electronics

MDPI

Article

Key Information Extraction for Crime Investigation by Hybrid Classification Model

Yerin Park, Ro Seop Park and Hansoo Kim *

Department of Forensic Information Science and Technology, Hallym University, Chuncheon-si 24252, Republic of Korea; iisabelle559@gmail.com (Y.P.); rspark@hallym.ac.kr (R.S.P.)
* Correspondence: kutestar@hallym.ac.kr

Abstract: The 2021 amendment to South Korea's Criminal Procedure Law has significantly enhanced the role of the police as investigative authorities. Consequently, there is a heightened demand for advanced investigative expertise among the police, driven by an increase in the number of cases each investigator handles and the extended time required for report preparation. This situation underscores the necessity for an artificial-intelligence-supported system to augment the efficiency of investigators. In response, this study designs a hybrid model that fine-tunes two Transformer-based pre-trained language models to automatically extract 18 key pieces of information from legal documents. To facilitate this, "The Major Information Frame of Homicide Criminal Facts" was developed, and a large-scale training dataset specialized in the criminal investigation field was constructed. The hybrid classification model proposed in this research achieved an F1 score of 87.75%, indicating superior performance compared to using a single machine reading model. Additionally, the model's top three predicted answers included the correct answer at a rate exceeding 98%, demonstrating a high accuracy level. These results suggest that the hybrid classification model designed in this study can play a crucial role in efficiently extracting essential information from complex legal and investigative documents. Based on these findings, it is confirmed that the hybrid classification model can be applied not only in drafting investigative reports but also in tasks such as searching for similar case precedents and constructing case timelines in various legal and investigative applications. The advancement is expected to provide a standardized approach that allows all investigators to perform objective investigations and hypothesis testing, thereby enhancing the fairness and efficiency of the investigative process.

Keywords: deep learning; natural language processing; criminal facts; key information of crime; information extraction

Citation: Park, Y.; Park, R.S.; Kim, H. Key Information Extraction for Crime Investigation by Hybrid Classification Model. *Electronics* **2024**, *13*, 1525. https://doi.org/10.3390/electronics13081525

Academic Editor: Arkaitz Zubiaga

Received: 15 March 2024
Revised: 9 April 2024
Accepted: 14 April 2024
Published: 17 April 2024

1. Introduction

The annual increase in the workload of investigators has consistently emerged as a significant social concern. Furthermore, the 2021 amendment to the Criminal Procedure Law in South Korea, which adjusted the investigative rights between the police and the prosecution, has expanded the police's investigative authority and scope. Now, the police can conclude most cases with a non-prosecution decision, except in instances where there is no suspicion of a crime. Such changes have heightened the demand for police investigative capability and transparency, leading to efforts to enhance investigative expertise through the establishment of a national investigation headquarters and improvements in the educational system [1]. Despite these institutional efforts, the problems of insufficient investigator capability and manpower shortages remain unresolved. Moreover, the number of cases handled by investigators has increased by approximately 26% compared to 2017, following the adjustment of investigative rights, leading to a heavier workload than in the past [2]. This situation adversely affects the efficiency and expertise of investigators, with delays in investigations due to heavy workloads posing a significant issue. These challenges cannot be resolved merely through the introduction of new systems, enhancement

of expertise, or staffing solutions but require the adoption of technologies that surpass human information processing limitations [3]. Particularly in the investigative environment, where complex relationships and vast amounts of information coexist, there is a high likelihood of errors, and investigators are at risk of confirmation bias, focusing on incorrect or irrelevant information [4]. Furthermore, investigators must perform tasks such as drafting investigation reports and administrative duties, where spending excessive time can lead to insufficient understanding of case information and logic, potentially resulting in errors. A survey conducted in late September 2021 revealed that investigators in the economic teams of 38 sample police stations affiliated with 18 regional police agencies nationwide spend an average of 72.377 h investigating a single case, with suspect interrogation and investigation reporting being the most time-consuming tasks, requiring an average of 9.262 h and 9.236 h, respectively [5]. Investigators must draft investigation reports based on a large volume of investigative documents, involving legal analysis based on legal requirements, sentencing standards, and investigative principles, as well as comparing similar precedents. This comprehensive process demands significant time and effort, making it challenging for investigators to thoroughly verify their hypotheses [6]. Thus, an optimized artificial intelligence (AI) support system is necessary for rapid task processing. AI support is essential for speeding up the analysis and discovery process in new paradigms [7]. With the recent introduction of AI technology in the legal field, there has been an increase in the use of deep learning models for tasks such as verdict searching, contract management, and legal document translation [8]. Government initiatives are also underway to apply AI technology to electronic litigation systems like Smart Court 4.0 [9]. However, while services like CLUE (Crime Layout Understanding Engine) [10] offer crime trend analysis and crime prediction, and there are AI-based research and projects providing case law search functions to police agencies, the sufficiency of their performance for actual use by investigators remains unclear [11].

Recent increases in research utilizing deep learning models like Transformer-based pre-trained language models for fine-tuning in legal document natural language processing tasks have led to automated contract review and analysis, legal document classification, voice phishing crime entity recognition, and the extraction and visualization of major crime information from legal documents that are challenging for humans to analyze automatically [12–17]. However, current public or ongoing research and projects have not yet clearly demonstrated their practical applicability in the field, and there is a scarcity of research supporting investigators in drafting investigation reports within the legal and investigative domain. Therefore, this study proposes a hybrid classification model that automatically extracts essential crime information from investigation documents, necessary for drafting investigation reports. For this purpose, this research defines the key information required in investigation reports, constructs and utilizes a large dataset suitable for the crime investigation domain, and fine-tunes two pre-trained language models. The ultimate goal is to improve the performance of the hybrid classification model and the data processing technology to enable a system capable of searching for similar case precedents based on extracted key information. Utilizing this system will enable more efficient and detailed drafting of investigation reports.

2. Materials and Methods

2.1. Data Collection

The objective of this study was to automate the extraction of key information essential for drafting investigation reports from a vast array of investigative documents using pre-trained language models. For the extraction of such critical information from investigative documents, this research has selected verdicts accessible to the public, including statements recorded in investigative documents and facts reported, as the learning data.

