Journal of the Korean Society of Safety
Vol. 36, No. 5, pp. 52-60, October 2021
https://doi.org/10.14346/JKOSOS.2021.36.5.52

ISSN 1738-3803 (Print)
ISSN 2383-9953 (Online)
http://www.kosos.or.kr/jkosos

사망사고와 부상사고의 산업재해분류를 위한 기계학습 접근법

강성식* · 장성록** · 서용윤***†

Machine Learning Approach to Classifying Fatal and Non-Fatal Accidents in Industries

Sungsik Kang* · Seong Rok Chang** · Yongyoon Suh***†

†Corresponding Author
Yongyoon Suh
Tel : +82-2-2260-3786
E-mail : ysuh@dgu.edu

Received : June 24, 2021
Revised : July 22, 2021
Accepted : August 5, 2021

Abstract : As the prevention of fatal accidents is considered an essential part of social responsibilities, both government and individual have devoted efforts to mitigate the unsafe conditions and behaviors that facilitate accidents. Several studies have analyzed the factors that cause fatal accidents and compared them to those of non-fatal accidents. However, studies on mathematical and systematic analysis techniques for identifying the features of fatal accidents are rare. Recently, various industrial fields have employed machine learning algorithms. This study aimed to apply machine learning algorithms for the classification of fatal and non-fatal accidents based on the features of each accident. These features were obtained by text mining literature on accidents. The classification was performed using four machine learning algorithms, which are widely used in industrial fields, including logistic regression, decision tree, neural network, and support vector machine algorithms. The results revealed that the machine learning algorithms exhibited a high accuracy for the classification of accidents into the two categories. In addition, the importance of comparing similar cases between fatal and non-fatal accidents was discussed. This study presented a method for classifying accidents using machine learning algorithms based on the reports on previous studies on accidents.

Key Words : machine learning, narrative texts, textmining, fatal accidents, non-fatal accidents, classification

1. 서 론

2019년 산업재해현황분석에 따르면 국내 업무상 사고사망자 수는 855명, 업무상 사고재해자 수는 94,047명으로 나타났다. 산업재해현황분석은 유사, 동종업종의 사고를 예방하기 위한 목적으로 시행되고 있다. 하인리히(H.W. Heinrich)는 사망사고와 같은 대형사고는 우연히 발생하지 않으며, 사망사고 발생 이전에 여러 번의 경미한 사고가 발생함을 지적하였다. 이에 따라 사망사고를 예방하기 위하여 사고분석을 통해 경미한 사고에 대한 원인을 파악하고 개선하여 대형사고 및 인명피해를 예방할 수 있다는 이론을 제시하였다[1]. 이는 사망사고로 이어질 수 있는 부상사고의 분석을 통해 사망사고를 예방하고, 이를 위해 안전관리를 통계적으로 확인하는 것이 중요하다는 것을 강조하고 있다.

사고분석을 위하여 활용될 수 있는 자료 중 비교적 쉽게 접근할 수 있는 자료가 산업재해조사표와 같은 재해보고문서이다. 재해보고문서는 유사 및 동종재해 예방을 위해 가장 기본적으로 활용되는 자료로서, 사고 일자와 업종·재해형태·재해자 인적사항 등 전반적인 사고의 정보와 발생 과정을 포함한 사고개요가 서술되어 있다. 사고개요는 사고 발생 시 현장의 작업자와 관리감독자, 안전관리자가 조사하여 서술형으로 작성되며, 수집기관의 담당자에 의해 관리되고 있다[2].

*부경대학교 안전공학과 박사과정 (Department of Safety Engineering, Pukyong National University)
**부경대학교 안전공학과 교수 (Department of Safety Engineering, Pukyong National University)
***동국대학교(서울캠퍼스) 산업시스템공학과 부교수 (Department of Industrial and Systems Engineering, Dongguk University(Seoul Campus))

이와 같은 재해보고문서의 유용성에 따라 학계에서도 이를 체계적으로 분석하는 다양한 연구들이 진행되고 있다. 특히, 재해보고문서 내에서 서술 내용 및 형태 즉, 정성적으로 작성된 사고개요를 정량적이고 체계적으로 분석하기 위하여 텍스트마이닝의 전처리와 기계학습을 활용하는 연구들이 진행되고 있다[3-5].

그러나 하인리히가 주장하고 지금까지 사고관리의 중요한 이슈인 사망사고와 부상사고의 관계에 초점을 맞추어, 사망사고와 부상사고를 분류하고 두 사고 간의 위험성 및 위험요인을 비교・분석한 연구는 아직까지 미비한 실정이다. 사고는 기인물이나 재해형태 등 다양한 요소들로 인해 발생하며, 사고의 결과는 사망 또는 부상으로 다르게 나타난다. 따라서, 두 사고 간에 위험요인의 차이를 파악하고 부상사고가 사망사고로 이어지지 않는 시사점을 파악해야 한다. 이를 위해, 재해보고문서에 포함된 사고개요 내용에 대해 텍스트마이닝과 기계학습을 활용하여 사망사고와 부상사고를 분류하고, 각 사고 결과에 영향을 미치는 위험요인 탐색과 키워드 분석이 요구된다.

본 연구에서는 사고문서의 분류를 위하여 기계학습 알고리즘을 적용하여 사망사고와 부상사고 분류 결과를 비교분석하고, 분류된 사고를 결정짓는 주요 위험요인 키워드를 도출하고자 한다. 먼저, 기계학습 알고리즘을 적용하기 위하여 비정형데이터인 사고개요를 텍스트마이닝을 활용하여 정형데이터로 변환하는 전처리 과정을 거친다[6]. 키워드와 문서로 이루어진 정형데이터는 문서의 양이 늘어남에 따라 도출되는 키워드가 급격하게 증가하여 분석의 어려움이 있다. 이를 해결하기 위해 주성분 분석(principal component analysis: PCA)을 통해 특성을 선정하고(feature selection)[7], 위험요인 키워드를 주요 특성으로 차원을 축소한다. 다음으로 지도학습 기반의 기계학습 알고리즘을 활용하여 특성에 따라 사망사고와 부상사고를 분류한다.

