Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2023.46.3.007

Vol. 46, No. 3 : 7-14, September 2023
ISSN : 2005-0461 / EISSN : 2287-7975

Developing and Evaluating Damage Information Classifier of High Impact Weather by Using News Big Data

Su-Ji Cho · Ki-Kwang Lee†

School of Business Administration, Dankook University

재해기상 언론기사 빅데이터를 활용한 피해정보 자동 분류기 개발

조수지 · 이기광†

단국대학교 경영학부

Recently, the importance of impact-based forecasting has increased along with the socio-economic impact of severe weather have emerged. As news articles contain unconstructed information closely related to the people's life, this study developed and evaluated a binary classification algorithm about snowfall damage information by using media articles text mining. We collected news articles during 2009 to 2021 which containing 'heavy snow' in its body context and labelled whether each article correspond to specific damage fields such as car accident. To develop a classifier, we proposed a probability-based classifier based on the ratio of the two conditional probabilities, which is defined as I/O Ratio in this study. During the construction process, we also adopted the n-gram approach to consider contextual meaning of each keyword. The accuracy of the classifier was 75%, supporting the possibility of application of news big data to the impact-based forecasting. We expect the performance of the classifier will be improve in the further research as the various training data is accumulated. The result of this study can be readily expanded by applying the same methodology to other disasters in the future. Furthermore, the result of this study can reduce social and economic damage of high impact weather by supporting the establishment of an integrated meteorological decision support system.

Keywords : Classification, Conditional Probability, Big Data, High Impact Weather, Impact-based Forecasting

1. 서 론

최근 지속적인 기후변화와 함께 도시화, 산업화 등의 사회적 변화는 재난 발생요소 간 상호작용을 증가시키는 원인으로 대두되고 있다[6, 31]. 따라서 과거 재해기상 피해정보의 분석을 통해 피해를 저감하고 선제적으로 대응할 수 있는 방안이 필요한 시점이다. 우리나라를 포함한 다수의 기상 선진국은 상당 수준의 정확도를 갖춘 예·특보 시스템을 갖추고 있음에도 불구하고, 기상현상의 물리적 특성에 집중하고 있다는 지적이 존재한다. 즉, 재해의 발생규모와 발생시점에 집중한 예보는 실질적으로 국민안전 및 재산에 미치는 영향을 반영하고 있지 못한다는 것이다. 일례로 2016년 1월 제주 폭설 당시 적시에 발행된 예·특보에도 불구하고 약 52.5억의 경제적 손실이 발생하였으며, 대규모 항공편 결항으로 약 9만여 명이 제주에 고립되는 사태가 발생한 바 있다[3]. 이는 정확한 예·특보와 더불어 기상현상이 미치는 사회·경제적 영향까지 고려한 '영향예보(Impact-based forecasting)'의 필요성을 역설하는

Received 6 September 2023; Finally Revised 15 September 2023; Accepted 15 September 2023
† Corresponding Author : kiklee@dankook.ac.kr

사례라고 할 수 있다.

우리나라 기상청 또한 2019년 폭염영향예보를 시작으로 2022년 현재 한파영향예보를 통해 피해 영향정보 및 대응요령을 제공하고 있다. 각 지자체에서도 기상재해의 영향예보 통합관리체계 구축 및 현행 영향예보서비스 지원 강화를 위한 연구가 다수 이루어지고 있으나, 국민 생활과 밀접한 피해정보를 포함하고 있는 언론기사 빅데이터를 활용한 연구는 미미한 실정이다. 재해기상으로 인한 피해분야 및 피해규모를 분류한 유사 사례로서 행정안전부의 재해연보가 있으나, 이는 피해 보상 및 행정관리 차원에서 사후적으로 집계된 자료라는 한계점이 있다. 또한 시간적으로도 연 단위 발행되기 때문에 실질적인 활용이 어렵다. 따라서 본 연구에서는 국민의 관심정도를 반영하고 있는 언론기사 빅데이터를 활용하여 재해기상 관련 정보를 추출하고 통합관리체계 구축을 지원하고자 하였다. 특히 대설 재해기상에 대하여 언론기사 본문 텍스트 마이닝을 통해 피해분야를 자동으로 분류하는 알고리즘을 개발하고 검증하였다. 재해기상 언론기사의 경우 한 기사가 다양한 피해분야를 언급할 수 있기 때문에 이러한 연관성을 효율적으로 통제할 수 있는 분류 알고리즘이 필요하다. 본 연구결과를 통해 기상 영향정보 통합 데이터베이스에서 언론기사 빅데이터의 활용 가능성을 확인할 수 있다. 또한 피해정보 분석 알고리즘을 통해 실시간에 준하는 기상정보 피해정보 추출이 가능할 것으로 기대된다.

