International Journal of Advanced Culture Technology Vol.8 No.3 61-72 (2020)
DOI https://doi.org/10.17703/IJACT.2020.8.3.61

IJACT 20-9-8

The Study on Implementation of Crime Terms Classification System for Crime Issues Response

[1] Inkyu Jeong, [2] Cheolhee Yoon, [3] Jang Mook Kang

[1] Researcher, Police Science Institute, Asan., Korea
E-mail ikjeong@police.go.kr
[2] Researcher, Police Science Institute, Asan., Korea
E-mail bertter@police.go.kr
[3] *Professor, Depart of AI Convergence, Global Cyber Univ., Korea*
E-mail Honukang@gmail.com[3] (Corresponding author)

Abstract

The fear of crime, discussed in the early 1960s in the United States, is a psychological response, such as anxiety or concern about crime, the potential victim of a crime. These anxiety factors lead to the burden of the individual in securing the psychological stability and indirect costs of the crime against the society. Fear of crime is not a good thing, and it is a part that needs to be adjusted so that it cannot be exaggerated and distorted by the policy together with the crime coping and resolution. This is because fear of crime has as much harm as damage caused by criminal act. Eric Pawson has argued that the popular impression of violent crime is not formed because of media reports, but by official statistics. Therefore, the police should watch and analyze news related to fear of crime to reduce the social cost of fear of crime and prepare a preemptive response policy before the people have 'fear of crime'. In this paper, we propose a deep - based news classification system that helps police cope with crimes related to crimes reported in the media efficiently and quickly and precisely. The goal is to establish a system that can quickly identify changes in security issues that are rapidly increasing by categorizing news related to crime among news articles. To construct the system, crime data was learned so that news could be classified according to the type of crime. Deep learning was applied by using Google tensor flow. In the future, it is necessary to continue research on the importance of keyword according to early detection of issues that are rapidly increasing by crime type and the power of the press, and it is also necessary to constantly supplement crime related corpus.

Keywords*: Multi-Layer Perceptron, Crime, Machine Learning, Classification, NLP*

1. Introduction

The crime arrest rate in South Korea has increased steadily since 2012, reaching 83.9% in 2016; the average arrest rate has been very high in the last 10 years (82.48%) [1]. In addition, according to an overseas statistics site, in 2016, the safety index of South Korea was 85.69, which was one of the highest among the 117 countries surveyed. In addition, Seoul was ranked 14th in the world city safety index in 2017 [2]. This is the outcome of

Received: July 29, 2020/Revised: August 22, 2020/Accepted: September 06, 2020
Corresponding Author: Honukang@gmail.com
Professor, Depart of AI Convergence, Global Cyber Univ., Korea

the crime response policy centered on prompt criminal investigations by the police.

However, media news reports show that Korea is not free of crime. People are often exposed to very dangerous and violent crime news in the media. Crime is always a favorite topic in the mass media because it defines the meaning of deviation and normal for many people [3]. Public anxiety increases with news about serial killers, such as Yoo Young-chul and Jeong Nam-gyu, and tragic events, such as the Gangnam murder (2016) and Molar Daddy (2017); such events are still being reported periodically [4].

The results of the National Statistical Office's 2016 survey show that 45.5% of the Korean people feel unsafe. In addition, crime (29.7%), national security (19.3%), and economic risk (15.5%) are among the main factors that cause concern in our society. In particular, Table 1 shows that factors such as national security, natural disasters, traffic accidents, fire, and crime risk are anxiety factors for people. Among these factors, crime is perceived as the greatest threat [5].

Table 1. Safety recognition index by type of social risk Year

Year / Type of risk	2008	2010	2012	2014	2016
National Security	24.6	14.9	22.0	14.9	19.8
Disasters	17.3	18.8	23.3	15.1	20.8
Traffic Accident	5.7	8.0	9.4	7.3	10.2
Fire	9.9	16.6	17.7	14.2	19.1
Crime risk	-	8.2	9.1	8.9	9.2

In fact, these anxiety factors may occur even if the person has never been a victim of the criminal act; therefore, the psychological stability of the individual suffers. Most individuals are indirect victims of crime, and society incurs indirect cost of crimes [6, 7, 8]. Thus, the fear of crime is treated as an important problem in society and among the academia because the consequences of the fear of crime and the damage caused by crime are measurable, severe, and negative at the individual and social levels [9]. A study conducted in Korea in 2008 showed that the social cost of all the crimes committed during the year was approximately 36.5 trillion Korean Republic Won (KRW). Among these, the social cost of the result of crime was approximately 15.5 trillion KRW, which was 42.2% of the social cost of all the crimes committed [10].