본 연구에서는 사망사고와 부상사고의 비교분석을 위하여 네 개의 대표적인 기계학습 알고리즘인 로지스틱 회귀분석(logistic regression), 의사결정나무분석(decision tree), 지지벡터 기계분석(support vector machine: SVM), 신경망 분석(neural network)을 사용하여 사고문서를 분류한다. 사망사고와 부상사고의 비교를 위해, 각 알고리즘의 문서분류 정확도를 도출하고 특히, 오분류된 문서를 통해 사망사고와 부상사고의 차이를 나타내는 핵심 위험요인을 도출한다. 이를 통해 부상사고에서 사망사고로 이어질 수 있는 위험요인을 도출하여, 사망사고를 감소시키고 이어 부상사고까지 관리할 수 있는 방안을 모색하고자 한다.

2. 기존연구 : 데이터 분석 기반 사고분석

사고개요의 분석은 서술형으로 작성되는 특성에 따라 텍스트마이닝을 활용하여 다양하게 분석되고 있다. 재해보고문서 내에서 서술 형태 즉, 정성적으로 작성된 사고개요를 정량적으로 분석하기 위하여 텍스트마이닝을 활용한 연구들이 진행되었다. 예를 들어, 재해보고문서에 자주 나타나는 키워드들을 도출하여 키워드들의 상관관계를 시각화하거나 사고 발생 유형을 체계화하는 연구들이 진행되었다. 특히, 텍스트마이닝을 기반으로 한 사고 분류 모형과 온톨로지 개발연구[4]는 안전보건공단에서 사용하는 사고 분류 기준을 참고하여 업종과 작업유형, 기인물, 사고유형, 사망자 수에 대한 계층 온톨로지를 개발하였다. 또한 SVM을 활용하여 사고개요를 분류하였으며, 이를 통해 사고 분류 모형을 제시하였다.

또한 데이터마이닝을 활용한 문서분류와 관련된 연구는 지도학습 기반의 기계학습 알고리즘이 주로 적용되고 있다[8,9]. 지도학습은 목적변수와 입력데이터를 활용하여 생성한 훈련 데이터를 학습하고 학습된 결과를 바탕으로 데이터를 분류하고 예측하기 위한 기계학습이며, 데이터에 목푯값이 개입되어 분석 정확도가 비지도 학습보다 높은 장점이 있다. 지도학습의 유형은 유추한 함수 중 연속적인 값을 출력하는 회귀분석(regression)과 주어진 입력데이터가 어떤 종류의 값인지 나타내는 분류(classification)로 구분할 수 있다. 지도학습 기반의 기계학습 알고리즘은 예측 모델을 생성하여 문서분류뿐만 아니라 패턴인식과 질병 진단, 주가 예측 등 다양한 분야에서 사용되고 있다[10-12].

3. 연구방법론

3.1. 연구절차

본 연구는 Fig. 1과 같은 절차로 연구를 진행하였다. 먼저, 부상사고와 사망사고의 사고개요를 수집하였으며, 텍스트마이닝 기법을 활용하여 위험요인 키워드를 추출하고, PCA를 활용하여 사고의 특성을 선정하여 매트릭스를 작성하였다. 이 정형화된 매트릭스 데이터를 k-fold cross validation을 통해 training data set와 test data set으로 분할하였으며, 총 4개의 분류모델을 사고문서에 적용하였다. 분류모델은 logistic regression과 decision tree, SVM, neural network를 활용하여 사고문서를 분류하는 모델을 작성하였다. 다음으로 통계적 가설검증 방법인 confusion matrix를 활용하여 정확도 지표인 정밀도와 재현율, 정확도를 통해 각 모델을 정

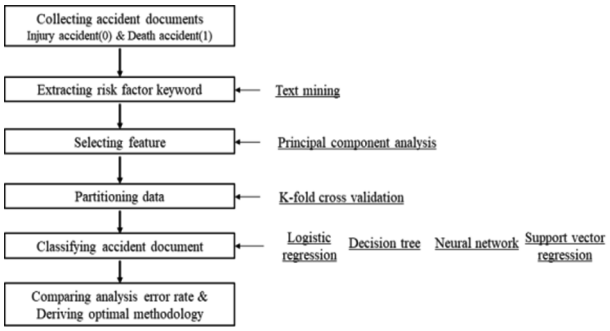

Fig. 1. Flowchart for the methodology.

확도를 평가하였다. 마지막으로 각 알고리즘에 대하여 사고를 분류하는 주요 키워드를 도출하였으며, 오분류된 문서에 대한 키워드를 분석하였다.

3.2. 분석방법론

3.2.1. Text mining

텍스트마이닝은 자연어처리 기술에 기반하여 비정형데이터에서 유용한 정보를 추출하고 가공하는 데이터분석 방법론이다[13]. 빈도가 높은 키워드를 위주로 문서와 키워드로 이루어진 정형데이터로 변환하여 분석한다. 이를 위해, 텍스트 문서를 수집하여 문서를 corpus(말뭉치)로 나누고 불용어의 전처리 과정을 실시하고 정형데이터로 변환하여 의미 있는 지식을 추출한다. 본 연구에서는 텍스트마이닝 방법론 중 하나인 TF-IDF를 사용하였다. 이 방법론은 문서 군에서 단어가 특정 문서 내에서 얼마나 중요한지를 나타내는 통계적 수치로 도출된 단어에 대하여 가중치를 표현하는 방법론이다. 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 지표인 TF(term frequency)와 한 단어가 문서 집합 전체에서 공통적으로 나타나는 정도인 IDF (inverse document frequency)를 곱하여 나타낸다.

