Identifying Business Trends from Stock Market Data utilizing Domain-Specialized Sentence-RoBERTa and BERTopic Technique*

Ye Lim Jung[1,2,*] and Hyoung Sun Yoo[1,3,*]

[1] Center for Global R&D Data Analysis, Korea Institute of Science and Technology Information (KISTI), Seoul, Republic of Korea

[2] Data and High Performance Computing Science, University of Science and Technology, Daejeon, Republic of Korea

[3] Science and Technology Management Policy, University of Science and Technology, Daejeon, Republic of Korea

Abstract

Stock market is expected to provide meaningful insights into promising businesses by reflecting public expectations and investments. However, it has been difficult to unveil specific business trends from stock market because the existing stock market analysis by sector or industry covers a wide range of businesses. In this study, we propose a method to identify business trends in detail by using the BERTopic technique with the Pro-SRoBERTa embedding model, which is devised for comparing similar business items. Using 246,468 business description documents for listed companies around the world over the past 10 years, we clustered companies with similar business themes by topic modelling using BERTopic with the Pro-SRoBERTa model, and then calculated the annual stock return of each cluster and the ratio of the cluster's market capitalization to the total market capitalization. By performing regression analysis of them by year, emerging or declining business topics have been identified. The proposed method enables cross-industry trend analysis for businesses spanning multiple industries due to the convergence of various technologies. It could also provide useful insights to business practitioners and investors by effectively and efficiently detecting changes in business trends at a granular level.

Keywords

Business trends, Stock market, Topic modelling, Sentence-RoBERTa, BERTopic

1. Introduction

Recently, business trends are changing more rapidly due to technological innovations and the variations of customer behavior patterns. For business trends identification, News, social network service (SNS), and patent data have been widely used over the past decades [1, 2]. However, when utilizing News or social media data, caution is required regarding bias of the results depending on the source of the data collected. In addition, in the case of using patent data, it may be difficult to sufficiently figure out their direct impacts on business or industry because it reflects more technological perspectives.

To address these limitations, this study proposes a new methodology for extracting business trends using stock market data and deep learning-based language model. Stock market information is expected to provide meaningful insights into business changes, especially in complex situations such as the emergence of a pandemic [3]. Even if the overall stock market is volatile, the relative stock price movements of different sectors or clusters can provide insights into promising or declining businesses.

For this purpose, the BERTopic algorithm, one of the latest topic modeling techniques, was employed to cluster companies with similar business topics [4]. In the procedure, a Pro-SRoBERTa

*Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT)@ CIKM 2025

[1,*] Corresponding author.

yelima@kisti.re.kr (Y. L. Jung); hsyoo@kisti.re.kr (H. S. Yoo)

0000-0003-0208-8870 (Y. L. Jung); 0000-0002-4010-3878 (H. S. Yoo)

(Product-specific Sentence-RoBERTa) that we specialized for comparing the similarity of products or services, was applied as an embedding model to cluster diverse businesses more effectively [5]. Afterwards, the stock market performance of these clusters was analyzed to discover rising or declining businesses. The regression analysis was performed on the stock performance of each cluster over the last 10 years. As a result, business topics with a statistically significant upward or downward trends have been identified (Fig. 1).

The proposed methodology facilitates the efficient detection of trends derived from stock market performance, specifically targeting business subjects at the intersection of multiple technological and/or industrial domains at a granular level.

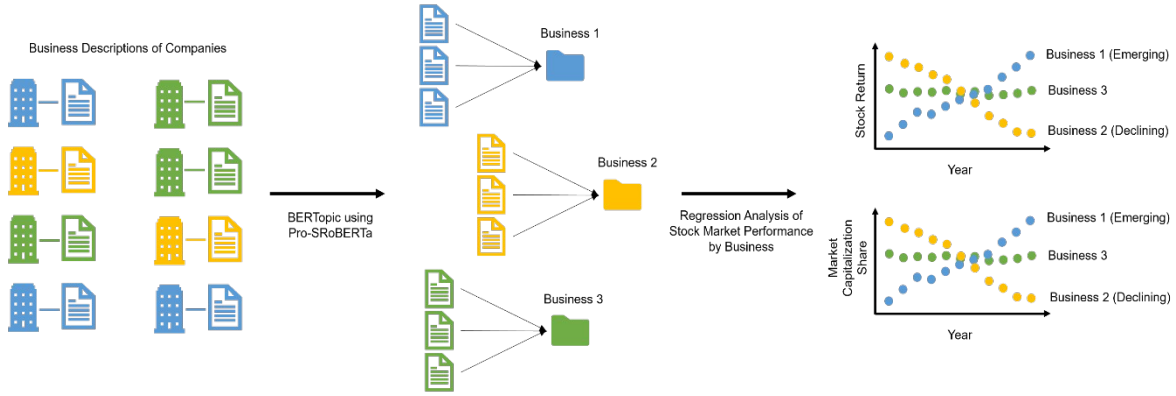

Figure 1: Framework of the proposed methodology for business trends identification

