now possible to adopt these new techniques to the text data and explore some new patterns that used to be difficult or impossible for humankind to discover.

In this background, in this project I'm interested in finding patterns between news scripts and stock index which to be specific is the patterns between Xinwen Lianbo and SSE Composite Index. From 2022/11, I gathered about one-year data for both Xinwen Lianbo news scripts and SSE index returns, and there are 268 valid trading days' data. For these 268 days' data, I will apply LDA (Latent Dirichlet allocation) topic modeling to the news scripts to divide each day of news script into several topics by how much proportion each topic occupies in the news script. After this division, I will use this proportion and SSE index returns to run several linear regressions and explore whether there is any pattern between news scripts and the SSE index returns.

Models and Methodologies:

I. Data set choosing and explaining:

In order to find patterns in news scripts and index, it's critical to choose the right pair of news and index. We don't want a rumorous news source, and the content of the news source should be aligned with the chosen index. So, for the index part, I choose SSE Composite Index since it's a general representation of the whole financial market in China. For the news part, to match the properties SSE holds, the news source should cover the whole domestic affairs instead of focusing on one particular field and we hope the news source is authentic and of a high quality as much as possible. For the above reasons, I choose Xinwen Lianbo as the news source in this project. Xinwen Lianbo is an important official comprehensive broadcaster that enjoys great significance domestically, and therefore can be used as the suitable news source fitting SSE both in content coverage and in this quality.

There are some properties I have to mention about Xinwen Lianbo. 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. The program reflects official positions of the Chinese government on a wide range of matters and may have leading effects on public behaviors including stock markets. To make this point clearer, 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, it authentically reports domestic affairs in many fields. Second, as it reflects official positions, it may have leading effects in public behaviors. 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 find patterns with SSE Composite Index.

II. Models used:

In this project, there are mainly two models used. They are the first LDA (Latent Dirichlet allocation) topic model and the second linear regression. 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. After the LDA, the time series will be used as explanatory variables to run several linear regressions with SSE index returns as the dependent variable to explore whether there are any capturable patterns.

The descriptions of the two 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.

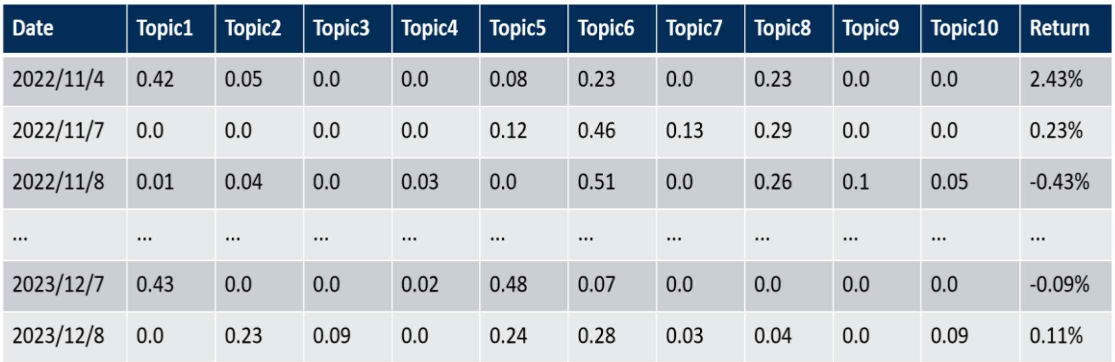
2. ordinary least squares (OLS) linear regression: To find explainable patterns between news and stock index, I choose linear regression to present these relations. As for the parameter estimation, ordinary least squares (OLS) is the most classic method. In this project, I will use the topic proportion time series obtained from LDA as explanatory variables and index returns as the dependent variable to run those linear regressions and try to find some patterns.

Data Analysis:

I.LDA 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.