3.2.2. 특성 추출: 주성분 분석

PCA는 다변량 데이터를 효과적으로 분석하기 위한 대표적인 분석 기법이며, 정형데이터에서 키워드의 특성을 나타내는 변수를 추출하거나 고차원 데이터를 저차원 데이터로 차원을 축소하기 위한 방법론으로 차원 축소와 시각화, 군집화, 압축 등 다양하게 활용된다. 또한 데이터의 분산을 최대한 보존하면서 서로 직교하는 새 축을 찾아 고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간으로 변환하여 데이터 분포에서 분산력이 가장 큰 직선(혹은 평면)을 찾는 방법론이다[7]. 본 연구에서는 수치화된 단어의 가중치를 활용하여 데이터의 차원을 축소하기 위해 사용하였다. PCA에서 사용되는 변수추출(feature extraction)은 기존 변수를 조합하여 새로운 변수를 만드는 기법으로, 기존의 변수를 선형결합(linear combination)하여 새로운 변수를 만들어 내는 방법을 사용한다. 연구자가 분석에 적합한 principal component(PC)의 개수를 선정하여 차원을 설정하며, 유사한 키워드들의 조합으로 특성이 선정된다.

3.2.3. 데이터 분할: k-fold cross validation

데이터 분할은 모델을 평가하여 검증하기 위해 사용되며, 전체 데이터를 training data set과 test data set으로 나누어 training data set을 사용하여 모델을 학습시키고, test data set으로 그 모델 성능을 평가한다. 그러나 전체 데이터의 일부인 test data set을 사용해 모델 성능을 평가하는 것의 문제는 test data set이 모델에 학습되지 않기 때문에, test data set의 선정에 따라 성능이 다르게 나타날 수 있어 test data set에 대한 성능 평가의 신뢰성이 떨어지게 된다. 이와 같은 문제를 해결하기 위하여 k-fold cross validation을 사용하였다. 이는 모든 데이터를 최소 한 번은 test data set으로 사용하여 모델 성능을 평가하는 방법론이며, 추가적인 데이터를 수집하기 어렵거나 비용이 많이 드는 경우에 사용된다[14]. 또한 기존의 training set과 validation set, test set 세 개의 집단으로 분류하는 것보다 training set과 test set으로만 분류할 때 학습 data set을 더 많이 생성할 수 있기 때문에 더 정확한 분류모델을 만들기 위하여 사용된다. K-fold cross validation은 Fig. 2와 같이 전체 데이터를 k개의 그룹으로 나누어 (k-1)개의 test fold와 1개의 validation fold로 지정하여 총 k회의 교차검증을 통해 각 hyperparameter를 지정하고 분할 정확도가 가장 높은 hyperparameter를 적용하여 분류모델을 평가한다. 또한 k값은 전체 데이터에 대하여 민감도 분석을 통해 설정한다. 민감도 분석은 fold의 개수에 따라 데이터의 분산과 편향을 계산하여 최소 및 최대 정확도를 도출할 수 있으며, 정확도가 가장 높은 k값으로 분석을 수행한다.

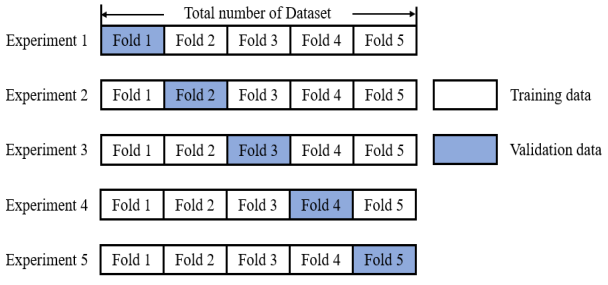

Fig. 2. Partition of data set.

3.2.4. 데이터 분류모델

3.2.4.1. Logistic regression

Logistic regression은 이진 목적 데이터(binary target data)의 분류를 위한 가장 전통적인 방법이며, 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하기 위해 활용되는 통계기법이다[15]. Logistic regression은 종속 변수와 독립변수 간의 관계를 독립변수의 선형 결합을 이용하여 사건의 발생 가능성에 대한 예측 모델에 사용한다. 선형 회귀분석과의 차이점으로는 종속 변수가 범주형 데이터를 대상으로 하며 예측값의 범위가 0과 1사이의 확률값으로 제한되고, 종속 변수가 이진으로 나타나기 때문에 조건부 확률의 분포가 정규분포 대신 이항 분포를 따른다. 이와 같이 이진종속변수에 대해 각 독립변수의 회귀계수의 표준 추정값을 최적화하여 도출한다.

3.2.4.2. Decision tree analysis

Decision tree analysis는 나무구조를 통해 요인 간의 관계를 분화시켜 분류 및 예측을 수행하는 방법론이다[16]. 이 방법론은 가장 위에 있는 부모 마디로부터 가지의 조건에 따라서 자식 마디로 분화하며 데이터를 분류한다. 데이터의 분화는 지니지수와 엔트로피, 카이제곱의 지표를 활용하여 데이터 분류의 순도를 계산하여 분류한다. 첫 번째로, 지니지수는 사회과학과 경제학에서 사용되는 척도이며, 무작위로 두 개의 데이터를 뽑았을 때 그 둘이 서로 같은 클래스일 확률을 나타낸다. 지니지수를 활용한 데이터의 분화는 지수를 최대화하는 방향으로 진행하며, CART(classification and regression trees) 알고리즘을 활용하여 분류모델을 형성한다. 두 번째로, 엔트로피는 정보이론에서 사용되며 메시지 당 평균 정보량을 측정하는 척도이다. 엔트로피는 불순도를 의미하며, 이를 최소화하는 방향으로 분화가 진행된다. 엔트로피를 활용한 decision tree 분석은 ID(iterative dichotomiser)3와 C4.5, C5.0 알고리즘을 주로 활용한다. 마지막으로 카이제곱은 실제 관찰된 관찰도수와 이론적 빈도와 기대도수의 차이가 있는지를 검정하는 척도이며, 최초 데이터 수의 비율과 관련이 있다. 데이터의 분화는 카이제곱값을 최대화하는 방향으로 진행하며, CHAID(chi-square automatic interaction detection) 알고리즘을 활용하여 분류모델을 작성한다.