Verdicts serve as documents that summarize and prove the contents of a trial, with both the first and second trials being fact-finding stages where factual aspects of a case and the corresponding legal judgments are determined. Notably, when a case is adjudicated

as guilty, the facts must be clearly marked, and the items and manner of recording are akin to those in an investigation report [17].

According to a survey among investigators, it was found that many of them refer to first and second trial precedents similar to the cases they are handling to understand the issues of the case and reference the criteria for guilty verdicts [18]. Particularly, the structure and content of first trial verdicts are most similar to investigation reports. Additionally, the criminal facts recorded in the verdicts resemble those included in investigation reports and contain most of the elements mandatorily recorded during investigations; hence, this study has chosen the criminal facts of first trial verdicts as the training data.

Up to the point of this research, the collected data include first trial criminal verdicts corresponding to serious crimes such as robbery, theft, sexual violence, fraud, and murder, with the data used in this study being first trial murder verdicts. Murder cases have a significant impact on society, and murder verdicts contain a variety of essential information based on elements of the crime, making them suitable for use; thus, 1500 first trial murder verdicts were utilized.

2.2. Definition of Key Information

As the goal of this study was to explore a model that can automatically extract key information from investigation documents, it was necessary to define what constitutes key information. Hence, 21 pieces of key information that can be extracted from investigation documents and verdicts were defined, although, due to the nature of verdict data, information about the defendant was not included, as opposed to victim information, which was included. Therefore, this study extracted only 18 pieces of key information, excluding three that pertain to the defendant (name, gender, age), which can be categorized into 'fundamental information', 'pre-crime information', and 'in-progress crime information'. These systematically reflect the progression of an incident and together form the 'Key Information Frame for Murder Case'. This frame was constructed with the assistance of police practitioners and legal experts, and it is crucial to record the actions and outcomes of the suspect in a chronological sequence, distinguishing between before and during the crime [17]. Each piece of key information was defined based on the essential items for drafting investigation reports as defined in investigation report writing techniques, including the crime subject, suspect's identity and criminal record, crime timing, crime location, motive/cause of the crime, victims/complainants/accusers/petitioners, means/method of crime, criminal act and result, intent/purpose, negligence, attempted/preparatory/conspiracy, and accomplices. Thus, based on the advice of police practitioners and legal experts and investigation report writing techniques, a total of 21 pieces of key information were defined, and the 'Major Information Frame of Homicide Criminal Facts' is illustrated in Figure 1.

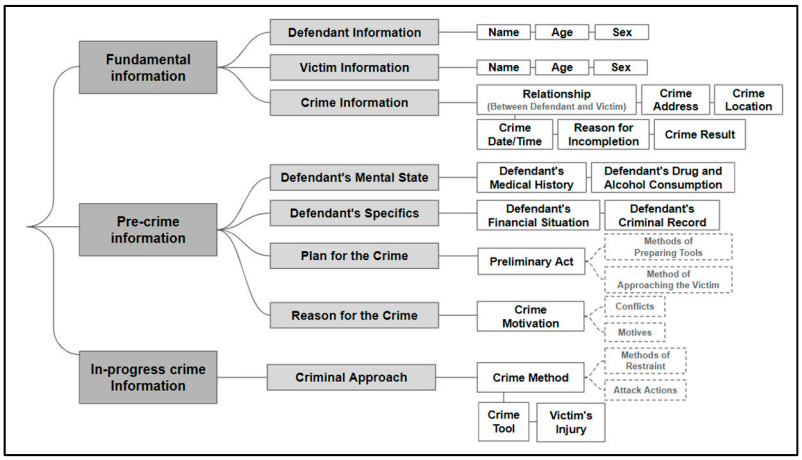

Figure 1. Key Information Frame for Murder Case.

2.3. Dataset Building

2.3.1. Building Procedure

This research aimed to automate the extraction of 18 out of the 21 key pieces of information mentioned earlier, excluding three pieces (suspect's name, gender, age) that are considered personal information and are redacted from verdict documents, from the criminal facts of 'murder' verdicts using pre-trained language models. Constructing a high-quality dataset is essential for training these models, which requires significant time, cost, and human resources. Specifically, building a dataset for specialized domains related to law or investigations demands even more resources and expertise [19]. To address these challenges, there has been a growing trend in using conversational AI like Chat GPT, recently released by OpenAI, for dataset construction. Among recent studies, one utilized the GPT-3 API (Application Programming Interface) to overcome the difficulties of data annotation within a limited budget, proposing a method that efficiently constructs datasets by leveraging both humans and GPT-3 [20]. Additionally, research has demonstrated that GPT-3.5 outperforms trained annotators or crowdsourcing in terms of accuracy and cost for document classification tasks [21]. Such studies employing pre-trained generative models like GPT for dataset construction underscore the importance of finding the optimal prompt for the task at hand to achieve high performance. A study that detailed the process of finding the optimal prompt used Chat GPT to solve law school exam questions and compared three types of prompts in an experiment [22]. The results showed significant performance differences depending on how the prompts were structured, proving that GPT could surpass human capabilities in various generative tasks with the optimal prompt.

In this study, the GPT-3.5 API (hereinafter referred to as GPT-3.5) was employed to significantly reduce the vast resources required for dataset construction and to build a high-quality dataset more efficiently. To ultimately generate the best output using GPT-3.5, utilizing the optimal prompt and subsequently verifying the work by humans are crucial. Therefore, this research constructed a golden dataset through a primary annotation with GPT-3.5 followed by a secondary human annotation and final review, as illustrated in Figure 2.

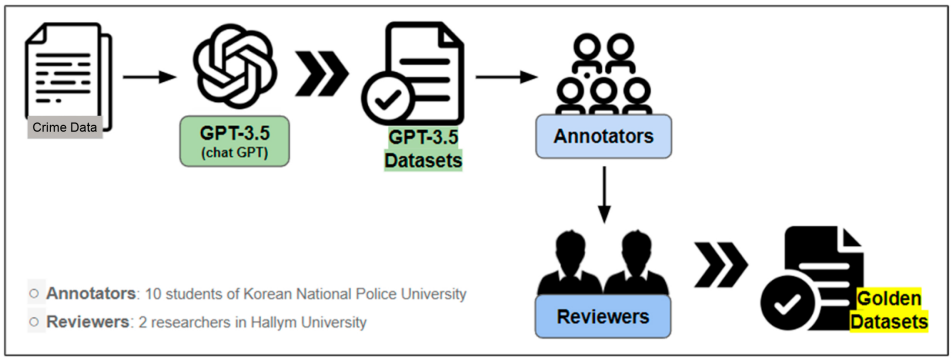

Figure 2. Procedures for building a dataset using GPT-3.5.

