JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 1: 1~7, Jan. 2021

코로나-19관련 웨이보 정서 분석을 통한 중국 주식시장의 주판 및 차스닥의 민감도 예측 기법

이가기[1] · 오하영[2*]

Sensitivity of abacus and Chasdaq in the Chinese stock market through analysis of Weibo sentiment related to Corona-19

Jiaqi Li[1] · Hayoung Oh[2*]

[1]Graduate Student, Department of Business Administration, Sungkyunkwan University, Seoul, 03063 Korea
[2*]Associative Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

요 약

최근 코로나 19발생과 동시에 소셜 미디어의 투자자 정서가 증시 가격 움직임을 주도해 관심을 모으고 있다. 본 연구는 행동금융 이론 기반 빅 데이터 분석을 활용하여 소셜 미디어에서 추출한 정서가 중국 증시의 실시간 및 단기적 가격 모멘텀을 예측하는데 활용될 수 있는 기법을 제안한다. 이를 위해, COVID-19와 관련 200만 건 이상의 시나 웨이보 빅 데이터를 키워드 방식으로 수집 및 분석하고 시간이 따른 영향력이 높은 감정 요인을 추출한다. 최종 결과 도출을 위해 다양한 지도 및 비지도 학습 모델을 다 각도에서 구현 및 성능평가를 비교 분석 후, BiLSTM mdoel이 최적의 결과를 낼 수 있음을 증명했다. 또한, 제안하는 기법을 통해 주가변동과 심리요인 간에도 비슷한 움직임을 보이고 있음을 제안했고 소셜미디어에서 추출한 공공분위기가 어느 정도 투자자들의 심리를 대변할 수 있고, 주식시장에 영향을 미칠 수 있는 특수행사에 몰두할 때 증시변동에 차이를 만들 수 있음을 증명했다.

ABSTRACT

Investor mood from social media is gaining increasing attention for leading a price movement in stock market. Based on the behavioral finance theory, this study argues that sentiment extracted from social media using big data technique can predict a real-time (short-run) price momentum in Chinese stock market. Collecting Sina Weibo posts that related to COVID-19 using keyword method, a daily influential weighted sentiment factors is extracted from the sizable raw data of over 2 millions of posts. We examine one supervised and 4 unsupervised sentiment analysis model, and use the best performed word-frequency and BiLSTM mdoel. The test result shows a similar movement between stock price change and sentiment factor. It indicates that public mood extracted from social media can in some extent represent the investors' sentiment and make a difference in stock market fluctuation when people are concentrating on a special events that can cause effect on the stock market.

키워드 : 소셜미디어, 투자자 정서, 텍스트 마이닝, 중국 주식시장

Keywords : Social media, Investor sentiment, Text mining, Chinese stock market

Received 2 October 2020, **Revised** 21 October 2020, **Accepted** 6 November 2020

*** Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com Tel:+82-2-583-8585)
Associative Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.1.1 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. Introduction

With the explosion of social media users in recent year, online forums, social network, and financial news websites become a substitution for the traditional media to acquire financial information and a widely used channel to express their opinions on financial markets [1]. According to the efficient market hypothesis, financial market movements are dependent on news and external events that have a significant impact on the market value of companies[2]. However, in the era of big data, traditional forms of information like dividend signaling, financial reports, and analysis report from famous analysts are not enough for catching timely and price-effective “news” to make extra profits on stocks. As reported by "Report of Investigation of Individual Investors Condition in 2017" by SHENZHEN STOCK EXCHANGE[3], social media that can be used on smart-phone occupied the top portion (47.6%) in all the information channels in China. Thus, the technique called sentiment analysis (SA), one of the many applications that have arisen to analyze opinions, sentiments, and emotions present in unstructured data on social media, leading many papers to address the influence of investors’ mood.

However, the profits of individual investors were generally lower than institutional investors even though they account for 99.76% of all the investors. As survey of stock market investors condition report in 2019 revealed[4], among the interviewees, 91.4% of professional institutional investors, 68.9% of general institutional investors, and 55.2% of individual investors made profits from their investments. A large gap of profitability is existed between institutional investors and individual investors. As referred by [5], it is broadly recognized that Chinese investors’ ability to evaluate the true value of listed companies is limited. Chinese market is particularly suitable for the research of the impact of social media sentiment on stock price movement.

