Journal of the Korean Society of Marine Environment & Safety
Research Paper
Vol. 29, No. 7, pp. 843-856, December 31, 2023, ISSN 1229-3431(Print) / ISSN 2287-3341(Online) https://doi.org/10.7837/kosomes.2023.29.7.843

SentenceBERT 모델을 활용한 해양안전심판 재결서 분석 방법에 대한 연구

윤보리* · 박세길** · 배혜림*** · 심성현****†

* 동의대학교 산업경영빅데이터공학과 학사과정, ** 한국해양과학기술원 부설선박해양플랜트연구소 책임연구원,
*** 부산대학교 산업공학과 교수, **** 동의대학교 산업경영빅데이터공학과 조교수

Maritime Safety Tribunal Ruling Analysis using SentenceBERT

Bori Yoon* · SeKil Park** · Hyerim Bae*** · Sunghyun Sim****†

* Undergraduate Student, Department of Industrial Management Big Data Engineering, Dong-eui University, Busan, 47340, Korea
** Principal Researcher, Maritime Digital Transformation Research Center, Korea Research Institute of Ships & Ocean Engineering, Deaejeon, 34103, Korea
*** Professor, Industrial Data Science & Engineering Major, Department of Industrial Engineering Pusan National University Busan, 46241, Korea
**** Assistant Professor, Department of Industrial Management Big Data Engineering, Dong-eui University, Busan, 47340, Korea

요 약 : 전 세계 선박 통행량의 증가에 따른 선박 충돌 사고의 증가는 큰 경제적, 환경적, 물리적 및 인간적 손해를 가져왔다. 선박 사고의 원인은 선원의 판단 오류나 부주의, 항로의 복잡성, 기상 조건, 선박의 기술적 결함 등 다양한 요인이 겹쳐 작용하여 사고를 유발하기 때문에 문장의 깊은 의미와 문맥 정보를 고려할 수 있는 방법론이 필요하다. 따라서, 본 연구는 부산해심 지역에서의 최근 20년 동안의 선박 충돌사고 데이터를 포함하고 있는 해양안전심판 재결서를 SentenceBERT 모델을 활용해 분석하였다. 분석 결과 사고의 주요 원인이 될 수 있는 키워드가 도출되었으며, 특정 키워드 출현 빈도를 바탕으로 군집 분석을 시행하고 시각화하였다. 추후 사고의 원인을 미리 파악함으로써, 이를 통해 선박 충돌 사고의 예방 및 사고 대응 전략 개발의 기초 자료로써 활용하고자 한다.

핵심용어 : 해양 사고, 해양 사고관리, 해양안전심판 재결서, 센텐스버트, 텍스트 마이닝

Abstract : *The global surge in maritime traffic has resulted in an increased number of ship collisions, leading to significant economic, environmental, physical, and human damage. The causes of these maritime accidents are multifaceted, often arising from a combination of crew judgment errors, negligence, complexity of navigation routes, weather conditions, and technical deficiencies in the vessels. Given the intricate nuances and contextual information inherent in each incident, a methodology capable of deeply understanding the semantics and context of sentences is imperative. Accordingly, this study utilized the SentenceBERT model to analyze maritime safety tribunal decisions over the last 20 years in the Busan Sea area, which encapsulated data on ship collision incidents. The analysis revealed important keywords potentially responsible for these incidents. Cluster analysis based on the frequency of specific keyword appearances was conducted and visualized. This information can serve as foundational data for the preemptive identification of accident causes and the development of strategies for collision prevention and response.*

Key Words : *Marine accident, Marine accident management, Maritime safety tribunal ruling, SentenceBERT, Text mining*

1. 서 론

한반도는 지형적으로 복잡하고 협소한 해안선을 가지고 있어 선박의 항로 설정 및 운행의 복잡성이 더해져서 해상 교통에 어려움을 초래한다(Cho et al., 2002). 동시에 태풍과 짙은 안개 같은 기상 악조건이 빈번하게 발생해 선박의 안전한 항로 유지를 어렵게 만든다(Kim and Kang, 2011). 최근 세계적인 해운 물동량이 급격하게 증가함에 따라 규모의 경제에 의해 선박의 대형화 및 고속화 추세가 가속화되고 있다. 하지만 이러한 변화는 해상 사고의 발생 빈도와 피해 규모를 증가시키는 직접적인 원인이 되고 있다(Kim and Kim, 2011). 2022년 기준 한반도에서 해양 사고가 발생한 건수는

* First Author : yoonbori12@gmail.com, 051-510-1482
† Corresponding Author : ssh@deu.ac.kr, 051-890-1485

2,863건, 선박 수는 3,167척, 인명피해를 입은 인원은 총 412명이다(KMST, 2022). 이처럼 선박 사고 발생 시 경제적, 환경적, 물질적 피해와 함께 인적 피해가 발생함으로 사고 예방의 중요성이 높아지고 있으며, 사고 예방을 위해서는 명확한 원인에 대한 파악이 선행되어야 한다(Chen et al., 2018; Huang et al., 2018).

해상 사고의 원인은 다양한 해상 사고 보고서에 기록이 되어 있다. 이를 분석하기 위해 최근 연구에서는 텍스트 마이닝 기반의 연구들이 활발히 진행되고 있다. 국내에서는 Lee et al.(2019)의 Electronic Chart Display and Information System (ECDIS) 사고 보고서를 대상으로 텍스트 마이닝 기법을 활용하여 주요 사고 원인과 관련 단어의 빈도를 파악하는 연구와 Jung et al.(2018)의 해양안전심판 재결서를 통해 선박의 크기별, 발생 해역 별 선박 사고의 원인을 분석한 연구가 있다. 두 연구 모두 사고의 원인을 통계적 방법론을 사용하여 파악하기 때문에 데이터의 양적 분석에는 강점을 보이나 사고 원인 간의 복잡한 상호 작용이나 깊이 있는 원인 분석에는 한계를 가질 수 있다. 단순 양적 분석에서 더 나아가 Kim et al.(2020)은 해양안전심판 재결서의 인명사상의 작업안전사고의 주요 원인을 분석하고, 원인들의 상관관계를 분석하여 어업 작업안전사고의 인명피해를 줄이기 위한 연구를 했고, Park and Park(2023)은 텍스트 마이닝 기법을 사용해 워드클라우드와 네트워크 분석을 통해 어선 사고의 원인별 위험요인을 도출했다. 또한, 재결서와 사고알림 문자의 기록을 기반으로 해양 사고를 분류하고, 베이지안 분석 기법을 사용하여 사고 발생의 사전 확률과 사후 추론을 통한 발생 확률을 평가하는 방법론을 연구 했다. 하지만 이러한 연구 또한 단일 변수에 초점을 맞추거나 몇 개의 변수 간 상호작용을 고려하는 데 그친다는 한계점이 있다.

해외 연구로 Fan et al.(2020)는 Tree Augmented Network (TAN) 모델을 활용하여 해상 사고 보고서에서 인간 요인에 따른 다양한 해상 사고 유형에 대한 영향을 분석하였다. Tirunagari et al.(2015)는 나이브 베이즈와 SVM(Support Vector Machine)을 사용하여 패턴 분석, 문장의 원인과 결과를 수동으로 추출하는 연결어 방법을 사용하여 해상 사고 조사 보고서의 원인 관계를 추출하는 연구를 진행하였다. Tirunagari et al.(2015)의 연구를 통해 사고 조사 보고서를 분석하여 사고의 직/간접적인 원인 관계를 파악할 수 있었지만, 나이브 베이즈와 같은 확률 기반의 전통적인 방법론만으로는 선박의 충돌사고에서 복합적인 인적 요인과 환경적 요인, 그리고 그들 사이의 복잡한 상호작용을 파악하는 데 한계가 있다(Choi et al., 2021).

따라서 선원의 판단 오류나 부주의, 항로의 복잡성, 기상조건, 선박의 기술적 결함 등 다양한 요인이 겹쳐 작용하여 사고를 유발하기 때문에 문장의 깊은 의미와 문맥 정보를 고려할 수 있는 방법론이 필요하다.

이에 따라 본 연구에서는 중앙해양안전심판원에서 제공하는 해양안전심판 재결서 중 충돌 사건에 해당하는 사건들을 딥러닝 기반의 텍스트 마이닝 방법론을 활용하여 분석하고자 한다. 분석에 사용된 모델은 최근 딥러닝 영역 텍스트 데이터에서 좋은 성능을 보이는 Bidirectional Encoder Representations from Transformers(BERT) 모형 중 SentenceBERT (SBERT) 모형을 활용하였다. SBERT 모형은 문장 간의 유사성을 측정하거나, 문장 임베딩을 사용하여 문서의 전반적인 의미를 파악할 수 있다고 알려져 있다(Reimers and Gurevych, 2019). 따라서 본 연구에서는 SBERT를 통해 비정형 텍스트 데이터인 해양안전심판 재결서 중 충돌사건에 초점을 맞춰 인과 패턴을 추출하고 분석하여 사고 패턴 및 시간/공간 관점에서 유의미한 정보를 추출하고 구조화하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구의 접근 방식과 이론을 소개한다. 3장에서는 본 연구에서 제안된 프레임워크에 대해 설명한다. 4장에서는 본 연구에 사용된 데이터와 SBERT 모델을 기반으로 학습된 정보를 군집화하여 항만 사고 패턴을 분석한 결과를 소개한다. 마지막으로 5장에서는 본 연구의 결론과 추후 연구 방향을 제시한다.