3.2.4.3. Neural network analysis

Neural network analysis는 인간의 학습체계를 비유하여 생물학적 두뇌의 신경세포(neuron)를 수학적으로 모델링한 인공지능의 한 분야이다[17]. 일반적으로 널리 사용되는 역전파 인공신경망 분석은 입력층(input layer)과 은닉층(hidden layer), 출력층(output layer)으로 구성되고 각 층은 노드들로 구성된다. 이와 같은 노드는 가중치와 바이어스 값의 조합을 포함하고 있으며, 전파되는 과정에서 가중치와 바이어스 값을 조절하여 최적화된 노드를 갖게 된다. 인공신경망 분석은 전진파와 역전파로 데이터를 분석하며, 각 뉴런의 가중치를 입력값에 따라 업데이트하면서 최종 결과치에 가장 유사한 값을 도출하도록 학습하는 모델이다. 또한 계산된 출력값과 실제 값의 차이를 계산하여 에러를 최소화하도록 역방향으로 가중치를 수정하여 모델의 정확도를 높인다[18].

3.2.4.4. Support vector machine (SVM)

SVM은 기계학습 중 하나로 패턴인식이나 자료 분석을 위한 지도학습 모델이며, 구조적 위험 최소화(structural risk minimization)를 기반으로 하여 일반화 오류의 상한을 최소화할 수 있는 머신러닝 기법이다[19]. 주어진 데이터 집합을 바탕으로 새로운 데이터가 어느 카테고리에 속할지 판단하기 위해 사용되며, 비확률적 이진 선형 분류모델을 작성하여 데이터를 분류할 수 있는 가장 큰 폭을 가진 경계, 즉 최적의 회귀식을 찾기 위한 방법론이다. SVM은 범주 혹은 수치 예측 문제에 대해 사용할 수 있으며, 노이즈 데이터에 영향을 크게 받지 않고 과적합화가 적게 나타나는 장점이 있다. 그러나 최적의 모델을 찾기 위해 커널과 모델에서 매개변수의 조합에 대한 테스트가 필요하며, 입력 데이터셋의 크기가 커질수록 다른 방법론들보다 훈련 속도가 크게 느려지는 단점이 있다.

3.3. 분류결과 검증

분류모델을 평가하는 방법은 training data set으로 분류모델을 작성하고 test data set으로 분류모델의 정확도를 평가한다. 각 분류방법론의 정확도는 Table 1과 같은 confusion matrix를 통해 정확도 지표들을 계산하여 평가한다. 모델을 평가하는 요소는 모델의 결괏값과 실제값의 관계로 정의할 수 있다[20]. 결괏값은 옳게 예측할 경우인 true와 예측에 틀렸을 경우인 false로 나누어져 있으며, 사망사고를 positive로, 부상사고를 negative로 이루어진 2*2 매트릭스로 표현할 수 있다. 또한 실제 발생한 부상사고를 귀무가설로 설정하여 분석을 수행하였다. Confusion matrix를 살펴보면 true positive(TP)는 실제 positive인 정답을 positive라고 예측하여 옳게 예측한 경우이다. 또한 false positive(FP)는 실제 negative인 정답을 positive라고 예측하여 오답으

Table 1. Confusion matrix

	Predicted : Positive	Predicted : Negative
Actual : Positive	True Positive	False Negative Type Ⅱ(Beta) Error
Actual : Negative (Null hypothesis)	False Positive Type Ⅰ(Alpha) Error	True Negative

로 분석된다. 다음으로 false negative(FN)는 실제 positive인 정답을 negative라고 예측하여 오답으로 나타난 경우이다. 마지막으로 true negative(TN)는 실제 negative인 정답을 negative라고 예측하여 정답으로 나타난 경우이다. 이와 같은 confusion matrix를 통해 정확도 지표를 활용하여 오류를 정량화하여 평가한다. 이는 통계의 1종 오류(Type I error)와 2종 오류(Type II error)와도 연결되는데, 귀무가설은 일반적으로 다수 표본인 negative로 인식하여, false positive가 1종 오류, false negative가 2종 오류를 설명하게 된다.

정확도 지표로는 정밀도(precision)와 재현율(recall), 정확도(accuracy)로 분류모델의 정확도를 평가할 수 있다. 정밀도(precision)란 실제 데이터에서 예측 데이터가 얼마나 같은지를 판단하는 자료로, 모델이 positive라고 분류한 것 중에서 실제 positive인 것의 비율로 수식 (1)과 같다. 재현율(recall)이란 실제 positive인 것 중에서 모델이 positive라고 예측한 것의 비율을 나타내며, 아래와 같은 수식 (2)로 계산한다. 정확도(accuracy)란 전체 데이터에서 올바르게 예측한 데이터의 비율을 의미하며, 수식 (3)으로 표현한다.

$$(\mathrm{Precision}) = \frac{TP}{TP+FP} \quad \cdots (1)$$

$$(Recall) = \frac{TP}{TP+FN} \quad \cdots (2)$$

$$(Accuracy) = \frac{TN+TP}{TN+FP+FN+TP} \quad \cdots (3)$$

4. 연구결과 : 건설업 사례분석

4.1. 데이터 수집

분석 데이터는 재해율이 가장 높은 업종인 건설업에 대하여 2012년부터 2017년까지 발생한 부상사고와 사망사고의 사고 발생 개요를 수집하였다. 그러나 단일 년도로는 사망과 부상사고의 케이스 불균형(imbalance)으로 인해 분류모델의 분석이 부적합하다. 이에 따라 2017년에 발생한 24,133개의 부상사고와 2012년부터 2017년까지 발생한 2,853개의 사망사고를 대상으로 분석하였다. 사고개요에서 도출할 수 있는 변수로는 연령, 업종, 사업장의 규모를 포함한 재해자의 기본 정보부터 재해 일자와 발생형태를 나타내는 사고의 전반적인 내용이 포함되어 있다. 이와 같은 사고개요 데이터를 활용하여 각 문서에서 부상사고는 0, 사망사고는 1로 표기하여 부상사고와 사망사고의 분류모델을 작성하였다.