In the first quarter financial report of Sina Weibo in 2020, the largest social platform similar to Twitter in China, the number of monthly active users was up to 550 million as reported by China Daily[6], nearly half of the population in China. In this context, the instant news related to financial market from influential posts in Sina Weibo can quickly spread through the whole society, where potential investors and analysts can easily catch and utilize them in investment decision, which following cause a fluctuation on the price of stocks. The empirical result of sentiment analysis based on Twitter [7] shows that social media sentiment can lead financial returns. They found that social media sentiment in a broad-based system such as Twitter is indicative of future market movements in some range of assets, in other words, in some specific stock market segmentation.

Chinese individual investors performed differently in regard to their blankets from main-board or growth enterprises market (GEM). Investors in GEM were reported to utilize more numbers of information channels than those who were in non-GEM, moreover, their usage rate of professional financial reports, institutional analysis reports, technical indexes analysis reports was 10% higher than others, as reported by “Report of Investigation of Individual Investors Condition in 2017”[3]. We choose to consider all the big-4 stock indexes that include both main-borad and GEM market for stock price fluctuation analysis in our study.

[8] showed that sentiment data can be used to predict the stock price only when the stock has high investor attention. Thus, this article is an event study based on the outbreak period of COVID-19 in China, which contains dramatic fluctuation both on investor’s sentiment and stock price.

Ⅱ. Data and Methodology

To ensure a reliable emotion extraction result from analysis, completeness and effectiveness are the two key factors for designing processes of data collection, processing, and sentiment extraction.

Data processing model is showed as Fig. 1 with screening processes, sentiment extraction, and sentiment factors building for each day in the sample period.

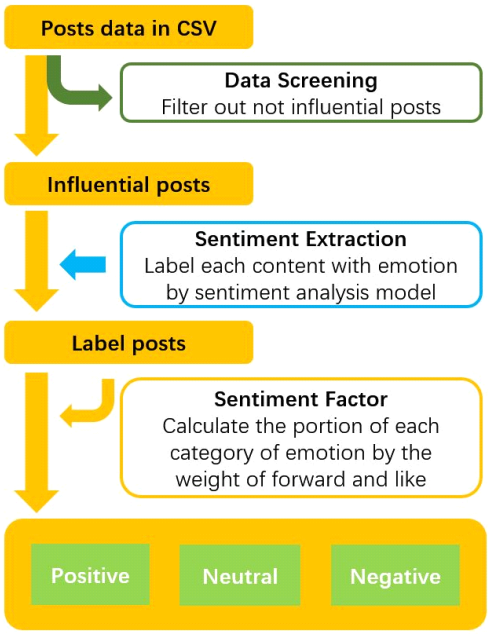
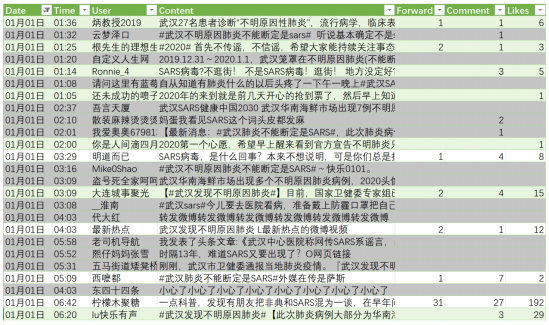


Fig. 1 Data Processing Model

2.1. Sina Weibo Posts

Considering the most severe outbreak period of COVID-19 in China, posts were crawled in daily frequency from January, 1st, 2020 to March 31st, 2020. In the early stage of outbreak of "COVID-19", people were still not knowing what exactly the virus was, the similar symptoms reported by news educated them that it was a virus that similar to SARS, therefore, sars is the keyword for referring COVID-19 in the first time period of data collection. Similarly, majority of COVID-19 patients were suffered from pneumonia, making people prefer to say COVID-19 as new pneumonia virus. Therefore, three COVID-19 related keywords ("sars", "新冠肺炎"→ pneumonia virus, "新型冠状病毒"→ corona virus) were selected for searching the event related posts. To avoid repeat collection of posts which includes two or three keywords, same posts will be deleted to get a clean dataset. From the crawling data, we got 2,315,971 posts from the three-month period, and after cleaning processes of repetitive posts, 2,221,539 posts were effective for the following screening processes.

