Original Article

Internet Gambling through Text Mining Analysis

Junghyun Choi[1], Daeho Seo[2*]

[1]Namseoul University. E-mail: jhc@nsu.ac.kr

[2*]Dagyeam. E-mail: seo_daeho@naver.com

Abstract

The current study aimed to identify the rising phenomenon of online gambling in news data using Big Kinds, and to check the connection structure of important words.This study used the text-mining method to analyze text data collected from the Naver News portal to identify important patternsof Internet gambling. The web-crawling technique was adopted for data collection from January 1, 2020 to December 31, 2021. Since the coronavirus began to spread across Korea in 2020, a two-year time window was established. All news articles within this timeframe containing the keyword "internet gambling" were identified. A total of 1,494 articles were included.The text mining analysis showed that there have been many cases of arrests due to crimes, mainly online gambling, or crimes being committed after the perpetrator became debt-riddendue to online gambling. It can be seen that those who gamble online are mostly linked to crimes. Journalists have brought these topics up in their provocative articles, but the fact that there are no positive articles about online gambling shows that people should be careful about online gambling.Moreover, many articles state that the number of people who enjoyed online gambling has increased significantly as the time spent using the Internet at home increasedrapidly due to COVID-19 restrictions. Online gamblers are increasingly having a negative impact on society; however. the government has not formulated any clear measures against this issue. The online gambling industry is expanding, leading to a negative impact on society, and the government needs to take an urgent action.

Keywords: Internet, Gambling, Big Data, Social, Text Mining.

DOI: 10.47750/pnr.2022.13.S03.093

INTRODUCTION

Although gambling is a relatively common activity, for a minority of people,it can lead to the development of a gambling problem, which is classified as a behavioral addiction [1]. Gambling problem poses great physical and psychological harm to humans, thereby attracting scholarly attention [2]. Gambling problems are experienced by 0.4–2.0% of adults worldwide [3]. The global online gambling market is expected to grow by 13.2% between 2019 and 2020, from USD$58.9 billion to USD$66.7 billion [4]. This growth appears to be due to COVID-19, which is limiting access to land-based gambling opportunities and resulting in more people gambling online [1].

Technology also has made a substantial impact on gambling over the past decade. There have been significant changes in the gambling landscape, particularly relating to gambling in the digital age. With daily use of the Internet and various smart devices, users have been able to communicate in real time and the existing communication style has changed. Online gambling provides convenience and the ability to gamble from the home and workplace [4]. Consequently, online gambling has seen major growth owing to the availability and convenience of the Internet and through innovative technology that has made remote gambling possible[5]. Some elements of online gambling, including isolation and continuous and easy access, have been argued to pose certain risks [4]. The rapid expansion of online gambling has meant that gambling policies have often lagged behind[5].This expanding landscape of online gambling may pose public health problems.

In contrast, due to the change of the information subject by the Internet, data has become more massive and caused the emergence of big data. These Big Data are viewed as a new opportunity to understand social issues [6]. It is labor intensive and time consuming to extract huge amounts of information from cumulated collections [7]. Text mining is an increasingly important research field because of the necessity of obtaining knowledge from the enormous number of text documents available, especially on the Web. Text

mining and data mining, both included in the field of information mining, are similar in some sense, and thus, it may seem that data mining techniques may be adapted in a straightforward way to mine text [8]. In this respect, text mining explores patterns using unstructured text data to find meaningful information. As text data exists in various places such as newspapers, books, and the web, the amount of data is diverse and large, making it suitable for understanding social reality. In recent years, there has been an increasing number of attempts to analyze texts from web such as SNS and blogs where gamblers can communicate freely. It is recognized as a useful method to immediately grasp online gamblers'and public'sopinions and thus, it can be used for issue research. Text mining has received considerable attention in examininggamblers'and public opinion as well as to investigate gamblers' potential feelings [6]. The increasing rates of social harm events in online gambling demand employing text mining techniques to not only better understand their causes but also develop optimal prevention strategies [9]. These results will provide interesting insights into the characteristics of online gambling among Koreans and important implications for online gambling in the education field [10]. The development and testing of a public health approach to addressing the harms associated with online gambling in these areas is needed [11].