However, the current structure of the police in Korea is unsuitable for complex policing activities that are essential for reducing the fear of crime, such as the rapid response to crime issues, crime prevention, investigation activities, and post-crime policing. Table 2 shows that the number of people under the charge of a police officer in Korea is considerably higher than that in other countries that are part of the Organization for Economic Cooperation and Development [12]. These factors have led to a steady increase in the private security industry in Korea, which currently meets the increasing demand for personal security. From 2005

onward, the private security industry has already exceeded the number of national police personnel, and the number of private security guards was 153,767 as of 2015. This number is more than 35.98% higher than the national police workforce of 113,077 in Korea [1, 11, 13].

In addition, rapid changes in the social structure have increased the burden on national security. The factors affecting the changes in the policing environment include the social, technological, environmental, demographic, political, and economic factors. The future policing environment is expected to be very complicated and unpredictable. Hostility, hatred, antagonism, jealousy, and violent discourses that easily spread through the social media trap people in a culture of distrust [14. 15]. The cultural crisis in society is expected to increase with the worldwide decay of social norms, values, legal systems, ethics, and morality.

Table 2. The number of population per police officer by countries in 2012

Country	The Number of Population
Korea	498
Japan	498
Australia	413
UK	381
USA	354
Germany	310
France	273
Hong Kong	252

A major reason for the future being so very complex and unpredictable is the data explosion. Technologies related to big data and artificial intelligence (AI) have become very advanced; intelligent technology has become accessible to common people because of the Internet of Things (IoT) technology. Society is now based on super connectivity. In developed countries such as the United States, Japan, and the United Kingdom, efforts are being made to find technological alternatives to overcome the structural limitations faced by the police force. For instance, developed countries use the Smart Policing initiative, which applies science and technology to the field of security; this has proved to be very effective.

The phenomenal technological changes that have occurred in the last few years have been termed the "Fourth Industrial Revolution." On January 21, 2016, the chairman of Klaus Schwab presented a paper titled Mastering the Fourth Industrial Revolution at the 2016 Davos Forum in Switzerland. The Davos Forum was significant because it paid attention to the revolutionary changes that could occur in the social structure through the Fourth Industrial Revolution. The Science and Technology Summit in October 2015 declared that the status of science and technology should be shifted from its peripheral, supplementary role to a central and core position [16].

The era of Fourth Industrial Revolution is marked by hyper-connectivity and hyper-intelligence. We need to determine the scientific preventive activities that the Korean police could perform in this scenario to resolve or reduce the fear of crime. In the domestic academic world, there have been discussions about the psychological approach and the interpretation of the fear of crime. Applications have been developed to solve the fear of specific crime types and analyze the social control phenomenon caused by distorted media reports on specific crimes [17, 18, 19]. Our society is already quickly becoming hyper-connected and hyper-intelligent because of IoT, cloud computing, and so on. The rapid development and diffusion of information communication technology has exponentially expanded the connectivity among human beings and objects, thereby further strengthening hyper-connectivity. We are already experiencing a hyper-intelligent society because of the linkage and fusion of AI and big data. To prevent the fear of crime that people feel during these

global changes, our police also needs to use the latest technological advances for crime prevention [11].

Therefore, it is necessary for the police to reduce the fear of crime that the public receives from exposure to the media. To minimize the social costs caused by the crime, it is necessary to monitor the news related to the crime and prepare a preemptive response policy before the fear takes hold of the person. In this paper, we propose a news classification system based on deep learning, which helps the police to efficiently analyze the news related to the crimes reported in the media and classify them by crime category; this can help the police to prepare a quick and detailed response policy.

2. System Flow and Pseudo Code

The proposed system is divided into two parts as shown in Figure 1. The first is the preprocessing process. The preprocessing process is a process of refining crime data and news data collected as unstructured text data into a form that can be analyzed. In the case of crime data, morphological analysis for NLP is performed as described above, and the process of performing feature extraction for machine learning is included. Through this process, crime data generates crime related corpus and learning data set for machine learning. In the case of news data, since there is very little news related to crime in all collected news articles, a validation data set is selected to test the precision of the classification model. After the validation data set is selected, the classification is performed using the remaining data.

