



ISSN: 2288-7709 © 2020 KODISA & ICMA.
JEMM website: <https://acoms.kisti.re.kr/jemm>
doi: <http://dx.doi.org/10.20482/jemm.2022.10.3.21>

Analysis of Business Performance of Local SMEs Based on Various Alternative Information and Corporate SCORE Index

Sun Hee HWANG¹ , Hee Jae KIM² , Dong Chul KWAK³

Received: May 11, 2022. Revised: June 3, 2022. Accepted: June 5, 2022.

Abstract

Purpose: The purpose of this study is to compare and analyze the enterprise's score index calculated from atypical data and corrected data. **Research design, data, and methodology:** In this study, news articles which are non-financial information but qualitative data were collected from 2,432 SMEs that has been extracted "square proportional stratification" out of 18,910 enterprises with fixed data and compared/analyzed each enterprise's score index through text mining analysis methodology. **Result:** The analysis showed that qualitative data can be quantitatively evaluated by region, industry and period by collecting news from SMEs, and that there are concerns that it could be an element of alternative credit evaluation. **Conclusion:** News data cannot be collected even if one of the small businesses is self-employed or small businesses has little or no news coverage. Data normalization or standardization should be considered to overcome the difference in scores due to the amount of reference. Furthermore, since keyword sentiment analysis may have different results depending on the researcher's point of view, it is also necessary to consider deep learning sentiment analysis, which is conducted by sentence.

Keywords: Credit Rating, Alternative Credit Rating, COVID-19 Impact, Text Mining, Sentiment Analysis, Sentiment Index

JEL Classification Code: C52, C80, G30, M10, R11

1. Introduction

If you search credit rating on Google, you can define it as a rating that evaluates the credit value of a person who issued a certain type of debt, specifically a debt issued by a enterprise or government. The institutions that evaluate these ratings will compile and evaluate the information. Enterprises, just like individuals, will receive credit ratings based on various financial data. COVID-19 has changed the global economy from two years ago. COVID-19 strengthened the distancing and caused severe damage to small business owners. In particular, low-credit small business owners are in a dilemma because they have difficulty making loans in the financial sector. To help them, the Ministry of SMEs and Startups received an application for a "HopeLoan" and supported them. This is because many middle and low credit holders are excluded from the existing credit evaluation.

¹ First Author. Researcher, Hyosung ITX, Seoul, Korea. Email: sunnyfg@naver.com

² Ph.D. Candidate, Department of Business Administration, Ewha Womans University, Korea. Email: kimhj_07@naver.com

³ Corresponding Author. Associate Professor, Department of Chinese Business and Economics, Hannam University, Korea. Email: korea7659@hnu.kr

© Copyright: The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Under the current credit rating system, there is no way to evaluate middle and low credit holders. The alternative credit evaluation that can make up for this is the one that is emerging recently. Alternative credit evaluation is a method in which the credit rating of the borrower is determined according to the credit score of the credit rating agency (CB), and credit is evaluated based on detailed criteria including various non-financial information. For example, various non-financial information such as transaction details of financial institutions, lifestyle patterns, and public charge payment history can be used as credit criteria. Non-financial information includes account transfer records, payment details such as fees and membership fees. Among Big Tech enterprises, Internet banks such as Kakao Bank are expanding and promoting loans to middle and low credit by 2011 by applying the new CSS (Advanced credit rating system based on cross-strait ratings). Explosive data growth has enabled the emergence of new digital technologies such as data proliferation and machine learning algorithms to easily understand and evaluate thin-failer customers without financial history. Another reason is that the existing credit rating system alone has limitations that could overestimate borrowers in financial difficulties due to unexpected situations like COVID-19 or underestimate borrowers in recovery.

According to the IMF, the new CSS actually complemented the traditional credit rating model and improved credit accessibility for economic players. In addition, the accuracy concerns proved to be effective for more accurate credit rating, such as predicting default risk, and the accuracy of predicting default was higher when combining existing and alternative information using machine learning. According to the U.S. Consumer Financial Protection Bureau, the new CSS is also possible for a small number of people who have difficulty benefiting from some financial sectors due to income, race and asset disparities. Blacks and Hispanics tend to lack more financial power than whites, and many of them have been found to rely on high-interest loans in non-financial areas. Morgan Stanley also evaluated that the latest information can be utilized in terms of timeliness to complement the limitations of traditional credit ratings. Traditional credit ratings reflect relatively backward indicators, but alternative credit ratings show the current situation relatively accurately in that they process huge amounts of data in real time. Therefore, the credit evaluation of SMEs in the COVID-19 era needs to be improved by designing a new credit evaluation model in a new independent institution that reflects the environment like COVID-19. Therefore, this study calculates SME Score, which is a news base for SMEs through sentiment analysis, and presents alternative credit evaluation criteria for credit evaluation of SMEs.

The structure of this study is as follows: Following the introduction of Chapter 1, Chapter 2 summarizes the prior research, and Chapter 3 analyzes the company score index while explaining the analysis methodology. In Chapter 4, the results of empirical analysis and the results of sentiment analysis are derived, and finally, the conclusions about the analysis were presented in Chapter 5.

2. Literature Review

Sentiment analysis can be used in various fields as it can find words related to sensibility in a huge amount of atypical data such as news. In particular, it is used as a methodology for analyzing policy demand. The following is a summary of the prior studies related to this. Hong, Yoo and Ahn (2019) collected 42,496 comments from Google Regional Review and analyzed using the Multi-channel CNN model combining morphemes, syllables, and self-claims, showing that positive sentiment about Gamcheon Culture Village's urban regeneration policy are growing at 8:2, and that it can be used as a basic data to determine whether a government project has succeeded or not as a new big data analysis evaluation method. As one of the sentiment analysis methodologies, text mining is an analysis method that finds words related to sentiment in data, analyzes them, and derives meaning. Massive data can be analyzed and used in various fields. In particular, it is used as a methodology for analyzing policy demand. Text mining techniques can now analyze themes that were difficult to analyze in conventional methodologies (Kim, Lee, Shin, & Park, 2019; Ko, 2021). This technique is used in various learning and fields, but it has the advantage of being able to draw out flow, trends, and influence. Choi (2021) used text mining and topic modeling techniques to check research trends in the field of expression art treatment in Korea based on papers published from 1999 to 2020, and to analyze related academic papers and academic journals. Jeong (2021) analyzed consumers' perception of food distribution platforms using LDA topic modeling techniques, one of text mining, and found that the platform app reviews created by consumers complement the limitations of existing questionnaire techniques. Won and Hong (2021) predicted bitcoin prices on the Korea-U.S. exchange using ARIMA and circulatory neural networks, and analyzed the separated RNA model based on separate learning. It will also be used to gather opinions before and after the implementation of policies and to form a foundation for more effective and efficient policies. Lee (2019) analyzed the tone of 24,079

newspaper articles before and after 152 Monetary Committees from March 2005 to November 2017, and found that monetary policy surprises measured changes in the tone of the article. This suggests that monetary policy surprises measured using text mining reflects policy expectations and market impact well.

Next, according to a preliminary study using stereotyped data, Yoo (2018) calculated and analyzed total factor productivity from 2005 to 2016 among the top 1,200 R&D investments in 2016 and found that R&D investment has a lower impact on productivity. Kwak (2018, 2021) analyzed the impact of COVID-19 on corporate R&D scores through corporate financial accounting data and found that South Jeolla Province, Gwangju, Sejong, Incheon, South Gyeongsang Province, and North Chungcheong Province, which has a low proportion of manufacturing industry by region, were the most affected. While manufacturing and rental services, which account for a large portion of small business owners and self-employed businesses, showed a large decline and suggested the need for active government policy measures such as restructuring. Kang (2019) presented an ecosystem model STI scoreboard system STI capability monitoring at the national level, and suggested the need to continuously and systematically conduct monitoring index system research, index data collection, DB construction, and operation based on the importance of detecting abnormal signs through monitoring.