2. 이론적 배경

2.1 영향예보와 재해기상 언론기사

초창기 연구는 영향예보의 필요성에 대한 사례연구와 영향예보 도입으로 인한 사회·경제적 편익 연구가 주를 이루었다[24, 30, 31]. 최근에는 재해기상과 관련한 빅데이터의 활용방안이 강조됨에 따라 관련 정책연구 또는 재난피해예측 및 피해방지 시스템 구축 연구가 이루어졌다. 일례로 Choi and Kim[1]은 재난 방재분야 빅데이터 활용 시스템 구축 현황을 정리하였으며, Choi et al.[2]은 재난 발생시 유관기관의 빅데이터 연계·분석을 통한 모니터링과 대국민과의 적극적 의사소통 체계의 중요성을 논의하였다. Lee[16]는 소셜미디어 분석을 통해 기상예보 신뢰도의 변화 추이와 원인 등을 파악하고 일반 국민의 예보 신뢰도를 제고할 수 있음을 강조하였다. Shin and Kim[27]은 재난관리 분야에서 공공 및 민간 부문의 정보개방이 핵심임을 밝히고 효과적 빅데이터 산업 활성화를 위한 정책과제를 제시하였다.

빅데이터를 보다 직접적인 피해정보의 원천(source)으로서 활용한 연구 또한 최근 다수 이루어지고 있다. 주로 국민 생활과 밀접한 정보를 담고 있는 언론기사 빅데이터 분석 연구가 증가하고 있는데, 재해기상 현상으로 인한 직·간접적 피해 정보를 체계적으로 수집하고 분류하고자 하는 연구가 수행되었다. Lee et al.[19]은 언론기사 빅데이터를 활용하여 강원지역의 과거 피해정보를 바탕으로 대설 영향예보를 위한 4개 위험수준 단계를 도출하였다. KMA[13]는 재해 발생시 피해정보를 연계한 영향예보 DB 구축을 위하여 2012년 태풍 카눈(Khanun) 발생당시 재해연보와 언론기사 피해정보를 정성적으로 비교·분석하였으며, 재해연보의 피해자료는 시공간적 해상도가 낮고 피해의 상세원인이 명시되어있지 않다는 한계점을 도출하였다. 특히 물리적인 피해정보는 행정절차를 통해 수집이 가능하지만, 사회·경제적 피해정보는 피해액 산정이 어렵기 때문에 언론기사의 활용도가 높다는 점을 강조하였다. NIMS[21]는 최근 약 10년간의 폭염 관련 언론기사를 수집하고 피해분야별 분류를 통해 지역별·분야별·연도별 피해정보를 분석하였다. 해당 연구에서는 언론기사수 대비 실제 온열질환자 수가 0.96의 상관계수를 보여 영향예보 DB구축을 위한 자료로서 언론기사의 타당성을 입증하였으며, 또한 현행 폭염 영향예보의 피해분야에 속하지 않는 기타 피해로서 수자원 피해(녹조, 적조, 저수율 등), 냉방기 및 배터리 화재사고, 차량 사고(질식사고, 영·유아 차량갇힘), 개학연기 등의 추가적인 피해분야를 언론기사를 통해 제시하였다. KEI[12]는 약 4년간의 폭염 및 한파 관련 언론기사를 수집하고 월단위 시간 흐름에 따라 누적 증가된 출현빈도를 통해 건강, 농·축·수산업 등의 피해 키워드 간 인과관계를 도출하였다. 이외에도 Lee[17]는 SNS 데이터를 중심으로 감성분석 및 연관규칙분석을 사용하여 기상예보서비스의 대국민 인식 및 만족도를 정량평가하였다. Park et al.[25]은 폭염으로 폐사한 가축 수와 축산 분야 언론기사간 증감 패턴과 그에 따른 주요 키워드 변화를 비교·분석하였으며, Jung et al.[8]은 폭염의 사회·경제적 영향 유형을 뉴스기사에 출현한 단어 기반 네트워크 분석을 통해 도출하였다. Seo and Kim[26]은 가뭄 영향의 발생 가능성을 예측하기 위하여 뉴스기사를 수집하고 모니터링에 활용하였다. Kim et al.[10]은 홍수로 인한 하천범람을 분석하고 재난안전지도 서비스를 구현하였다. 재해기상 이외에도 구제역[22], 조류독감[20], 식품안전[7] 등 기타 재난안전 분야에서 비정형 자료로서 언론기사 데이터를 활용하였다.