2. 이론적 배경

본 장에서는 본 논문의 이론적 배경이 되는 해양안전심판 재결서와 SBERT에 관련된 연구를 소개한 후, 해당 연구들과 본 논문과의 관련성을 간략하게 설명한다. 또한, 본 연구에서 사용된 방법론인 군집화와 텍스트 군집화를 소개한다.

2.1 해양안전심판 재결서

해양안전심판 재결서는 중앙해양안전심판원(2심) 및 각 지방해양안전심판원(1심)에서 심판 청구된 해양 사고에 대한 조사와 심판의 결과를 포함한 문서로유사 사건의 재발을 방지하는 것을 목적으로 하고 있다(Cho et al., 2002). 해양안전심판 재결서를 통해, 사고 원인과 그 결과를 깊이 있게 분석하며, 이를 공개함으로써 선박 운항자나 관계자들에게 같은 실수를 반복하지 않도록 경고하고 교육하는 자료로 사용된다(Park et al., 2020).

해양안전심판 재결서에는 Fig. 1과 같이 사고의 일시, 장소와 같은 기본적인 정보부터 사고의 상세한 경위, 선박의 제원, 원인, 그 결과, 그리고 이를 통한 판결 내용 및 미래 사고 예방을 위한 권고사항까지 포함하고 있다. 이들 정보는 비정형 텍스트 데이터로 표현되어 사고 상황, 환경, 인적 요인 등 다양한 정보를 제공하며, 해양 사고 분석 및 예방

방안 수립에 활용할 수 있다. 각 해양안전심판 재결서에는 일시, 장소, 사건 개요, 사고 경위 등의 특정 키워드가 중점적으로 포함되어 있으며, 이러한 키워드를 통해 사고의 상세한 경위, 관련 법적 근거, 심판원의 판단 및 권고사항 등의 중요 정보를 추출할 수 있다.

재 결 요 약 서

사 건 명	[redacted]
해양사고관련자 (직책, 해기면허)	1. A[redacted] 기관장, 4급기관사)
판 시 사 항	1. 판시요지 가. 기관부품의 잠재적 하자 등의 원인으로 발생한 기관손상사고에 대하여는 기관장의 직무상 과실 책임이 없다고 판시
주 제 어	기관부품의 잠재적 하자, 릴레이, 기관손상
사 건 개 요	□ 관련선박 선명 / 용도 / 총톤수/길이(gt/m) / 소유자/선적항 / 피해 [redacted] / 유선 / 29/20.13 / B/거제시 / -
	□ 일 시 : 2016. 6. 11. 15:00경
	□ 장 소 : 북위 34도 44분 01초 · 동경 128도 40분 35초 (경남 거제시 갈곶도 서방 0.2마일 해상)
	□ 사고경위 [redacted]는 총톤수 29톤, 길이 20.13미터, 디젤기관 446kw 1기를 장치한 강화플라스틱조 유선으로 2016년 6월 11일 14시 25분경 선장을 포함한 선원 2명과 여객 95명 총 97명이 승선한 가운데 거제 장승포항에서 출항하여 거제시 외도로 가기 위해 항해하던 중 상기 일시 및 장소에서 갑자기 주기관이 정지된 후 시동이 걸리지 않자 구조요청을 한 사건임.
주 문	이 기관손상사건은 [redacted]가 정상적으로 운항하고 있던 중 기관부품의 고장으로 인하여 발생한 것이다.

Fig. 1. Example of Maritime safety tribunal ruling (Korean Maritime Safety Tribunal, 2016b).

2.2 SentenceBERT (SBERT)

SBERT는 BERT 아키텍처를 기반으로 한 자연어 처리(NLP) 기법으로 문장 레벨의 임베딩을 효과적으로 생성하는 방법론이다(Reimers and Gurevych, 2019). 기존 BERT 모델은 문장이나 단어를 고차원의 벡터로 변환하는 임베딩 과정에서 뛰어난 성능을 보였다. 하지만 BERT는 고정된 길이의 입력을 받기 때문에 긴 문장의 경우 의미를 파악하기 어렵다. 또한, 여러 문장 또는 문단 간의 관계를 파악하는 데 제한적이기 때문에 문장 간의 의미적 유사도를 계산하는 데에는 한계가 있다(Devlin et al., 2018). SBERT는 기존 BERT모형의 구조를 확장하여 문장의 전체적인 의미 구조를 고려한 임베딩을 할 수 있도록 개선된 모형이다. 특히 SBERT는 Siamese와 Triplet Network 구조를 적용함으로써, 문장 간의 유사도 측정의 정확도를 높인다(Reimers and Gurevych, 2019).

Siamese Network는 두 개의 병렬 입력을 받아 동일한 서브 네트워크 구조를 통해 각각을 처리한다(He et al., 2018). 이후 생성된 임베딩 결과를 비교하여 문장 간의 유사도를 측정한다(Reimers and Gurevych, 2019). 이 구조는 문장의 구조적 특징뿐 아니라 의미적 특성을 잘 포착하며, 문장 간의 유사도 계산에 뛰어난 성능을 제공한다. Triplet Network는 앵커(anchor) 문장, 긍정(positive) 문장, 부정(negative) 문장의 세 가지 입력을 처리한다. 이 네트워크는 앵커 문장과 긍정 문장 사이의 유사도가 앵커 문장과 부정 문장 사이의 유사도보다 높도록 학습되며, 이를 통해 더욱 정밀한 유사도 측정이 가능하다(Devlin et al., 2018). 이러한 구조적 특징 덕분에 SBERT는 문장 간의 의미적 관계를 더욱 정밀하게 분석하고 빠르게 미세조정 되며, 유사도 측정, 군집화, 검색 등의 작업에서 기존 BERT 방법론보다 뛰어난 성능을 보여준다(Joshi et al., 2023). 또한 SBERT의 특성은 해양안전심판 재결서과 같은 비정형 텍스트 데이터 분석에 있어, 문장 간 의미적 유사도가 높은 부분을 빠르게 식별하고, 유사 사고 사례나 패턴을 효율적으로 군집화하고 분석하기 때문에 유용하게 사용된다(Devlin et al., 2018).

2.3 군집화

군집화는 비지도 학습(Unsupervised Learning)기법 중 하나로 데이터의 내재적인 패턴이나 구조를 찾아 유사한 데이터끼리 군집으로 묶는다(Xie and Jiang, 2010). 군집화는 주로 계층적(Hierarchical) 군집화와 비계층적 또는 분리형 군집화(Partitional, Non-hierarchical)라는 두 가지 방법으로 분류된다. 첫째, 계층적 군집화는 군집 구성 과정에서 이전 단계의 군집을 참조하여 군집을 형성하며, 집합적(agglomerative, bottom-up) 방법과 분할적(divisive, top-down) 방법으로 나뉜다. 집합적 방법에는 단일 연결법, 완전 연결법, 평균 연결법 등 다양한 연결법이 포함되며, 각각은 군집 간의 거리를 측정하는 방법에 차이가 있다(Soni and Ganatra, 2012). 반면 분할적 방법은 데이터 세트를 분할하여 분석하며, 다이애나 방법 등이 특징적인 방법으로 사용된다(Madhulatha, 2012). 둘째, 분리형 군집화는 데이터를 군집에 할당할 때 전체 집합 개수를 기준으로 작동하며, K-Means 군집화, DBSCAN(Density Based Spatial Clustering of Applications with Noise), EM(Expectation Maximize), PAM(Partitioning Around Medoids) 등이 대표적인 방법론으로 속한다.

텍스트 군집화는 문서의 유사도 또는 목적 함수를 기준으로 거리를 측정하여 유사한 텍스트 데이터끼리 군집화하게 된다(Vijaymeena and Kavitha, 2016). 텍스트 군집화는 크게 문서의 구조와 내용을 분석하여 텍스트 데이터 간의 유사도를 측정하고, 유사한 문서들을 하나의 군집으로 분류하는 단계를 포함한다(Abualigah et al., 2016). 텍스트 데이터의 전처리는 이 과정에서 중요한 역할을 하며, 텍스트의 노이즈 제거,

토큰화, 어간 추출 등이 이루어진다(Kalra and Aggarwal, 2017). 전처리 후에는 텍스트 데이터의 특징을 추출하고, 이를 토대로 다양한 군집화 알고리즘을 적용하여 문서를 군집화한다(Kadhim et al., 2014). 군집화 알고리즘을 선택할 때는 데이터의 특성과 분석 목표에 맞는 알고리즘을 고려해야 한다. 또한, 텍스트 군집화의 성공은 군집의 수가 올바르게 결정되는 것에 큰 영향을 받는다(Jee et al., 2007). 군집의 수를 과다하게 설정하면 과적합의 위험이 있고, 반대로 너무 적게 설정하면 군집화의 효과가 줄어들 수 있다. 이를 해결하기 위해 Elbow Method, Silhouette Analysis 등 다양한 방법들이 사용되어 최적의 군집 수를 찾을 수 있다(Ashari et al., 2023). 군집화의 성능은 군집 내 및 군집 간의 데이터 차이를 기준으로 평가된다. 군집 내 거리를 측정하는 방법으로는 Complete Diameter Distance, Average Diameter Distance, Centroid Diameter Distance 등이 있으며, 군집 간 거리를 측정하는 방법으로는 Single Linkage Distance, Complete Linkage Distance, Average Linkage Distance, Centroid Linkage Distance 등이 사용된다(Han, 2022).

