Analysis of Sentiment Analysis Research Trends using Text Mining

Seohyun Kim

Abstract

Recently, sentiment analysis is actively researched in the field of natural language processing because of the expansion of computer science and social media. However, sentiment analysis has a wide range of applications and it might be challenging for individual researchers to review and comprehend the trends of the studies. Therefore, this study suggests trends in sentiment analysis research to help sentiment analysis researchers. The sources of data collection are Google Scholar and Scopus. As a result, studies using various modalities, researching the change from a particular event and based on language itself have been explored actively in recent years.

1. Introduction

Text mining is a method for extracting significant information from text data. Despite its usage as an analytical technique in research in domains such as medical informatics, social science, and design, there is a scarcity of research on text mining itself [1]. Text mining has many applications, including text classification and question answering.

Sentiment analysis is the process of analyzing subjective information, such as opinions, sentiments, evaluations, and attitudes contained in texts, using a computer. It is actively researched in the field of natural language processing, and is also widely researched in data mining, text mining, and web mining. Additionally, as it becomes more significant in management and society at large, sentiment analysis is extending its reach beyond computer science to include social science and management science as well. The significance of sentiment analysis is especially highlighted by the expansion of social media, such as reviews and Twitter [2]. It is now being used in research in industries like finance and pharmaceuticals. People are increasingly expressing their ideas and finding information on social media or comments thanks to the advent of the Internet. According to this trend, sentiment analysis is being regularly investigated, and it is quite likely that research will be conducted in more interdisciplinary areas.

Since sentiment analysis has such a wide range of applications, it might be challenging for individual researchers to review and comprehend all of the studies. Research trend analysis for sentiment analysis is thus required in order to aid researchers in understanding the research flow [3]. But research on text mining itself is not active, as was already indicated. Current developments in sentiment analysis research have mostly centered on research methodology and models. Since the study methodology and models are the main topics, the methodology and its connection to the topic are not being revealed. Consequently, more comprehensive research is required.

The objective of this study is to investigate trends in sentiment analysis research without favoring certain areas, such as methodology or subjects. To this purpose, we will analyze sentiment analysis research articles in academic databases and utilize text mining to determine which aspects of previously completed sentiment analysis studies have been emphasized and discussed. This study attempts to assess the research trend in a neutral and impartial manner, avoiding as much as possible the researcher's personal preference or point of view using artificial intelligence technologies. This may provide information and clues regarding research to sentiment analysis researchers, as well as the possibility to investigate hitherto unexplored research fields.

2. Literature Review

Previous studies can be divided into two groups. The first group consists of studies that analyzed trends of research using text mining. Kim & Delen (2018) examined the evolution of important domains within the field of health informatics. The Term-by-Document Matrix (TDM) was used to validate the association between words and documents, clustering was performed, and each cluster was interpreted. Biomedical, algorithmic, and statistical approaches, adoption of HIT, Internet-enabled research, and knowledge representation comprised a total of six clusters. This research allows for a full understanding of the developments in the area of medical informatics from 2002 to 2013 and which academic fields have matured [4].

Minaee et al. (2021) evaluated the deep learning-based text categorization research trend. Text classification involves sentiment analysis and categorization of news articles. We examine over 150 deep learning models, ranging from LSTM to reinforcement learning, in terms of their technical contributions, similarities, and benefits. In addition, they investigated frequently used datasets and deep learning model performance in each text classification field. Minaee et al. (2021) generally reviewed the methodologies and general data commonly used in the relevant fields, but it is difficult to understand the current research trends [5].

Kim et al. (2022) investigated domestic economics education research trends using text embedding and clustering. SentenceBERT, which specializes in sentence embedding, is used to overcome BERT's inability to adequately describe the meaning of sentences. The analysis results were divided into six clusters: analytical findings, empirical analysis of students' economic education achievements, comprehension and application of economic education contents, and the need for financial education and financial literacy. In addition, by analyzing important patterns in economics education research within each cluster, they provided data for researchers and teachers interested in economics education to discover the most recent research trends [6].