The key to constructing a dataset using GPT-3.5 lies in enhancing the accuracy of GPT-3.5 to minimize the time required for secondary annotation and verification. Therefore, to conduct the key information extraction task with GPT-3.5 under optimal conditions, three prompts were compared to determine the most effective one. The first prompt constructed in this research was a ① Simple prompt, which lists brief task information and key information. The second prompt was a ② One-Shot prompt, which includes task information, an example for each key information, and conditions to consider during extraction. The final prompt, a ③ Few-Shots prompt, was developed with task information and more than two examples for each piece of key information, along with specified conditions. Some examples of the three prompts are as shown in Tables 1–3.

Table 1. Simple prompt.

<Task Description> I want you to extract information from crime-related text in Korean based on the key information I give you. <Key Information> 1. motive of crime 2. injury of victim … If you can't find matching information, don't infer anything and simply add "없음 (None)".

Table 2. One-Shot prompt.

<Task Description> I want you to extract information from crime-related text in Korean based on the key information and an example I give you. <Key Information and Example> 1. motive of crime—e.g., "decided", 2. injury of victim—e.g., "multiple lacerations, etc." … <Conditions> If you can't find matching information, don't infer anything and simply add "없음 (None)". ✔Tag must be in English and extracted information must be in Korean. ✔It is VERY important that you extract the information word by word. ✔NEVER summarize, rephrase or translate the given text. Tag the following text:

Table 3. Few-Shots prompt.

<Task Description> I want you to extract information from crime-related text in Korean based on the key information and examples I give you. <Key Information and Examples> 1. motive of crime—e.g., "enraged by the thought that the victim was ignoring him/her", "decided to commit murder, fueled by anger when the victim got mad and refused to listen to him/her", etc. 2. injury of victim—e.g., "multiple lacerations, etc.", "asphyxiation due to neck compression", "severe head injury" … <Conditions> If you can't find matching information, don't infer anything and simply add "None". ✔Tag must be in English and extracted information must be in Korean. ✔It is VERY important that you extract the information word by word. ✔NEVER summarize, rephrase or translate the given text. Tag the following text:

To identify the most effective prompt among the three, the annotations derived from each prompt were compared with a truth dataset to evaluate their performance. The truth dataset used for the prompt experiment consisted of criminal facts from 200 first trial murder verdicts, annotated initially by 10 researchers from the Police University, followed by a secondary annotation and verification process conducted by two reviewers with substantial domain knowledge and research experience. The performance of the three prompts was assessed using ROUGE-L, which measures the longest common subsequence that is not necessarily contiguous, along with recall, precision, and the F1 score. The results showed that the Few-Shots prompt achieved the highest overall performance with a recall of 61.85%, a precision of 78.01%, and an F1 score of 68.97% as shown in Table 4. Consequently, the primary annotation in this study was conducted using the Few-Shots prompt.

Table 4. Prompt performance comparison.

ROUGE-L	Simple	One-Shot	Few-Shots
Recall	54.59%	54.85%	61.85%
Precision	57.11%	71.66%	78.01%
F1 score	50.35%	57.48%	68.97%

This study utilized Doccano [23], an annotation tool, for the secondary annotation and verification of the primary annotation results generated using GPT-3.5. Following the upload of the GPT-3.5 annotated dataset to Doccano, twelve researchers carried out the secondary annotation and verification, culminating in the creation of the final training dataset.

Employing GPT-3.5 for annotating 1500 verdict documents resulted in an F1 score of 68.97%. Although this score might not be considered high from a quantitative metric standpoint, it facilitated a significant reduction in both time (51.67% reduction, from 375 h to 181.25 h) and cost (49.53% reduction, from KRW 3,697,500 to KRW 1,866,025), as shown in Table 5. This demonstrates the efficacy of integrating advanced AI tools like GPT-3.5 with human annotation processes, achieving considerable efficiency gains in terms of both time and cost without compromising the quality of the dataset significantly.

Table 5. Effects of GPT.

Annotation without GPT		Annotation with GPT	
Time	**Total 375 h** (time for case × total number of cases) ⇒ 15 min × 1500 cases = 22,500 min (375 h)	**Total 181.25 h** (1st annotation time + 2nd annotation time) ⇒ 375 min + 10,500 min = 10,875 min	
		1st Annotation Time (time for case × total number of cases) ⇒ 0.25 min × 1500 cases = 375 min	**2nd Annotation Time** (time for case × total number of cases) ⇒ 7 min × 1500 cases = 10,500 min
Cost	**Total KRW 3,697,500** (minimum wage × annotation time) ⇒ KRW 9860 × 375 h	**Total KRW 1,866,025** (GPT-3.5 Fee + Annotator Minimum Wage Cost) ⇒ KRW 78,900 + KRW 1,787,125	
		GPT-3.5 Fee (cost for case × total number of cases) ⇒ KRW 52.60 (USD 0.04) × 1500 cases = KRW 78,900	**Annotator Wage Cost** (minimum wage × annotation time) ⇒ KRW 9860 × 181.25 h = KRW 1,787,125

2.3.2. Token and Sequence Datasets

In this study, a hybrid model was designed by combining a token classification model and a sequence classification model. The key information extractable by these two models is categorized as shown in Table 6. The eight token-type pieces of key information could be nouns or noun phrases, including victim's name, age, perpetrator–victim relationship, crime address, date and time, and crime tool. The remaining ten sequence-type pieces of key information, such as preparatory acts, criminal acts, motives for the crime, and outcomes, could be extracted as phrases.

Table 6. Token and sequence key information.

Token (8)	• Victim Name; • Victim Age; • Victim Sex; • Relationship (Between Defendant and Victim); • Crime Tool; • Crime Address; • Crime Place; • Crime Date/Time.
Sequence (10)	• Defendant's Medical History; • Defendant's Financial Situation; • Defendant's Criminal Record; • Defendant's Drug and Alcohol Consumption; • Preliminary Act; • Crime Method; • Crime Motivation; • Crime Result; • Reason for Incompletion; • Victim's Injury.