Finally, various studies have been actively conducted on the credit evaluation methods of enterprises. In particular, research to complement the problems with the credit evaluation methodology currently in place is a representative example. Park (2019) proposed a privacy credit evaluation method in which individuals calculate credit scores themselves and prove that credit scores have been successfully calculated with English knowledge and blockchain to solve the current credit evaluation system. Furthermore, in order to confirm whether the information used is actually the value provided by the financial institution through the blockchain, a zero-knowledge proof technique capable of efficiently proving committed inputs is presented. It was confirmed that the method was actually available by providing perfect zero-knowledge, fast verification process, and applying credit score algorithms similar to the actual environment. Kim (2022) optimized ensemble models for improvement research of corporate credit evaluation models and compared the discrimination power of each model to derive optimized models for each model. The empirical analysis confirmed that the error ensemble model had a limit of superior performance in some indicators compared to random forest and gradient boosting, but improved overall performance over logistics regression and could be used as a rating model. Research is also underway to complement the limitations of existing methodologies. Hong (2019) proposed an intelligent personal credit evaluation model based on deep learning that takes into account the lack of customer numbers, a characteristic of fintech businesses. In order to utilize Lending Club data to reflect the shortage of customer numbers, five of 10,000 datasets were randomly sampled, and insufficient bad-rated customer data were generated using SMOTE, GAN techniques to solve the imbalance data problem. Cheon (2021) proposed an artificial intelligence-based credit evaluation algorithm using various credit information data, and presented an algorithm to derive what characteristics of the data affected the result derivation. Expanding this, it was confirmed that measures to explain the results of changes in the results derived by artificial intelligence can be applied to financial data to provide the ability to explain when artificial intelligence is introduced into financial services. Park (2019) searched for human resource management and development factors that affect the credit status of the enterprise and applied them to build a model of corporate credit evaluation. As a result of the model construction, it was concluded that the response rate, which is an evaluation measure, decreases gradually from a good credit rating interval to a bad credit rating interval, and that the C statistic is 0.732, which is very high. However, the disadvantage is that the integrity of the survey respondents is essential for accurate credit rating measurement. There are many studies that utilize such diverse analysis methodologies, but it is also true that there are areas that cannot be analyzed yet. As mentioned above, there is no way to evaluate middle and low credit holders under the existing rating system. Therefore, in this study, we present alternative credit evaluation methods such as SMEs that process news and other atypical data in real time and utilize enterprise score index and sentiment analysis.

3. Methodology

3.1. Subjects of Analysis

The standard data of this study was based on 49,746 enterprises with continuous time series information from 2004 to 2020, provided by KoDATA (Korea Rating & Data). As shown in <Table 1>, large and medium-sized enterprises were removed from the entire list. Among the data considered to be SMEs, the remaining 18,9

10 enterprises that deleted 12 delisted companies were selected as reference data, except for 21,544 enterprises with overlapping enterprise names and business registration numbers.

Table 1 : SME selection process for reference data

Reference Data		Number of companies
'04~'Full list with 20 consecutive years of time series information(A)		49,746
Removal of Large and Medium Businesses(B)		9,280
SMEs(C=A-B)		40,466
Deduplication of company's name and business registration number(D)		21,544
delisted companies(E)		12
SMEs to be analyzed(F=C-D-E)		18,910

The sampling was done by stratification on regions and industries, but the method of “square proportional stratification extraction” was chosen because less than five minor industries, such as “mining” or “agriculture, forestry and fisheries,” were very unlikely to be extracted. First of all, out of 18,910 SMEs subjected to analysis, sampling was conducted by region and industry, and then “square proportional stratification extraction” was performed on it, and 2,432 were confirmed. The region is based on 17 cities and provinces, including Seoul, Gyeonggi, Incheon, Gangwon, Daejeon, South Chungcheong Province, North Chungcheong Province, Sejong, Gwangju, South Jeolla Province, North Jeolla Province, Daegu, Ulsan, Busan, South Gyeongsang Province, North Gyeongsang Province, and Jeju. The industry was classified based on construction, mining, wholesale and retail industries, especially in the case of Korean industries, the standard was extracted and integrated to prevent the concentration of some manufacturing sectors due to the high proportion of manufacturing industries.

The distribution of 18,910 SMEs subject to analysis by region and industry is shown in <Table 2>. About 50 percent of the industries were extracted from Seoul and Gyeonggi Province, and about half of the industries were extracted from manufacturing, while wholesale and retail, real estate, and construction accounted for about 10 percent, respectively. The regions and industries sampled are shown in <Table 3>.

Table 2 : Phase(Step) 1: Out of 18,910 SMEs subject to analyze, sampling by region and industry

Industries (in General Category)	Gang won	Gyeong gi	Gyeong nam	Gyeong buk	Gwang ju	Dae gu	Dae jeon	Bu san	Seo ul	Se jong	Ul san	In cheon	Jeon nam	Jeon buk	Jeju	Chung nam	Chung buk	Total
Construction	36	353	95	63	88	60	37	200	423	4	47	103	84	44	47	50	43	1,777
Mining	9	12	3	3	0	0	0	1	9	0	4	7	3	1	1	9	3	65
Educational Service	0	3	1	0	1	1	0	1	24	0	0	0	0	0	0	0	0	31
Agriculture, Forestry and Fisheries	4	22	6	15	0	1	0	16	6	2	0	0	10	2	5	15	6	110
Wholesale and Retail Business	22	607	76	26	40	73	37	246	1,089	3	25	134	21	20	19	38	41	2,517
Health and Social Welfare Services	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	3
Real Estate	38	587	83	38	84	104	63	146	516	46	34	118	34	21	45	48	28	2,033
Business Facilities Management/Business Support/ Rental Services	2	29	4	12	5	8	1	16	155	0	3	7	9	4	9	2	5	271
Water supply/sewage/ waste disposal/raw material regeneration	2	62	12	15	3	6	1	17	9	1	10	20	8	4	1	18	7	196
Accommodation and restaurant business	11	24	10	5	5	6	7	15	81	0	2	8	3	3	18	6	5	209
Arts, sports and leisure- related services	10	55	16	16	2	3	3	5	27	1	3	7	16	6	21	9	13	213
Transportation and Warehousing	9	121	36	24	7	14	10	136	179	2	21	46	28	7	16	27	10	693
Electricity, Gas, Steam and Air Conditioning Supply Businesses	7	16	6	3	5	3	0	2	6	0	3	2	16	8	1	2	6	86
Professional, Scientific and Technical Services	3	74	11	5	4	5	7	15	189	1	5	10	6	0	2	10	9	356
Information and	4	99	8	4	3	11	9	7	363	1	0	4	2	5	1	4	5	530

Communications																		
Manufacturing	106	3,268	1,027	632	168	404	140	645	925	54	248	714	165	187	19	592	411	9,705
Associations/Organizations, Repairs and Other Personal Services	0	26	6	6	10	8	3	11	26	0	0	6	2	4	3	2	2	115
Total	263	5,358	1,400	867	425	707	318	1,479	4,030	115	405	1,186	407	316	208	832	594	18,910

Table 3 : Phase(Step) 2 : Extract 2,432 out of 18,910 companies by using ‘square proportional stratification extraction’ method