4.2. 데이터 전처리 및 특성 추출

먼저, 수집된 텍스트데이터에서 프로그램 R의 tm 패키지와 KoNLP 패키지로 키워드를 추출하였다. 또한 위험요인의 효율적인 분석을 위하여 글자 수 2~5개로 이루어진 명사를 추출하였다. 다음으로 분석에 필요하지 않은 데이터를 전처리하였다. 정확한 단어로 이루어져 있지 않은 불완전한 데이터와 분석에 부적절한 단어를 제거하고, 조사나 기호와 같은 불용어를 제거하였다. 전처리한 데이터를 TF-IDF를 활용하여 문서와 키워드로 이루어진 document-term matrix로 변환하였으며, 27,285개의 문서와 100개의 키워드로 이루어진 행렬을 작성하였다. 도출 빈도가 높은 상위 키워드 100개를 추출하였으며, TF-IDF를 통해 도출된 가중치 값을 계산하여 해당 키워드를 포함하고 있지 않은 문서를 제거하고 분석을 수행하였다. 이를 통해 최종적으로 총 100개의 키워드와 25,837개의 문서에 대하여 분석을 수행하였다.

키워드의 특성을 선정하고 행렬의 크기를 줄이기 위하여 PCA를 활용하였다. 데이터를 보존하면서, 분석에 적절한 PC의 수를 선정하기 위하여, explained variance ratio를 계산하여 전체 키워드의 90%를 대표할 수 있는 88개의 PC를 선정하였다. 사고문서의 특성을 가장 많이 포함하고 있는 PC1의 경우, '절단'과 '손가락', '톱날', '엄지', '합판' 등의 키워드들이 분포되어있으며, 절단 작업 시 톱날로 인해 발생할 수 있는 위험요인들이 포함되어 있다. PC2의 경우 '해체', '거푸집', '파이프', '비계', '형틀' 등의 키워드들이 분포되어있으며, 가시설물의 해체작업 시 발생할 수 있는 위험요인들이 포함되어 있다. 이와 같이 PCA를 활용하여 키워드를 대상으로 설명력(분산력) 90%를 나타내는 88개의 PC와 문서로 이루어진 매트릭스로 축소하였다.

다음으로, 분류모델의 평가를 위하여 생성된 매트릭스를 k-fold cross validation을 사용하여 training data set과 test data set으로 분할하였다. 전체 데이터를 2에서 30까지의 k값에 대하여 민감도 분석을 수행하여, 데이터의 분산과 편향이 가장 적게 나타난 5개의 fold로 나누어 교차검증을 수행하여 data set을 작성하였다. 그 결과, 18,479개의 부상사고문서와 2,268개의 사망사고

문서로 이루어진 training data set과, 4,505개의 부상사고문서와 585개의 사망사고문서로 이루어진 test data set으로 분할하였다. 이를 통해, 20,747개의 training data set으로 4개의 분류모델을 작성하였으며, 5,090개의 test data set으로 분류모델을 평가하였다.

4.3. 모형별 분류결과

분류모델별 정확도 분석 결과와 confusion matrix는 Table 2와 같다. 먼저, logistic regression의 모델은 R의 generalized linear model을 사용하여 작성하였다. 전체 5,090개의 Test data set의 분류는 1종 오류가 18개, 2종 오류가 26개로 나타났으며, 모델의 정확도 지표 결과는 정밀도는 0.9688, 재현율이 0.9538, 정확도가 0.9914로 나타났다. 다음으로, 다음으로, decision tree 모델은 R의 rpart package 중 지니지수로 순도를 계산하는 CART를 사용하였다. 전체 5,090개의 test data set의 분류는 1종 오류가 108개, 2종 오류가 301개로 나타났으며, decision tree 모델의 정확도 지표 결과는 정밀도가 0.7244, 재현율이 0.4855, 정확도가 0.9196으로 나타났다. 또한, neural network 분류모델은 R의 neuralnet package를 사용하였으며, tolerable error가 가장 낮게 나타난 2개의 hidden layer로 총 node = (20,5)로 분석하였다. 전체 5,090개의 test data set의 분류는 1종 오류가 14개, 2종 오류가 9개로 나타났으며, 정확도 지표 결과는 정밀도가 0.9763, 재현율이 0.9846, 정확도가 0.9955로 나타났으며, 4개의 모델 중 정확도가 가장 높게 나타났다. 마지막으로, SVM 모델은 R의 e1071 package를 사용하였다. 먼저, 2-5부터 1까지의 gamma값과 1부터 24까지의 cost값을 각각 대입하여 모델의 최적값을 계산하였다. 그 결과, gamma = 0.03125, cost = 4의 최적값을 도출하였으며, support vector의 수는 13,349개로 나타났다. 본 연구에서는 kernel function 중 RBF(radial basis function) - gaussian kernel을 사용하여 training data set으로 분류모델을 작성하였다. 전체 5,090개의 test data set의 분류는 1종 오류가 17개, 2종 오류가 163개로 나타났으며, SVM 모델의 정확도 지표 결과는 정밀도는 0.9613, 재현율은 0.7213, 정확도는 0.9646으로 나타났다.

각 모델별 정확도 지표를 살펴보면, neural network 분류모델이 전체적으로 가장 높은 정확도를 나타냈으며, 다음으로 logistic regression 분류모델이 높게 나타났다. 다음으로 SVM 분류모델과 decision tree 모델의 순서로 정확도가 높게 나타났다. Neural network와 logistic regression의 경우, 모든 정확도 지표에서 95% 이상의 정확도를 보이고 있으며, 가장 낮은 정확도를 보인 모델인 decision tree는 재현율에서 50% 미만의 정확도가 나타났다.