There will be 6 kinds of information be collected in the crawling process: date of the post, poster name (user name), specific time of the post, content of the post, number of forward, number of comment, number of like. (Fig. 2 & 3)



Date	Time	User	Content	Forward	Comment	Likes
01月01日	01:36	炀教授2019	武汉27名患者诊断"不明原因性肺炎"，流行病学，临床表	1	1	6
01月01日	01:32	云梦泽口	#武汉不明原因肺炎不能断定是sars# 听说基本确定不是		1	
01月01日	01:25	根先生的理想生	#2020# 首先不传谣，不信谣，希望大家能持续关注事态	2	1	3
01月01日	01:20	自定义人生网	2019.12.31～2020.1.1，武汉发现在不明原因肺炎(不能断			
01月01日	01:14	Ronnie_4	SARS病毒?不是街！ 不是SARS病毒！ 是街！ 地方没定好		3	5
01月01日	01:08	请问这里有蓝莓自从知道有肺炎什么的以后头疼了一下午一晚上#武汉SA				
01月01日	01:05	还未成功的喂子	2020年的来到就是前几天开心的抢到票了，然后早上知道			1
01月01日	02:37	吾言天厦	武汉SARS健康中国2030 武汉华南海鲜市场出现7例不明肺			
01月01日	02:10	散装麻辣烫烫烫奶蛋我看见SARS这个词头皮都发麻			2	
01月01日	02:01	我爱奥奥67981	【最新消息：#武汉肺炎不能断定是SARS#，此次肺炎病		1	
01月01日	02:00	你是人间逃四月	2020第一个心愿，希望早上醒来看到官方宣告不明肺炎			1
01月01日	03:29	明道而已	SARS病毒，是什么回事？本来不想说明，可是你们总是	1	4	8
01月01日	03:16	MikeGShao	#武汉不明原因肺炎不能断定是SARS#～快乐0101。			
01月01日	03:09	盗号死全家呵呵	武汉华南海鲜市场出现多个不明原因肺炎病例，2020头			
01月01日	03:09	大连城事聚光	【#武汉发现不明原因肺炎#】目前，国家卫健委专家组	2	4	15
01月01日	03:08	_淮南	#武汉sars#今儿要去医院看病，准备戴上防霾口罩把自己			
01月01日	04:03	代大红	转发微博转发微博转发微博转发微博转发微博			
01月01日	04:03	最新热点	武汉发现不明原因肺炎 L最新热点的微博视频	2	1	12
01月01日	05:58	老司机导航	我发表了头条文章:《武汉中心医院称网传SARS系谣言，			
01月01日	05:52	照仔妈妈张雪	时隔13年，难道SARS又要出现了？O网页链接			
01月01日	05:31	五马街退矮凳机	刚刚，武汉市卫健委通报当地肺炎疫情。」武汉发现不明			
01月01日	05:09	西嘟都	#武汉肺炎不能断定是SARS#外媒在传是萨斯	1	7	2
01月01日	04:03	东四十四条	小心了小心了小心了小心了小心了小心了小心了			
01月01日	06:42	柠檬木聚糖	一点科普，发现有朋友把非典和SARS混为一谈，在早年	31	27	192
01月01日	06:20	lu快乐有声	#武汉发现不明原因肺炎#【此次肺炎病例大部分为华南		3	29

Fig. 2 Crawling Data

Fig. 3 Elements in a Sina Weibo

2.2. Data Screening

Since the crawling data from Sina Weibo posts with one keyword can up to about 12 thousands a day, some of them are not influential reflected by the number of forward, comment, and like. Here, posts without any number in both forward, and like will be filtered out from the raw posts data. (As show in grey in Fig. 3) Leaving posts that at least have one number in forward, or like which indicates their influence in Sina Weibo network. And this process will be accomplished in

calculating the weight of each post in a day. (As show in not-grey in Fig. 3)

2.3. Sentiment Extraction

In order to get the most accurate classification results, four models were tested with a manually formed training set from Sina Weibo. This training set is consisted of 6,000 of label posts sampled from the collected dataset as referred. There are 3 categories of sentiment in the train set with label 0 (null), 1 (positive), 2 (negative) of 2,000 respectively. Null means neutral sentiment of the posts related to COVID-19, positive means an overall positive attitude toward COVID-19, negative means an overall negative attitude toward COVID-19.