The format of the token classification dataset was inspired by the WNUT 17 (Workshop on Noisy User-Generated Text 2017) NER dataset and consists of 'Tokens' and 'Tags', as illustrated in Table 7. 'Tokens' refer to the text of criminal facts, which are lists of tokens separated by spaces. 'Tags' are sets of entities assigned to the tokens composing the criminal facts. For all entities, BIO tagging (Beginning, Inside, Outside tagging) was performed to delineate the boundaries of each entity. BIO tagging assigns a B (Beginning) tag at the start of an entity, an I (Inside) tag within the entity, and an O (Outside) tag to non-entity tokens, enabling the model to clearly recognize the boundaries and types of entities.

Table 7. Token classification dataset structure.

Tokens	Tags
"Crime Details: On 31 March 2017, around 8:00 p.m., at 'E Bar' located on the underground first floor of Building D in Gangnam-gu, Seoul, the defendant, along with four colleagues including the victim F G (37 years old), was attending the defendant's farewell party. While intoxicated, the defendant got into an argument with the victim F outside the bar, grabbing each other by the collar. When the victim G, who is F's younger brother, tried to intervene and stop the altercation, the defendant, in a fit of rage, decided to stab the victims whom he had premeditated to kill. Subsequently, at approximately 11:48 p.m. on 31 March 2017, in front of 'E Bar' on the street, the defendant took out a recreational knife from the back pocket of his pants and stabbed the victim G's abdomen and left side of the neck four times each, and then stabbed the left side of the victim F's waist four times with the same knife as he tried to intervene. However, due to the intervention of other colleagues who restrained the defendant and immediately called emergency services (119), the victims were promptly transported to the hospital, preventing fatalities, and the crime remained attempted."	O O O O O O O crime_addr_B crime_addr_I crime_addr_I crime_addr_I crime_addr_I O O O v_name_B v_name_I v_name_I O v_age_B O O a_v_relation_B a_v_relation_I O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O crime_datetime_B crime_datetime_I crime_datetime_I crime_location_B crime_ location_I crime_ location_I crime_ location_I O O O O attack_tool_B attack_tool_I attack_tool_I attack_tool_I attack_tool_I attack_tool_I O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

The token classification dataset comprises a total of 269,905 token–entity pairs, which have been divided into training, evaluation, and test data at a ratio of 8:1:1. As a result, 214,688 instances were allocated to the training set, 28,097 to the evaluation set, and 27,120 to the test set. The distribution of entities (token-type key information) constituting this dataset is as detailed in Table 8.

Table 8. Statistics in token classification dataset.

Key Information	Entity Name	Training	Valid	Test
Victim's Name	v_name_B, v_name_I	2413	314	298
Victim's Sex	v_sex_B, v_sex_I	533	79	71
Victim's Age	v_age_B, v_age_I	1188	152	149
Relationship	a_v_relation_B, a_v_relation_I	1337	184	187
Crime Date/Time	crime_datetime_B, crime_datetime_I	4988	613	602
Crime Location (Place)	crime_place_B, crime_place_I	2792	350	325
Crime Address	crime_addr_B, crime_addr_I	4291	543	519
Crime Tool	attack_tool_B, attack_tool_I	5535	846	800
None	O	191,611	25,016	24,169
Sum	-	214,688	28,097	27,120

The sequence classification dataset is composed of sequence–label pairs, as illustrated in Table 9. In sequence classification, each sequence constituting the criminal facts is separated into phrases or sentences in order to assign a class to each sequence. This structure facilitates the model's ability to understand and categorize the context and content of each sequence within the criminal facts.

In this study, rule-based sentence and phrase segmentation was performed for sequence separation. Sentence and phrase segmentation consists of primary sentence segmentation using the sentence splitting function of the Kiwi library and secondary phrase segmentation through major information morpheme patterns. Kiwi is an open-source library for Korean natural language processing, providing functions primarily for morphological analysis, part-of-speech tagging, and sentence splitting of Korean texts. Particularly, the 'kiwi.split_into_sents' function is a sentence splitting feature provided by the Kiwi library, used to split the given text into individual sentences by identifying sentence boundaries within the text and returning them as separate sentences. Therefore, it can be effectively used in texts composed of languages like Korean where sentence boundaries are not clearly defined.

Once the sentence-level separation is completed using the 'kiwi.split_into_sents' function, phrase segmentation is conducted through morpheme patterns utilizing the OKT morphological analyzer. OKT is one of the morphological analysis tools provided by KoNLPy, a Python library for Korean natural language processing, offering various functionalities including morphological analysis, part-of-speech tagging, and noun extraction of Korean texts. The 'okt.pos' function used for phrase segmentation is a function that performs morphological analysis and part-of-speech tagging for the given text, breaking down the input text into morphemes and tagging the corresponding parts of speech for each morpheme. Morpheme patterns are constructed through part-of-speech analysis tagged to the endings of major information texts, consisting of morpheme patterns for endings that should be segmented into phrases such as "let's", "only did", and "after doing", as shown in Table 10. The parts of speech constituting morpheme patterns, such as "Noun", "Josa", and "Verb", are types of speech provided by the OKT morphological analyzer, distinguishing the roles of each word within the sentence.

Table 9. Sequence classification dataset structure.


```
[
  {
  "sequence": "Wearing a prepared dagger (total length 22 cm, blade length 12 cm) concealed in the vest pocket, I entered the restaurant counter.",
  "label": "prep_act"
  },
  {
  "sequence": "Asking the victim, 'Where's the restroom, are you the manager?' When the victim didn't respond immediately, feeling displeased with the attitude,",
  "label": "motive"
  },
  {
  "sequence": "immediately pulled out the dagger and swung it towards the victim's face.",
  "label": "crime_method"
  },
  {
  "sequence": "Then stabbed the victim's abdomen and slashed the victim's hands and neck as the victim defended.",
  "label": "crime_method"
  },
  {
  "sequence": "Attempted to kill the victim, but the victim resisted by grappling with the defendant and struggling, preventing the intention from being fulfilled.",
"label": "reason_incmpl"
  },
  {
  "sequence": "The attempt failed.",
  "label": "crime_result"
}
]
```


Table 10. Morpheme patterns for phrase segmentation.