2. Literature Review

In recent years, many studies have been conducted to analyze large amounts of text data using embedding techniques to detect emerging trends in various fields. Dridi et al. proposed a methodology called "Leap2Trend" for early detection of new scientific research trends [6]. To overcome the time delay of existing citation-based analysis and the limitations of content analysis, the authors employed temporal word embedding using Word2Vec techniques to track dynamic changes in keyword relationships within scientific corpora.

In addition, Saritas et al. developed a method to identify emerging business trends by using globally accumulated text data in mobile commerce (m-commerce) area [7]. The large amounts of unstructured data such as news reports and patents in m-commerce were processed by text mining, semantic analysis, machine learning, and bibliometric analysis to extract key technologies and business trends. The authors proposed a business intelligence data mart that provides an overview of the whole landscape of m-commerce.

Another emerging trends detection method called "BERTrend" has been devised by Boutaleb et al [8]. The authors presented a framework that detects and monitors weak signals from large-scale text corpora to find emerging trends. The neural topic modeling (BERTopic) technique was applied in the form of online learning to track topic evolution over time. A metric to quantify topic popularity was introduced in the study by considering the number of documents and the update frequency within a certain period. As a result, the metric classified topics as "strong signal," "weak signal," and "noise" based on the empirically chosen thresholds from two types of large text corpora: the arXiv dataset and the New York Times (NYT) news dataset.

Although there have been studies to detect trends using various data and cutting-edge techniques, there have been few attempts specialized in detecting business trends with a multimodal approach that combines stock market time series data and text data about a company's business activities.

3. Methods

3.1. Datasets Collection and Preprocessing

Stock market data (stock prices and market capitalizations (MC)) of listed companies around the world for the past 10 years (2014-2023) and text data describing the companies' businesses and products/services were collected from Moody's analytics (Table 1).

The business description documents consist of an overview of the company, primary business lines, main activities, and main products and services. They were pre-processed including conventional stop word processing. Since they are business-specific documents, words such as 'product', 'development', 'business', 'company', 'department' were additionally used as stop words to refine the documents.

Table 1
Descriptive statistics for the dataset

Year	N	MC total[a]	SR mean	SR std. dev.	SR median
2014	21,088	69,329	0.139	1.033	0.030
2015	21,894	62,170	0.181	6.632	-0.025
2016	23,025	66,211	0.506	39.490	0.011
2017	23,969	75,045	0.196	2.007	0.024
2018	25,013	65,857	-0.089	2.308	-0.193
2019	25,694	80,484	0.177	1.318	0.055
2020	26,503	89,186	0.540	26.025	0.001
2021	27,245	107,307	0.300	2.132	0.086
2022	29,214	93,903	-0.104	0.892	-0.184
2023	22,823	96,901	0.125	1.397	0.003

[a]Billion USD, MC: market capitalization, SR: annual stock return

3.2. Topic Modelling and Stock Market Performance Analysis

Topic modelling using the BERTopic algorithm was conducted to cluster the companies with similar business topics. We used the Pro-SRoBERTa as an embedding model of the BERTopic for effective clustering. The Pro-SRoBERTa is a domain-specific model we have devised for calculating product or service similarity that was built by fine-tuning data on approximately 26 million pairs of product/service similarity relationships [5]. The dimensionality reduction was conducted using UMAP (Uniform Manifold Approximation and Projection) (n_neighbors= 15, n_components= 5, metric= cosine), followed by clustering with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (min_cluster_size= 20, min_samples= 10, metric= Euclidean). The number of topics were determined automatically to incorporate semantically similar topics.

For each topic cluster, the annual stock return (SR) in year t and the share of the cluster's market capitalization to the total market capitalization of all companies (MCS) in year t were calculated as shown below.

$$SR_t = \frac{Stock\ \ Price_t - Stock\ \ Price_{t-1}}{Stock\ \ Price_{t-1}} \quad (1)$$

$$MCS_t = \frac{\sum_{i \in C} MC_{i,t}}{\sum_{j \in A} MC_{j,t}} \quad (2)$$

where C denotes the set of companies belonging to a specific cluster, and A denotes the set of all companies in the entire stock markets.