2.3.1. Model-1 Supervised Emotion Dictionary

Supervised machine learning is one of the simple algorithm in text classification. It utilized a dictionary with categories of words labeled in one kind of emotion. Here, Chinese emotional vocabulary ontology library developed by Dalian University of Technology [9] was used as the emotion dictionary. It was widely used by text-mining projects in Chinese because it contained 27,466 emotional words in 7 categories: anger (388), disgust (10,282), fear (1,179), sadness (2,314), surprise (228), good (11,107), happy (1,967).

From the result of training process applied to our training set, this algorithm performed weakly in classifying the sentiment inside each post. It only got an overall accuracy of 0.44, especially poor performance in predicting neutral sentiment with only recall of 0.27 and F1 score of 0.31. (Fig. 4) This is consistent with the view of [10][11] that short text like the posts from social media can have semantic sparseness problem. In order words, the complexity of the structure of text will confuse machine learning seriously because of the inverse meanings in different parts of one posts.

model	score	sentiment			Accuracy
		null	positive	negative	
Dalian dictionary	precision	0.38	0.43	0.51	0.44
	recall	0.27	0.59	0.47	
	f1-score	0.31	0.50	0.67	
Word frequency + LSTM	precision	0.64	0.75	0.76	0.68
	recall	0.65	0.66	0.71	
	f1-score	0.64	0.70	0.75	
Word frequency+ BiLSTM	precision	0.65	0.75	0.74	0.72
	recall	0.65	0.78	0.71	
	f1-score	0.65	0.77	0.72	
Word2Vec + LSTM	precision	0.52	0.54	0.59	0.55
	recall	0.45	0.56	0.65	
	f1-score	0.48	0.55	0.62	
Word2Vec + BiLSTM	precision	0.53	0.60	0.59	0.57
	recall	0.57	0.47	0.67	
	f1-score	0.55	0.53	0.63	

Fig. 4 Training result of 4 text classification models

2.3.2. Model-2 Word frequency & LSTM and BiLSTM

A manually labeled dataset from Sina Weibo which contains about 40,133 of posts in 6 categories: 13,993 in null (neutral), 6,697 in like, 5,348 in sad, and 5,978 in disgust, 3,167 in angry, 4950 in happy.

In consideration of the impact of each type of emotion that can affect stock price, digust was dropped from the final classification model building. The reason why to remain null type is that there are still many posts without distinct emotion such as posts about news or announcement release. Also, except for null type, the remained emotional dataset for building model is reasonable balanced, which can avoid classification error from extreme dataset. That is to say, total of 34,155 manually labeled posts from Sina Weibo were used to build the classification model with the usage of One-Hot and LSTM& biLSTM. However, this dataset was trained to get a low accuracy that not suitable for the following research. Therefore, the self-constructed training set was also used in the this model and the following two models in training processes.

This model used One-Hot as word vector translator with consideration of word frequency. It calculated a post sentence into a vector which considered maximum of 100 words. The classification model was basically using LSTM (a classical recurrent neutral network called long short term memory) and BiLSTM (bidirectional long short term memory). It consisted of input layer with embedding size of 256, a layer of "SpatialDropout" to avoid over-fitting, 128 hidden layers in long short term memory, 3 connection layers for 3 sentiments with activation layers of softmax before and after it, and finally a loss pool with function of categorical

crossentroy that suitable for multi-class classification. (Fig. 5)