The second is the machine learning process. In the machine learning process, learning is performed using a learning data set. To validate the performance of the learned classification model, we use the validation data set selected in the previous step. Classifying the test data set through a validated model and verifying the precision and recall of the classification results to verify the classification results. However, since the crime data used in the learning data set in this study is not data that includes the crime type of the whole crime, it is impossible to classify all the crime types. Therefore, it is necessary to acquire a learning data set of this type in order to classify the news that is related to a specific type of crime type such as violent crime such as murder or assault or political crime. Since the classification system constructed through this study includes the data processing process from data collection to classification, it can be continuously used if only the learning data set is replaced in the future. The python module was created for each unit. The Python module used to drive the entire system can be divided into four parts. Each is divided into data collection, data pre-processing, data learning and data classification.

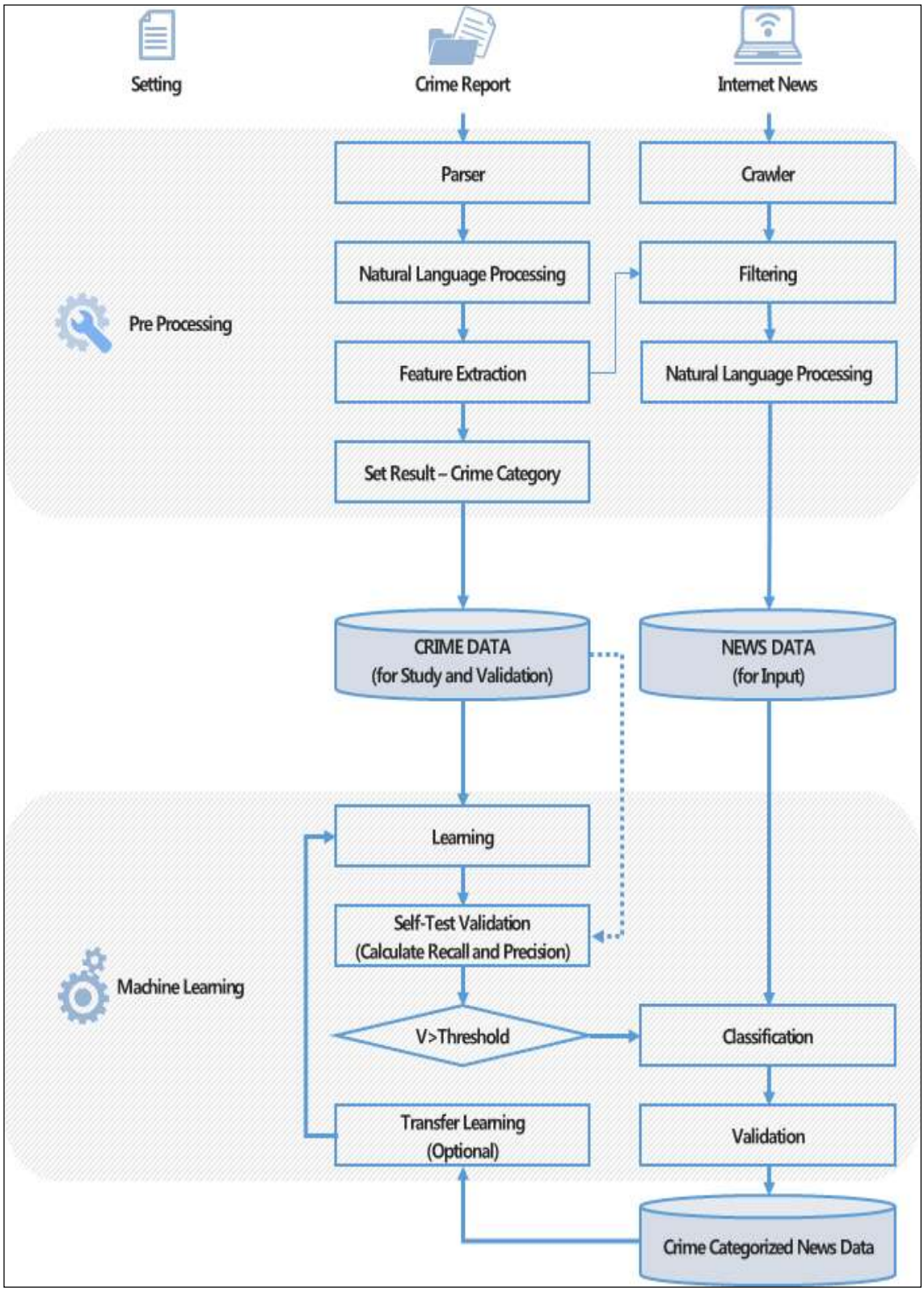

Figure 1. System flowchart

2.3 Used Data and Pre-Processing

To allow research to be conducted within a reasonable range, some of the crime data from 2006 to 2016, in which the values related to sensitivity information such as personal information was removed, were used for learning. The structure of the data was grasped prior to learning the full classification model. Data for 223,552 criminal cases were used in the study for a total of 11 years. Among the data used in the study, crime data is a total of 216,761, and the data is used as learning data to learn models for classifying news articles by crime type. To learn the model, it is necessary to refine the data into a form that can be learned. First, since there are 100 fields in raw data, only the data fields necessary for learning are extracted. The fields used in the study are nine crime categories of major crime type fields and crime report fields in which the crime facts are described as unstructured texts.