Industries (in General Category)	Gangwon	Gyeonggi	Gyeongnam	Gyeongbuk	Gwangju	Daejeon	Daejeon	Busan	Seoul	Sejong	Ulsan	Incheon	Jeonnam	Jeonbuk	Jeju	Chungnam	Chungbuk	Total
Construction	6	19	10	8	9	8	6	14	21	2	7	10	9	7	7	7	7	156
Mining	3	3	2	2	0	0	0	1	3	0	2	3	2	1	1	3	2	27
Educational Service	0	2	1	0	1	1	0	1	5	0	0	0	0	0	0	0	0	11
Agriculture, Forestry and Fisheries	2	5	2	4	0	1	0	4	2	1	0	0	3	1	2	4	2	35
Wholesale and Retail Business	5	25	9	5	6	9	6	16	33	2	5	12	5	4	4	6	6	157
Health and Social Welfare Services	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2
Real Estate	6	24	9	6	9	10	8	12	23	7	6	11	6	5	7	7	5	161
Business Facilities Management/Business Support/Rental Services	1	5	2	3	2	3	1	4	12	0	2	3	3	2	3	1	2	51
Water supply/sewage/waste disposal/raw material regeneration	1	8	3	4	2	2	1	4	3	1	3	4	3	2	1	4	3	50
Accommodation and restaurant business	3	5	3	2	2	2	3	4	9	0	1	3	2	2	4	2	2	50
Arts, sports and leisure-related services	3	7	4	4	1	2	2	2	5	1	2	3	4	2	5	3	4	54
Transportation and Warehousing	3	11	6	5	3	4	3	12	13	1	5	7	5	3	4	5	3	93
Electricity, Gas, Steam and Air Conditioning Supply Businesses	3	4	2	2	2	2	0	1	2	0	2	1	4	3	1	1	2	33
Professional, Scientific and Technical Services	2	9	3	2	2	2	3	4	14	1	2	3	2	0	1	3	3	57
Information and Communications	2	10	3	2	2	3	3	3	19	1	0	2	1	2	1	2	2	58
Manufacturing	38	245	126	100	47	77	46	106	129	24	57	109	49	55	11	96	83	1,401
Associations/Organizations, Repairs and Other Personal Services	0	5	2	2	3	3	2	3	5	0	0	2	1	2	2	1	1	37
Total	79	387	189	152	92	129	83	191	301	41	93	173	100	91	54	148	129	2,432

3.2. Method of Analysis

The enterprise score index through atypical data is calculated through text purification process, sentiment analysis, and frequency analysis. First, the text analysis stage consists of five major stages. The first step was to collect articles, and we searched on Naver News for the names of enterprises on the actual list of SMEs. News searches collected approximately 11 years of data from December 2011 to December 2021 and categorized them according to the time of the COVID-19 outbreak.



Figure 1 : News article text crawling operation (Example)

The second step is the process of refining articles. The process of refining articles is to exclude articles that have no content, such as overlapping news articles, breaking news, and articles that have nothing to do with corporate management. We then proceeded with Word Cleaning (NLP), a process that excludes Natural Language Process and meaningless postpositional particle and indistinguishable words. Examples of results are shown in <Table 4>.

Table 4 : Article Refining (Example)

Keywords	Frequency	Processing
Choseon(조선)	3	-
Improvement(개선)	1	-
It(것)	1	Exclude
Management(경영)	1	-
Public Corporation(공사)	1	-
Discussion(논의)	1	-
Creditor Bank(채권은행)	1	Creditors
Month(월)	1	Exclude
Words(말)	1	Exclude
Justification(명분)	1	-
Yeouido(여의도)	1	Exclude

Third step is the Word Cleansing process (NLP). The data is collected by enterprise’s news by region and industry in a single code, and the total number of articles collected through this process is 23,306 as shown in <Table 5>. There are 19,457 cases, excluding URL deduplication and the same article content of data collected as shown in <Table 6>.

Table 5 : Current status of the number of articles collected through portal sites, etc.

Industries (in General Category)	Gang won	Gyeong gi	Gyeong nam	Gyeong buk	Gwang ju	Dae gu	Dae jeon	Bu san	Se oul	Se jong	Ul san	In cheon	Jeon nam	Jeon buk	Jeju	Chung nam	Chung buk	Total
Construction	79	98	123	95	107	56	89	159	180	32	72	132	165	94	57	60	71	1,669
Mining	18	0	19	3	0	0	0	0	7	0	8	11	10	13	0	21	1	111
Educational Service	0	1	0	0	0	0	0	0	114	0	0	0	0	0	0	0	0	115
Agriculture, Forestry and Fisheries	8	44	11	20	0	13	0	23	25	4	0	0	5	26	1	26	35	241
Wholesale and Retail Business	76	210	71	40	43	58	47	115	411	0	70	107	53	4	22	29	10	1,366
Health and Social Welfare Services	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	28
Real Estate	67	231	114	53	98	109	104	116	266	53	66	74	89	45	73	156	56	1,770
Business Facilities Management/Bus iness Support/ Rental Services	0	79	8	18	6	30	0	28	106	0	16	30	1	39	61	0	3	425
Waters supply/ sewage/waste	41	29	38	45	4	3	0	33	30	0	21	17	65	31	0	48	15	420

disposal/raw material regeneration																		
Accommodation and restaurant business	84	27	48	14	12	20	29	65	207	0	1	34	0	11	75	38	9	674
Arts, sports and leisure-related services	79	25	12	22	4	23	34	6	37	17	0	35	24	59	61	28	4	470
Transportation and Warehousing	47	51	28	34	11	10	17	132	120	20	57	78	98	33	30	39	12	817
Electricity, Gas, Steam and Air Conditioning Supply Businesses	4	53	13	26	12	12	0	30	32	0	55	18	15	28	0	0	29	327
Professional, Scientific and Technical Services	1	106	4	8	2	27	19	91	138	14	19	27	25	0	21	30	23	555
Information and Communications	73	61	29	16	4	22	64	20	183	3	0	35	10	47	0	7	36	610
Manufacturing	580	2,018	1,092	928	554	759	317	1,231	1,690	180	607	1,032	575	627	56	736	444	13,426
Associations/Organizations, Repairs and Other Personal Services	0	18	10	9	29	23	14	37	84	0	0	4	6	22	8	0	18	282
Total	1,157	3,051	1,620	1,331	886	1,165	734	2,086	3,658	323	992	1,634	1,141	1,079	465	1,218	766	23,306

Table 6 : URL deduplication of collected data (same article content excluded from analysis)

Industries (in General Category)	Gangwon	Gyeonggi	Gyeongnam	Gyeongbuk	Gwangju	Daejeon	Daejeon	Busan	Seoul	Sejong	Ulsan	Incheon	Jeonnam	Jeonbuk	Jeju	Chungnam	Chungbuk	Total
Construction	58	75	110	95	81	48	52	147	168	32	58	80	129	82	47	58	70	1,390
Mining	9	0	17	3	0	0	0	0	7	0	8	10	10	12	0	18	1	95
Educational Service	0	1	0	0	0	0	0	0	114	0	0	0	0	0	0	0	0	115
Agriculture, Forestry and Fisheries	5	39	6	20	0	13	0	16	20	3	0	0	5	26	1	20	23	197
Wholesale and Retail Business	61	174	57	40	30	50	39	105	340	0	47	56	53	4	20	29	8	1,113
Health and Social Welfare Services	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	25
Real Estate	51	168	93	53	91	107	99	104	208	44	47	74	69	45	65	156	56	1,530
Business Facilities Management/Business Support/Rental Services	0	42	8	18	6	30	0	20	92	0	16	29	1	39	45	0	3	349
Water supply/sewage/waste disposal/raw material regeneration	23	24	38	45	4	3	0	26	24	0	11	17	36	31	0	44	8	334
Accommodation and restaurant business	52	27	44	14	11	20	29	57	149	0	1	34	0	11	68	33	8	558
Arts, sports and leisure-related services	66	25	12	22	4	23	33	6	37	17	0	20	23	57	60	28	4	437
Transportation and Warehousing	31	29	27	31	10	10	17	112	96	9	45	70	97	21	29	34	10	678
Electricity, Gas, Steam and Air Conditioning Supply Businesses	4	46	11	26	12	8	0	30	32	0	31	15	7	27	0	0	29	278
Professional, Scientific and Technical Services	1	83	4	8	2	21	16	53	137	5	12	22	15	0	21	18	21	439
Information and Communications	73	61	29	16	4	22	61	17	178	3	0	35	10	44	0	6	36	595
Manufacturing	454	1,717	907	805	347	655	280	1,078	1,453	151	458	812	455	482	42	567	404	11,067
Associations/Organizations, Repairs and Other Personal Services	0	18	10	8	29	21	10	35	74	0	0	4	5	22	8	0	13	257
Total	888	2,529	1,373	1,204	631	1,031	636	1,806	3,154	264	734	1,278	915	903	406	1,011	694	19,457

The fourth is the creation of a sentiment dictionary on positivity and negativity. We made a list to create a positive/negative dictionary and conducted morphological analysis of news titles and articles. A list of keywords having positive and negative meanings among the extracted keywords was created by extracting only meaningful nouns (NNG/NP), verbs (VV), and adjectives (VA) as morphemes. For example, a dictionary including “contract”, “development”, “profit”, “certification”, “enhancement”, etc. is generated as a positive word, and “regulation”, “prosecution”, “damage”, “drop”, “loss”, etc. as an negative word. Among the extracted nouns, verbs, and adjectives, keywords that are meaningless in analysis are treated as unnecessary words and excluded.