Recently, several studies related to Internet addiction and problematic smartphone using data mining have been reported [12,13]. A study of Korean Internet Addiction Disorder applied network analysis and decision making-tree analysis to the social big data collected from online news sites, blogs, Internet cafes, social network services, and Internet message boards [12]. This study suggested data mining analysis and network analysis of Internet Social Big Data presented as a prediction model for Internet addiction risk factor, which was considered significant in both policy and analysis methodology. Lee [13] investigated problematic smartphone use in news data using Big Kinds, andchecked the connection structure of important words. Through the analysis of news data, it was possible to confirm the semanticconnection structure surrounding the recently rising problematic smartphone use[13].However, although some text mining studies on game addiction and sports have also been reported, there is scant data research on online gambling.

In this study, we practiced text mining to analyze news about online gambling to discover hidden textual patterns [7]. We then performed trend analysis to explore the dynamics of the proportions of topics and hierarchical clustering analysis to cluster similar topics [7]. Therefore, the current study aimed to identify the rising phenomenon of online gambling in news data using Big Kinds, and to check the connection structure of important words. We assume that a good understanding of text ming with online gambling is important for developing regulatory initiatives, awareness, and prevention programmes for responsible online use [14].

MATERIALS AND METHODS

1. Data collection

The web-crawling technique was used to collect data from the Naver News portal between January 1, 2020, and December 31, 2021. Since the coronavirus began to spread across Korea in 2020, a two-year time window was established, and all news articles containing the keyword "internet gambling" in the main body were identified. Finally, a total of 1,494 articles were included.

2. Data processing

First, stop words were removed from the collected data. Many words appearing in articles contained stop words. For example, articles contained words that were unrelated to the main text of the article, such as the reporter's name, the newspaper's name, and a disclaimer about unauthorized distribution. Therefore, such information was filtered in advance to avoid the production of inappropriate analysis results (Choi and Park, 2017). In this study, a list of relevant stop words was prepared in advance and all tweets containing such stop words were deleted. Subsequently, MeCab, a Python-based Korean morphological analysis engine, was used for part-of-speech tagging. In Korean information retrieval, nouns are used as index words or keywords representing the article (Shin and Lee, 2009). Accordingly, we extracted and analyzed only nouns(Table 1).

Table 1. Data collection information

Data collection	Information
Target Period	• Naver News • January 1, 2020 toDecember 31, 2020
Data scale	• 1,494 articles
Method	• Python 3.0 • Selenium package • MeCab package
Elements	• Post title • Post timestamp • Body of post

3. TF, TF-IDF

Keywords were derived by term frequency (TF), and term frequency-inverse document frequency (TF-IDF) analyses, which were performed on the main text of the articles from Naver News. TF simply calculates the frequency of a term, assuming that terms with higher frequency are more important. Subsequently, TF-IDF calculates the importance of a term by multiplying the term frequency and inverse document frequency. It assumes that terms appearing in all documents inversely have low importance. For example, terms such as "newspaper" and "today" may appear at a higher frequency in newspaper articles, but their importance is low because such terms are more likely to be included in most articles. TF and TF-IDF analyses were performed with each newspaper article, which was considered an individual document. Subsequently, a word cloud was created for a visual representation to allow the keywords to be viewed and examined in a single view. Word cloud is a method of forming a cloud of words with words having higher importance indices appearing more centrally in a larger size.