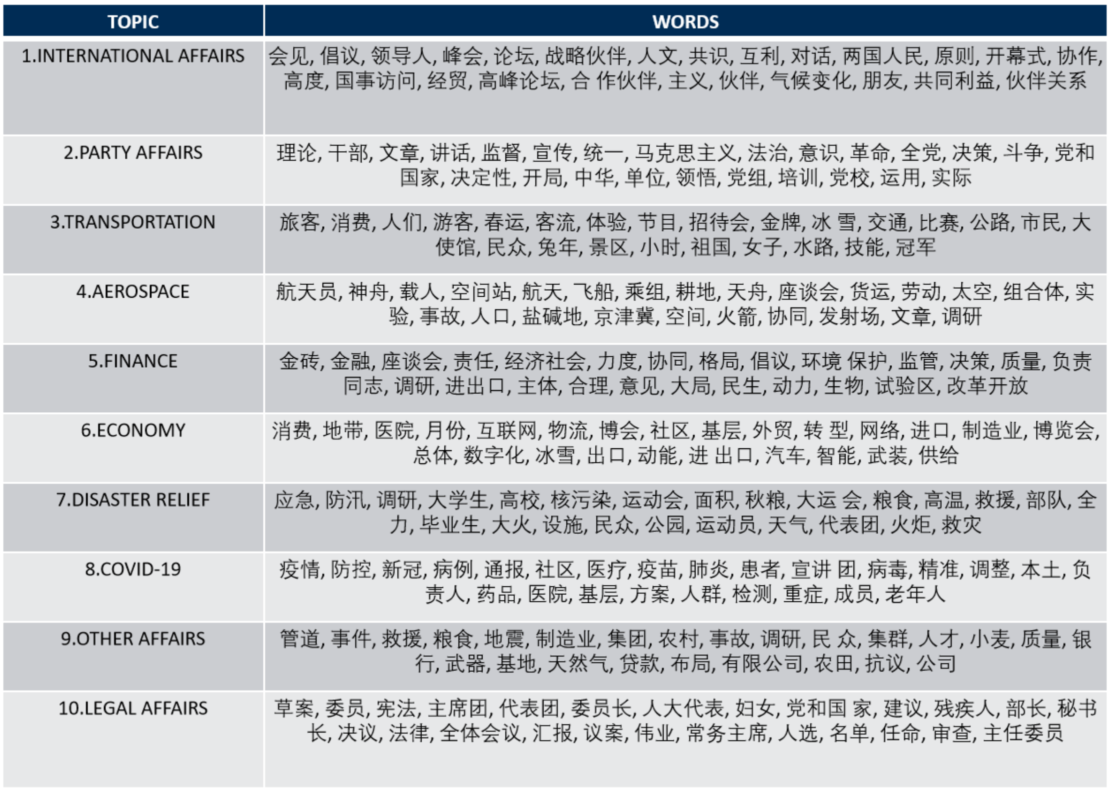
Table 1 Topic Proportion Sheet for News Scripts



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 follows. 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

II.Linear regression:

In Table 1, there are ten topic time series, and one index returns time series. Our linear regressions focus on the relationships between those topics and the index. And because the sum of all topic proportions is 1, there is Perfect Multicollinearity Problem, therefore it’s impossible to regress the index returns on all topic proportions at once. One thing to notice is that the market closing time is earlier than the time of the news broadcast, so if you use topic proportions to fit the returns, it is not a prediction.

By fitting the index returns on each topic with its proportion time series one by one, I try to find some linear patterns between index returns and topics. Most of the regressions are not significant at the significant level of 5%, except the one with topic finance as the regressor which is not surprising.

Here is the figure for index returns and topic finance occupation in news scripts. The regression table is also attached.