이상에서 정리한 다수의 선행연구를 통해 재해기상 분야에서 빅데이터 분석의 중요성을 확인할 수 있으나, 보다 실용적인 측면에서 언론기사의 영향예보 DB에의 활용 가능성을 검증한 연구가 부족하다. 또한 비정형자료 텍스트 마이닝으로 다수 활용되는 기계학습(Machine Learning)

분류기법을 적용한 연구가 미미한 실정이다. 따라서 본 연구는 다양한 기계학습 분류기법을 살펴보고 향후 재학습이 용이한 실용적 측면의 피해정보 분류 알고리즘을 설계・검증하고자 하였다.

2.2 텍스트 마이닝을 통한 문서 분류

전 세계 실존하는 데이터의 약 80%가 2025년까지 비정형 데이터(Unstructured Data) 즉, 텍스트, 사진, 영상과 같은 데이터로 이루어질 것으로 예상된다[28]. 최근 컴퓨터 처리기법의 발전으로 인해 이 같은 비정형 데이터를 보다 효율적으로 분석할 수 있게 되었고, 따라서 비즈니스 인텔리전스(Business Intelligence) 관점에서 정형 데이터가 제공하는 것 이상의 정보 추출 및 의사결정 지원이 가능해졌다[18]. 특히 텍스트 마이닝은 대표적인 비정형 데이터인 텍스트를 인식하고 분석이 가능한 알고리즘을 연구하는 빅데이터 처리 기법이다[5]. 주로 문서를 특성별로 분류하거나 또는 문서에 내재된 특정 주제를 추출하기 위해 사용하는데, 본 연구의 목적에 따라 텍스트 마이닝을 활용한 문서 분류에 대해 살펴보고자 한다.

문서 분류는 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning) 방법을 활용할 수 있으며, 그 중 지도학습 과정은 학습 데이터(training data)와 검정 데이터(test data)의 분할을 통해 이루어진다.[1] 학습 데이터를 통해 분류 모델을 설계한 뒤, 이를 검정 데이터에 적용함으로써 분류 즉, 모델의 예측이 적절한지 여부를 판단한다.

지도학습을 통한 문서 분류 모델은 로지스틱 회귀[4], 서포트 벡터 머신(SVM, Support Vector Machine)[29], 나이브 베이즈 분류(Naive Bayes Classifier)를 가장 많이 사용한다. 이 같은 전통적인 통계기법을 활용하는 방식 외에도, 최근에는 인공지능의 대두와 함께 신경망을 활용하는 방법 또한 연구되고 있다[14].

텍스트 마이닝을 활용한 문서 분류는 정치학, 사회학, 공학, 교육학, 문헌정보학 등 다양한 연구 분야에서 수행되어 왔다. 공공 분야에서도 Oh et al.[23]은 부정적 여론과 같이 언론기사에 나타난 사회적 리스크(risk) 유형을 정의하고 서포트벡터머신 기법을 통한 분류기를 구축하여 자동으로 분류 및 검증하였다. Kang and Ko[9]는 나이브 베이즈 분류모형을 통해 정부 부처별 인사에 활용되는 직무기술서를 9개 업무 특성으로 자동 분류하는 모델을 개발하였다. 재난안전 분야에서는 Kwon et al.[15]이 119 신고접수 음성 데이터를 텍스트로 변환한 뒤 화재, 구급, 구조, 기타신고 4개 유형으로 분류하는 기계학습 모델을 통해 재난상황에서 효율적인 대응 체계를 제시하였다.

재해기상 분야에서는 빅데이터를 활용한 연구는 이전 절에서 정리한 바와 같이 최근 증가하고 있는 추세이나, 아직까지 텍스트 마이닝을 활용한 문서 분류 연구는 충분히 이루어지지 않았다. NIMS[21]가 폭염 및 대설 재해와 관련한 언론기사를 나이브 베이즈 분류기 등 머신러닝 분류기를 통해 자동으로 수집・정제・분류하는 알고리즘을 제시하였으며, 크게 피해 관련 및 비관련기사에 대한 분류, 중복기사 판별, 피해분야 분류의 세 가지 분류모델을 개발하고 검증하였다. 해당 연구는 대량의 뉴스기사를 수집하여 피해분야 분류 모델을 생성하였다는 데에 의의가 있으나, 다소 낮은 분류 정확도로 인해 실질적인 활용에는 어려움이 있다. 특히 재해기상의 피해분야 분류 시에는 하나의 기사에서 여러 가지 피해를 언급할 가능성이 높고, 주요 피해분야 간 발생 가능성이 각기 다르기 때문에 분류기 설계 시 이에 대한 적절한 고려가 필요할 것으로 판단된다. 따라서 본 연구에서는 보다 실용적인 관점에서 대량의 언론기사를 효율적으로 분류하고 향후에도 지속적으로 자동 갱신이 가능한 분류 알고리즘을 제안하고자 하였다.