Patterns	Examples
Verb +,	"let's", "however", "while doing", "but"
Josa +,	"as", "by means of"
Adjective +,	"while", "in fact", "despite", "since"
Verb + Noun +,	"after doing", "only after", "on the other hand"
Josa + Noun +,	"in the state of"
Adjective + Noun +,	"on the other hand", "during", "in the midst of"
Verb + Adjective +,	"after doing", "while", "after"

Once both primary separation using the sentence splitting function of the Kiwi library and secondary separation through major information morpheme patterns are completed, the sequence classification dataset is finalized. The sequence classification dataset comprises a total of 17,999 instances, which were divided into training, evaluation, and testing sets at an 8:1:1 ratio. As a result, 14,399 instances were allocated to the training set, 1800 instances to the evaluation set, and 1800 instances to the testing set. The distribution of the sequence-type major information (labels) composing this dataset is presented in Table 11.

Table 11. Statistics in sequence classification dataset.

Key Information	Train	Valid	Test
a_eco_bg	87	11	11
a_crim_rec	130	16	16
a_mental_con	378	48	47
a_med_rec	299	37	38
motive	3925	490	491
prep_act	693	86	87
crime_method	2231	279	279
reason_incmpl	747	94	93
crime_result	1347	168	169
v_injury	861	108	107
nan	3701	463	462
Sum	14,999	1800	1800

2.4. Model Design

The procedure for extracting (classifying) key information through a hybrid classification model is as shown in Figure 3. The model receives tokens or sequences constituting criminal facts as input and proceeds to assign one of the labels specified by the user to each token and sequence.

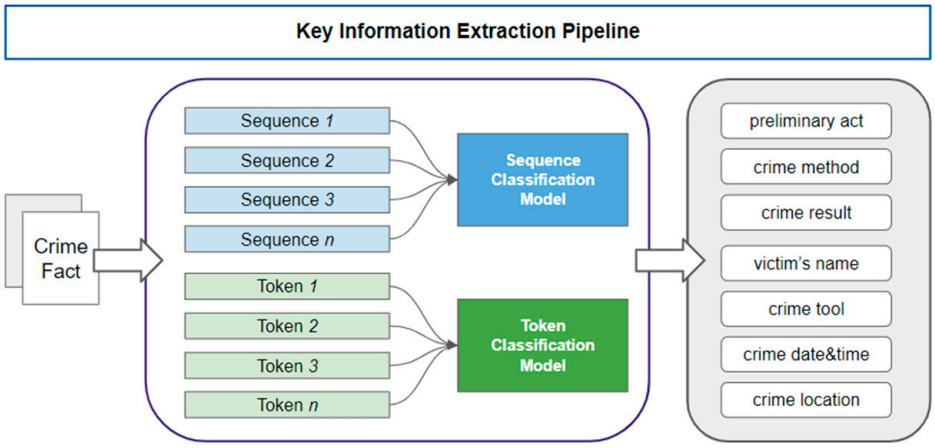

Figure 3. Key information extraction pipeline.

In token classification tasks, when the text "The defendant lived with the victim, Ms. Lee (female, 50 years old)…" is input, the model first splits the entire text based on spaces and learns the role of each token within sentences or paragraphs and its relationship with surrounding words. Then, the model calculates the probability of each token being key information and predicts the highest-probability key information. For example, it classifies "피해자 이" as "name", "여" as "gender", and "50 세" as "age". If consecutive tokens are classified as the same key information, the BIO tagging scheme is used to identify the Beginning (Begin), Inside (Inside), and Outside (Outside) of each entity.

In sequence classification tasks, when a sequence segmented based on rules is input into the model, a [CLS] token is inserted at the beginning of each sequence. Then, the model calculates the probability of each sequence being key information by understanding the contextual meaning of individual tokens composing the sequence. Finally, based on the probabilities, the sequences are classified into classes such as "motive of crime", "criminal act", and "unfulfilled intention".

3. Results

3.1. Experimental Results of Benchmark Models

The baseline for this study was a model trained on open-source datasets during both training and evaluation processes. The dataset used for the large classification baseline was the KLUE-NER dataset (Korean Language Understanding Evaluation Dataset for Named Entity Recognition) [24], which is a massive Korean dataset constructed for named entity recognition tasks. KLUE-NER includes both formal (WikiTree) and informal (Naver movie reviews) texts and consists of 21,008 instances in the training set and 5000 instances in the evaluation set, each containing entity–label pairs. KLUE-NER encompasses six types of named entities, namely, person (PS), location (LC), organization (OG), date (DT), time (TI), and quantity (QT), annotated with BIO tagging. These named entities are largely similar to the key information defined in this study. For example, person (PS) is related to 'victim's name', location (LC) to 'crime address', organization (OG) to 'crime scene', and date (DT) to 'date of crime'. Similarly, quantity (QT) is associated with information such as 'victim's age' or 'crime tool'. This relevance suggested that the KLUE-NER dataset could be effectively utilized for extracting important information in this study. Therefore, token classification experiments were conducted using the KLUE-NER dataset.

For sequence classification tasks, the dataset used was the K-MHaS dataset (Korean Multi-Label Hate Speech Dataset) [25], which is a multi-label dataset for detecting increasing hate speech in Korean online news comments. K-MHaS contains 109,692 sequence–label pairs, with each sequence classified into 1 to 4 labels. The data composing K-MHaS were collected from Twitter, Wikipedia, and Korean online news comments from January 2018 to June 2020, consisting of 78,977 instances in the training set and 8776 instances in the evaluation set, each with sequence–label pairs. The classification system of this dataset is composed of eight detailed hate speech categories, politics, origin, appearance, age, gender, religion, race, and insult. Since the sequence classification task conducted in this study was not about multiple classes but rather classification for a single class, single sequence–label pairs with only one label assigned (training 69,138, evaluation 7762) were utilized for sequence classification baseline experiments.

The models used in this study were KLUE-BERT-base, Multilingual-BERT-base, XLM-RoBERTa-base, KLUE-RoBERTa-base, and KoELECTRA-base. Commonly, all models used AdamW as the optimizer, the learning rate was set to 2×10^{-5}, the number of epochs was set to 20, and the batch size for both training and validation was set to 32 and 16, respectively. Tokenizers were loaded and used according to each model. As a result, it was confirmed that the designed models trained on the key information extraction dataset outperformed the baseline models trained on publicly available datasets during both training and evaluation processes, as shown in Table 12.