Natural language processing involves obtaining a TF-IDF to extract important words from a set of documents. In this study, two major steps are performed to select words with high TF-IDF in the crime report. First, POS tagging process. When POS tagging is performed through the morpheme analyzer, the sentence is decomposed into morpheme units, and meaningful words are extracted. First, after selecting the words tagged with nouns after going through the morpheme analyzer, 90,618 nouns are selected. The second step is to calculate the TF-IDF to determine how important the selected words are within the document set. Among the 90,618 words for which the TF-IDF was calculated, the words that best represent the corpus characteristics of the crime report group were largely compressed to the upper 2,000 words. The reason why the top 2,000 words can be compressed is as follows. The histogram of the calculated TF-IDF distribution shows that 98% or more of the words are distributed at a significantly lower TF-IDF value. This phenomenon can be interpreted to mean that a large number of words appear once in a document for each crime report. For example, the victim, the name of the suspect, or the address of a specific place are very rare, so the TF-IDF value is relatively small, and these words do not represent the features. Therefore, the top 2%, that is, about 2,000 words, is a feature that describes the behavior of crime reports.

However, among the top 2,000 keywords, unnecessary keywords still exist. For example, TF-IDF is somewhat higher because of the large-scale geographical distribution (such as Seoul and Gyeonggi-do), but it is difficult to represent the characteristics of crime types. Another example is the words of instruction pronoun, company name, and so on. These words were removed according to certain rules, and 2,000 features processed through these exception processing procedures were used for the learning process. Figure 2 is a visualization of the top 100 words of the 2,000 selected words.

Figure 2. Among the heuristically selected crime-related features, the top 100 keywords with high TF-IDF

2.4 Machine Learning

In order to create a classification model for crime types, it is necessary to transform the crime data into a form that can be learned. The process of vectorizing 223,552 crime fact documents by fixing the width of the input layer by the number of Feature 2,000 extracted above. After that, the data used for learning and the data to be used for the self-validation are classified, and 25% of the data is used as a verification dataset at random. Figure 3 shows the structured and visualized ratio of the learning dataset to the validation dataset. As a result of testing the classification of nine crime types with actual crime data, the average precision was 89% and recall was 86%. However, since this study aims to classify news data, we have included a part of news data in learning.