3. 제안 방법론

본 연구에서는 해양 사고의 원인 도출을 위해 해양안전심판 재결서 데이터를 분석하여 패턴을 도출하는 프레임워크를 제안한다. 제안된 방법은 Fig. 2과 같이 세 단계로 구성되며 각 단계는 다음과 같다. 단계 1에서는 텍스트 데이터인 해양안전심판 재결서의 전처리를 수행한다. 단계 2에서는 SBERT 모델의 학습 수행 및 임베딩 된 결과를 활용하여 서로 다른 해양안전심판 재결서의 유사도를 정량적으로 출력한다. 단계 3에서는 임베딩 된 결과를 활용하여 K-Means 군집화 방법으로 군집화 및 워드 클라우드를 활용한 통한 군집 별 주요 사고 인자를 도출한다. 본 연구에서 제안하는 프레임워크를 통해 다른 해양안전심판원의 재결서 중 유사한 문서를 묶어 주요 키워드와 패턴을 도출하는 것을 목표로 한다.

3.1 데이터 전처리

단계 1의 데이터 전처리 과정은 Fig. 3과 같이 구성된다. Input 데이터는 해양안전심판 재결서 데이터이고, Text Extraction 단계에서 연구에 필요한 주요 정보인 본문 텍스트 정보를 각 HWP 파일 내 'BodyText' 디렉토리에서 추출한다. 추출된 데이터는 연속된 긴 텍스트 형태를 가지기 때문에 문장 내용 파악이 어렵다. 따라서 Sentence Splitting 단계에서 문장의 끝을 나타내는 구분자('.', '?', '!')를 기준으로 전체 텍스트를 문장 단위로 세분화하는 분할 작업을 한다. 그 후 Sentence Filtering 단계에서 세분된 문장 중에서 '사건 개요', '일시', '장소', '사고 경위'를 핵심 키워드로 선별한다. 선별된 키워드 이후에 이어지는 문장들을 추출하는 문장 필터링 과정을 통해 재결요약서 내의 주요 정보만을 추출한다.

필터링된 텍스트 내에서 여러 위치에 존재하는 줄 바꿈 문자(\n)는 문장의 연속성을 방해하며, 텍스트 분석 시 노이즈로 작용할 수 있으므로 Sentence Cleaning 단계에서 이를 제거하는 문장 정제 작업을 한다. 또한 연속된 문자열 형태의 문장에서는 문장의 끝과 다음 문장의 시작 사이에 적절한 공백이나 구분자가 필요하기 때문에 문장 끝을 나타내는 구분자('.', '?', '!') 뒤에 공백을 추가해 문장 간의 구분을 뚜렷하게 한다. 텍스트 내의 특수 문자들도 적절한 문자나 공백으로 치환하거나 제거하여 문장 내의 불필요한 요소를 최소화한다. 이러한 문장 정제 과정을 통해 원시 텍스트 데이터가 구조화된 형태로 변환된다. 추출 및 정제 과정을 거친 텍스트와 원본 파일명을 딕셔너리 형태로 저장한다. 이 딕셔너리는 '사건 개요', '일시', '장소', '사고 경위'와 같은 핵심 키워드 다음에 나오는 문장들로 구성된 재결서의 중요 정보를 담고 있어, 후속 분석과 처리를 위한 최종 데이터셋으로서의 역할을 한다.

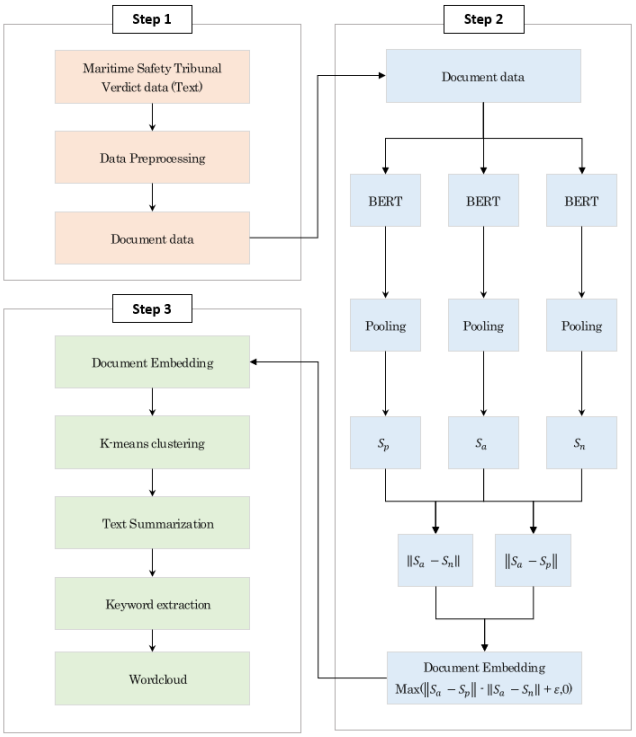


Fig. 2. Proposed Framework.

3.2 모델 학습 수행 및 유사도 출력

이 단계에서는 SBERT 모델을 활용하여 문장 간의 의미적 유사도와 문서 전체의 의미를 정확하게 파악한다. SBERT 모델은 한국어 문장 간의 유사도가 표시된 KorSTS 데이터셋을 학습에 사용하며, 이 데이터셋은 문장 간의 의미적 유사성 학습에 적합하다(Ham at al., 2020). SBERT 모델은 이 데이터셋을 통해 문장 간의 미묘한 의미 차이와 구조적 특성을 학습한다. 학습 모델은 klue/roberta-base를 기반으로 하는 Sentence Transformer의 구조이다. 이 구조에서 Transformer 모델은 문장을 토큰으로 분해하고, 각 토큰에 대한 임베딩을 생성한다(Liu at al., 2019). 생성된 임베딩은 문장의 각 토큰의 평균을 취하여 고차원의 문장 임베딩으로 변환되며, 이를 통해 문장의 전반적인 의미를 파악한다(Liu at al., 2019). 학습 과정에서는 Cosine Similarity Loss(CSL)를 활용하여 모델의 출력값과 실제 유사도 라벨 간의 차이를 최소화한다(Tan et al., 2005; Wolfram, 2007). CSL은 적은 데이터 셋에서도 벡터 간의 방향성을 기반으로 한 유사도 계산을 가능하게 하여, 복잡한 기술 용어와 다양한 언어 패턴이 포함된 해양안전심판 재결서 데이터를 효과적으로 처리할 수 있다(Wolfram, 2007). CSL 함수는 식(2)와 같이 정의되며, 문장 간의 유사도를 측정하는 데 있어 효과적이다(Tan et al., 2005; Wolfram, 2007). 식(1)에서 $\cos(\theta)$는 Cosine Similarity를 의미하며, $\|A\|, \|B\|$는 벡터의 놈(norm)을 의미한다. i는 벡터의 원소, n은 벡터의 차원 수, $\cdot$은 두 벡터의 내적을 의미한다. cos은 Cosine 함수, θ은 벡터 사이의 각도를 의미한다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \ \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

$$\text{Cosine Similarity Loss} = 1 - \cos(\theta) \quad (2)$$

모델의 성능을 주기적으로 확인하기 위한 평가 척도는 Embedding Similarity를 사용한다. 이때 두 문장 간의 임베딩 유사도를 정량화하며, 학습 과정에서 모델의 성능을 지속적해서 모니터링하는 역할을 한다(Faruqui at al., 2016). 또한 초기 학습 단계에서 전체 훈련 데이터의 10%에 해당하는 스텝을 warm-up으로 설정함으로써, 모델의 안정적인 수렴을 만든다. 이렇게 학습된 SBERT 모델을 통해 서로 다른 재결서가 입력으로 주어졌을 때, 두 재결서 사이의 의미적 유사도를 정량적으로 출력할 수 있게 된다(Reimers and Gurevych, 2019).

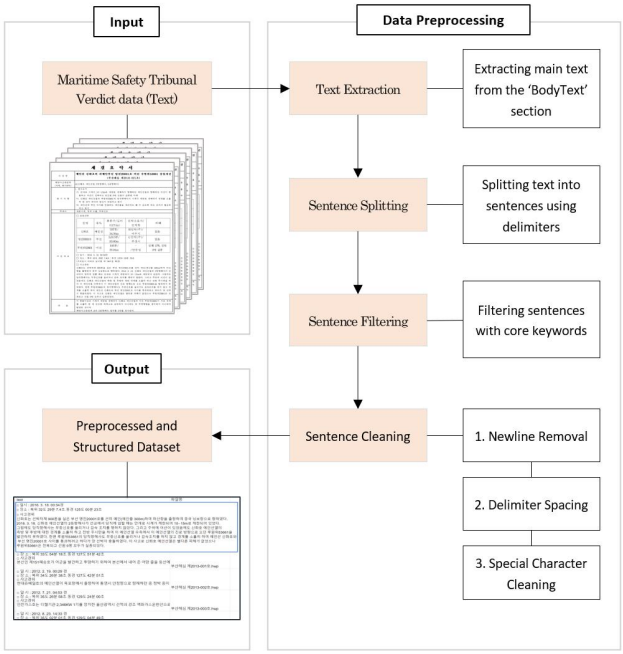


Fig. 3. Data Preprocessing.