Table 12. Performance comparison between baseline and designed model.

Model	Token Base	Token Ours	Sequence Base	Sequence Ours
K-BERT	40.98	**84.74**	35.61	**88.11**
M-BERT	38.17	**84.15**	33.2	**86.05**
M-RoBERTa	32.58	**86.93**	31.56	**86.66**
KoELECTRA	44.3	**85.75**	45.21	**86.83**
K-RoBERTa	43.52	**85.44**	32.51	**87.27**

3.2. Performance Comparison of Models

Figures 4 and 5 depict the performance changes of the KLUE-BERT-base, KLUE-RoBERTa-base, Multilingual-BERT-base, XLM-RoBERTa-base, and KoELECTRA-base models over 20 fine-tuning epochs when using the key information extraction dataset for text

classification. Figure 4 shows the F1 score for token classification, while Figure 5 shows the F1 score for sequence classification.

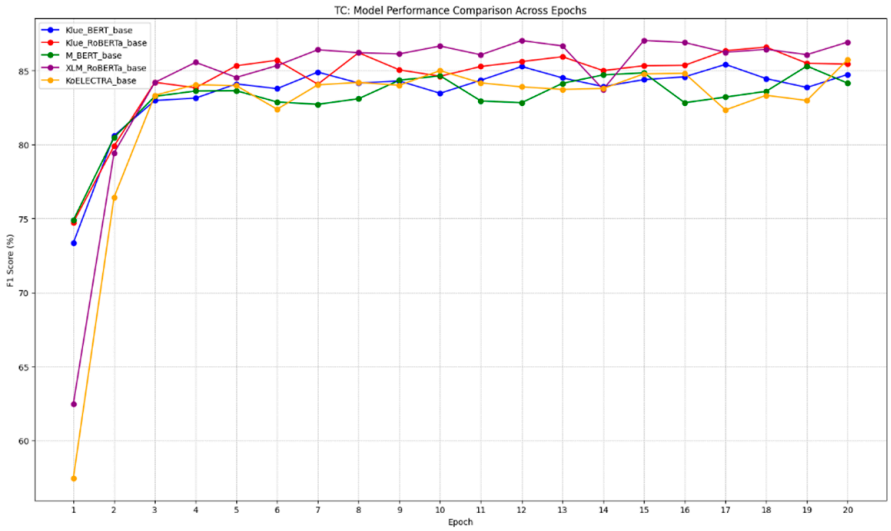

Figure 4. F1 score change graph of token classification.

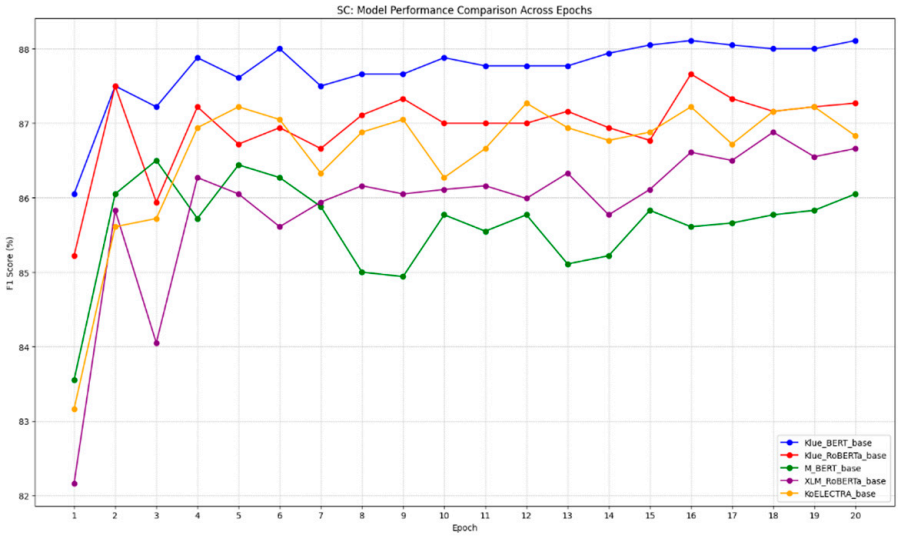

Figure 5. F1 score change graph of sequence classification.

Ultimately, XLM-RoBERTa-base exhibited the best performance in token classification, achieving an F1 score of 87% after the 15th training epoch, while KLUE-BERT-base achieved the highest performance in sequence classification, attaining an F1 score of 89.2%. The performance of the hybrid classification model for each key information category is detailed in Table 13.

Table 13. Performance by key information.

Key Information	F1 Score
Victim's Name	99.98
Victim's Age	98.6
Victim's Name	98.12
Relationship	75.04

Table 13. *Cont.*

Key Information	F1 Score
Crime Address	81.24
Crime Location	70.06
Crime Date/Time	85.36
Crime Result	92.08
Reason for Incompletion	86.17
Defendant's Medical History	91.55
Defendant's Financial Situation	78.11
Defendant's Criminal Record	95.68
Defendant's Drug and Alcohol Consumption	94.16
Preliminary Act	83.08
Crime Method	**92.04**
Crime Motivation	**88.11**
Crime Tool	88.01
Victim's Injury	91.38
Avg.	**87.75**

4. Discussion

In this study, we evaluated the performance of the hybrid classification model by measuring the F1 score using only the answers predicted with the highest probability. However, since this method has limitations in fully assessing the predictive ability of the model, we further analyzed the top three answers corresponding to the highest probabilities to evaluate the model's predictive ability more precisely. As a result, the proportion of correct answers matching the ground truth for the top-ranked answer in the token classification model was 97.5%, for the second-ranked answer it was 2.2%, and for the third-ranked answer it was 0.1%. The proportion of instances where the model failed to predict the correct answer was 0.03%. The proportion of correct answers within the top three ranks was over 99%, indicating the significantly high accuracy rate of the token classification model (see Figure 6).