Prior studies on sentiment analysis research trends mostly focused on reviewing previously published studies. Particularly, many studies examined the methodologies used.

Lee (2018) investigated many domains of sentiment analysis, including application areas and methodologies. However, there is a limitation in that it only examined specific fields, such as movie reviews, product evaluations, and social media, and specific methodologies such as machine learning techniques such as SVM and Nave-Bayes.

Birjali et al. (2021) assessed studies in a variety of aspects, including application fields, data, and methodologies. In the case of data, the data sources, input data formats, and feature extraction were described in depth [7].

Mantyla et al. (2018) reviewed research trends in sentiment analysis using text mining. To explore the size and annual trend of the relevant research area, the number of sentiment analysis-related papers

published each year and the number of citations were assessed, together with the published journals and the ratio of sentiment analysis-related articles published in each journal. The outcome of applying Latent Dirichlet Allocation (LDA) to extract relevant study topics included terms such as summarization, Twitter, and aspect. It is evident that these terms are associated with sentiment analysis, but it is challenging for researchers to gain in-depth information for their studies from this. Through the classification tree, papers were categorized by data, analysis, objective, etc. It is a method of categorizing application areas according to research topics such as travel, security, and medical care. Even if a subject is picked from every branch of each classification tree, it may be difficult for a researcher to comprehend research trends using simply the keywords. For instance, words such as health and disease, which are picked as topics in the medical sector, are only widely used terms and cannot be considered unique topics. Furthermore, when one topic emerges in connection to another topic, it is difficult to evaluate and comprehend them just by extracting the topics [3].

Existing studies on sentiment analysis research trends have aided scholars with an interest in the topic, but they have limitations in analyzing the papers because they simply analyzed the methodologies or just extracted topics. This study aims to contribute to future sentiment analysis research by conducting a thorough examination of research trends related to sentiment analysis.

3. Data

The sources of data collection are Google Scholar and Scopus. Google Scholar is an academic search engine that Google provides. You may search papers, academic journals, and publications using this tool. Furthermore, Google Scholar is excellent for locating early sentiment analysis research publications that do not overlap with Scopus [3].

Elsevier developed Scopus in 2004 as an academic database that provides a comprehensive collection of scholarly citations [8]. Moreover, Scopus is a dependable database that does not search duplicate results and has the least discrepancy in terms of content certifications and quality [9]. As a result, we consider it to be the ideal area to explore scholarly publications. We gathered the titles of publications that were retrieved using "sentiment analysis" in Google Scholar and Scopus.

4. Research Model

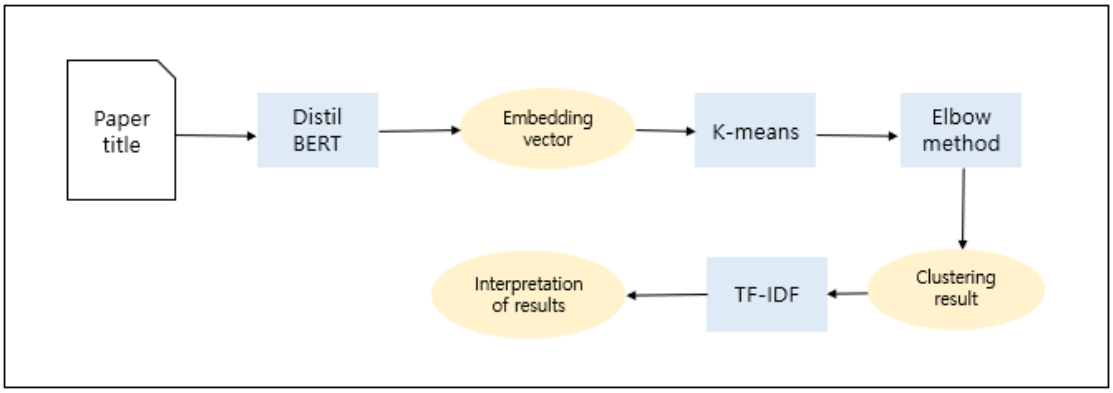

Figure 1. Research Model

The research model converts the collected data into vectors embedded with DistilBERT, identifies ideal clusters with the elbow technique, and clusters the embedding vectors with K-means according to the optimal number of clusters (Figure 1).