3.3 군집화 및 시각화

군집화 알고리즘은 다양한 데이터셋에서 유용한 패턴과 구조를 식별하기 위한 기법 중 하나로 2.3장에서 언급된 바와 같이 대표적인 군집화 알고리즘으로는 계층적 군집화(Hierarchical Clustering), 밀도 기반 군집화(DBSCAN), 기대값 최대화(EM Clustering), 그리고 K-Means 군집화 등이 있다(Soni and Ganatra, 2012). 단계 3에서는 여러 군집화 알고리즘 중에서 K-Means 군집화 알고리즘을 선택하여 사용하였다. K-Means 군집화 알고리즘은 대규모 데이터셋에 대해서도 빠르고 효과적인 성능을 보이는 장점이 있다(Soni and Ganatra, 2012). 또한, 다른 군집화 방법과 비교하였을 때 계산 복잡성이 낮고 직관적인 결과를 제공한다(Kodinariya and Makwana, 2013). 이는 다차원 공간에서의 데이터 포인트를 이해하고 해석하는 데에 있어서 용이하게 작용한다. 본 연구에서는 SBERT 모델을 통해 얻은 고차원 특징 벡터에 기반하여 선박 사고 데이터를 군집화하고자 하였으며, K-Means 알고리즘은 이러한 벡터 공간에서 군집의 중심을 효율적으로 찾아내는 데 매우 적합하다(Soni and Ganatra, 2012). 본 연구에서는 이러한 이점을 최대화하기 위해 K-Means 군집화 알고리즘을 채택하여 분석을 수행하였다.

이때 군집의 개수를 지정하는 데에 모호함이 존재하므로 군집 내 거리의 제곱합 변화를 분석하는 Elbow-method를 활용한다. 군집 수가 증가함에 따라 군집 내 거리의 제곱합이 줄어드는 양상을 보이며, 이 감소율이 급격히 변화하는 지점을 'Elbow'라 칭하며, 이 지점을 군집 수로 선정한다. 감소

율의 변화가 일정해지면 사용자의 판단으로 군집 수가 결정된다(Kassambara, 2017). 그 후 각 군집 내에 포함된 재결서의 내용을 Text Summarization 기법을 이용하여 문장을 요약한 다음 주요한 Keyword를 추출한다. 이렇게 추출한 키워드를 바탕으로 군집 유형별 워드 클라우드를 생성하여 시각화한다.

4. 실 험

4.1 데이터

분석 대상 데이터는 2003년부터 2022년까지의 20년간 해양안전심판원에서 제공된 '충돌' 사건 중, 부산 해심에 해당하는 737건의 재결서로 구성되어 있다. 이 데이터는 한글 워드 프로세싱 파일(HWP) 형식의 비정형 텍스트 형태로 제공되었다. HWP 파일은 복잡한 구조를 가진 바이너리 형식의 파일이므로, 텍스트 데이터 추출을 위해 특정 라이브러리와 메서드를 사용하였다. 우선, Python의 'olefile' 라이브러리를 사용하여 HWP 파일을 읽은 후, 해당 파일 내에서 'BodyText' 디렉토리에 위치한 본문 텍스트 정보를 추출하였다. 예를 들면 '일시', '장소', '사건 개요', '사고 경위'와 같은 특정 정보를 중점적으로 추출하였다. 특히, '사고 경위' 키워드 다음의 텍스트는 사고의 상세한 경위를 설명하므로, 해당 부분을 별도로 추출하였다. 만약 '사고 경위'에 해당하는 텍스트가 없다면, '사건 개요'를 대신하여 사용하였다. 각 재결서는 사고의 일시, 장소, 원인, 결과, 그리고 판결 내용 등 다양한 정보를 포함하고 있으며, 사건에 대한 상세 설명, 관련 법적 근거, 심판원의 판단 및 권고사항 등으로 구성되어 있다. 이 연구는 이러한 데이터의 특성을 정밀하게 파악하며, 딥러닝 모델에 맞게 데이터를 구조화하였다. 이 데이터는 재결서에서 추출한 핵심 정보(일시, 장소, 사고경위 등)를 나열하고, 각 사건의 특징과 패턴을 분석할 수 있게 한다.

4.2 모델 학습 결과 시각화

SBERT를 통해 문장 간의 의미적 유사도 학습을 진행하고 그 결과를 시각화하였다. SBERT 모델은 사전에 훈련된 BERT 모델을 기반으로 하며, 이를 용도에 적합하게 파인튜닝함으로써 모델 성능을 향상시킨다(Reimers and Gurevych, 2019). 사전에 훈련된 모델을 기반으로 파인튜닝을 진행하기 때문에 Loss 또한 비교적 높은 값에서 시작한다. 학습 과정에서는 CSL을 사용하여 문장 간 의미적 유사도를 정확히 학습할 수 있도록 돕는다. 따라서 Loss 그래프가 점점 증가하는 것은 문장 간의 의미적 유사도가 높아진다는 것을 의미하며, 이는 모델이 데이터셋의 의미적 특성을 잘 파악하고 있음을 나타낸다(Tan et al., 2005).

Fig. 4는 훈련 데이터 셋에 대한 CSL의 변화를 시각화한 그래프이고, Fig. 5는 테스트 데이터 셋에 대한 손실 변화를 시각화한 그래프이다. 모델의 Test Loss는 초기에 0.832로 시작하여 학습이 진행됨에 따라 0.897로 증가하였고, 이는 모델이 데이터의 복잡한 의미 구조를 점차적으로 인코딩하고 있음을 나타낸다. 반면, 테스트 로스는 0.9969에서 시작하여 최종적으로 0.9976에 도달함으로써 학습이 마무리되었다. 이는 테스트 데이터셋에 대해 모델이 일관된 성능을 보이고 있으며 과적합을 피하면서 의미적 유사도를 잘 포착하고 있음을 의미한다. 두 그래프 모두 처음 몇 단계에서는 모델이 데이터의 패턴을 파악하기 시작하는 초기 학습 단계로 이 단계에서는 손실 값이 크게 변동하는 것을 확인할 수 있다. 그러나 이후 훈련 단계에서는 모델이 점점 데이터의 특성을 잘 학습하게 되어 손실 값이 점진적으로 줄어들며 안정화되는 것을 확인할 수 있다.

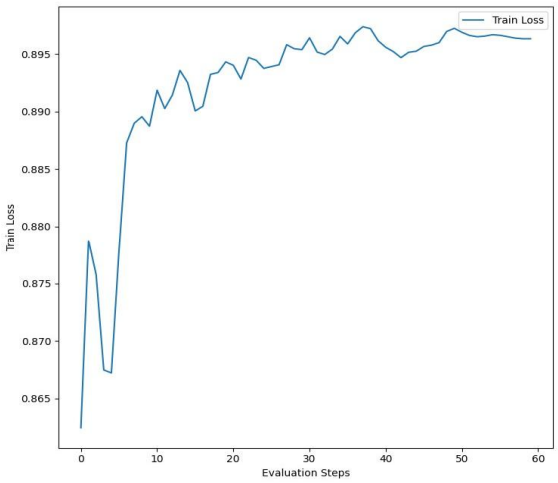


Fig. 4. Result of Train Loss.

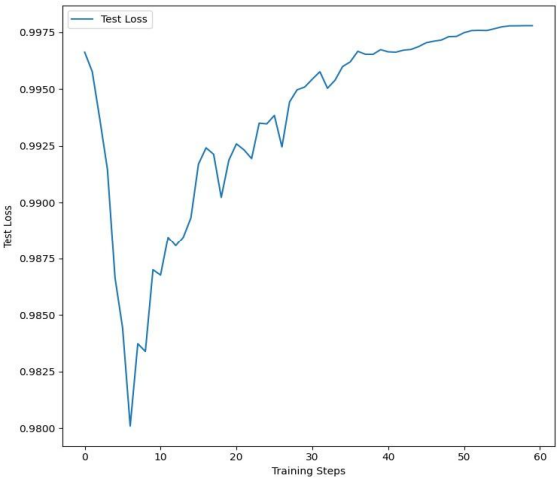


Fig. 5. Test Loss.

4.3 군집화 결과 시각화

본 연구에서는 문장 임베딩의 패턴을 K-Means 군집화 방법으로 군집화한다. 이 과정을 통해 문서나 문장들의 특성을 기반으로 유사한 패턴을 동일한 군집에 할당하였다. 군집 중심과 군집에 속한 값의 거리 제곱합 감소 비율이 가장 큰 군집 수를 채택하거나, 수준 이하의 거리 제곱합을 보이는 군집 수를 채택하는 elbow-method 결과 도표는 Fig. 6과 같다. 다양한 군집 수에 따른 내부 군집 내 제곱 오차의 합을 분석하여 적정 군집의 개수로 10개를 선정하였다. 그 결과 특징이 유사한 10개의 군집을 형성한다. 이러한 군집화를 통해 해양안전심판원의 재결서의 주요 주제와 패턴을 도출하였다. 또한, 고차원의 데이터를 2D나 3D 공간에 표현하기 위해 T-SNE 방법을 사용하였다. T-SNE 결과는 Fig. 7과 같다. 이를 통해 군집화의 효율성과 군집 간의 경계, 특성 등을 명확하게 확인하였다. 또한 한글 텍스트 처리를 위해서는 Komoran 형태소 분석기를 활용하였으며, 각 군집별 주요 키워드를 워드 클라우드로 표현하였다. 워드 클라우드에서는 키워드의 빈도수에 따라 다양한 글자 크기로 주요 키워드를 강조하였다.