4.4. 분류 키워드 분석 결과

사망사고와 부상사고를 분류하는 위험요인을 분석하기 위하여, 사고문서를 분류하는 PC를 파악하고, PC에서 위험요인 키워드를 도출하였다. 본 연구에서 작성한 분류모델 중 정확도가 높고 문서의 분류과정을 파악할 수 있는 방법론인 logistic regression의 분류모델을 활용하여 PC에 따른 회귀계수의 표준 추정값을 도출하였으며, 도출된 PC에 대하여 키워드-PC 매트릭스를 통해 PC가 포함하고 있는 키워드를 Table 3과 같이 분석하였다.

사망사고와 부상사고에서 도출된 PC와 키워드를 살펴보면, 사망사고를 분류하는데 있어서 영향을 미치는 표준 추정값이 큰 PC는 PC4과 PC6, PC23의 순서로 도출되었다. PC4는 '해체', '거푸집', '비계', '설치', '추락'의 키워드를 포함하고 있으며, 가시설물의 설치・해체 시 발생하는 추락사고에 대한 특성을 가지고 있다. PC6은 '상부', '배관', '하부', '철골', '용접', '크레인'의 위험요인을 포함하는 PC로 나타났으며, 배관작업 및

Table 2. Confusion matrix and degree of accuracy for each methodology

	Confusion matrix			Validation		
		Predicted : Fatal (Positive)	Predicted : Non-Fatal (Negative)	Precision	Recall	Accuracy
Logistic regression	Actual : Fatal(Positive)	**559**	**26**	0.9688	0.9538	0.9914
	Actual : Non-Fatal(Negative)	**18**	**4487**			
Decision tree	Actual : Fatal	**284**	**301**	0.7244	0.4855	0.9196
	Actual : Non-Fatal	**108**	**4397**			
Neural network	Actual : Fatal	**576**	**9**	0.9763	0.9846	0.9955
	Actual : Non-Fatal	**14**	**4491**			
Support vector machine	Actual : Fatal	**422**	**163**	0.9613	0.7213	0.9646
	Actual : Non-Fatal	**17**	**4488**			

Table 3. Classification factors of logistic regression

	Number of PC	Overview of disaster characteristics
High weight for fatal accidents	PC4	Fall Disasters that occur during installation and dismantling of temporary facilities
	PC6	Accidents during piping work and welding
	PC23	Accidents caused by machinery during wood working and form working
High weight for non-fatal accidents	PC1	Hazards from saw blades when cutting
	PC9	Hazards associated with welding and assembly operations
	PC5	Fractures due to piping and demolition work

용접 시 발생하는 사고의 특성을 보이고 있다. PC23은 '목공', '형틀', '용접', '아파트', '철골', '기계' 등의 키워드를 포함하고 있으며, 목공 및 형틀 작업과 기계로 인해 발생한 사고의 특성이 도출되었다. 부상사고에 대해 표준 추정값이 큰 PC는 PC1과 PC9, PC5의 순서로 도출되었다. PC1에 포함되었는 키워드를 살펴보면, '절단', '손가락', '톱날', '엄지', '합판', '전기톱'을 포함하고 있으며, 절단 작업 시 톱날로 인한 위험요인들이 도출되었다. PC9는 '철근', '조립', '골절', '사다리', '용접', '설치'의 키워드가 도출되었으며, 용접 및 조립 작업과 관련된 위험요인이 포함되어 있다. PC5는 '골절', '해체', '화장실', '손목', '배관', '철거'의 키워드가 포함되어 있으며, 배관 및 철거작업으로 인한 골절이 위험요인으로 나타났다. 이와 같이 각 사고에 대해 분류에 대한 표준 추정값이 PC에 포함된 키워드를 바타응로 사망사고와 부상사고를 분류하며, PC에서 도출한 키워드를 통해 사고의 위험성을 의미하는 위험요인을 도출하였다.

5. 토 의

Heinrich는 부상사고에서 도출되는 위험요인을 파악하고 개선하여 부상사고 뿐만 아니라 사망사고까지 예방할 수 있다는 이론을 제창하였다. 이와 같은 이론을 바탕으로 본 연구에서는 국내 건설업에서 발생한 사망사고와 부상사고의 관계를 파악하기 위하여 사고문서를 분류하여 비교 · 분석하였다. 사고문서의 분류는 위험요인 키워드를 포함하고 있는 PC에 따라 분류되며, 각 사고에서 많이 발생하는 PC가 문서에 포함되어 있는가에 따라 문서를 분류한다. 사고문서의 오분류는 문서가 실제와는 다른 예측치를 나타냈다는 점에서 유사한 PC가 포함되어 있을 수 있다. 따라서 오분류된 사고문서를 분석하여, 사망사고와 부상사고의 경계에서 나타나는 위험요인 키워드와 사고 프로세스를 분석할 필요가 있다. 오분류는 2가지로 1종 오류와 2종 오류로 나눌 수 있으며, 그 정의와 예측 및 의미는 Table 4와 같다.

1종 오류에 해당하는 문서는 실제 부상사고인 문서를 사망사고로 분류한 오류를 말한다. 이는 사망사고에서 나타나는 특성을 가지며, 사망사고와 연관성이 높다고 볼 수 있다. 따라서, 사고의 결과가 부상으로 나타났지만, 주의를 기울일 필요가 있는 문서로 판단할 수 있다. 또한 2종 오류는 실제 사망사고인 문서를 부상사고로 분류한 오류를 말한다. 또한 부상사고에서 주로 나타나는 키워드들로 구성된 문서이며, 이와 같은 오류는 사고 결과의 우연성에 의해 다양하게 나타날 수 있다. 1종 오류와 같이 사망사고와 연관성이 높은 키워드를 포함한 PC가 도출된 경우, 키워드뿐만 아니라 같은 PC에서 도출된 키워드들에 대한 관리가 필요하다. 부상사고와 사망사고의 특성은 다르게 나타나며 사망사고에 대해 집중적으로 관리할 필요가 있지만, 2종 오류로 분류되는 사고들에 대한 관리도 필요하다고 생각된다.