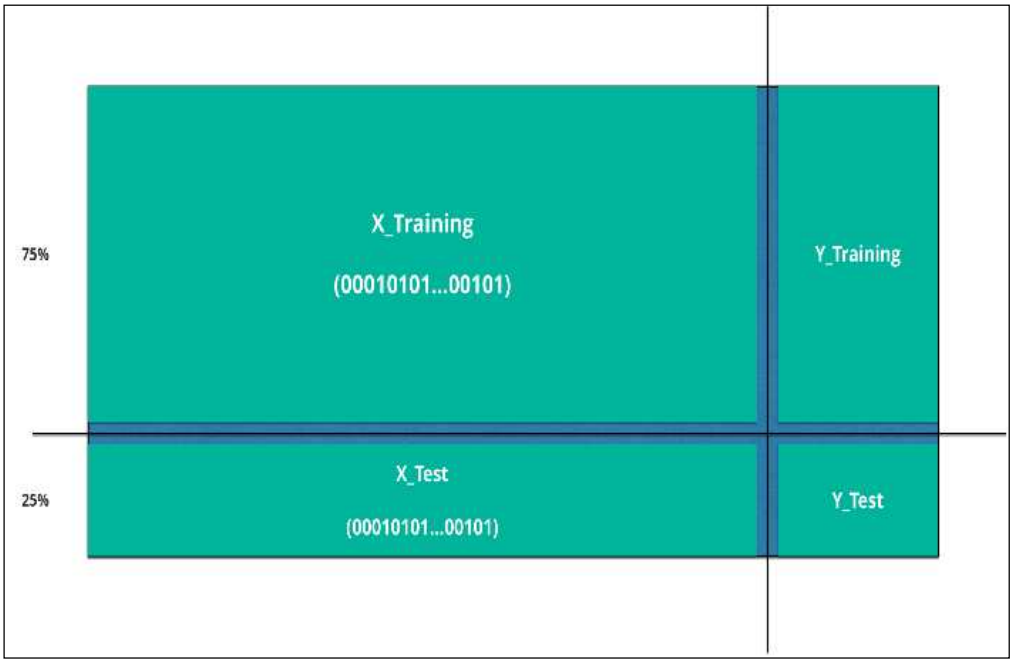

Figure 3. The ratio of the learning dataset to the verification dataset

However, since the number of articles in the newspaper articles is larger on weekdays than on weekends, weekends are excluded, and monthly, mid and late are included evenly in consideration of the repeatability of newspaper articles. The date selected to select the validation dataset based on these criteria is shown in Figure 4. Heuristic analysis among news articles of selected date again selects news corresponding to nine crime categories. The data set for screening has been reduced to 10%, but since the frequency of news related to crime is still low, it is necessary to handle the crime related news in a roughly searchable way. Therefore, we used the crime-related vocabulary extracted from the feature extraction process to select only news scripts related to crime. Using the generated vocabulary, I created a branch statement, skipped if the news word appeared less than 5 times in the news script, and stopped when it appeared more than 5 times. After the researcher has read the news selected as being related to the crime through the module, he or she will select which of the 9 categories belong to the category.

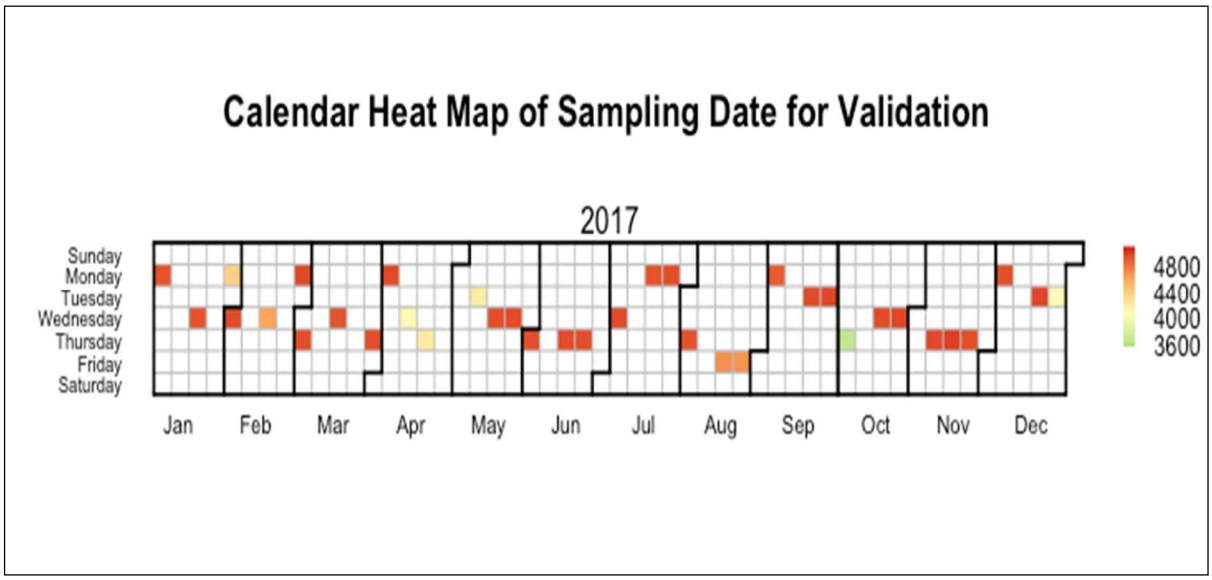

Figure 4. Calendar Heat Map of Sampling Date for Validation

A deep learning model was constructed to learn the generated dataset. There are three major factors for model configuration. The first is the activation function of the input layer. The input layer activation function selected ReLU (Rectified Linear Unit) to solve the vanishing gradient problem. The connections between layers were fully connected, and the dropout was adjusted to 0.2. The activation function for outputting the predicted value of the machine learning was selected by selecting softmax. Figure 5 is a visualization of the designed learning layer.