4. Centrality Analysis

Centrality analysis is a method used to calculate the centrality index of a word by examining its importance. It is used to calculate importance by establishing relationships between words. Centrality indices typically used include degree centrality (Cd), closeness centrality (Cc), and betweenness centrality (Cb), and this study used these three centrality indices. Cd refers to the sum of all other nodes to which a node is connected within a network. With this, it is possible to identify how many other nodes a node is related to and how many relationships the node is involved in. A node with a high Cd is more likely to be a core node within the network. Thus, more information can be acquired by using the relationships of this node with other nodes in the network. Cc is a method used to measure centrality based on the distance between nodes. While Cd identifies only the number of nodes that are directly connected to a node, Cc measures centrality based on the distances between all nodes that are even indirectly connected. Cc also identifies the relationships between indirect nodes that can be approached through people who are directly connected and can measure centrality in a broader range as compared to Cd. Higher indirect centrality refers to the many relationships with core nodes, which would indicate a higher likelihood of being a core node within that network through relations with many core nodes. Cb is an indicator of the position between a node and other nodes within a network. In other words, the amount of influence a node has can be determined by identifying its position within the network. A node with a high Cb within a network is highly likely to function as a core node and can act as a broker between other core nodes.

5. Word network analysis

Word network analysis is a graphic analysis technique for expressing the relationships among words within a network. The co-occurrence frequency of words within the same document and higher weights are assigned to edges with higher co-occurrence frequency for the visualization of the network. In this study, Cc was used to indicate the weight of each node (word). The importance of edges can be easily visualized through network analysis.

6. Topic modeling

LDA(Latent Dirichlet Allocation) topic modeling was conducted. For topic modeling, each article was considered a document, and nouns extracted from each document were used to construct a document–term frequency matrix. Subsequently, Python's Gensim package was used to derive the LDA model. The learning parameters for modeling were set to batch size = 300, iteration = 50, and gamma = 0.001. Moreover, hyperparameters were designated as alpha = 0.03 and eta = 0.01.

ANALYSIS RESULTS

1. TF(Term Frequency) and TF-IDF(Term Frequency-Inverse Document Frequency) analyses results

The results of the TF analysis, which is a simple frequency analysis, showed that terms such as "gambling," "criminal investigation," "charges," "Internet," "police," "site," "coast guard," and "defection to North Korea" appeared frequently in that order (from first to last). "COVID-19," which was a subject of interest in the study, ranked 36th. The terms "coast guard" and "defection to North Korea" appeared in high frequency because of many articles being published in September 2020 on a South Korean government worker being shot dead while attempting to defect to North Korea to escape from internet gambling debt. The term "COVID-19" appeared frequently in articles on surveys that reported an increase in illegal internet gambling because of more time spent on computers or smartphones during the pandemic(Figure 1).

Figure 1. Words clouds of TF analysis

The TF-IDF analysis results showed that terms such as “gambling,” “charges,” “criminal investigation,” “police,” “coast guard,” “site,” “defection to North Korea,” “mask,” and “Internet” appeared frequently in that order. Interestingly, the term “COVID-19” ranked 23rd. Unlike in TF analysis, the term “COVID-19” also moved from the 36th position to 23rd and the term “mask” ranked among the most frequent terms. “Mask” appeared frequently because of the inclusion of articles on a group of criminals who made over 110 million won from posting false internet ads of inexpensive masks during the mask shortage crisis in February 2020 and then lost all of their money on internet gambling(Figure 2).

Figure 2.Words clouds of TF-IDF analysis

2. Centrality analysis

Top 50 Cd, Cc, and Cb terms were extracted(Table2).

Table 2. The most 50 probable words of Cd, Cc, and Cb

Degree centrality (Cd)	Closeness centrality (Cc)	Betweenness centrality (Cb)
('Internet', 1.0)	('Internet', 1.0)	('Internet', 0.3759436815793122)
('gambling', 0.9858356940509916)	('gambling', 0.9860335195530726)	('gambling', 0.34777602770344296)
('criminal investigation', 0.4929178470254958)	('criminal investigation', 0.6635338345864662)	('charges', 0.04205909526907374)
('charges', 0.4900849858356941)	('charges', 0.6622889305816135)	('criminal investigation', 0.029145603938170953)
('distribution', 0.4334277620396601)	('distribution', 0.6383363471971067)	('distribution', 0.021327689602649053)
('confirmation', 0.4107648725212465)	('confirmation', 0.6292335115864528)	('confirmation', 0.015459462549048064)
('related to', 0.3767705382436261)	('related to', 0.6160558464223386)	('related to', 0.012704502479407784)
('investigation', 0.3654390934844193)	('investigation', 0.6117850953206239)	('investigation', 0.011231163583684414)