DistilBERT is a model that HuggingFace suggested in 2019. By using the distilling knowledge approach to the conventional BERT model, its size and speed are reduced, but its performance is comparable to that of the conventional BERT. In the conventional NLP area, pre-trained models showed a growing tendency toward bigger models. The benefit of a large-scale pre-trained model is that it improves performance, but the downside is that it is difficult to employ in real time on a small device. DistilBERT utilizes knowledge distillation technology [10] to compensate for these shortcomings. Knowledge distillation is a compression technique whereby a small model (the student) is trained to replicate the behavior of a large model (the instructor) [11, 12]. When this knowledge distillation technique is applied to the BERT model, knowledge is transmitted to a small BERT (student) based on the pretrained BERT-base (teacher). Comparing the model's performance to the conventional BERT using the GLUE benchmark, it demonstrated a comparable performance of around 97%. Moreover, it solved inference problems with fewer parameters and at a quicker rate than BERT-base. Due to the model's tiny size and fast processing speed, it is suitable for usage on portable devices [10].

Clustering, which is part of unsupervised learning using unlabeled data, is a useful data science method. K-means is the most popular and well-known clustering method [13]. K-means first marks N central points and determines the location of the central point n that minimizes the sum of the distances between each point and the center. Then it groups points close to this central point around the central point [14].

K-means clustering is a local optimization technique that is sensitive to locating a starting point from the cluster center. Therefore, if the beginning point is put at the center of a faulty cluster, the clustering algorithm generates a large number of errors and wrong cluster results. Therefore, the appropriate number of K-means clusters must be determined using the elbow method [15]. Visualizing the total sum of squares (WSS) inside a cluster on a graph based on the number of clusters and visually determining the location of the elbow is how the elbow method determines the best number of clusters. The clusters are generated based on the result of the elbow method.

Term Frequency-Inverse Document Frequency (TF-IDF) is a technique for allocating a substantial weight to each word in a document word matrix based on the frequency of the term and the inverse document frequency. Consequently, this methodology is mostly used for assessing the similarity of documents and the significance of search results in a search engine. Term Frequency (TF) is the number of times a certain term occurs in a document. Inverse Document Frequency (IDF) determines the frequency of words in a document. For instance, common but unimportant words, such as "of", are assigned less weight [16]. In this study, TF-IDF is used to assist the analysis of clustering results. We evaluate the word composition of clustered papers by applying TF-IDF to the clustering results.

5. Analysis Result

The result of clustering with 7 clusters was turned into a data frame with paper titles. Then, in order to visually check whether the clusters were well formed, they were made two-dimensional through t-distributed stochastic neighbor embedding (t-SNE) and then visualized. t-SNE is an algorithm that lowers the dimensions of high-dimensional data to two. It is one of the manifold learning techniques used to visualize complicated data.

Based on the cluster results using t-SNE, it was determined that clusters 0, 2, 3, 4, and 5 were significant among the seven clusters. Due to the enormous number of papers, it is difficult to comprehend the themes of the papers in the cluster by examining merely the cluster results. Hence, we compared the results of clustering with the results of TF-IDF extraction of frequently occurring keywords for each cluster. Checking the keywords revealed that prepositions were overused, making it hard to determine the real subject of the paper. Therefore, nltk was used to extract noun-only portions. Excluding sentiment and analysis, which appear often because they are keywords, words such as "multimodal", "twitter", "tweets", and "multi" occurred frequently in the search results. Checking the cluster results centered on these words revealed that cluster 0 was a multimodal cluster that employs many modalities, which have been explored extensively in recent years.