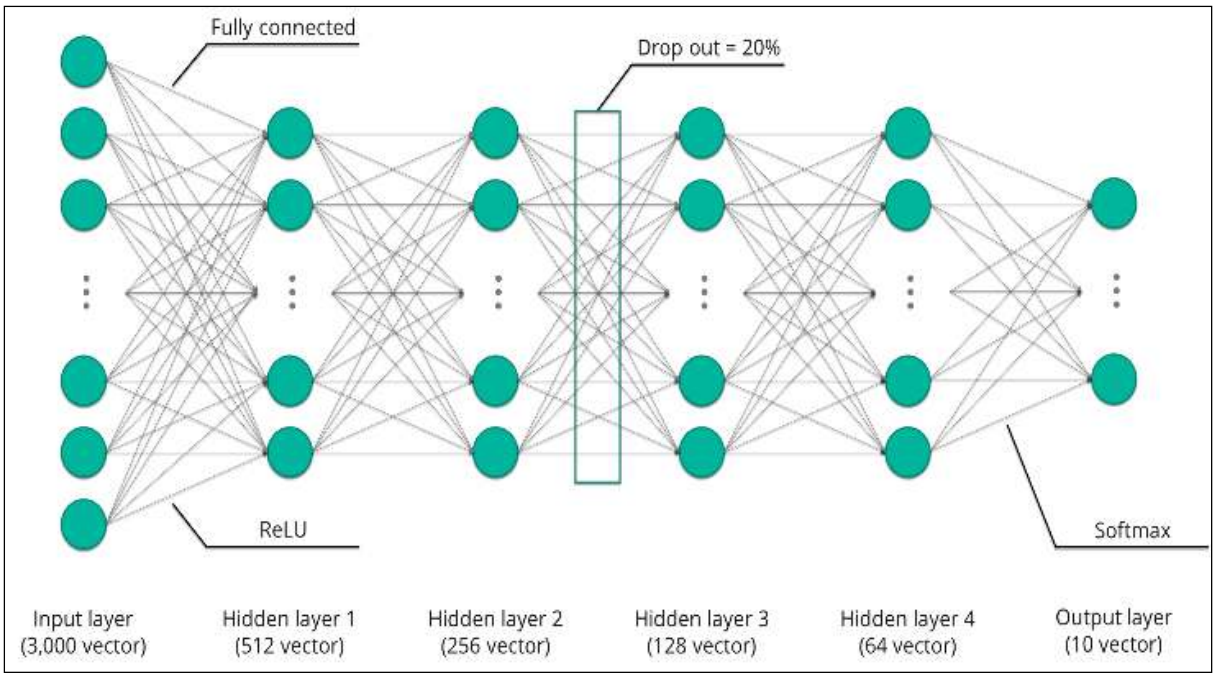

Figure 5. Visualization of the designed learning layer

3. Verification

We conducted the learning through the deep learning model which designed the selected learning data. The learning was repeated up to 100 generations of the epoch, and the results of each epoch are shown in Figure 6.