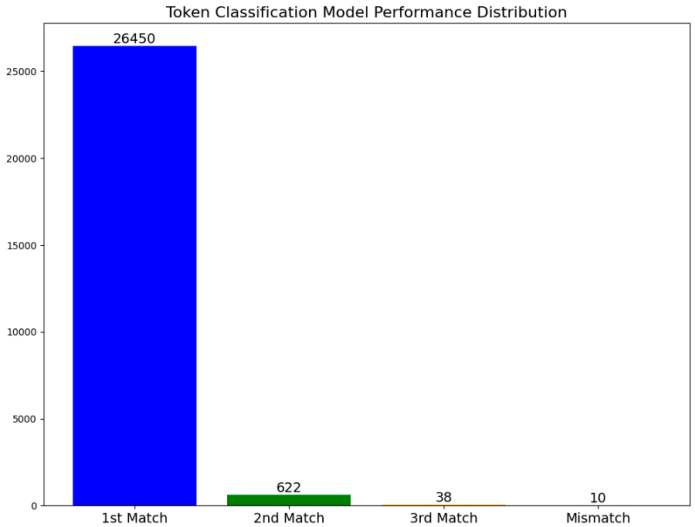

Figure 6. Prediction distribution of token classification model.

Figure 7 illustrates the distribution of the top three answers from the sequence classification model. The proportion of correct answers matching the ground truth for the top-ranked answer in the sequence classification model was 88.72%, which is the highest proportion. For the second-ranked answer, it was 7.1%, for the third-ranked answer, it was 2.2%, and, for instances where the model failed to predict the correct answer, it was 1.8%. Therefore, it was confirmed that the sequence classification model had a proportion of over 98% for correctly predicting the answer within the top three ranks, indicating the high accuracy rate of the sequence classification model.

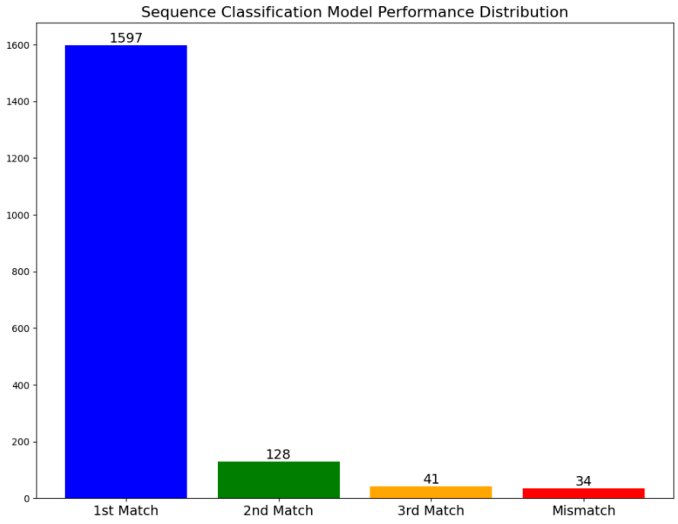

Figure 7. Prediction distribution of sequence classification model.

Next, based on cases where the model failed to predict the correct answer, we conducted an analysis of misclassification types. Table 14 reveals a total of three identified misclassification types. Firstly, in the token classification model, there was a total of 10 instances where the model failed to predict the correct answer, with 6 instances of misclassification type 1 and 4 instances of misclassification type 2. Secondly, the sequence classification model exhibited 34 instances of inconsistency, with 19 instances of misclassification type 1, 12 instances of misclassification type 2, and 3 instances of misclassification type 3.

Table 14. Misclassification types and statistics.

Type	Criteria	Frequency (Tok)	Frequency (Seq)
Type 1	When the target is key information but the predicted value is not key information	6	19
Type 2	When the target is not key information but the predicted value is key information	3	12
Type 3	When both the target and the predicted value are key information but are not identical	0	3

Upon examining the cases of the most frequently occurring misclassification type 1 in token classification, it was observed that the model mostly failed to predict 'victim's age' as key information. In the case of 'victim's age', it is typically mentioned in patterns including terms such as "years old" in Korean. However, expressions like those in Examples 1 and 2 in Table 15 deviate from the typical age pattern, which appears to be the reason why the model failed to predict them. This issue arises due to the nature of the judgment document data, where victim information is often expressed in standardized patterns. It is expected

that this problem can be addressed by refining and augmenting judgment document data in various forms for use as training datasets in the future.

Table 15. Detailed misclassification types and targets.

Idx	Token	Predicted	Target	Type
1	after birth	O	v_age	Type 1
2	one month	O	v_age	Type 1

When examining the misclassification types in Table 16 for sequence classification, we observed that Examples 1 and 2 should correspond to 'criminal acts', but the model failed to classify these sequences as key information. This issue arises from sequences being incorrectly split during the sequence segmentation process. To address this problem, it will be necessary to refine the morpheme-based sequence segmentation rules to ensure that sequences such as the actions of taking out a tool and committing an assault are not erroneously separated.

Table 16. Examples of misclassification in sequence classification.

Idx	Sequence	Predicted	Target	Type
1	Keeping the tip of the knife facing downwards After grabbing the kitchen knife	etc.	crime_method	Type 1
2	After retrieving the knife from the belt	etc.	crime_method	Type 1
3	Due to the mental stress caused by debt obligations and living in hiding Wondering what I should do next. Should I even consider going to prison?	a_eco_bg	motive	Type 3

Additionally, Example 3 should be classified as 'motive of crime', but the model misclassified it as 'suspect's economic situation'. This misclassification seems to have occurred because the sequence includes keywords commonly associated with 'suspect's economic situation' such as 'loan' and 'debt'. This problem can be addressed by adding diverse types of motive data, such as financial motives, motives arising from resentment or revenge, etc., to enable the model to learn about various types of motives more comprehensively.

5. Conclusions

In this study, we designed a hybrid classification model by combining token classification and sequence classification models and conducted experiments to extract 18 key pieces of information from the criminal first-instance murder verdicts. Analyzing the model's accuracy by extracting the top three predictions made by the hybrid classification model, we confirmed that the model correctly predicted the answer within the top three ranks over 98% of the time. These results indicate that the hybrid classification model can serve as an important tool for extracting key information from complex legal and investigative documents. Furthermore, the high proportion of correct answers among the top three predictions demonstrates the model's ability to effectively grasp and predict essential content even among diverse information. These findings provide a crucial foundation for the development of AI-based automation systems in the legal field and contribute to technological advancements in the legal and investigative domains.