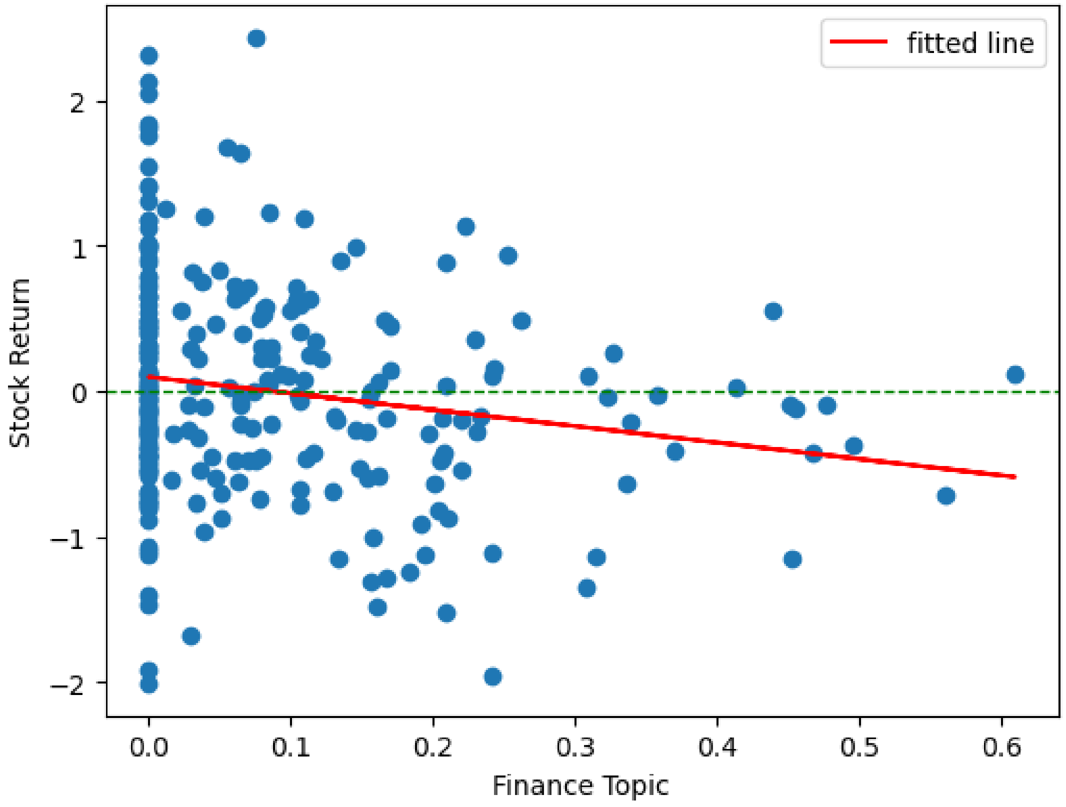


Figure 1 Return vs Topic proportion

Table 3 Regression Table for Index and Topic Finance

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.028
Dependent Variable: returns          AIC:                596.7910
Date:               2024-11-24 11:16 BIC:                603.9655
No. Observations:   267              Log-Likelihood:     -296.40
Df Model:           1                F-statistic:        8.677
Df Residuals:       265              Prob (F-statistic): 0.00351
R-squared:          0.032            Scale:              0.54326
--------------------------------------------------------------------
           Coef.    Std.Err.      t      P>|t|    [0.025    0.975]
--------------------------------------------------------------------
const      0.1000     0.0553    1.8091   0.0716   -0.0088   0.2089
x1        -1.1291     0.3833   -2.9457   0.0035   -1.8838  -0.3744
------------------------------------------------------------------
Omnibus:               6.721        Durbin-Watson:           2.014
Prob(Omnibus):         0.035        Jarque-Bera (JB):        7.205
Skew:                  0.273        Prob(JB):                0.027
Kurtosis:              3.592        Condition No.:           9
==================================================================
```

From the regression results, we can learn that this regression is significant even at 1% significant level which shows there is a very strong relationship between the two variables and Durbin-Watson statistic is very close to 2 suggesting there is no autocorrelation problem in the error terms.

The most important findings are in Figure 1 and the coefficient of finance topic proportion. The coefficient is significantly negative at 0.5% significant level leaving Top-right and Bottom-left two empty spaces and the Top-right blank seems bigger.

Here are my interpretations for these phenomena:

1. When the market is very hot, as a news broadcast with official background, Xinwen Lianbo tends to be reluctant to mention the financial market in big chunks in order to prevent over-heating. In a country whose financial market is not that mature, this action is very reasonable.
2. Similarly, when the market crashes, Xinwen Lianbo also tends to bring some good news. (it’s extremely rare for Xinwen Lianbo to bring some bad news deliberately, so the increase in topic proportion is very likely to be a positive reference)
3. The fact that the Top-right blank is bigger suggests that the attitude to avoid over-heating is very strong.

Conclusion and Discussion:

In this project, by using LDA topic model and linear regression, I explored the hidden patterns between news reports and stock index. As results, I found Xinwen Lianbo’s attitudes as an official background media towards finance topic when the market is hot or cold. It seems that Xinwen Lianbo may act in a manner for Counter-market Regulation, avoiding this topic when market is too hot and mentioning this topic positively when it crashes.

There are one limitation which may cause problems. As a TV news broadcast program, the strategy of its reports may vary through time which brings question to the stability of the findings in this project. A possible solution is to expand our data size. If the findings persist through out the expanding, then the stability of this strategy can be more trustworthy.