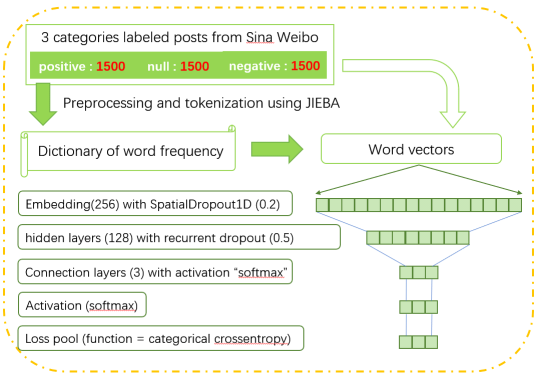


Fig. 5 Sentiment Extraction Model 1- Word frequency & LSTM and BiLSTM

2.3.3. Model-3 Word2vec & LSTM and biLSTM

This model used Word2vec as word vector translator with consideration of a window of 2 words in neighbor position. It calculated a post sentence into a vector with 100 dimensions. That caused the different embedding size in classification model in 100.(Fig. 6)

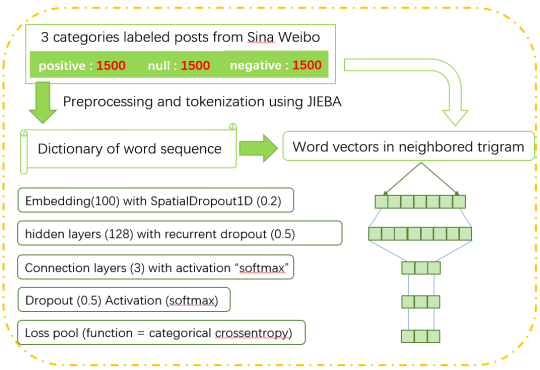


Fig. 6 Sentiment Extraction Model 2- Word2vec & LSTM and BiLSTM

From the training result (Fig.4), the accuracies of using Word2vec with LSTM or BiLSTM are not satisfied as they are only around 56%. This is due to the non-sentiment related content in most of the posts among period of COVID-19. For example, posts that related to the notice of the situation of COVID-19 such as announcement of the number of new patients or seeking of potential patients. In this context, the information point or keywords in the posts are realted to number rather than text itself which cannot be figured out by text mining method. Also, the word windows used in building Word2vec prediction model is hardly effective in catching the keywords in one posts since there is only weak or even no relationship between the textual content without numbers who really contain information of the epidemic situation. These kinds of posts domain a large portion of all the data we collected from the sample period, and therefore, generate the key problem in extraction of sentiment.

Ⅲ. Results Analysis

3.1. Best-fitted model selection

Training results (Fig. 5) show that Word frequency and BiLSTM model performed best with self-constructed training set in the sentiment extraction processes. Therefore, we used this model in the sentiment factor building processes to get the sentiment score of each sample day.

3.2. Stock price indexes

We chose the big-4 stock indexes in mainland China as showed in Fig. 7.

Stock index name	Index number
SSE Composite Index	000001
Shenzhen Component	399001
SZSE SME PRICE INDEX	399005
CHINEXT PRICE INDEX	399006

Fig. 7 Big-4 stock indexes in mainland China

The first 3 indexes represent mostly the mainboard market, while the CHINEXT PRICE INDEX fully reflects the innovation and entrepreneurship characteristics of Shenzhen Market, and selects 100 companies with large market value and good liquidity as samples, which is the benchmark and product index of GEM. The

change of index prices of the big-4 indexes at each trading day in the sample period from January 1st to March 31st were showed in Fig. 8.

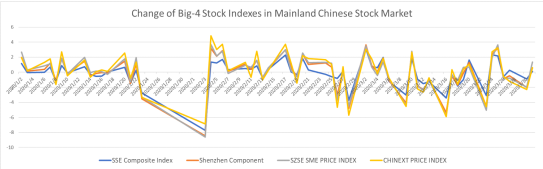


Fig. 8 Change of Big-4 Stock Indexes

A similar tendency of all of Big-4 indexes can be observed from the price change. 3 indexes from Shenzhen Stock exchange fluctuated more in range compared to SSE Composite from Index because the later one contains a large quantity of stocks that price change of each stock inside the index can eliminate with each other according to the portfolio theory in finance. [12] collected financial articles in Sina News for sentiment analysis, and use the results to predict stock prices change. The evidence showed that investor sentiment had a particularly strong effect for value stocks relative to growth stocks in China due to the immature financial market. In the consideration of the Big-4 stock indexes we used in our study, CHINEXT PRICE INDEX is more according with this characteristic since the top 100 stocks with large market value are selected to build this index.