('at the time', 0.3371104815864023)	('at the time', 0.6013628620102215)	('at the time', 0.007987102528816865)
('result', 0.29178470254957506)	('result', 0.5854063018242123)	('site', 0.005566212413075417)
('case', 0.2776203966005666)	('case', 0.5805921052631579)	('illegal', 0.00516933107293067)
('illegal', 0.2691218130311615)	('illegal', 0.5777414075286416)	('police', 0.004962878969983726)
('site', 0.2691218130311615)	('site', 0.5777414075286416)	('arrest', 0.0048925626418405916)
('police', 0.2492917847025496)	('police', 0.5711974110032363)	('result', 0.0040723110080044984)
('arrest', 0.2464589235127479)	('arrest', 0.5702746365105008)	('crime', 0.003975263765048313)
('crime', 0.2464589235127479)	('crime', 0.5702746365105008)	('case', 0.0037427523811399406)
('National Police Agency', 0.24362606232294617)	('National Police Agency', 0.5693548387096774)	('government worker', 0.0025865875167798555)
('explanation', 0.24362606232294617)	('explanation', 0.5693548387096774)	('prosecution', 0.002370269536880109)
('photograph', 0.22946175637393768)	('photograph', 0.5648)	('National Police Agency', 0.0023311964019241543)
('government worker', 0.2237960339943343)	('government worker', 0.5629984051036683)	('Defection to North Korea', 0.0022444890164285023)
('fact', 0.2181303116147309)	('fact', 0.56120826709062)	('explanation', 0.0022257069774117284)
('determination', 0.21529745042492918)	('determination', 0.5603174603174603)	('photograph', 0.0020566211051075596)
('defection to North Korea', 0.21529745042492918)	('defection to North Korea', 0.5603174603174603)	('missing', 0.001948222784513148)
('possible', 0.2096317280453258)	('possible', 0.5585443037974683)	('Seoul', 0.0018773684111032148)
('North Korea', 0.2096317280453258)	('North Korea', 0.5585443037974683)	('North Korea', 0.001870874788352518)

('missing', 0.2067988668555241)	('missing', 0.5576619273301737)	('previous year', 0.0018060641594137628)
('prosecution', 0.2039660056657224)	('prosecution', 0.556782334384858)	('criminal act', 0.0017828179130764384)
('Seoul', 0.20113314447592068)	('Seoul', 0.5559055118110237)	('operation', 0.0017488723165147181)
('coast guard', 0.1954674220963173)	('coast guard', 0.554160125588697)	('determination', 0.001673261194952447)
('previous year', 0.1926345609065156)	('previous year', 0.5532915360501567)	('fact', 0.0016597357654349195)
('operation', 0.1926345609065156)	('operation', 0.5532915360501567)	('claim', 0.0013677301886599342)
('claim', 0.1898016997167139)	('claim', 0.5524256651017214)	('sentence', 0.0013552837679347853)
('self', 0.18696883852691218)	('self', 0.5515625)	('coast guard', 0.0013404414387643217)
('criminal act', 0.1813031161473088)	('criminal act', 0.5498442367601246)	('possible', 0.0012300200157656877)
('information', 0.1813031161473088)	('information', 0.5498442367601246)	('information', 0.0010421986910732527)
('condition', 0.1671388101983003)	('condition', 0.5455950540958269)	('self', 0.0010397976064613875)
('situation', 0.1643059490084986)	('situation', 0.5447530864197531)	('citizen', 0.0007788336275594472)
('debt', 0.15864022662889518)	('debt', 0.5430769230769231)	('prosecutor', 0.0007371300259162791)
('Ministry of Oceans and Fisheries', 0.15864022662889518)	('Ministry of Oceans and Fisheries', 0.5430769230769231)	('imprisonment', 0.0007013729814827481)
('sentence', 0.1558073654390935)	('sentence', 0.5422427035330261)	('condition', 0.0006796927522037785)
('citizen', 0.1558073654390935)	('citizen', 0.5422427035330261)	('debt', 0.0006076912758188854)
('discovery',	('discovery',	('Ministry of Oceans and