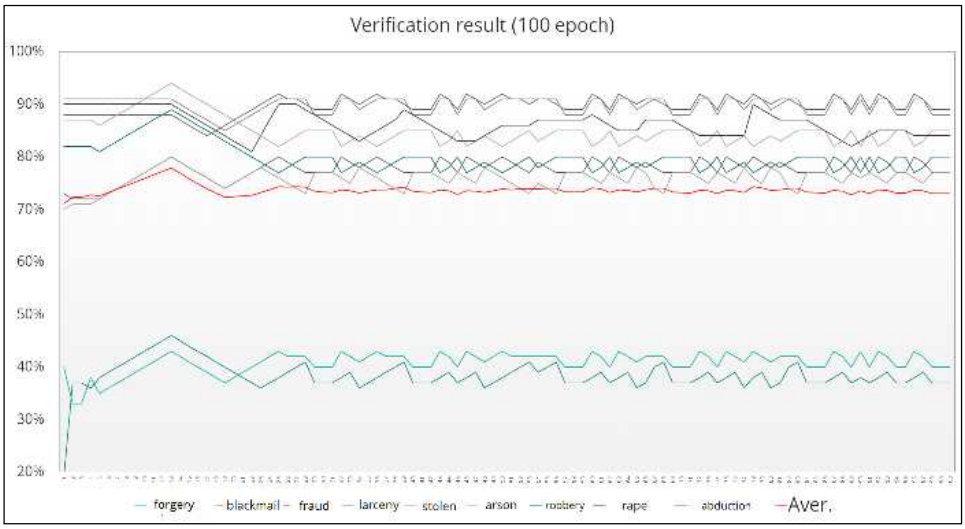

Figure 6. Verification result (100 epoch)

Because of learning up to 100 epoch, we confirmed that performance converged. Table 3 shows the classification performance of the models learned up to 100 epochs. In the case of classification accuracy, only the order of similarity is judged. The average F1-score is 84%, but forgery and abduction are 20% lower than the average. Additional learning of 10 epochs was performed separately to improve classification performance for both types.

Table 3. 100 epoch results in precision and recall

	Category	Precision	Recall	F1-score
1	forgery	0.61	0.59	0.60
2	blackmail	0.82	0.84	0.83
3	fraud	0.95	0.88	0.91
4	larceny	0.94	0.94	0.94
5	stolen	0.93	0.72	0.81
6	arson	0.99	0.88	0.93
7	robbery	0.88	0.92	0.90
8	rape	0.93	0.94	0.94
9	abduction	0.80	0.57	0.67
Total		0.87	0.81	0.84

4. Classification Result

The classification of the news articles collected during the first year of 2017 was carried out through the classification model for each type of crime that we created. We have implemented the system to classify news

articles collected daily, and the classified results are shown in Figure 7.

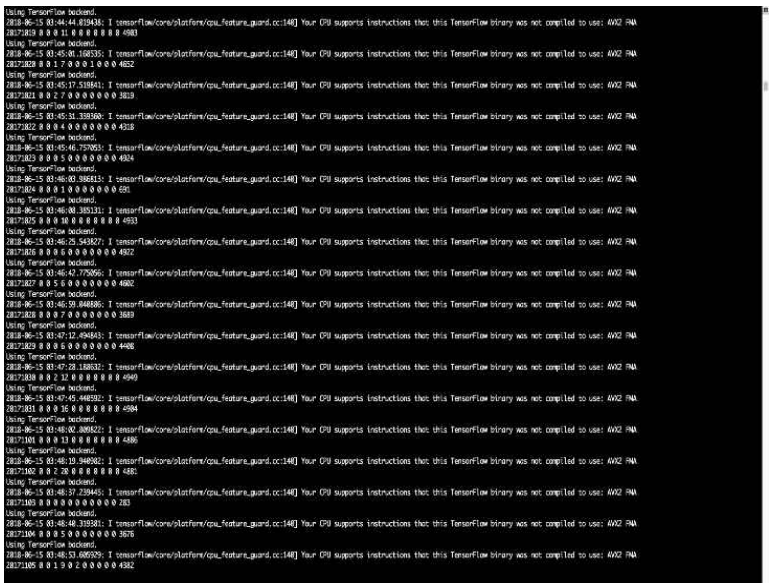

Figure 7. News article classification screen by date

Table 4 shows the results of counting the similarity rankings among the total classification results. In total, 3,533 news stories related to the crime were found among 1,593,938 news articles collected over one year. As a result of analyzing classified articles, there have been cases where articles that introduce crimes were introduced, dramas, films, and so on. In case of two or more crime features at the same time, classification accuracy is significantly lowered.

Table 3. Classification results of news articles collected in 2017 (similarity rank 1)

	Category	Count
1	forgery	20
2	blackmail	17
3	fraud	325
4	larceny	3041
5	stolen	12
6	arson	60
7	robbery	11
8	rape	37
9	abduction	10
10	NONE	3,533

5. Conclusion

In this paper, we propose a system to classify news articles based on the crime type by using unstructured data generated by the police. We have not been able to use learning data that includes all types of crimes because of the structure of the Korean police force and the legal and institutional limitations. However, we have implemented a pilot system for collecting, refining, and analyzing data; therefore, this system can be applied to various areas if the data can be secured in the future. The results of the classification system using limited data shows that the police force uses only the information it requires to carry out the policing activities.