Through regression analysis of SR and MCS depending on the year, it was classified as an emerging (hot) business topic when both were statistically significant ($p<0.05$) and positive, and it was classified as a declining (cold) business topic when both were statistically significant and negative. Such analytical approach has been widely used in the field of scientometrics to detect hot/cold topics in scientific literature by utilizing topic extraction using Latent Dirichlet Allocation (LDA) and its regression analysis over time [9, 10]. We adopted this methodology but utilized SR and MCS as indicators (dependent variables) to identify business trends.

4. Results

A total of 224 topics were derived as a result of topic modelling by the BERTopic using the Pro-SRoBERTa model. Among the topics, the most representative topics in which the SR and MCS are both significantly positive or negative are shown in Table 2 and Fig. 2. In the case of the insurance business ('insurance_life_reinsurance_accident'), both the SR and MCS have tended to decrease over the past 10 years. In the case of the robot business ('robot_linear_intelligent_robotic'), both the SR and MCS have increased over the past 10 years and have shown a particularly drastic rise recently.

Table 2
Examples of regression analysis results for business topics

Topic number	Business themes	Coefficient for SR	Coefficient for MCS	Trend
8	insurance_life_reinsurance_accident	-0.564*	-0.001**	Declining
195	robot_linear_intelligent_robotic	0.411*	0.00006*	Emerging

* p<0.05, ** p<0.01

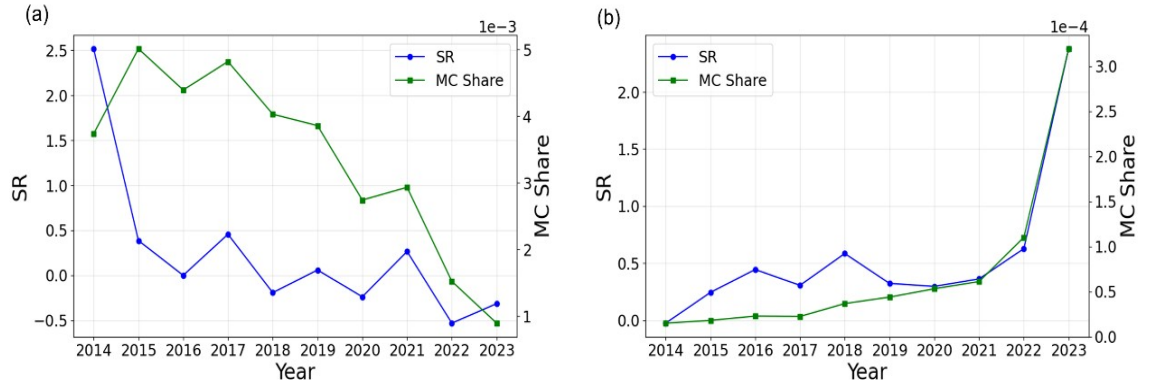

Figure 2: Time series of SR and MCS by business topic (a) Topic 8 ('insurance_life_reinsurance_accident'), (b) Topic 195 ('robot_linear_intelligent_robotic')

The results suggest that public expectations and investments have become more concentrated or marginalized in these businesses. To compare our analysis results with other existing methods, we compared them with Google Trends [11]. The results were also consistent with Google Trends search results for the business & industrial category (Fig. 3). While Google Trends requires prior knowledge of search terms to identify trends, the method proposed in this study has the advantage of being able to extract trends in an unsupervised manner from the unseen data.

Although this study presented long-term trends using a decade of data, short-term trends at a specific period can also be investigated by adjusting the analysis window (ongoing work). For example, if the analysis period is specified as 2019 to 2021 and monthly SR and MCS are used, trends during the COVID-19 pandemic period can be identified. According to the preliminary experiment results, the business themes that received the attention by the stock market during that period were "blockchain, bitcoin, cryptocurrency, ecosystem", "games, game, gaming, online", "pharmaceutical, medical, care, treatment", etc., while the businesses that did not receive attention from the market were "travel, hotel, hotels, rooms", "parts, vehicles, automotive, motor", "construction engineering, buildings, real", and so on.

Therefore, the new methodology proposed in this study has the potential to be utilized to automatically detect short- and long-term business trends.