정부 세계적 기업 300개 육성현재 230곳 낙점 ... 도 1곳 불과산경원 4곳 선정 역량강화 나서도내 기업의 글로벌 역량을 키우는 사업이 절실한 것으로 나타났다.		Re-enter
정부	_080002/NNG	정부_080002/NNG
세계적	/MMA	세계적/MMA
기업	_010000/NNG	기업_010000/NNG
300개	300/SN + 개_100001/NNB	300개_300/SN + 개_100001/NNB
육성현재	육성_040000/NNG + 현재_020001/NNG	육성현재_육성_040000/NNG + 현재_020001/NNG
230곳	230/SN + 곳_010002/NNG	230곳_230/SN + 곳_010002/NNG
낙점	_010001/NNG + .../SE	낙점_010001/NNG + .../SE
도	/JX	도/JX
1곳	1/SN + 곳_010002/NNG	1곳_1/SN + 곳_010002/NNG
불과산경원	불과산경원/NNG	불과산경원/NNG
4곳	4/SN + 곳_010002/NNG	4곳_4/SN + 곳_010002/NNG
선정	선정_070000/NNG	선정_070000/NNG
역량강화	역량_010000/NNG + 강화_040002/NNG	역량강화_역량_010000/NNG + 강화_040002/NNG
나서도내	나서_000102/VV + 어도/EC + 나_030100/NP + 의/JKG	나서도내_나서_000102/VV + 어도/EC + 나_030100/NP + 의/JKG
기업의	_010000/NNG + 의/JKG	기업의_010000/NNG + 의/JKG
글로벌	/NNG	글로벌/NNG
역량을	_010000/NNG + 를/JKO	역량을_010000/NNG + 를/JKO
키우는	_000001/VV + 는/ETM	키우는_000001/VV + 는/ETM
사업이	_040000/NNG + 이/JKS	사업이_040000/NNG + 이/JKS
절실한	_000002/VA + 는/ETM	절실한_000002/VA + 는/ETM
것으로	_010001/NNB + 으로/JKB	것으로_010001/NNB + 으로/JKB
나타났다.	나타나_000102/VV + 았/EP + 다/EF + ./SF	나타났다_나타나_000102/VV + 았/EP + 다/EF + ./SF

Figure 2 : Calculation of word-by-word similarity

Final step is the calculation of the sentiment index. Based on the data extracted through the first to fourth stages, the sentiment index is calculated through points. The sentiment index measurement is set to calculate positive keywords as +1 point and negative keywords at -1 point, and constructed to calculate before/after COVID-19 by article/enterprise unit, region and industry. Frequency analysis is performed around the extracted keywords, and frequency keywords by region and period can be compared through frequency analysis of the extracted keywords. It is also possible to perform positive and negative sentiment analysis among keywords, calculate the sentiment index, and analyze the impact of COVID-19.

4. Results

4.1. Frequency Analysis

As a result of keyword frequency analysis extracted from all articles in 17 cities/provinces and industries was followed by “enterprise” (12,501 cases), “representative” (6,970), and “industry” (6,386). When collecting data by enterprise’s name, there are many terms related to the enterprise. In order to understand the characteristics of each region, the analysis showed that “enterprise” (3,862 cases), “representative” (2,472) and “industry” (1,896) were high in the case of “metropolitan areas.” It was found that “enterprise” (8,639 cases), “representative” (4,498), and “industry” (4,490) appeared the same in “non-metropolitan area.” This means that there is not much difference between regions. Keywords in the “metropolitan area” and “non-metropolitan area” appeared in almost the same order, with “Seoul” being mentioned in the “metropolitan area” and “Busan” being mentioned in the “non-metropolitan area.” According to a detailed analysis of the metropolitan area as “Seoul”, “Gyeonggi”, and “Incheon,” “Seoul” had keywords mentioned in the order of “enterprise” (1,675 cases), “representative” (1,211), “Seoul” (984), “market” (859), and “business” (785). In “Gyeonggi,” it appeared in the order of “enterprise” (1,548 cases), “representative” (895), “industry” (831), “company” (706) and

“business” (675), while “Incheon” was followed by “enterprise” (639 cases), “Incheon” (470), “company” (454), “industry” (382), and “representative” (366).

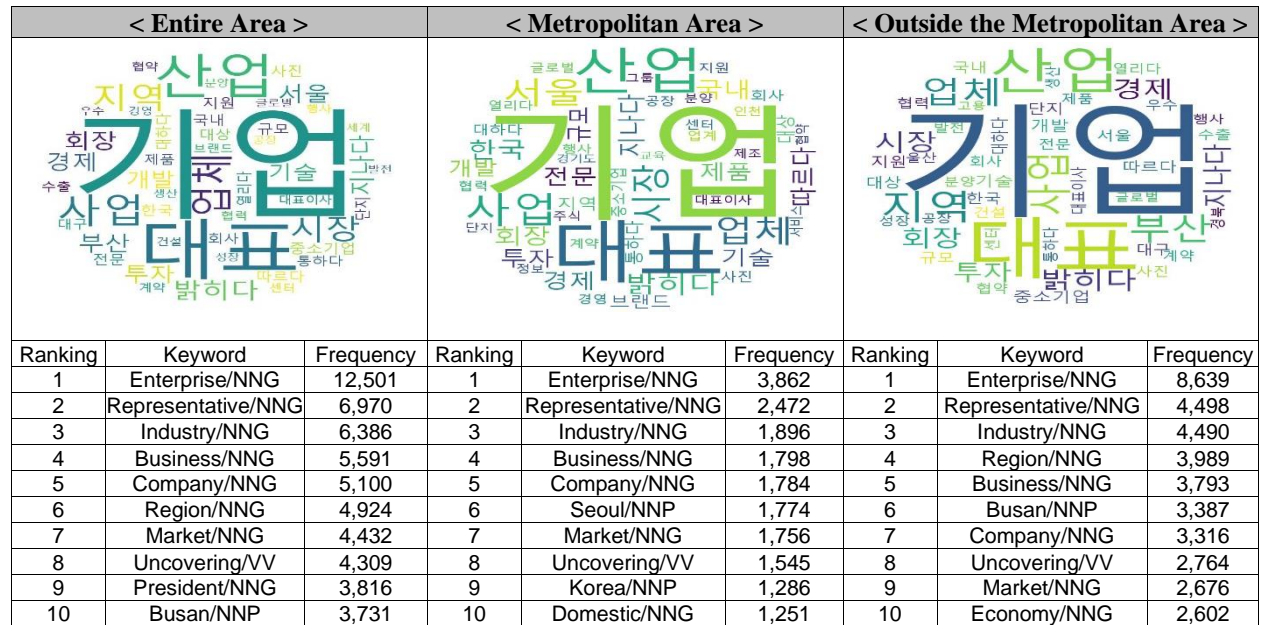
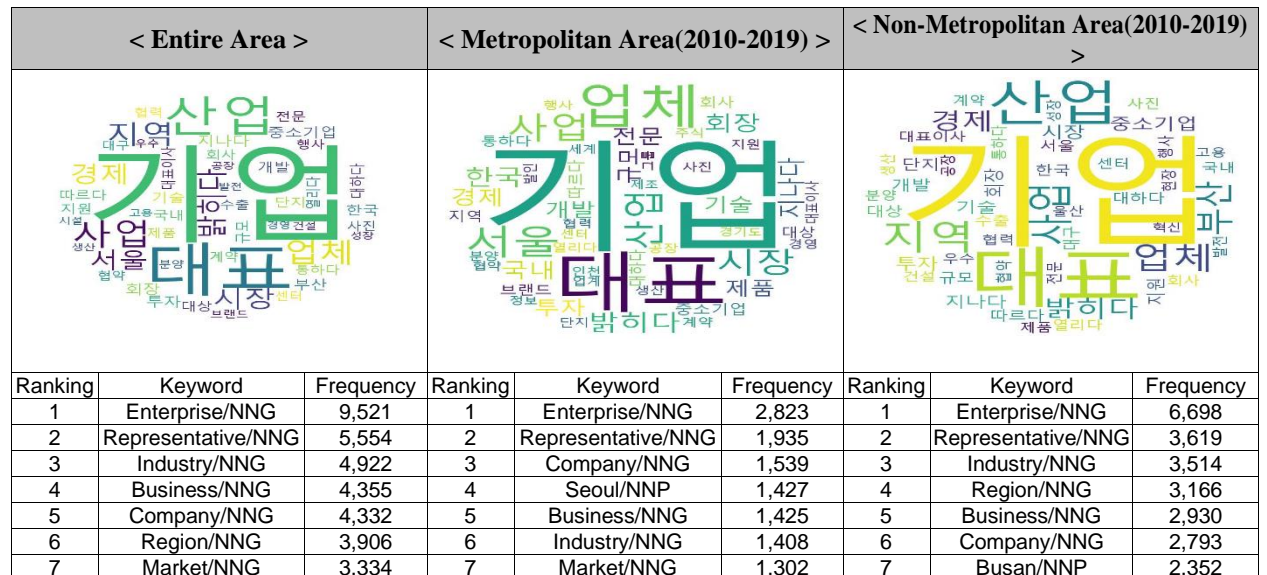