0.1501416430594901)	0.5405819295558959)	Fisheries', 0.0005420421813551844)
('prosecutor', 0.1473087818696884)	('prosecutor', 0.5397553516819572)	('COVID-19', 0.0004614289580387656)
('content', 0.1473087818696884)	('content', 0.5397553516819572)	('situation', 0.00042076635365345113)
('this day', 0.1359773371104816)	('this day', 0.5364741641337386)	('detective', 0.0004174272191828966)
('announcement', 0.1359773371104816)	('announcement', 0.5364741641337386)	('content', 0.0003768304273287368)
('to be shot', 0.1359773371104816)	('to be shot', 0.5364741641337386)	('funds', 0.0003452023469442194)
('afternoon', 0.1331444759206799)	('afternoon', 0.5356600910470409)	('victim', 0.00034239434400212674)
('official', 0.13031161473087818)	('official', 0.5348484848484848)	('Lee Jae-Myung', 0.00033176574924825207)
('victim', 0.1274787535410765)	('victim', 0.5340393343419062)	('official', 0.0003279691444210383)

Terms related to the police investigation of cases involving internet gambling were ranked at the top. In particular, terms related to the government worker who was shot while trying to defect to North Korea to escape from internet gambling debt appeared most frequently. The term "COVID-19" ranked within the top 50 terms for only Cb. It was determined that terms related to internet gambling were concentrated mostly in news related to a specific issue.

3. Network analysis

The terms were visualized through network analysis. The search keywords "Internet" and "gambling" were positioned at the very center. Terms related to the shooting of a government worker who tried to defect to North Korea appeared in the upper part of the figure; terms related to police and prosecutor investigations of cases involving illegal internet gambling appeared on the left side; and terms related to organized gangs, such as "threat," "gang member," "assault," and "KakaoTalk" appeared at the bottom(Figure 3).

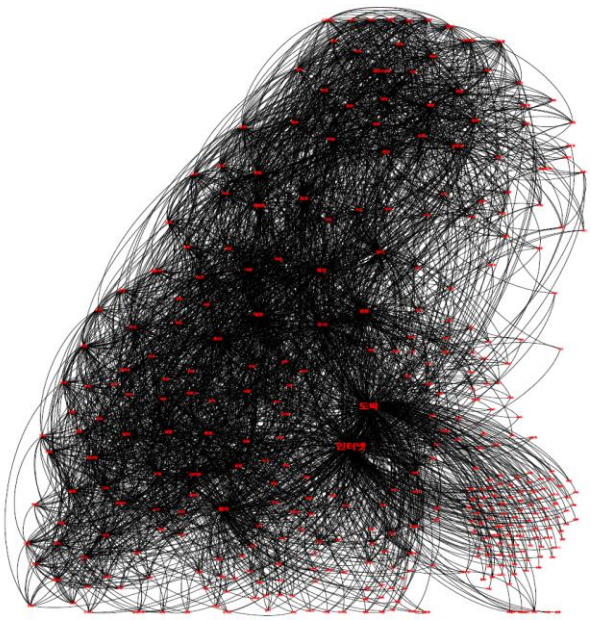

Table 3.

DISCUSSION

This study searched for domestic online articles related to online gambling and used them to analyze major issues. Simply put, we looked at the main frequent words usingtf and tf-idfanalyses, and then, we looked at the relationship between words using centrality analysis and network analysis.

The text mining analysis results show that there have been many cases of arrests for specific crimes, mainly through online gambling, or crimes being committed after the prepatratorbecame debt-ridden through online gambling. Many analysis articles state that the number of people who gamble on the Internet has increased due to the increased time indoors due to COVID-19 restrictions.