Looking at cluster 2, it is evident that "vaccine" and "covid" appear more often than in other clusters. Since the breakout of COVID-19 in 2020, there have been more studies on sentiment analysis of people's reactions or opinions regarding vaccinations, and these studies have been actively conducted to the degree that clusters appeared.

The terms "literary", "semantic", "word", and "turkish" occur often in cluster 3. This cluster is a sentiment analysis research cluster focused on review papers that investigated the area of sentiment analysis research and language itself. In cluster 4, there are several terms such as "stock," "price," and "product." This cluster consists of research papers on product review data sentiment analysis and stock price forecasting. Cluster 5 included many terms like "online" and "research". When the extracted terms and cluster results were examined together, it was evident that there are several studies about online media or online learners.

6. Conclusions

This study gathered papers titles from Google Scholar and Scopus. Utilizing DistilBERT, K-means clustering, and TF-IDF, we aimed to provide information to other researchers in the area of sentiment analysis using this data. As a result of the analysis, studies pertaining to sentiment analysis were categorized into seven clusters, and these clusters represent the significant research trends of studies pertaining to sentiment analysis. The following are the findings and inferences drawn from this study.

First, not only research topics such as stock price prediction, but also Twitter and product analysis reviews seem to occupy a crucial part in sentiment analysis research. A search for novel sentiment analysis target texts might be a good direction of research.

In addition, it was found that sentiment analysis research, which was once restricted to natural language processing, is developing into research using various modalities and features, such as voice, video, and image, and that its weight is rising. Instead of seeing images and natural language as different fields of study, researchers have discovered that they are inextricably interconnected and have begun to investigate their interactions. On the basis of this observation of change, sentiment analysis study should be undertaken in conjunction with the search for clues in other research domains.

The flow of people's emotions after a particular event or change, such as COVID-19, may also be a key subject of research. When people encounter specific events or changes, their thoughts and emotions change, and we should pay close attention to the research of these changes and their implications.

In the current trend analysis of sentiment analysis research, several studies were assessed just for their methodology. Furthermore, it was difficult to determine the relationships between the linked topics. By reviewing the clustering results and TF-IDF results exhaustively, it was feasible to generate

analytical results that included the linkages and features between themes in this study. This analysis technique is applicable to other studies, including research trend analysis and content analysis in different domains.

This study can help sentiment analysis researchers to find current research trends and undertake follow-up studies. However, there are a few limitations. Data was collected only from Google Scholar and Scopus, but Web of Science, a large academic database, and RISS, a domestic academic paper database, were omitted from the data collection and analysis in this study. It is vital, therefore, to increase the scope of data collection to provide more extensive information for future scholars. In addition, there were overlapping terms in each cluster as a result of this study. In future research, it is necessary to exclude these terms before analyzing them. If research trend analyses are developed further by improving these limitations, it will be able to offer more detailed sentiment analysis research directions to future scholars.

References

[1] Jung, H., & Lee, B. G. (2020). Research trends in text mining: Semantic network and main path analysis of selected journals. Expert Systems with Applications, 162, 113851.

[2] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[3] Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32.

[4] Kim, Y. M., & Delen, D. (2018). Medical informatics research trend analysis: A text mining approach. Health informatics journal, 24(4), 432-452.

[5] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. ACM Computing Surveys (CSUR), 54(3), 1-40.

[6] Dong Jin Kim, Ha Rim Lee, and Gil Jae Lee. "Analysis on Korean Economic Education Research Trends Using Text Imbedding and Clustering based on AI BERT model". The Journal of Learner-Centered Curriculum and Instruction 22.18 (2022): 931-947.

[7] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226,* 107134.

[8] Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. Journal of informetrics, 1(2), 161-169.

[9] Adriaanse, L. S., & Rensleigh, C. (2013). Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. The Electronic Library

[10] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[11] Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006, August). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535-541).

[12] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).

[13] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access* , *8*, 80716-80727.

[14] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

[15] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering* (Vol. 336, No. 1, p. 012017). IOP Publishing.

[16] Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications* , *181*(1), 25-29.