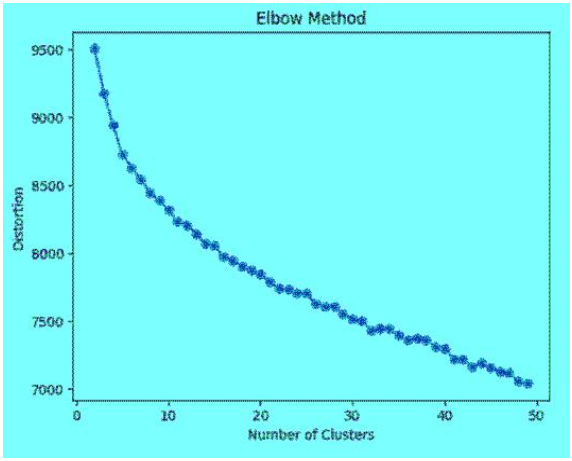


Fig. 6. Elbow Method for Determining the Number of Clusters.

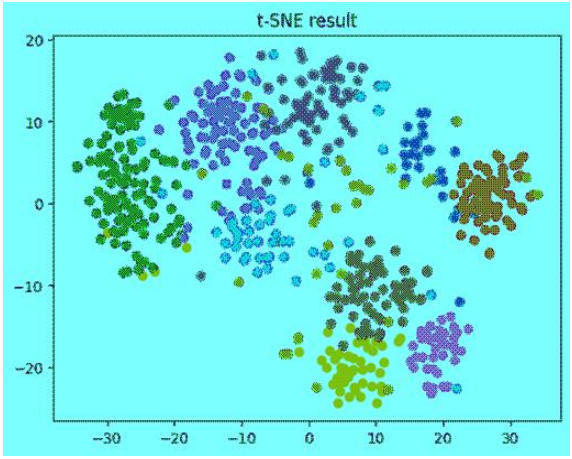


Fig. 7. T-SNE Results for 10 Clusters.

4.4 결과 분석

본 연구를 통해 SBERT를 학습한 다음, 총 20년간의 데이터에 포함된 충돌사건 중 부산 해심에 해당되는 737개의 재결서를 입력값으로 사용하여 학습된 모형에 입력시키고, 얻어지는 스코어 값을 이용하여 재결서 사이에 유사성을 분석하였다. 그런 다음 K-Means 방법을 활용하여 유사한 Score 값을 가지는 10개의 군집을 도출하였다. 그 후 각 군집 내에 포함된 재결서의 내용을 Text Summarization 기법을 이용하여 문장을 요약한 다음 주요한 Keyword를 도출하였으며 그 결과 각 군집은 해양 사고 유형, 원인 및 특징을 포함하는 것을 확인할 수 있었다. 부산 해심에서 발생한 '충돌' 사고의 군집화 결과 워드 클라우드 그림은 Fig. 8과 같다. 군집의 번호는 단순 구분을 위하여 임의로 부여된 번호이다. 이하에서 각 군집별 주요 키워드와 그것이 시사하는 사고 유형과 원인에 대해 상세히 살펴본다.

Fig. 8 (a) Cluster 0은 '경계', '발견', '상대', '협력', '일인' 등의 키워드가 주를 이루며, 이들 키워드는 해양 네비게이션과 관련된 사고의 내용을 보여준다. '경계'와 '발견'이라는 단어의 빈도가 높은 것은 선박 운행 중 경계 소홀과 늦은 발견이 주된 사고 원인임을 시사한다. 이러한 상황에서 '협력'이라는 단어의 비교적 높은 빈도는 충돌 사고에서의 피항협력동작의 부족이 큰 부분을 차지하고 있음을 나타낸다.

Fig. 8 (b) Cluster 1은 '선적', '장치', '경계', '강조', '거리' 등의 키워드가 두드러진다. 이러한 키워드들은 선박의 톤수나 장비, 장치의 상태나 기기에 관련된 사항들이 사고의 주요 요인이 됨을 의미한다. 특히 '거리'와 관련된 키워드의 빈도가 높은 것은 충돌 사고에서 상대 선박 간의 거리가 중요함을 나타낸다.

Fig. 8 (c) Cluster 2에서는 '경계', '일인', '협력', '일시', '장소' 등이 주요 키워드로 등장한다. 통발 설치나 양망과 같은 작업 중 충돌 사고가 발생하는 경우가 많이 포함되었으며 이러한 키워드들은 작업 중의 경계 소홀이나 피항협력동작의 부족에서 비롯된 사고를 의미한다.

Fig. 8 (d) Cluster 3에서는 '발견', '경계', '일인', '항행' 등이 주된 키워드로 분석되며, 항해 중 경계 소홀이 사고 원인으로 추출된다. '일인'이라는 키워드 또한 높은 빈도로 등장하며, 선장 혹은 항해사의 일인이 사고 원인이 되는 경우를 의미한다. 또한 선박 간의 상호작용 및 접근 제한과 관련된 문서가 포함된다.

Fig. 8 (e) Cluster 4에서는 '강조', '선적', '선수', '장치', '방향', '거리'와 같은 주요 키워드들이 도드라져 나타난다. 이 군집은 주로 선박의 적재와 선수 충돌에 관련된 사고를 중점으로 다루며, 거리와 방향과 같은 물리적 요소들이 사고의 원인으로 추출된다.

Fig. 8 (f) Cluster 5에서는 '계류', '예인', '접근', '정박' 등이

주요 키워드로 등장하며, 이러한 키워드들은 주로 선박이 정박 중에 발생하는 다양한 사고 유형을 나타낸다. 이 군집은 정박 중인 선박과의 접근 충돌, 정박용 줄 관리 실패로 인한 부두에서의 충돌 등의 사고를 중점적으로 탐구한다. '정박'이라는 키워드와 그 빈도는 선박이 부두에 정박 중이거나 정박 후 작업 중 발생하는 사고들이 주요 사고 유형을 형성하고 있음을 나타낸다. 특히, 정박용 줄을 거두어들이는 과정과 후진 시 시각적 제한으로 인한 충돌 사고가 두드러진다.

Fig. 8 (g) Cluster 6에서는 '상대', '선수', '접근' 등의 키워드가 빈도가 높게 나타남으로써, 상대 선박과의 상호 작용이 사고의 주요 원인이 되는 것을 나타낸다. 특히 '접근'이라는 단어의 높은 빈도는 선박 간의 접근 관리의 중요성을 강조하며, 상대 선박의 접근을 늦게 발견하여 사고가 발생하는 상황을 다룬다.

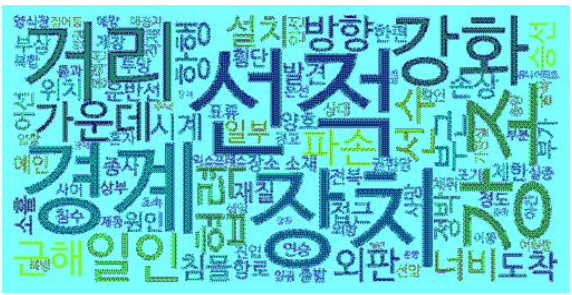
Fig. 8 (h) Cluster 7에서는 '적재', '선수', '경계', '장치', '부가' 등의 키워드가 주요하게 도출되며, 이들 키워드는 선박 적재 과정에서 발생하는 사고의 복합적인 특성을 보여준다. 사고 사례 중 다수는 선박의 적재 중에 충돌이 발생하며, 특히 선수부 일부가 파손되는 경우가 두드러진다. 이러한 사고 상황은 대체로 경계의 소홀함과 밀접한 관련이 있는 것으로 분석되었다. '경계'라는 키워드의 빈도가 상당히 높게 나타나며, 이는 선원들의 경계 소홀이 사고의 주된 원인임을 시사한다. '장치'와 '부가'와 같은 키워드도 주목할 만한 빈도로 나타나며, 선박의 장치 오류나 부가적인 요소들이 사고에 영향을 미칠 수 있음을 나타낸다.

Fig. 8 (i) Cluster 8에서는 '입항', '계류', '하역', '항해사', '거리', '강풍' 등의 주요 키워드로 구성된다. 이 군집은 주로 계류 중인 선박과의 충돌 사고에 초점을 맞추고 있으며 특정 기상 조건을 포함하는 경우가 많다. '계류'와 '입항'이라는 키워드가 고빈도로 나타나는 것은 선박이 항구에 입항하거나 계류 중일 때 발생하는 사고의 높은 빈도를 보여준다. 특히, 하역 작업 후 급하게 입항하는 과정에서 발생한 사고가 많이 포함된다. '강풍'이라는 키워드 또한 사고 발생에 영향을 미치는 주요 요인으로 도출되며, 강풍 조건에서 선박이 충돌하는 사례가 포함된다.

마지막으로 Fig. 8 (j) Cluster 9에서는 '경계', '일인', '협력', '선적', '선수', '강화', '부근'등의 키워드가 주로 등장하며, 해양 네비게이션과 관련된 사고 중 특히 선수 충돌 사고에 중점을 둔 내용을 보여준다.