First of all, the analysis highlighted a lot of important words such as gangster, police, and crime. If you looked at articles that contain these words, a lot of them were very negative articles, as online gambling is mainly related to gangster and crime. It can be seen that those who gamble online are mostly linked to crime. Of course, journalists may have brought these topics up to write provocative articles, but the fact that there are no positive articles about online gambling shows that people should be careful of it.

Next, there are articles that stated that the number of people who enjoyed online gambling has increased significantly as the time spent using the Internet at home rapidly increased due to COVID-19. Online gamblers are increasingly having a negative impact on society; yet, the government has not come up with any clear measures. None of the articles were about any action the government took to eradicate online gambling. The online gambling industry is expanding, which has a negative impact on society, and the government needs to take urgent measures against it.

Author Contributions: All authors were responsible for the study conception and design and obtaining funding. D.S. performed data analysis and provided statistical expertise. J.C. and D.S. drafted the manuscript and will make critical revisions to the paper for important intellectual content. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Non applicable.

Funding: Thisworkwas supported by the National Research Foundation ofKorea (NRF-2020R1I1A3052207).

Acknowledgments: Non applicable.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

Gainsbury, Sally M., et al. "Reducing Internet Gambling Harms Using Behavioral Science: A Stakeholder Framework." Frontiers in Psychiatry 11 (2020): 1477.

Chou, Li-Wei, Kang-Ming Chang, and Ira Puspitasari. "Drug abuse research trend investigation with text mining." Computational and Mathematical Methods in Medicine 2020.

Potenza MN, Balodis IM, Derevensky J, Grant JE, Petry NM, VerdejoGarcia A, et al. Gambling disorder. Nat. Rev. Dis. Primers. (2019) 5:1–21. doi: 10.1038/s41572-019-0099-7

The Business Research Company. Online Gambling Global Market Report 2020-30: COVID-19 Growth and Change (2020). Report No.: 5024091

Bonello, Maris, and M. D. Griffiths. "Behavioural tracking, responsible gambling tools, and online voluntary self-exclusion: implications for the gambling industry." Casino and Gaming International 38 (2019): 41-45.

Seo, D. H., Jim, J. H., Kim, C. K. Issue tracking and voting rate prediction for 19th Korean president election candidates. Journal of intelligence and information systems, 24(3), 199-219.

Wang, S. H., Ding, Y., Zhao, W., Huang, Y. H., Perkins, R., Zou, W., & Chen, J. J. (2016). Text mining for identifying topics in the literatures about adolescent substance use and depression. BMC public health, 16(1), 1-8.

Delgado, M., Martín-Bautista, M. J., Sánchez, D., & Vila, M. A. (2000). Mining text data: special features and patterns. In Pattern detection and discovery (pp. 140-153). Springer, Berlin, Heidelberg.

Pandey, R. (2020). Text Mining for Social Harm and Criminal Justice Applications (Doctoral dissertation).

Lee, K., Lee, D., & Hong, H. J. (2020). Text mining analysis of teachers' reports on student suicide in South Korea. European Child & Adolescent Psychiatry, 29(4), 453-465.

Lawn, Sharon, et al. "A literature review and gap analysis of emerging technologies and new trends in gambling." International journal of environmental research and public health 17.3 (2020): 744.

Song, T. M., Song, J. Y., Jing, D. L.(2014)Risk Prediction of Internet Addiction Disorder by Using Social Big Data. Health and Social Welfare Review,34(3), 106-134.

Lee, D. S., Kim, Y. J. (2021). A Study on the COVID-19 Before and After Trends of Problematic Smartphone Use through Text Mining. The Journal of Social Science, 6(1), 65-82.

Ineme, Mfon E., et al. "The roles of nicotine dependence and demographic variables on internet gambling addiction among youths in a Nigerian City." African Journal of Drug and Alcohol Studies 19.2 (2020): 101-115.