가장 높은 정확도를 나타낸 neural network 분류모델에 대해서 오분류된 문서를 탐색하였으며, 1종 오류에 해당하는 14개의 문서와 2종 오류에 해당하는 9개의 문서로 나타났다. 1종 오류로 분류된 사고문서를 살펴보면, 2,854번 문서의 경우, 위험요인 키워드는 '부라

Table 4. Definition of misclassification and implications of prediction and management

Incorrect prediction	Definition	Implications of prediction and management
Type Ⅰ error	Error in predicting a non-fatal accident as a fatal accident	• Non-fatal disaster that almost led to fatal accident • Non-fatal accident, but disaster that will strengthen corrective action • Needs to be recognized and highlighted as fatal accident
Type Ⅱ error	Error in predicting a fatal accident as a non-fatal accident	• Fatal disaster that was close to non-fatal accident • Disaster that resulted in fatal accident due to abnormal accidental loss • Disaster to analyze the special and contingent reasons that led to fatal accident

캣', '설치', '추락'으로 나타났다. 이 문서의 경우, 부라캣 설치작업 중 추락으로 인해 부상사고가 발생하였으며, 추락으로 인한 사고의 경우 사망사고로 이어질 수 있으며, 추락에 대한 안전 대책이 필요한 것을 알 수 있다. 다음으로 2882번 문서는 '리모델링', '철거작업', '벽체', '머리', '충격'의 단어로 위험요인 키워드가 도출되었으며, 벽체의 무너짐과 작업자의 보호구에 대한 대책이 필요하다. 또한 2,889번 문서의 경우, '콘크리트', '철근배근', '철근', '벤딩기', '장갑', '검지', '손가락', '기계'와 같은 위험요인 키워드가 도출되었으며, 2913번 문서의 위험요인 키워드는 '철근', '설치', '발판', '높이', '추락'으로 나타났다. 1종 오류에 해당하는 문서의 경우, 사망사고에서 자주 도출되는 키워드들로 구성되어 있으며, 특히, 2882번 문서와 2,913번 문서는 사망사고문서에서 가장 많이 도출되는 '철거작업'과 '머리', '설치', '발판', '높이', '추락'이 포함되어 있다. 이처럼 부상사고에서 도출된 키워드임에도 불구하고 사망사고로 발생할 위험이 있는 것으로 나타난다. 이를 통해 1종 오류에 해당하는 문서들은 도출된 키워드의 위험성을 확인하고, 해당 키워드의 관계 분석을 통해 위험성을 탐색해야 할 필요가 있다. 2종 오류로 분류된 2712번 문서를 살펴보면, 위험요인 키워드로는 '단독주택', '신축공사장' '비계작업', '바닥', '응급실', '중환자실' 등으로 나타났다. 이와 같은 키워드들을 통해, 비계작업 시 안전관리가 필요하다고 볼 수 있다. 2,793번 문서의 경우, '가스통', '토치', '스티로폼', '호스', '밸브', '응급처치', '구급차' 등이 주요 키워드로 도출되었으며, 키워드들을 통해 토치 사용 시 발생한 화재로 호스나 밸브에 대한 관리가 필요하며, 화재가 발생할 수 있는 주변 환경에 대해서도 관리가 필요할 것으로 사료된다. 2종 오류로 분류된 2,793번 문서의 경우, 가스통과 토치, 밸브 등의 부상사고와 관련성이 크게 나타난 키워드들이 분포되었지만 3~4초 동안 불에 노출되어 사망사고로 이어진 사고개요이다. 특히, 2,793번 문서의 위험요인 키워드를 살펴보면, 토치, 스티로폼, 호스 등 화상으로 이어지는 부상사고와 관련이 깊은 키워드들이 도출되었다. 이처럼 부상사고와 연관성이 큰 위험요인 키워드들도 사망사고로 이어질 수 있다.

따라서 두 종류의 오분류는 업종과 작업내용, 사고의 기인물의 특성에 따라 비슷한 사고 프로세스에서 부상사고와 사망사고가 다르게 발생할 수 있으며, 사고의 분류뿐만 아니라 오분류에 대한 분석이 필요하다는 것을 나타낸다. 사망사고가 가장 많이 발생하는 업종인 건설업과 관련된 사고문서의 경우는 1종 오류가 많이 발생할 수 있으며, 특히 고소작업의 경우, 낮은 높이에서 작업 중 추락으로 인한 사고의 경우, 실제로 부상사고인 사고 프로세스에서 사망사고가 많이 발생하는 '추락'과 같은 키워드로 인해 1종 오류가 발생할 수 있다. 이와 반대로 2종 오류에 대해 도출된 기인물 키워드를 분석하여 사고의 예방에 도움이 될 것으로 사료된다.

6. 결 론

본 연구는 2017년 국내 건설업에서 발생한 부상사고와 사망사고를 비교・분석하여 두 사고 간에 위험요인의 차이를 파악하고 부상사고가 사망사고로 이어지지 않는 시사점을 파악하기 위하여 다음과 같이 연구를 진행하였다. 먼저, 사고개요를 수집하여 비정형데이터의 정형화와 특성 선정을 텍스트마이닝과 PCA를 활용하여 행렬 데이터로 작성하였다. 다음으로 k-fold cross validation을 활용하여 전체 데이터를 training data set과 test data set으로 분할하였으며, 분할된 training data set으로 각 분류모델을 작성하고, test data set으로 분류모델의 정확도를 평가하였다. 평가된 분류모델은 neural network와 logistic regression, SVM, decision tree의 순서로 정확도가 높게 나타났으며, 특히 neural network 분류모델은 사용된 3개의 정확도 지표 모두 95% 이상의 정확도를 나타내었다. 마지막으로 작성한 분류모델에서 위험요인 키워드를 도출하고, 오분류된 문서를 분석하여 사고의 분류에 영향을 미치는 위험요인을 도출하였다.