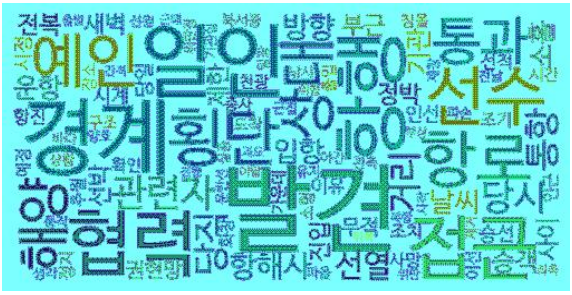
Table 1은 각 군집의 주요 해양안전심판 재결서 키워드와 내용을 요약한 것이다. 본 분석을 통해 각 군집의 키워드와 빈도를 통계적으로 분석하여 사고 원인, 경향, 변화 추이를 연구하고 이를 바탕으로 사고 예방 및 대응 전략을 개발할 수 있다. 예를 들어, 군집 0번의 분석 결과에서는 경계, 발견, 상대, 협력, 일인의 키워드가 발생하며 선박의 경계 의무 미이행이나 부주의가 주된 사고 원인임을 밝혀낼 수 있다. 더 나아가 이러한 사고가 발생하는 시간대, 선박 유형, 항로 특성 등을 파악하여 추가적인 주의가 필요한 구체적인 안전 지침을 개발할 수 있다. 본 군집화 분석은 새로운 재결서가 추가될 때 기존의 군집화 방법을 이용하여 유사한 군집으로 자동 분류하는 것이 가능하며, 이를 통해 지속적으로 사고 분석과 대응 전략을 업데이트할 수 있다. 또한, 각 군집에서도 세부 분류를 통해 사고 원인 및 유형을 더욱 구체적으로 파악하는 것이 가능하다.

(a) Cluster 0



(b) Cluster 1

(c) Cluster 2



(d) Cluster 3

(e) Cluster 4

(f) Cluster 5

(g) Cluster 6

(h) Cluster 7

(i) Cluster 8

(j) Cluster 9

Fig. 8. Word Cloud of Clustering Results.

Table 1. Summary of Key Maritime Incident Keywords by Cluster

군집	주요 키워드	내용
0	경계, 발견, 상대, 협력, 일인	주위에 어선이 있었음에도 oo호 예인선열의 측방 및 후방에 대한 경계를 소홀히 하고 전방 주시만을 하여 이 예인선열 우측에서 이 예인선열의 진로 방향으로 오던 oooooo을 발견하지 못하였다(Korean Maritime Safety Tribunal, 2018a).
		이 충돌사건은 ooooo호 항해당직자가 조타실을 이탈하여 조종불능선 oooo 호를 조기에 발견하지 못함으로써 피하지 아니하여 발생한 것이나 oooo호측이 주위경계를 소홀히 하여 접근하는 oooo호를 발견하지 못함으로써 주의환기신호를 하지 아니한 것도 일인이 된다(Korean Maritime Safety Tribunal, 2007).
1	선적, 장치, 경계, 강조, 거리	출력 882킬로와트 디젤기관 1기를 장치한 부산광역시 선적의 강조 화물선 ooo호가 선장을 포함한 선원 8명이 승선한 가운데 2011년 11월 26일 01시 10분경 일본국 카고가와(加古川)항에서 와이어 로드(Wire Rod) 1,209톤을 적재하고 출항하여 한국의 마산항으로 항해하던 중이던 2011년 11월 27일 05시 01분경 일본국 오키노시마 북방 약 15마일 거리인 북위 34도 29분 30초・동경 130도 04분 36초 해상에서 선장을 포함한 선원 2명이 승선하고 오징어조업 후 항해하던 총톤수 16.00톤인 일본 어선 ooooo와 충돌한 사건이다(Korean Maritime Safety Tribunal, 2012).
		총톤수 1,833톤, 길이 85.65미터, 출력

		1,600마력인 디젤기관 1기를 장치한 강조 화물선 oooo이 2009. 3. 2. 19:33경 마산항 20번 등부표 부근에서 도선사 B의 지휘를 받으며 마산항 4부두 9번선석으로 입항하던 중, 2009. 3. 2. 20:34경 마산항 포스코부두 남측 잔교로부터 226도 방향, 약 350미터 거리인 북위 35도 10분 44초 · 동경 128도 35분 18초 해상에서 화물선 oooo과 피예인부선 ooooo호의 선수 좌현이 충돌한 사건임(Korean Maritime Safety Tribunal, 2009b)
2	경계, 일인, 협력, 일시, 장소	oooo호는 2020. 6. 30. 출항하여 조업하다가, 투망장소를 찾기 위해 항해하던 중 위 사고일시와 장소에서, 2020. 6. 30. 출항하여 외끌이저인망 투망작업 중인 oooo호와 충돌하였다(Korean Maritime Safety Tribunal, 2021).
		해상에 도착하여 전어자망을 투망하고 양망작업에 몰두하느라 주위경계를 소홀히 하여 자선의 후방에서 충돌이 위험을 안고 접근하는 oooo호의 피예인부선 oooo호의 우현선수와 자선의 선미가 충돌한 사건임(Korean Maritime Safety Tribunal, 2008).
3	발견, 경계, 일인, 항행	시계가 양호한 야간에 여수시 거문도 남방 해상에서 항행 중인석유제품운반선과 어선이 충돌한 사안, ooooo호가 경계를 소홀히 하여 충돌의 위험이 발생한 뒤에 금지되어 있는 좌현 변침을 한 것도 일인이 된다(Korean Maritime Safety Tribunal, 2020a).
		이 충돌사건은 제한된 시계 상태에서 oooooo호가 안전한 속력으로 감속하지 않은 채 경계를 소홀히 하여 좌현 변침해 발생한 것이나, ooo이 안전한 속력으로 감속하지 않고 또 경계를 소홀히 하여 상대선과 충분한 거리를 두지 아니한 것도 일인이 된다(Korean Maritime Safety Tribunal, 2020b).
4	강조, 선적, 선수, 장치, 방향, 거리	oo호는 총톤수 1.98톤의 연안자망어업 어선으로 2017. 5. 3. 18:00경 통영시 견유항에서 선장 혼자 승선하고 출항하여 조업을 하다가 다음 날 00:30경 조업지 이동을 위하여 항해등을 점등하지 않고 이동을 하던 중, oo호는 총톤수 2.99톤의 연안복합어업 어선으로 같은 달 3일 23:30경 조업을 마치고 도남동으로 귀항을 하던 중 oo호를 발견하지 못하고 충돌한 사건임(Korean Maritime Safety Tribunal, 2017a).
		05시 05분경 부산광역시 다대포항 경도등대로부터 017도 방향, 약 575미터 거리인 북위 35도 02분 30초 · 동경 128도 59분 03초 해상에서 정박하고 있던 총톤수 569톤, 길이 54.74미터, 너비 16.00미터, 깊이 3.30미터인 부산광역시 선적의 강조 부선 oooooo호와 충돌한 사건으로, oo호는 상갑판 상부 정선수가 파손되었고 낚시승객 4명이 부상을 입었으며, ooooooo호는 선미 좌현에 50센티미터 정도의 굴곡손상을 입었다(Korean Maritime Safety Tribunal, 2009c).
5	계류, 예인, 접근, 정박, 인선	oooo호는 ooo 정선수에서 내어준 길이 30미터의 계류삭(직경 80mm의 합성수지로프)을 예인삭으로 선미에 잡았으며, ooooo호 선장은 감천항 도선점에 도착하자 ooo가 약 3.5노트의 전진타력이 있는 상태에서 예인삭을 풀기 위해 정선하여 예인삭을 풀었으나 같은 날 11시 39분경 북위 35도 02분 48초 · 동경 129도 01분 13초 해상에서 ooo의 정선수와 자선의 우현 선미가 충돌하였다(Korean Maritime Safety Tribunal, 2020c).
		항해 중인 oooo호측이 경계를 소홀히 하여 닻정박 중인 ooooo호를 충돌에 이르기까지 발견하지 못하여 발생한 사건.선박은 정박중에도 경계를 철저히 하여 충돌의 위험을 안고 접근하는 선박이 피항동작을 취하지 않을 때 스스로 충돌을 피하기 위하여 적절한 협력동작을 취해야 한다(Korean Maritime Safety Tribunal, 2010b).
6	상대, 선수, 접근, 거리	충돌한 ooooo호와 ooooo호는 모두 기선권현망어선이다. 이 선박은 다른 기선권현망어선들과 함께 새벽 조업을 위해 사천 삼천포신항에서 30초 간격을 두고 출항하였다. 당시 약 100척의 선박들이 아주 짧은 간격으로 출항하였는데, ooooo호의 뒤에서 바로 ooooo호가 출항하였다. 양 선박의 거리는 짧았고 oooo호의 우현 뒤쪽에 ooooo호는 항해를 하고 있었다. 그러나 속