Figure 3 : Regional Frequency Analysis

The collection period is from 2010 to 2021, and based on 2020 when “COVID-19” hit the world, we looked at how keywords will change by dividing them into pre-“COVID-19” (2010-2019) and post-“COVID-19” (2020-2021). As a result, pre-“COVID-19”, keywords were mentioned in the order of “enterprise” (9,521 cases), “representative” (5,554), and “industry” (4,922). It is understood that the reason is that the number of data before “COVID-19”, which was longer than actual “COVID-19” era, is higher. In order to identify regional characteristics before “COVID-19” (2010-2019), the “metropolitan area” was followed by “enterprise” (2,823 cases), “representative” (1,935), “company” (1,539), “Seoul” (1,427) and “business” (1,425). In “non-metropolitan area,” keywords such as “enterprise” (6,698 cases), “representative” (3,619), “industry” (3,514), “region” (3,166) and “business” (2,930) are mentioned.



8	Uncovering/VV	3,303	8	Uncovering/VV	1,157	8	Uncovering/VV	2,146
9	Economy/NNG	3,039	9	Korea/NNP	1,060	9	Economy/NNG	2,107
10	Seoul/NNP	2,833	10	President/NNG	986	10	SME/NNG	2,052

Note: Category Period : ① (before COVID-19 outbreak) from January 1, 2010 to December 31, 2019, ② (after COVID-19 outbreak) from January 1, 2020 to November 30, 2021.

Figure 4 : Frequency Analysis before COVID-19 Outbreak

When comparing after “COVID-19”(2020-2021), the word “COVID”(1,279 cases) appeared, followed by “enterprise” (2,980 cases), “industry” (1,464) and “representative” (1,416). In order to understand the characteristics of each region, we analyzed the whole keyword separately, and according to the analysis, the metropolitan area had the order of “enterprise” (1,039 cases), “representative” (537), “industry” (488), “corona” (471), and “market” (454), while non-metropolitan area had “enterprise” (1,941 cases), “Busan” (1,035), “industry” (976), “representative” (879), and “business” (863). The keyword “follow” (299 cases) and “invest” (298) appeared in the metropolitan area and “region” (823 cases) appeared in the non-metropolitan area. The following is before and after COVID-19 being divided into the metropolitan area and non-metropolitan area. As a result of collecting news by enterprise name, keywords such as “enterprise”, “industry” and “representative” were most mentioned before and after COVID-19. While keywords mentioned only in the metropolitan area before COVID-19 were “Incheon,” “Gyeonggi Province,” and “manufacturing”, keywords that appeared only after COVID-19 were “services,” “education,” “smart,” “online,” and “games”, shows that the COVID-19 has affected our lives. Keywords mentioned only in “non-metropolitan area” before COVID-19 includes “Ulsan,” “excellent,” “employment,” and “growth”, while eco-relate keywords like “hydrogen,” “Jeju,” “U.S.” and “environment” stands out after COVID-19. We can see the difference by excluding the keywords that were commonly mentioned. Prior to COVID-19, manufacturing industries in Incheon and Gyeonggi Province were characterized in the metropolitan area, while employment and growth in Ulsan was the center of Korea’s SME industry in the non-metropolitan area. After COVID-19, however, services such as online, games, and education have become the center of the metropolitan area, and non-metropolitan area has become an eco-friendly industry with keywords like “hydrogen” and “environment” mentioned. This is where we can confirm what kind of industries will Korean SMEs invest in to achieve technological development.

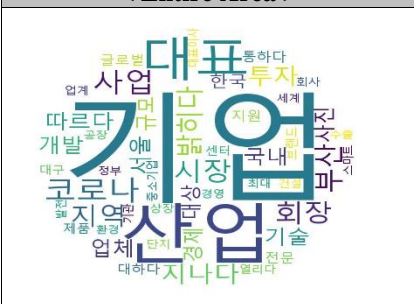

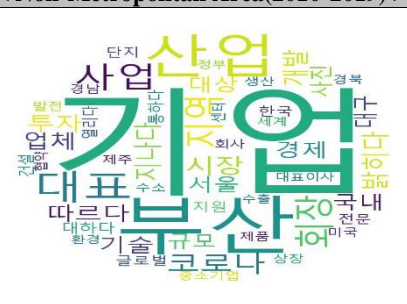
< Entire Area >			< Metropolitan Area(2010-2019) >			< Non-Metropolitan Area(2010-2019) >		
								
Ranking	Keyword	Frequency	Ranking	Keyword	Frequency	Ranking	Keyword	Frequency
1	Enterprise/NNG	2,980	1	Enterprise/NNG	1,039	1	Enterprise/NNG	1,941
2	Industry/NNG	1,464	2	Representative/NNG	537	2	Busan/NNP	1,035
3	Representative/NNG	1,416	3	Industry/NNG	488	3	Industry/NNG	976
4	COVID-19/NNG	1,279	4	COVID-19/NNG	471	4	Representative/NNG	879
5	Business/NNG	1,236	5	Market/NNG	454	5	Business/NNG	863
6	Busan/NNP	1,123	6	Uncovering/VV	388	6	President/NNG	824
7	Market/NNG	1,098	7	Business/NNG	373	7	Region/NNG	823
8	President/NNG	1,059	8	Seoul/NNP	347	8	COVID-19/NNG	808
9	Region/NNG	1,018	9	Follow/VV	299	9	Market/NNG	644
10	Uncovering/VV	1,006	10	Invest/NNG	298	10	Uncovering/VV	618

Figure 5 : Frequency Analysis after COVID-19 Outbreak

4.2. Sentiment Analysis

Sentiment analysis is an area of text mining analysis that is a way of speculating and categorizing feelings about affirmation and injustice in a particular document. Keywords are selected and analyzed, and even make predictions by creating deep learning models by learning on sentence. There are various methods for analyzing emotions, but in this study, we select keywords and calculate scores. This is interpreted differently from the positive and negative feelings in general sentences in the news articles of SME news articles. The keyword “certification” or “patent” by the government or the institution seems to have no sentiment, but enterprises have positive sensitivity on it. On the contrary, the word “court” or “prosecution” has a negative sensibility. More than 50 mentioned keywords were extracted from nouns (NNG/NNP), verbs (VV), and adjectives (VA) included in the analysis to create a sentiment dictionary of positivity/negativity. In the sentiment analysis, adjectives expressing sensibility were likely to be used the most, but in fact, nouns and verbs were more often expressed sensibility. 3,632 keywords were mentioned more than 50 times. Among them, 96 keywords were evaluated positively and 87 keywords were evaluated negatively. These were created as a list of sentiment dictionary. The table below shows the 30 most mentioned positive and negative keywords.