본 연구는 위험성이 큰 요인을 탐색하기 위하여 부상사고와 사망사고에 대한 분류모델을 구축하고 평가하였다. 또한 각 분류모델의 정확도를 계산하여 사고문서의 분석에 적합한 모델을 탐색하여 사고문서의 분류에 영향을 미치는 위험요인을 분석하였다. 2017년 국내 건설업에서 발생한 사고의 경우 neural network 분석모델이 사고의 위험요인을 탐색하는 것에 적합하다고 판단된다. 또한 오분류된 문서들을 탐색하여 부상사고로 발생하였지만 비슷한 사고가 발생하였을 때 사망사고로 이어질 수 있는 위험요인 키워드들과 부상사고의 주요 키워드 중 사망사고로 발생할 수 있는 키워드를 탐색하였다. 재해문서의 분류를 활용한 키워드 분석을 통해 위험성에 따른 요인을 도출하고 이를 활용하여 현장의 위험요인에 따른 안전관리에 활용될 수 있을 것이라고 생각된다.

추후 연구는 부상사고와 사망사고의 불균형으로 인해 연도별 분석이 불가능하였으며, 본 연구에서 사용

한 k-fold cross validation 외에 데이터의 편향을 해결하는 방법에 대한 연구가 필요하다. 또한 새롭게 발생하는 사고 데이터를 추가적으로 분류하기 위해서 2017년 이후의 데이터에 대한 분석이 필요할 것으로 사료되며, 다른 업종의 위험요인을 탐색하기 위해서 다양한 업종의 사고 데이터를 분석하여 위험요인을 도출할 필요가 있다.

Acknowledgement: 이 성과는 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1C1C1007302).

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT: Ministry of Science and ICT) (NRF-2020R1C1C1007302)

References

1) H. W. Heinrich, “Industrial Accident Prevention: A Scientific Approach”, McGraw-Hill, 1931.
2) KOSHA, “Statistical Survey and Analysis of Industrial Disasters”, 2019.
3) S. Kang and Y. Suh, “On the Development of Risk Factor Map for Accident Analysis using Textmining and Self-Organizing Map(SOM) Algorithms”, J. Korean Soc. Saf., Vol. 33, No. 6, pp. 77-84, 2018.
4) G. Ahn, M. Seo and S. Hur, “Development of Accident Classification Model and Ontology for Effective Industrial Accident Analysis based on Textmining”, Journal of The Korean Society of Safety, Vol. 32, No. 5, pp. 179-185, 2017.
5) Y. M. Goh and C. U. Ubeynarayana, “Construction Accident Narrative Classification: An Evaluation of Text Mining Techniques”, Accident Analysis and Prevention, Vol. 108, pp. 122-130, 2017.
6) G. Salton and M. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, NY, 1983.
7) S. Wold, K. Esbensen and P. Geladi, “Principal Component Analysis”, Chemometrics and Intelligent Laboratory Systems, Vol. 2, No. 1-3, pp. 37-52, 1987.
8) MOEL, “Classification of Causes of Deaths in Industrial Accidents”, 2017.
9) B. Kim, S. Chang and Y. Suh, “Text Analytics for Classifying Types of Accident Occurrence Using Accident Report Documents”, J. Korean Soc. Saf., Vol. 33, No. 3, pp. 58-64, 2018.
10) J. Jeong, M. Jee, M. Go, H. Kim, H. Lim, Y. Lee, and W. Kim, “Related Documents Classification System by Similarity between Documents”, Journal of Broadcast Engineering, Vol. 24, No. 1, pp. 77-86, 2019.
11) D. Pak, M. Hwang, M. Lee, S. Woo, S. Hahn, Y. J. Lee, and J. Hwang, “Application of Text-Classification Based Machine Learning in Predicting Psychiatric Diagnosis”, Korean Journal of Biological Psychiatry, Vol. 27, No. 1, pp. 18-26, 2020.
12) W. Lee, “A deep learning analysis of the KOSPI's directions”, Journal of the Korean Data and Information Science Society, Vol. 28, No. 2, pp. 287-295, 2017.
13) R. Kostoff, D. Toothman, H. Eberhart, and J. Jumenik, “Textmining Using Database Tomography and Bibliometrics: A Review”, Technological Forecasting and Social Change, Vol. 68, No. 3, pp. 223-252, 2001.
14) M. Leblanc and R. Tibshirani, “Combining Estimates in Regression and Classification”, Journal of the American Statistical Association, Vol. 91, No. 436, pp. 1641-1650, 1994.
15) S. Dreiseitl and L. Ohno-Machado, “Logistic Regression and Artificial Neural Network Classification Models”, Journal of Biomedical Informatics, Vol. 35, pp. 352-359, 2002.
16) Y. Song and Y. LU, “Decision Tree Methods: Applications for Classification and Prediction”, Shanghai Arch Psychiatry, Vol. 25, No. 2, pp. 130-135, 2015.
17) R. Feraud and F Clerot, “A Methodology to Explain Neural Network Classification”, Neural Networks, Vol. 15, No. 2, pp. 237-246, 2002.
18) J. P. Bigus, “Data Mining with Neural Networks”, Macgraw-Hill, pp. 61-97, 1996.
19) A. Mathur and G. M. Foody, “Multiclass and Binary SVM Classification: Implications for Training and Classification Users”, IEEE Geoscience and Remote sensing Letters, Vol. 5, No. 2, pp. 241-245, 2008.
20) J. Han, J. Pei and M. Kamber, “Data Mining: Concepts and Techniques”, Elsevier, 2011.