		력은 ooooo호가 1노트 정도 더 빨랐고 이러한 이유로 ooooo호의 뒤를 ooooo호의 선수로 추돌하였다. 이 충돌로 ooooo호는 별다른 피해가 없었으나 ooooo호는 좌현으로 전복되었다. ooooo호에는 선장 포함 선원 3명이 승선하고 있었으나 모두 무사히 구조되었다(Korean Maritime Safety Tribunal, 2017c).
		이 충돌사건은 해상교통량이 많은 항계 부근에서 oo호 측이 경계를 소홀히 한 상태로 갑자기 속력을 내며 돌진하여 발생한 것이나, oo호가 경계를 소홀히 한 것도 일인이 된다(Korean Maritime Safety Tribunal, 2016a).
		이 충돌사건은 ooooo호가 발전기의 정비·점검을 소홀히 하여 출항 중 갑자기 발전기가 정지되어 조타장치 등에 전원이 공급되지 않아 발생한 것이나 선장이 조타불능상태에서 적절한 비상조치를 하지 않는 것도 일인이 된다(Korean Maritime Safety Tribunal, 2018b).
7	적재, 선수, 경계, 장치, 부가	낚시객을 태우고 출발항으로 귀항하는 선박의 선장이 자동조타로 항해 중 갑자기 기관 경고 알람이 울리자 당황하면서 그 원인을 찾기 위해 선교에 설치된 기관실 CCTV를 살펴보다가 전방 주시를 태만히 하였고 이로 인해 레이더로 초인한 한성호의 존재를 잊어버린 상태로 항행하다가 어로 중인 어선을 추돌하였다. 한편 어로에 종사하던 oo호는 충돌 직전에 oo호를 발견하기는 했으나 유효한 피항협력동작을 취하지 못하였다(Korean Maritime Safety Tribunal, 2017b).
		이 충돌사건은 oooo호가 화물을 실은 부선을 예인하여 전곡항 물양장에서 이안하는 중, 조선부주의로 계류되어 있던 oo호와 oo호를 피하지 못하여 발생한 것이다(Korean Maritime Safety Tribunal, 2019a).
8	입항, 계류, 하역, 항해사, 거리, 강풍	어선 oooooo호는 오징어를 잡는 오징어채낚기 어선이다. 이 선박은 흑산도 인근 해상에서 조업을 하다가 마치고 모항인 구룡포항을 향해 출항하였다. ooooo호 선장은 자동조차 기능을 이용하여 항해를 하면서 모처럼의 모항 입항이라 선교 청소를 하기로 마음먹었다. 이후 선교 청소를 하다가 경계를 소홀히 하였고, 이 선박 진로 앞에서 정류한 채 기관 후진 클러치에 문제가 생겨 정비를 하던 상대선을 발견하지 못하였다. 한편 ooo호는 조업예정지로 항해하다가 기관 후진 클러치에 이상이 있음을 발견하였다. 선장은 해상에 정류한 채 기관 후진 클러치 수리를 하게 하였으며 사고 무렵에는 정비를 마친 채 고장 부위에 대해 기관장과 이야기를 하느라 기관실에 내려가 있었다. 이런 이유로 oooo호 선장 또한 상대선을 발견하지 못하였고, 양 선박은 충돌하였다. 이 사고로 사람은 다치지 않았으며 사고 후 양 선박은 자력항해를 하였다(Korean Maritime Safety Tribunal, 2019b).
		이 충돌사건은 항행중인 동력선 ooooo호가 조타장치의 고장으로 표류하고 있던 oooo호를 피하지 아니함으로써 발생한 것이나 ooo호가 경계를 소홀히 하여 충돌을 피하기 위한 적절한 협력동작을 취하지 못한 것도 일부 원인이 된다.해상에서 서로 충돌하여 ooooo호 선수부가 일부 파손되고, oooo호의 우현 기관실 외판이 파손되어 기관실이 침수된 사건임(Korean Maritime Safety Tribunal, 2009a).
9	경계, 일인, 협력, 선적, 선수	oo호는 총톤수 493.00톤, 길이 59.58미터, 출력 1,029킬로와트 디젤기관 1기를 장치한 부산광역시 선적의 강조 유조선으로 2010년 2월 4일 20시 45분경 부산광역시 남구 용당동 소재 현대저유소부두를 출항하여 여수항으로 가기위해 매물도 북쪽 해상을 항해하던 중, 2010년 2월 4일 23시 51분경 석문도 등대기점 110도 방향 약 1,070미터 거리인 북위 34도 41분 26초·동경 128도 36분 38초 해상에서 총톤수 141.00톤, 길이 26.24미터, 출력 1,470킬로와트 디젤기관 1기를 장치한 부산광역시 선적의 강조 예인선 oo호에 의하여 선미 예인하는 상태로 총중량 미상의 선창 덮개를 선적한 총톤수 1,827.00톤, 길이 72.02미터, 너비 22.00미터, 깊이 5.20미터인 인천광역시 선적의 강조 부선 ooooo호와 충돌한 사건으로 oo호의 예인선열 사이로 진입한 oo호에 의해 예인

삭이 끊어지고 전진타력이 있던 피예인부선 oooooo호의 좌현 선수 모서리 부분과 oo호의 4번 화물창 좌현부분이 충돌하였다 (Korean Maritime Safety Tribunal, 2010a).

5. 결 론

본 연구에서는 선박 충돌사고의 원인을 깊이 분석하기 위해 부산 해심 지역의 최근 20년간 사고 재결서 데이터를 SBERT 모델을 이용하여 분석하였다. 이 연구를 통해 사고 원인 및 유형별 문장 간의 의미와 문맥 정보를 고려하여 군집을 나눌 수 있었다. 각 군집은 특정 키워드의 출현 빈도를 통해 선박 운행 중에 일어날 수 있는 주된 사고 유형과 그 원인(예: 경계 소홀, 기기의 오류 등)을 선별하고 분석한다. 군집 내에서 주목할 만한 키워드들은 높은 빈도로 나타나며, 관련 사고 유형의 주요 특징과 발생 요인을 명확하게 설명해준다. 이러한 분석을 통해, 선박 운항 중 경계 유지의 중요성, 상호 작용과 기상 조건이 사고 발생에 어떻게 영향을 미치는지에 대해 이해할 수 있게 된다. 또한 이 연구는 다양한 선박 충돌 사고 유형을 분류하고 원인을 상세히 조사함으로써, 더욱 효과적인 예방 방안을 도출할 수 있는 기반을 제공한다.

제안한 프레임워크는 선박 운행 중에 일어날 수 있는 주된 사고 유형과 그 원인을 선별하고 분석할 수 있는 장점이 있음에도 불구하고, 다음과 같은 한계점이 존재한다. 첫째, 본 연구는 각각의 시간대에 따른 사고 패턴과 해양안전심판 재결서의 변동성을 고려하지 못했다. 이는 특정 시간대에 나타나는 잠재적인 위험 요소를 정밀하게 식별하는 데에 한계가 있다. 둘째, 본 연구에서는 오직 부산해심 지역의 충돌 사고 재결서를 중심으로 조사 및 분석을 했기 때문에 연구 결과를 모든 해상 사고에 적용해 일반화하기에는 무리가 있다. 또한, 사용 가능한 데이터의 양이 제한적이며, 주로 소형 선박 사고가 재결서로 기록되는 경향이 있어, 이는 본 연구의 결과를 전체 해양 사고에 적용하는 데 한계로 작용한다. 게다가 수집된 데이터는 20년에 걸쳐 축적된 것이므로, 시간이 지남에 따라 변화하는 해양 교통환경을 완전히 반영하지 못할 수도 있다는 점도 고려해야 한다. 셋째, 본 연구에서는 키워드 간 상호작용이나 상관 관계 분석을 통해 변수들 사이의 연관성을 고려하지 않았다.결론적으로 본 연구는 부산해심 지역뿐만 아니라 다른 지역에서도 해양 사고 예방 및 대응 전략의 효율성을 향상시킬 수 있는 기반을 제공한다. 이를 통해 부산해심 지역에서의 사고 빈도 감소뿐만 아니라, 글로벌 규모에서의 사고 예방에도 중요한 역할을 할 수 있을 것으로 기대된다.

본 연구의 한계점을 극복하기 위해 향후 연구에서는 다음과 같은 3가지 추후 연구가 진행되어야한다. 먼저, 본 연구는 각각의 시간대에 따른 사고 패턴과 해양안전심판 재결서의 변동성을 고하기 위해 시간대별 사고 패턴의 변화와 해양안전심판 재결서의 변화를 조사하며 분석해야 할 필요가 있다. 다음으로, 지역적 특성에 관계없이 제안 방법을 일반화하기 위해 부산해역을 넘어 다른 지역 및 해외의 충돌 사고 재결서를 포함하여 분석을 확장할 필요성이 있다. 마지막으로 키워드 간 상호작용이나 상관 관계 분석을 통해 변수들 사이의 연관성을 고려한 분석을 통해 각 군집의 특성을 단일 단어 출현 빈도에만 의존하지 않고 군집별 주요 키워드들의 연결성을 분석하면 더 심층적인 해석 도출이 가능할 것이다. 이러한 향후 연구를 통해서 폭넓은 관점에서 해상 충돌 사고의 원인을 조사하고, 국내외 사고 원인을 비교 연구함으로써 국내외 해양 안전을 강화할 수 있는 방안을 제시할 수 있을 것으로 기대된다.