Table 7: Major Positive and Negative Keywords

Positive Keywords(top 30)			Negative Keywords(top 30)		
Keyword10	Keyword20	Keyword30	Keyword10	Keyword20	Keyword30
development/NNG	enhancement/NNG	happiness/NNG	sale/NNG	difficulty/NNG	prosecution/NNG
first/NNG	profit/NNG	be approved/VV	regulation/NNG	drop/NNG	illegal/NNG
new/VA	success/NNG	excellent enterprise/NNP	crisis/NNG	court/NNG	police/NNG
creation/NNG	reward/NNG	authorization/NNG	problem/NNG	suspicion/NNG	low/VA
improve/NNG	increase/VV	be built/VV	accident/NNG	situation/NNG	deficit/NNG
certification/NNG	superb/VA	gain/NNG	cost/NNG	decline/NNG	loss/NNG
good/VA	grow/VV	being stable/NNG	damage/NNG	depression/NNG	fall into/VV
develop/VV	benefit/NNG	new technology/NNG	decrease/VV	litigation/NNG	caution/NNG
spotlight/NNG	succeed/VV	lead the way/VV	concern/NNG	block/VV	penalty/NNG
contract/NNG	approval/NNG	enhance/VV	dispose/NNG	objection/NNG	leave/VV

Note: NNG(general noun), NNP(proper noun), VV(verb), VA(adjective).

The overall positive keywords are “development/NNG”, “first/NNG”, “new/VA”, “creation/NNG”, “improve/NNG”, and the negative keywords are “sale/NNG”, “regulation/NNG”, “crisis/NNG”, and “problem/NNG”. A positive Keyword is calculated as +1, a negative keyword as -1, and a sentiment index is calculated for each article. Therefore, the index is calculated by region and industry, and it can be identified by year to see what kind of time series characteristics it actually has. By region, we can understand the impact of COVID-19 on which industries were hit and which were benefited. Let's check the distribution of positive and negative keywords before checking the sentiment index. Sentiment analysis of the collected articles showed that the most positive keywords were “development” (1,535 cases), “first” (805), and “new” (779), while the most negative keywords were “sale” (519 cases), “regulation” (444), and “crisis” (419). In order to understand the characteristics of each region, the positive keywords for the metropolitan area were “development” (379 cases), “new” (296), and “first” (272), while the negative keywords were “sale” (206 cases), “problem” (154), and “accident” (144). In the “metropolitan area,” the keyword “enhancement” appeared higher than the “non-metropolitan area,” and more negative keywords were mentioned as “decrease”, “drop”, and “recall”. According to the title of the news which mentioned “recall,” was confirmed as “recall recommendations for 18 products including electric carpet and mats by the Korean Agency for Technology and Standards,” “recall’ up to 467 times more harmful substances in 23 products including school supplies, toys, and textiles.” “Harley-Davidson recalled 27 models due to the risk of ‘engine stopping while driving’,” “recall of 403,128 units of 38 models including Hyundai Motor, Maserati and Ford was carried out.” While keywords like “development” (1,156 cases), “creation” (542), and “first” (533) appeared positive in non-metropolitan area, “regulation” (338), “sale” (313), and “crisis” (295) appeared negative. The positive keywords that appeared high in the “non-metropolitan area” were “creation,” and the negative keywords were “damage,” “concern,” and “difficulty.” If you look at the title of the news where “creation” was mentioned, “Gyeonggi Province started to support job creation companies,” “KORAIL President Oh Young-sik led the creation of youth jobs at the railway station job fair,” “Venture companies have more job creation than the four major enterprises,” “Create value while

communication with consumers... Innovative management between companies and brands shines”, there are also references to value creation as well as job creation.

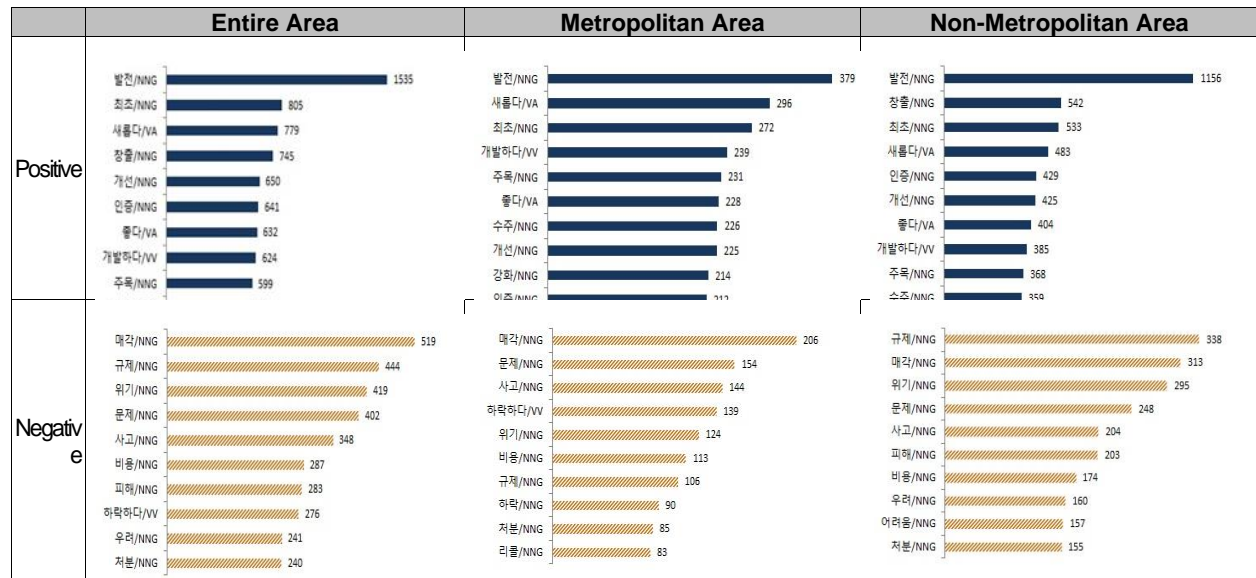


Figure 6: Sentiment Analysis of Keywords by Region

According to a sentiment analysis of news articles before the outbreak of “COVID-19”, “development” (1,168 cases), “creation” (653), and “first” (642) appeared as positive keywords, while “sale” (399 cases), “regulation” (316), and “problem” (279). Before “COVID-19”, if “creation”, “contract” and “profit” were mentioned as positive keywords, then “spotlight”, “increase” and “enhancement” are high. According to sentiment analysis of keywords after “COVID-19”, “development” (367 cases), “spotlight” (187 cases) and “new” (186) appeared as positive keywords, while negative keywords were “crisis” (178 cases), “situation” (130), and “regulation” (128). After “COVID-19”, it is noticeable that the most negative keyword is “crisis.” Before “COVID-19”, “accident,” “cost,” and “dispose” were mentioned as negative keywords, while after “COVID-19”, keywords such as “situation,” “difficulty,” “concern,” and “suspicion” appeared high. In the case of enterprises, before “COVID-19” there were economic issues such as “profit” and “cost”, while after “COVID-19” psychological factors, such as “difficulty”, “concern”, and “suspicion”, have been significant.