The hybrid classification model designed and trained in this study can be utilized as a system to extract and recommend key information necessary for composing investigation reports. By presenting the top three predicted answers corresponding to each key information category, the model can support investigators in selecting one of them to include

in the report, thereby assisting in decision-making and saving time during report writing. Moreover, this system can contribute to improving the quality of investigations by ensuring accurate and consistent information inclusion in reports, and potentially pave the way for standardization and automation of the investigation process. Future utilization of this model in developing various investigation support systems such as key-information-based similar case search systems or event timeline construction systems is expected. Such advancements are anticipated to provide an opportunity to enhance the fairness and efficiency of the investigation process by offering a standardized approach that transcends individual investigators' capabilities, enabling everyone to conduct objective investigations and hypothesis testing. Our future works include the comparative analysis of the proposed system and other AI-based investigative systems, the expansion of the proposed core algorithm to other language processing areas, and the integration of the system with existing law enforcement and management systems, providing appropriate user convenience.

Author Contributions: Writing—original draft, Y.P.; Supervision, R.S.P. and H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was conducted with the support of Dongsimwoo Co., Ltd.

Data Availability Statement: Restrictions apply to the availability of these data. Data are obtained from the Supreme Court of Korea and can be issued upon payment of a fee with the permission of the Supreme Court of Korea.

Conflicts of Interest: The authors declare that this study received funding from Dongsimwoo Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

References

1. Kim, D.G. *Status and Improvement Plans for Investigation Closure after the Adjustment of Investigative Powers*; Occasional Research Report 21-AB-04; Korean Institute of Criminal Justice Policy: Seoul, Republic of Korea, 2022.
2. Kang, J.G. Piling Investigations… One Year After the Adjustment of Investigative Powers 'No One Was Satisfied'. *Hankyoreh*, 22 February 2022.
3. Noh, Y.J. 'Insufficient Follow-up Measures' Adjustment of Prosecutorial and Police Investigative Powers, The Damage Is on the People? *LIFEIN*, 30 March 2023.
4. Park, N.S. A Study on the Reconstruction of Criminal Facts and the Role of Hypothetical Reasoning (Abduction). *Police Sci. Res.* **2012**, *12*, 3–22. [CrossRef]
5. Jung, W. *Analysis of Increased Workload for the Economic Team Due to the Establishment of a Police Responsibility Investigation System*; Responsibility Research Report 11-1332522-000117-01; Police University Security Policy Institute: Asan, Republic of Korea, 2021.
6. Won, G.J. A Rational Crime Analysis and Fact-Acknowledgment Inference Visualization Model. Master's Thesis, Hallym University Graduate School, Chooncheon, Republic of Korea, 2018.
7. Ranaldi, L.; Pucci, G. Knowing Knowledge: Epistemological Study of Knowledge in Transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]
8. Lee, S.J. The Future of Legal Services Depending on How to Utilize Artificial Intelligence AI. *Legal Journal*, 18 August 2022.
9. Jung, C.Y. *Policy Study for the Introduction and Acceptance of Artificial Intelligence Technology in Judicial Procedures and Judicial Services*; [JPRI] Research Report 32-9741568-001430-01; Judicial Policy Research Institute: Seoul, Republic of Korea, 2021.
10. Bang, J.S.; Park, W.J.; Yoon, S.Y.; Sin, J.H.; Lee, Y.T. Trends of Intelligent Public Safety Service Technologies. *Electron. Telecommun. Trends* **2019**, *34*, 111–112. [CrossRef]
11. Park, S.; Lee, Y.; Choi, A.; Ahn, J.M.J. The 'Online Access to Judgment' Service in Korea: A Study on Improving Judgment Data for the Development of Legal AI (Artificial Intelligence). *J. Police Law* **2021**, *19*, 3–36. [CrossRef]
12. Hendrycks, D.; Burns, C.; Chen, A.; Ball, S. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *arXiv* **2021**, arXiv:2103.06268.
13. Chalkidis, I.; Androutsopoulos, I.; Michos, A. Extracting contract elements. In Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, London, UK, 12–16 June 2017; ACM: London, UK, 2017; pp. 19–28. [CrossRef]
14. Shaheen, Z.; Wohlgenannt, G.; Filtz, E. Large Scale Legal Text Classification Using Transformer Models. *arXiv* **2020**, arXiv:2010.12871.
15. Kim, H.D.; Hong, S.; Kim, D.H.; Kim, J.Y. Analysis on Voice Phishing using Artificial Intelligence Named Entity Recognition Model for Information Search. *J. Police Sci.* **2020**, *20*, 255–283. [CrossRef]
16. Kim, H.-D.; Lim, H. A Named Entity Recognition Model in Criminal Investigation Domain using Pretrained Language Model. *J. Korea Converg. Soc.* **2022**, *13*, 13–20. [CrossRef]

17. Korean Institute of Criminal Justice Policy Investigation Reform Team. Investigation Report Writing Technique. 2016. Available online: https://www.yes24.com/Product/Goods/117752063 (accessed on 20 August 2023).
18. Lee, Y. A Study on Extracting Crime Information from Criminal Judgments Using Machine Reading Comprehension. Master's Thesis, Hallym University Graduate School, Chooncheon, Republic of Korea, 2021.
19. Park, Y.; Park, R.-S.; Won, G. A Plan for Building a Criminal Judgment Information Extraction Dataset—Focusing on the Use of GPT-3.5 Prompts. In Proceedings of the 2023 Korea Computer Congress, Jeju, Republic of Korea, 18–20 June 2023.
20. Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; Zeng, M. Want to Reduce Labeling Cost? GPT-3 Can Help. *arXiv* **2021**, arXiv:2108.13487.
21. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [CrossRef] [PubMed]
22. Choi, J.H.; Hickman, K.E.; Monahan, A.; Schwarcz, D.B. ChatGPT Goes to Law School. *SSRN Electron. J.* **2023**. [CrossRef]
23. Doccano: Open Source Annotation Tool. Available online: https://github.com/doccano/doccano (accessed on 8 April 2024).
24. Park, S.; Moon, J.; Kim, S.; Cho, W.I.; Han, J.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; et al. KLUE: Korean Language Understanding Evaluation. *arXiv* **2021**, arXiv:2105.09680.
25. Lee, J.; Lim, T.; Lee, H.; Jo, B.; Kim, Y.; Yoon, H.; Han, S.C. K-MHaS: A Multi-label Hate Speech Detection Dataset in Korean Online News Comment. *arXiv* **2023**, arXiv:2208.10684.