Of the 1.5 million news stories on the Internet, only a few news stories are related to crime. Given the ratio of the news related to the crime, our model can be applied to police activities if the model is improved by securing future data.

However, there is a need for further study on the exception handling in classifying news data. When we look at actual classified results, it is often difficult to see them as news related to crime. First, the crime news is also related to life and culture, and often it introduces works such as novels, movies, and dramas related to crime. This type of classification that emerges primarily from the life and cultural aspect needs to be filtered using a different exception handling. Second, the classification of crime news is a concern when various incidents occur simultaneously. In actual crime cases, more than two types of crimes occur more frequently than only one type of crime. Therefore, follow-up studies are required on various methods, such as classification priority or duplicate tags.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2020-0-00901, Information tracking technology related with cyber crime activity including illegal virtual asset transactions)

Reference

[1] Korea National Police Agency, *Police Statistical Yearbook 2016*, Korea National Police Agency, vol. 60. pp. 106-111, 2017.
[2] S. Murray and C. Clague, *SAFE CITIES INDEX 2017: Security in a rapidly urbanising world*, The Economist Intelligence Unit Limited 2017, pp. 5-7, June 2017.
[3] J. Yang, "News Representation of Crime: The Case of Sexual Violence against Children", *Journal of Communication Science*, Vol. 10, No. 2, pp. 343-379, 2010.
[4] E. Hwang and J. Cheong, "A Macro-Level Study on the Cause of Homicide Rate: Nationwide Analysis Using Spatial Regression Model", *Korean Criminological Review*, Vol. 22, No. 1, pp. 157-184, 2010.
[5] Statistics Korea, *REPORT ON THE SOCIAL SURVEY 2016*, Statistics Korea, Vol. 22, 2016.
[6] J. Garofalo, "The Fear of Crime: Causes and Consequences", *The Journal of Criminal Law and Criminology*, Vol. 71, No. 2, pp. 839-857, 1981.
[7] C. Hale, "Fear of Crime: A Review of the Literature", *International Review of Victimology*, Vol. 4, No. 2, pp. 79-150, 1996.
[8] J. E. Conklin, *The impact of crime*, New York: Macmillan, pp. 294,1975
[9] M. Warr, "Fear of Rape among Urban Women", *Social Problems*, Vol. 32, No. 3, pp. 691-702, 1985.
[10] H. Kim, H. Cho, and W. Min, "Estimates of the Social Costs of Crime in Korea: considering the seriousness index of crime", *Korea Social Policy Review*, Vol. 17, No. 2, pp. 163-199, 2010.
[11] H. Kim, "Development Plan of Korea's Private Security System: The Comparison of UK's Private Security System", *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol. 6, No. 11, pp. 463-470, 2016.
[12] J. Kim, "Seeking strategic countermeasures against future social change in the era of the 4th Industrial Revolution", *KISTEP InI(Inside and Insight)*, Vol. 15, pp. 45-58, 2016.
[13] T. Jones and T. Newburn, *Plural policing: A comparative perspective*, Psychology Press, pp. 34-54, 2006.
[14] L. Health and K. Gilbert, "Mass media and fear of crime", *American Behavioral Scientist*, Vol. 39, No. 4, pp. 379-386, 1996.
[15] D. Romer, K. H. Jamieson, and S. Aday, "Television news and the cultivation of fear of crime", *Journal of communication*, Vol. 53, No. 1, pp. 88-104, 2003.
[16] P. Jang, "2016 Davos Forum: What are our strategies for the forthcoming 4th industrial revolution", *Science & Technology Policy*, Vol. 26, No. 2, pp. 12-15, 2016.

[17] H. Cha, “The Theoretical Implications of on the Fear of Crime”, *Korean Criminal Psychology Review*, Vol. 10, No. 2, pp. 241-257, 2014.
[18] S. Jun, Y. Kang, and E. Ko, “Research of Application for Reduce Women's ‘Fear of Crime’: Focused on Street harassment”, *Proceedings of HCI Korea 2018*, pp. 740-744, 2018.
[19] I. Kwon and H. Lee, “Fear of Sexual Violence and Social Control: Focusing on the media's treatment of child sexual violence cases”, *The Journal of Asian Women*, Vol. 50, No. 2, pp. 85-118, 2011.