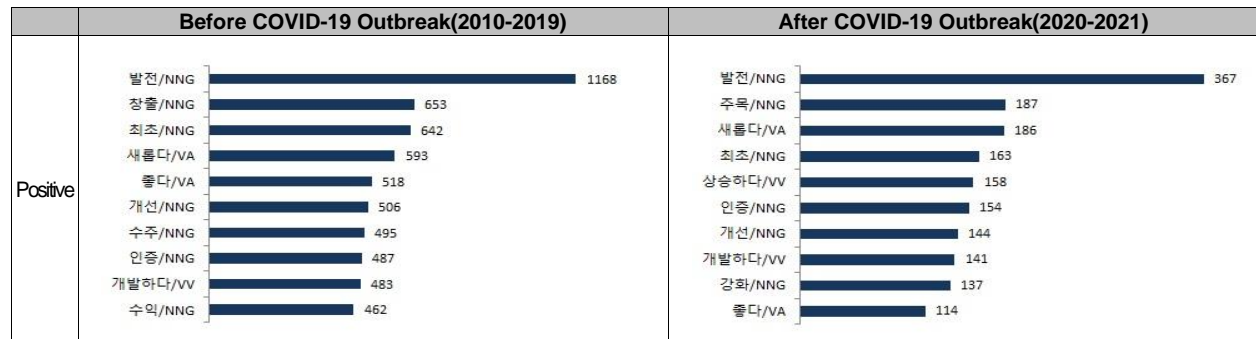




Figure 7: Sentiment Analysis of before and after the Outbreak of COVID-19 (2010-2019)

4.3. Index Calculation of the SME Score(SME News Sentiment Index) based on SME News

Although policies and systems for SMEs have been greatly improved, research is needed to develop effective policies by country, region, and industry. In order to establish effective development and support policies, related statistical data are essential, but the reality is that there is a huge shortage of data available in real-time use. Recently, research on language processing has increased, and the development of NLP (Natural Language Process) technology for processing text can be utilized by searching, collecting, and processing news from SMEs. There is advantage of being able to reflect trends depending on the characteristics of the industry and to check the contents according to the industry group by region. While frequency analysis and sentiment analysis previously examined can examine characteristics of each region and industry focused on keywords, quantitative and qualitative data are all available for exponential calculation when sentiment analysis is calculated by quantitative data, combined with existing evaluation model items. Sentiment analysis gave positive keywords +1, negative keywords -1 and calculated by quantitative scores of collected articles. This is called SME Score (SME News Sentiment Index) based on SME News. The advantage is that it is not only possible to disclose the simple index, but also qualitative analysis that can confirm the cause of the index through morphological analysis, one of the natural language processing techniques that can identify the main drivers of the index. For example, if the index falls sharply compared to the previous year, we can see what keywords are mentioned and what keywords have decreased compared to the previous year. Keywords can also be used to understand which industries are attracting attention and which industries will turn into declining industries. The article collected using sentiment score as SME news scores is represented by SME Score based on SME news, and the calculation procedure is summarized as shown in <Table 8>.

Table 8: SME Score calculation procedure based on SME news

STEP 1	(Collect news articles) Collect news articles related to SME by enterprises from web portal
↓	
STEP 2	(Sentiment score) Give ±1 to the positive/negative keywords of the article, and calculate the sentiment index
↓	
STEP 3	(Index calculation) Derive index by averaging sentiment index by period

As shown in <Figure 7>, since the outbreak of COVID-19 by region in 17 cities and provinces, it has been showing an average upward trend in 2021. Regardless of the impact of COVID-19, Seoul showed a higher positive value after 2019 and continued to maintain at the 2020 level in 2021. However, Jeju and Jeollanam-do were found to have deteriorated in two regions as many illegal keywords were mentioned based on sound values. Generally speaking, 2021 shows that it is recovering compared to 2020.

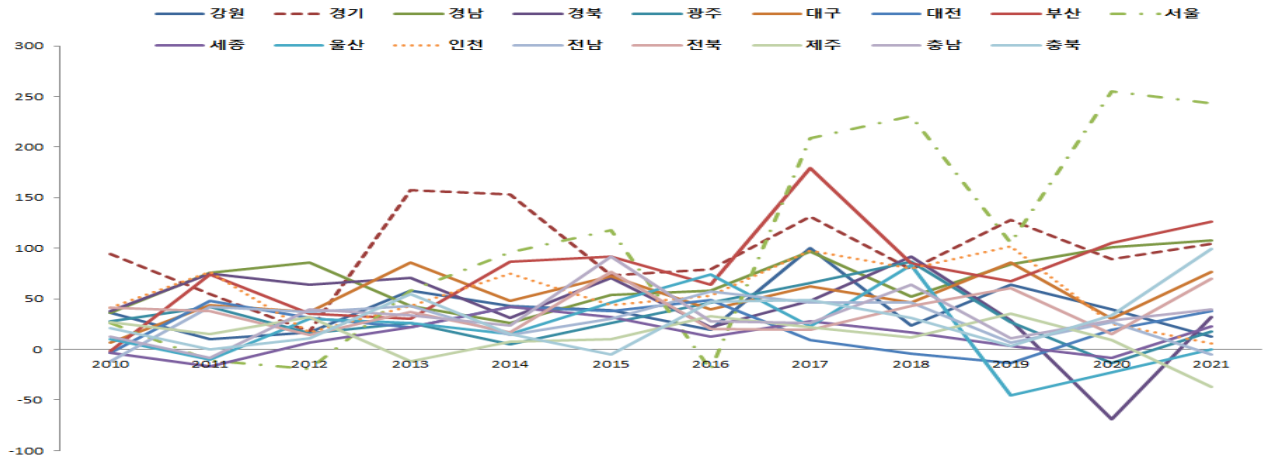


Figure 8: SME Score index based on SME News by local government

If you categorize Seoul, Gyeonggi, and Incheon into “metropolitan areas,” and other regions into “non-metropolitan areas,” the “metropolitan area” showed poor economic conditions in 2012 and 2016, but not much difference since 2017. Rather, it can be confirmed that COVID-19 has come by without knowing it. On the other hand, “non-metropolitan areas” showed low scores in 2020, but are showing signs of recovery in 2021. If the metropolitan area is subdivided by region, Seoul will not show any decline in 2020 unlike other regions. On the other hand, the index of Gyeonggi and Incheon will fall due to COVID-19 while the index of 2020 will rise higher than the previous year. Incheon showed a larger decline than Gyeonggi, and did not turn upward in 2021.



Figure 9: SME Score based on News of SMEs in Metropolitan Area

By industry, the manufacturing industry has a wide variety of industries and a large collection of articles, showing high sentiment scores. The manufacturing industry is on the rise in 2020 and 2021, but the transportation and warehousing industries are on the decline in sentiment scores in 2020 and 2021. This means that the transportation and warehousing industries have not yet recovered. Looking closely at the four industries that have been affected by COVID-19 in addition to manufacturing, The “construction industry” did not show much difference from the previous year in 2020, but decreased in 2021. Education services, wholesale and retail industries, and real estate industries decreased in 2020, but recovered to the 2019 level in 2021.