Ⅳ. Conclusion

We do an event study based on the progress of COVID-19 in China from January 1st to March 31st, and observe the movement of public mood regarding to COVID-19 and the change of big-4 stock price indexes in mainland Chinese stock market. We collect sentiment related dataset from Sina Weibo, which is the most commonly used social media platform in mainland China and can represent public mood in some extent. The empirical result shows that no significant correlation between stock price change and public mood when the sentiment factor is the only factor in consideration. Regardless of the strict linear regression results, we still find similar tendency exists between stock price change and sentiment factor, which indicates that investors can get the momentum of stock price change from the movement of public mood of social media (Sina Weibo).

We have contribution to the sentiment analysis on stock market in the following two ways. Firstly, we confirm that a specific training set in sentiment extraction process is important when we are applying sentiment analysis to a special event with specific areas of words just like COVID-19. A generally used training set can get a high accuracy in in-sample prediction, but perform very bad in a special event-related corpus because the training set does not contain most of the words that show up in the event. Secondly, we get a similar tendency between stock price change in mainland China and public mood movement, which indicates the influence of investor sentiment on stock price in the individual dominated stock market. Even though sentiment factor cannot be homogeneously used in predicting stock price change, it is proved that sentiment factor is one of the heterogeneous factors that can be added to the classical asset pricing models to see if it can improve their prediction ability in further study.

References

[1] C. Xie and Y. Wang, “Does Online Investor Sentiment Affect the Asset Price Movement? Evidence from the Chinese Stock Market,” *Mathematical Problems in Engineering*, pp. 1 – 11, 2017.

[2] A. Carosia, G. P. Coelho, and A. E. A. Silva, “Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media,” *Applied Artificial Intelligence*, vol. 34, no. 1, pp. 1 – 19, 2020.

[3] Report of Investigation of Individual Investors Condition in 2017 by SHENZHEN STOCK EXCHANGE [Internet] Available: https://www.sac.net.cn/hyfw/hydt/201803/t20180319_134756.html.

[4] Survey of stock market investors condition report in 2019 [Internet] Available: https://www.sac.net.cn/hyfw/hydt/202003/

t20200330_142269.html.

[5] Q. Lin, "Noisy prices and the Fama–French five-factor asset pricing model in China," *Emerging Markets Review*, vol. 31, pp. 141–163, 2017.

[6] Weibo's monthly active users over 550 million, revenue surpassed Wall Street's expectations, 2020, June, 4th [Internet] Available: https://caijing.chinadaily.com.cn/a/202005/20/WS5ec4bf1ea310eec9c72ba2ec.html.

[7] I. Zheludev, R. Smith, and T. Aste, "When Can Social Media Lead Financial Markets?," *Scientific Reports*, pp. 1–12, 2014.

[8] K. Guo, Y. Sun, and X. Qian, "Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market," *Physica A*, vol. 469, pp. 390–396, 2017.

[9] L. Xu, H. Lin, and J. Zhao, "Construction and analysis of emotional corpus," *Journal of Chinese Information Processing*, vol. 22, no. 1, pp. 116–122, 2008.

[10] J. Li, G. Huang, C. Fan, Z. Sun, and H. Zhu, "Key word extraction for short text via word2vec, doc2vec, and textrank," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 27, no. 3, pp. 1794–1805, 2019.

[11] H. Zhou, M. Huang, X. Zhu, and B. Liu. "Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory," *AAAI* 2018, New Orleans, Louisiana, USA, 2018.

[12] D. D. Wu, L. Zheng, and D. L. Olson, "A Decision Support Approach for Online Stock Forum Sentiment Analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 8, pp. 1077–1087, 2014.

이가기(Jiaqi Li)

Master student in Frontier Business (Finance track) at Sungkyunkwan University (2020~)
B. B. A. in Accounting at Shantou University, China (2017)

오하영(Hayoung Oh)

Sungkyunkwan University Professor (2019~)
Ajou University Professor (2016~2019)
Soongsil University Professor (2013~2016)
Ph.D. in computer engineering at Seoul National University (2013)