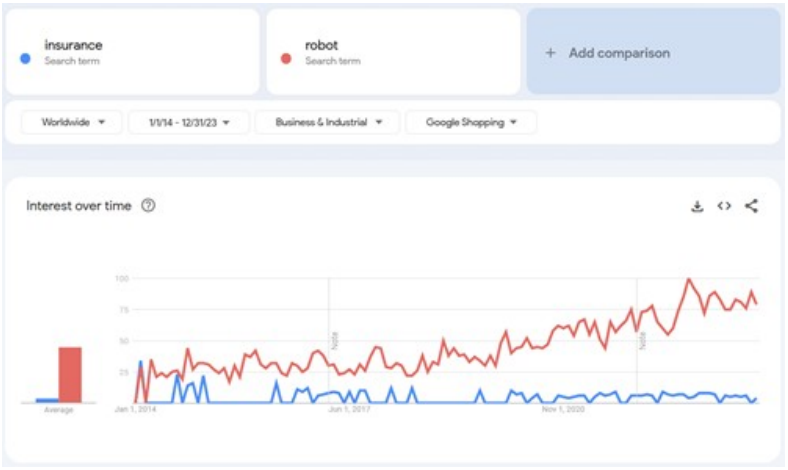

Figure 3: Google Shopping search results for "Insurance" and "Robot" in the Business & Industrial category of Google Trends

5. Conclusions

In this study, we have devised a method to identify emerging or declining businesses by clustering companies with similar business topics using the BERTopic algorithm with the Pro-SRoBERTa model. Through regression analysis on the SR and MCS of each cluster by year, business topics that were receiving attention or being alienated were identified. This study has limitations in that if the number of companies assigned to a topic is too small, it might be influenced by the performance of a few unique companies. Additionally, there have been cases where topics derived that are difficult to clearly identify the business subject. In follow-up studies, the results should be further improved through optimization of the number of topics and the number of companies assigned to each business topic.

This study has several contributions to preceding literature. First, a new method was developed to identify business trends by combining heterogeneous types of data: structured stock market time-series data and unstructured corporate business description data. In addition, it has an advantage to extract business trends in an efficient and automated manner by leveraging small and efficient fine-tuned LLM models. Third, this study addresses the challenge that it was difficult to understand trends by detailed business topic in existing industry level analysis. Further studies should ensure the

robustness of the model and increase its practical applicability. It can finally help corporate managers establish new product development strategies and investors adjust their investment portfolios.

Acknowledgements

This work was supported by the Research Program funded by Korea Institute of Science and Technology Information (KISTI) (No. K25L4M2C3) and the National Research Foundation of Korea (NRF) (No. 2022R1A2C1010387).

Declaration on Generative AI

Generative AI software tools have been used only for simple proofreading tasks, such as correcting grammatical errors in the manuscript.

References

[1] Wang, J., Zhao, W. X., Wei, H., Yan, H. and Li, X. Mining new business opportunities: Identifying trend related products by leveraging commercial intents from microblogs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1337-1347)*, 2013.

[2] Ena, O., Mikova, N., Saritas, O. and Sokolova, A. A methodology for technology trend monitoring: the case of semantic technologies. *Scientometrics*, 108 (2016), 1013-1041.

[3] Wagner, A. F. What the stock market tells us about the post-COVID-19 world. *Nature Human Behaviour*, 4, 5 (2020), 440-440.

[4] Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[5] Jung, Y. L. Market intelligence applications leveraging a product-specific Sentence-RoBERTa model. *Appl. Soft. Comput.*, 165 (2024), 112077.

[6] Dridi, A., Gaber, M. M., Azad, R. M. A. and Bhogal, J. Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. *IEEE Access*, 7 (2019), 176414-176428.

[7] Saritas, O., Bakhtin, P., Kuzminov, I. and Khabirova, E. Big data augmentated business trend identification: the case of mobile commerce. *Scientometrics*, 126, 2 (2021), 1553-1579.

[8] Boutaleb, A., Picault, J. and Grosjean, G. BERTrend: Neural Topic Modeling for Emerging Trends Detection. *arXiv preprint arXiv:2411.05930* (2024).

[9] Griffiths, T. L. and Steyvers, M. Finding scientific topics. *Proceedings of the National academy of Sciences,* 101, suppl_1 (2004), 5228-5235.

[10] Mane, K. K. and Börner, K. Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences,* 101, suppl_1 (2004), 5287-5290.

[11] Google. Retrived from https://trends.google.com/.