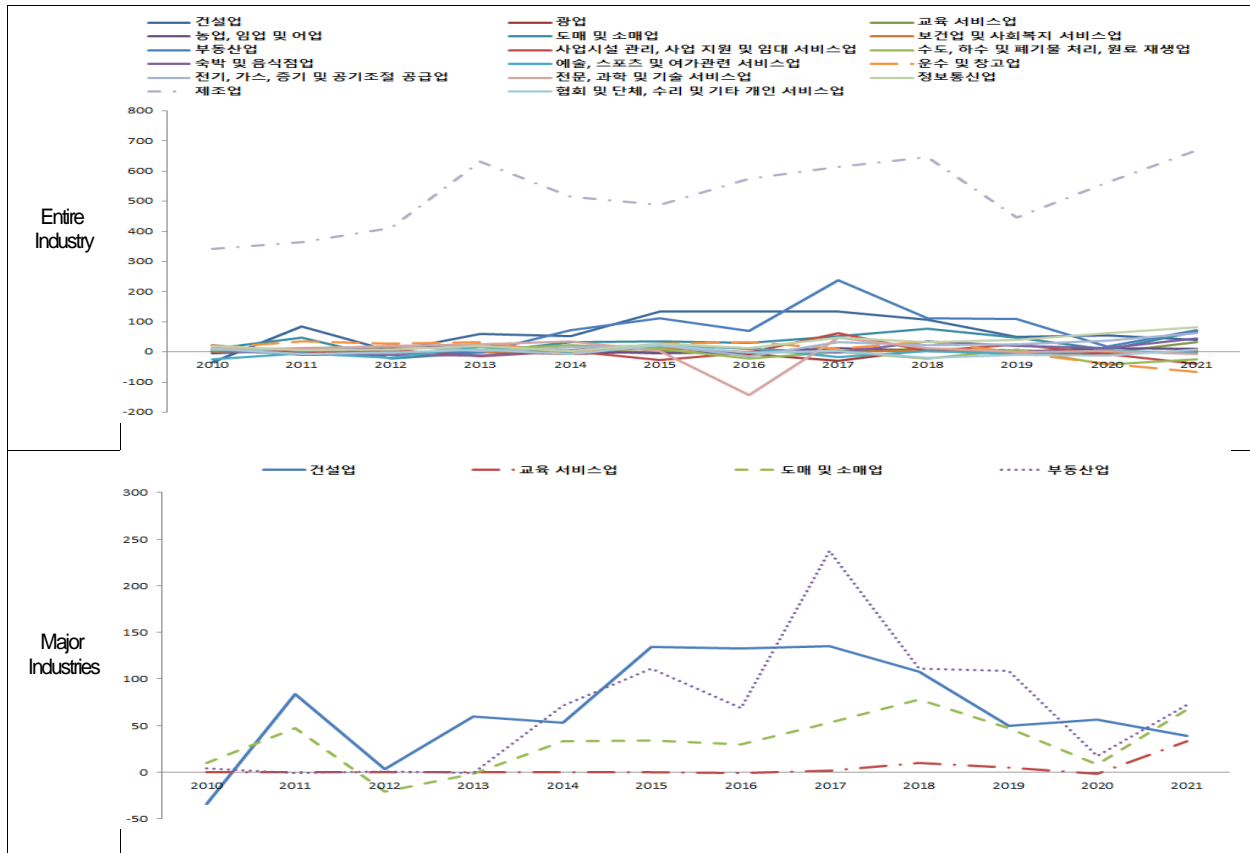


Figure 10: SME Score based on SME news by industry

5. Conclusion

This study shows that SMEs can collect news and quantify qualitative data by region, industry and period. There were concerns that it could be an element of alternative credit evaluation, but it was confirmed that it could be applied by dispelling it.

Therefore, instead of being excluded from the assessment on the grounds that materials and data are insufficient, we hope that the assessment will serve as a supplement to the missing data. Of course, there is a limit to the fact that news data cannot be collected even if there is a small amount of reference in the case of a small business or a small business that is single-person self-employment. The size of the company and the amount of news references are proportional, and the larger the number of news articles, the phenomenon of rich get richer and the poor get more poor appears, so that data normalization or standardization should be considered. Furthermore, since keyword sentiment analysis may differ depending on the researcher's point of view, it is also necessary to consider deep learning sentiment analysis, which is conducted by sentence. You can choose the advantages of the two methods and use deep learning and pre-analyzing at the same time, or you can apply weights by enterprise size or industry.

Through this study, SME Score (SME News Sentiment Index) based on SME news expects to provide the necessary information in a timely and reliable manner.

References

- Becker, G. S. (1993). Nobel Lecture: The Economic Way of Looking at Behavior. *Journal of Political Economy*, 101(3), 385-409.
- Cee. (2006). *Financial Fitness for Life-High School Test Examiner's Manual*, New York, NY.
- Cee. (2013). *National Standards for Financial Literacy*, <http://www.councilforeconed.org>.
- Chen, H. & R. P. Volpe. (1998). An Analysis of Personal Financial Literacy among College Students. *Financial Services Review*, 7(2), 107-128.
- Cheon, Y. E., Kim S. B., Lee, J. Y., & Woo, J. H. (2021). *Research on rating models using explainable AI technology*.
- Choi, E. J. (2021). Research Trends Analysis in Expression Art Therapy Field Using Text Mining and Topic Modeling (1999-2020). *Art Psychotherapy Research*, 17(2), 375-402.
- Hahn, J. & K. Jang. (2012). The Effects of a Translation Bias on the Scores for the Basic Economics Test. *Korean Journal of Economic Education*, 43(2), 133-148.
- Hahn, J. S. (2013). Financial Literacy of Korean Elementary School Students: Levels and Determinants. *Korean Journal of Economic Education*, 20(2), 39-63.
- Hong, S. G., Yoo, S. W., & Ahn, S. J. (2019). Sentimental Analysis on Urban Revitalization Policy: Focusing on Gamcheon Culture Village's Visitor Reviews.
- Hong, T. H., Kim, S. H., & Kim, E. M. (2019). Deep Learning-Based Intelligent Personal Credit Evaluation Model Using GAN and DNN.
- Jang, K., Hahn, K., & Kim, K. (2010). Comparative Korean Results of TUCE with U.S. and Japan. *Comparative Studies on Economic Education in Asia-Pacific Region*, Japan: Shumpusha Publishing, 53-77.
- Jumpstart. (2010). *2008 Survey with Results – High School*, <http://rijumpstart.org>.
- Jumpstart. (2006). *National Standards in K-12 Personal Financial Education*, <http://rijumpstart.org>.
- Jumpstart Coalition for Personal Finance. (2008). Financial Literacy Still Declining Financial Behavior, *Journal of Financial Counseling and Planning*, 20(1).
- Jumpstart Coalition for Personal Financial Literacy. (2006/2008). *The Financial Literacy of Young American Adults*.
- Jung, M. K., Kwon, J. Y., Lee, J. W., Lee, Y. A., & Lee, S. B. (2021). Consumer Awareness Analysis of Food Distribution Platforms Using Text Mining: Focusing on Topic Modeling Techniques. *Study on Restaurant Management*, 24(7), 71-100.
- Jung, Y. S., Cho, Y. D., & Park, H. J. (2013). The Factors Influence the Rational Financial Behaviors of Korean Adolescents. *Theory and Research in Citizenship Education*, 45(3), 201-227.
- Kim, S. H., Lee, Y. J., Shin, J. Y., & Park, K. Y. (2019). *Text Mining for Macroeconomic Analysis*.
- Kim, S. K. (2008). International Comparison of University Students' Knowledge of Economics: Korea, the United States and Japan. *Korean Journal of Economic Education*, 15(2), 33-61.
- Kim, K., and K. Jang. (2013). International Comparison of the Economic Knowledge from the Result of the Test of Economic Knowledge in Korea. *Korean Journal of Economic Education*, 20(2), 109-133.
- Ko, H. J. (2021). Analyzing the topic of supply chain risk management using text mining. *Electronic Trade Research* 19(3), 65-83.
- Kim, Y. H., Kim, D. H., Heo, J. H., & Kim, K. Y. (2022). *Comparative study of corporate credit evaluation models using error ensemble models*.
- Lee, S. H., Choi, J., Kim, J. W. (2016). Sentiment analysis on movie review through building modified sentiment dictionary by movie genre, *J Intell Inform Syst*, 22(2): 97-113.
- Lee, Y. J., Kim, S. H., & Park, K. Y. (2019). *A surprise monetary policy measured by text mining*.
- Oh, Y. S. (2005). Financial Quotient and Financial Education Contents for High School Students in Daegu. *Secondary Education Research*, 53(1), 283-300.
- OECD INFF. (2010). *Financial Literacy Measurement Questions and Socio-demographics*.
- OECD. (2012). Economics Assessment Framework. *Assessment of Higher Education Learning Outcomes*, 6-18.
- OECD. (2014). *PISA 2012 Financial Literacy*, <http://www.oecd.org>.
- Oh, Y. S. & Park, S. E. (2012). A Comparative Analysis between Korea and U.S. High School Students' Economic Literacy. *Korean Journal of Economic Education*, 19(2), 139-162.
- Park, C., Kim, J. H., & Lee, D. H. (2019). *Privacy Assessment Method Using English Knowledge Certification*.
- Park, J. W. (2019). *Research on the Application of Corporate Credit Assessment of Human Resource Management and Development Variables: Using Human Capital Enterprise Panel Materials*.
- Park, M. L. (2002). The Results of Test of Economic Literacy of High School Students. *Korean Journal of Economic Education*, 9, 33-61.
- Peng, Tzu-Chin Martina, Suzanne Bartholomae, Jonathan J. Fox, Garrett Cravener. (2007). The Impact of Personal Finance Education Delivered in High School and College Courses. *Journal of Family Economic Issues*, 28, 265-284.
- Won, J. K., & Hong, T. H. (2021). *Virtual Currency Price Forecast Using Text Mining and Deep Learning: Comparison of Korean and U.S. Markets*.