Acknowledgement

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구 (No.RS-2023-00218913)와 해양수산부 재원으로 선박해양플랜트연구소의 기본사업인 “스마트 해양안전 및 기업지원을 위한 오픈플랫폼 기술개발”에 의해 수행되었습니다 (1525014880, PES4880).

References

[1] Abualigah, L. M., A. T. Khader, and M. A. Al-Betar(2016), Multi-objectives-based text clustering technique using K-mean algorithm. In 2016 7th International Conference on Computer Science and Information Technology (CSIT), IEEE, pp. 1-6.

[2] Ashari, I. F., E. D. Nugroho, R. Baraku, I. N. Yanda, and R. Liwardana(2023), Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. Journal of Applied Informatics and Computing, 7(1), 95-103.

[3] Chen, P., Y. Huang, J. Mou, and P. Van Gelder(2018), Ship collision candidate detection method: A velocity obstacle approach. Ocean Engineering, 170, pp. 186-198.

[4] Cho, D. O., J. Y. Mok, and Y. U. Park(2002), The direction of development for the maritime safety tribunal system in Korea. Han'guk Haeyang Susan Kaebarwŏn.

[5] Choi, C. W., Y. N. Roh, D. S. Shin, H. M. Kim, and H. C. Park(2021), Identifying Risk Factors of Marine Accidents in

Coastal Area by Marine Accident Types. Journal of the Korean Society of Transportation, 39(4), 540-554.

[6] Devlin, J., M. W. Chang, K. Lee, and K. Toutanova(2018), Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7] Fan, S., E. Blanco-Davis, Z. Yang, J. Zhang, and X. Yan (2020), Incorporation of human factors into maritime accident analysis using a data-driven Bayesian network. Reliability Engineering & System Safety, 203, 107070.

[8] Faruqui, M., Y. Tsvetkov, P. Rastogi, and C. Dyer(2016), Problems with evaluation of word embeddings using word similarity tasks. ACL 2016, 30.

[9] Ham, J., Y. J. Choe, K. Park, I. Choi, and H. Soh(2020), KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. arXiv preprint arXiv:2004.03289.

[10] Han, Y. J.(2022), Development of risk leading indicators by sea area based on ship operation characteristics (Master's thesis). Pusan National University.

[11] He, A., C. Luo, X. Tian, and W. Zeng(2018), A twofold siamese network for real-time object tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4834-4843.

[12] Huang, Y., P. Van Gelder, and Y. Wen(2018), Velocity obstacle algorithms for collision prevention at sea. Ocean Engineering, 151, pp. 308-321.

[13] Jee, T. C., H. J. Lee, and Y. B. Lee(2007), Determining the number of Clusters in On-Line Document Clustering Algorithm. The KIPS Transactions: PartB, 14(7), 513-522.

[14] Joshi, A., A. Kajale, J. Gadre, S.Deode, and R. Joshi(2023), L3Cube-MahaSBERT and HindSBERT: Sentence BERT Models and Benchmarking BERT Sentence Representations for Hindi and Marathi. In Science and Information Conference (pp. 1184-1199). Cham: Springer Nature Switzerland.

[15] Jung, C. H.(2018), A study on the improvement of safety by accidents analysis of fishing vessels, J. Fish. Mar. Sci. Educ, Vol. 30, pp. 179-186.

[16] Kadhim, A. I., Y. N. Cheah, and N. H. Ahamed(2014), Text document preprocessing and dimension reduction techniques for text document clustering. In 2014 4th international conference on artificial intelligence with applications in engineering and technology, IEEE, pp. 69-73.

[17] Kalra, V. and R. Aggarwal(2017), Importance of Text Data Preprocessing & Implementation in RapidMiner. ICITKM, 14, pp. 71-75.

[18] Kassambara, A.(2017), Practical guide to cluster analysis in R: Unsupervised machine learning, Vol. 1.

[19] Kim, G. and H. Kim(2011), Development of ship safety navigation supporting equipmentusing infrared LED. Journal of the Korea Institute of Information Technology, 9(2), pp. 27-32.

[20] Kim, S.-K. and J.-P. Kang(2011), A Study on the Relationships between the Casualties of Fishing Boats and Meteorological Factors (Doctoral dissertation).

[21] Kim, W.-S., Y.-K. Hyun, and Y.-W. Lee(2020), Risk factors of fisher on stow net fishing vessel using analysis of adjudication, Journal of the Korean Society of Fisheries and Ocean Technology, Vol. 56, pp. 155-162.

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

[23] Korean Maritime Safety Tribunal(2007), Busan Regional Maritime Safety Tribunal Decision 2007-036.

[24] Korean Maritime Safety Tribunal(2008), Busan Regional Maritime Safety Tribunal Decision 2008-002.

[25] Korean Maritime Safety Tribunal(2009a), Busan Regional Maritime Safety Tribunal Decision 2009-032.

[26] Korean Maritime Safety Tribunal(2009b), Busan Regional Maritime Safety Tribunal Decision 2009-039.

[27] Korean Maritime Safety Tribunal(2009c), Busan Regional Maritime Safety Tribunal Decision 2009-059.

[28] Korean Maritime Safety Tribunal(2010a), Busan Regional Maritime Safety Tribunal Decision 2010-020.

[29] Korean Maritime Safety Tribunal(2010b), Busan Regional Maritime Safety Tribunal Decision 2010-062.

[30] Korean Maritime Safety Tribunal(2012), Busan Regional Maritime Safety Tribunal Decision 2012-046.

[31] Korean Maritime Safety Tribunal(2016a), Busan Regional Maritime Safety Tribunal Decision 2016-052.

[32] Korean Maritime Safety Tribunal(2016b), Busan Regional Maritime Safety Tribunal Decision 2016-061.

[33] Korean Maritime Safety Tribunal(2017a), Busan Regional Maritime Safety Tribunal Decision 2017-054: Summary of the Collision Case between Fishing Vessels Deukyongho and Buyeongho.

[34] Korean Maritime Safety Tribunal(2017b), Busan Regional Maritime Safety Tribunal Decision 2017-058: Summary of the Collision Case between Fishing Vessel Geoseongho and Hansungho.

[35] Korean Maritime Safety Tribunal(2017c), Busan Regional

Maritime Safety Tribunal Decision 2017-069: Summary.
[36] Korean Maritime Safety Tribunal(2018a), Busan Regional Maritime Safety Tribunal Decision 2018-021: Summary.
[37] Korean Maritime Safety Tribunal(2018b), Busan Regional Maritime Safety Tribunal Decision 2018-069: Summary.
[38] Korean Maritime Safety Tribunal(2019a), Busan Regional Maritime Safety Tribunal Decision 2019-007: Summary.
[39] Korean Maritime Safety Tribunal(2019b), Busan Regional Maritime Safety Tribunal Decision 2019-018: Summary.
[40] Korean Maritime Safety Tribunal(2020a), Busan Regional Maritime Safety Tribunal Decision 2020-008: Summary.
[41] Korean Maritime Safety Tribunal(2020b), Busan Regional Maritime Safety Tribunal Decision 2020-024: Summary.
[42] Korean Maritime Safety Tribunal(2020c), Busan Regional Maritime Safety Tribunal Decision 2020-086: Summary.
[43] Korean Maritime Safety Tribunal(2021), Busan Regional Maritime Safety Tribunal Decision 2021-024: Summary.
[44] KMST. (2022). Marine accident statistics and casebook, 12.
[45] Lee, J. S., B. K. Lee, and I. S. Cho(2019), Text Mining Analysis Technique on ECDIS Accident Report. Journal of the Korean Society of Marine Environment & Safety, 25(4), 405-412.
[46] Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov(2019), Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
[47] Madhulatha, T. S.(2012), An overview on clustering methods. arXiv preprint arXiv:1205.1117.
[48] Park, H., M. A. Cheon, Y. Namgung, H. Yoon, M. S. Choi, J. G. Kim, and J. H. Kim(2020), Classification of vessel accidents according to word and sentence embedding. Proceedings of the Korean Institute of Information Scientists and Engineers Conference, 413-415.
[49] Park, S. -A. and D. -J. Park(2023), A study on the analysis of marine accidents on fishing ships using accident cause data, Journal of Korean Navigation and Port Research, Vol. 47-1, pp. 1-9.
[50] Reimers, N. and I. Gurevych(2019), Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
[51] Soni, N. and A. Ganatra(2012), Categorization of several clustering algorithms from different perspective: a review. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 8, pp. 63-68.
[52] Tan, P. -N., M. Steinbach, and V. Kumar(2005), Introduction to Data Mining, Addison-Wesley, ISBN 0-321-32136-7, Chapter 8, page 500.
[53] Tirunagari, S., N. Poh, D. Windridge, A.Iorliam, N. Suki, and A. T. Ho(2015), Detection of face spoofing using visual dynamics. IEEE transactions on information forensics and security, 10(4), pp. 762-777.
[54] Vijaymeena, M. K. and K. Kavitha(2016), A survey on similarity measures in text mining. Machine Learning and Applications: An International Journal, 3(2), pp. 19-28.
[55] Wolfram Research(2007), CosineDistance - Wolfram Language & System Documentation Center, wolfram.com.
[56] Xie, J. and S. Jiang(2010), A simple and fast algorithm for global k-means clustering. 2010 Second International Workshop on Education Technology and Computer Science.

Received : 2023. 10. 19.
Revised : 2023. 11. 09.
Accepted : 2023. 12. 29.