3. 분 석

3.1 분류기 설계 및 검증

본 연구에서는 기계학습 기반 분류문제에 가장 많이 사용되는 나이브 베이즈를 바탕으로 하되, 본 연구목적에 적합하게 수정된 확률기반 분류기를 설계하여 언론기사의 피해분야를 분류하였다. 일반적으로 나이브 베이즈 분류기는 문서 i의 n가지 특성을 나타내는 특성벡터 $x_i = [x_{i1}, \cdots, x_{in}]$와 범주 y가 주어졌을 때 사전확률(prior probability) $P(y_i)$와 사후확률(posterior probability) $P(y_i \mid x_i)$를 도출한다. 이 때 베이즈 법칙에 의해 다음과 같이 변형된 사후확률을 극대화하는 값으로 $\hat{y_i}$을 추정하고 문서를 분류한다.

$$P(y_i \mid x_{i1}, \cdots, x_{in}) \propto P(y_i) \prod_{j=1}^{n} P(x_{ij} \mid y_i)$$

본 연구에서 특성벡터는 각 언론기사에서 출현한 단어 토큰(token)에 해당하며, 범주는 14개 대설 피해분야에 해당한다. 여기서 단어 토큰이란, 명사 등 문장 내에서 의미를 가지는 단어를 의미한다. 따라서 특정 단어 w_i가 특정 피해분야 $Class_k$에서 출현할 확률인 $P(w_i \mid Class_k)$가 우도(likelihood)에 기반하여 기사를 분류할 수 있다.

1) 필요에 따라 검증 데이터(validation data)를 사용하기도 한다.

그러나 일반적 경우와 달리 재해기상 언론기사의 피해분야 분류 문제는 다음과 같은 세 가지 특성을 가진다. 첫째, 한 문서가 단일 범주가 아닌 다중 범주에 속할 수 있다. 즉, 하나의 언론기사에서 도로교통과 학교휴교 관련 피해를 동시에 언급할 수 있다. 따라서 사후확률을 극대화하는 단 하나의 범주에 문서를 할당하는 것이 아니라, 개별 문서에 따라 가변적인 상수 c개($1 \le c \le 14$)의 범주에 문서를 할당한다. 둘째, 한 문서가 여러 피해분야를 언급하는 경우 분류기 학습 과정에서 각 단어의 확률 도출 시 노이즈(noise)를 생성할 가능성이 있다. 즉 학습데이터로서 한 기사 내에 도로교통 피해와 학교휴교 관련 피해를 동시에 언급한 경우, 필연적으로 학교휴교와 관련한 단어가 도로교통 피해분류에 해당할 확률이 증가한다. 셋째, 대설로 인한 주 피해분야와 비주류 피해분야 간 학습 기사 수 불균형이 존재한다. 예를 들어 조난사고와 같이 피해분야에 해당하는 기사 수가 적은 경우 분모의 크기로 인해 $P(w_i \mid Class_k)$를 증가시킨다. 이는 실제 분류 상황에서 조난사고 피해가 아닌 기사를 조난사고 피해 기사로 분류하는 위양성(false positive) 비율이 높아진다. 따라서 각 피해분야 집단의 크기 편차를 표준화할 필요가 있다.

따라서 본 연구에서는 언론기사의 이 같은 특성을 고려하여 다음과 같이 피해분야 분류기를 설계하였다. 먼저 문서의 특성벡터를 구성하는 단어를 선정할 시, 전체 표본기사 중 특정 분야에 대한 단일 피해만을 언급하고 있는 기사를 대상으로 하였다. 다음으로 각 피해분야 범주별 집단 크기를 보정하기 위하여 단어의 피해분야 내외 출현비를 사용하였다. 즉 $P(w_i \mid Class_k)$와 $P(w_i \mid \sim Class_k)$ 비율인 I/O Ratio(In/Out Ratio)를 정의함으로써 단어의 I/O Ratio가 높을수록 해당 피해분야로 분류할 확률이 높아지도록 분류기를 설계하였다. 마지막으로 각 기사의 다중분류를 위하여 I/O Ratio의 평균값이 c순위 대비 c+1순위에서 가장 크게 차이 나는 구간을 기준으로 상위 c개 피해분야에 모두 해당할 수 있도록 유동적으로 설계하였다.

<Table 1> Performance Index for a Classifier

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$\Pr ecision = \frac{TP}{TP + FP}$$

$$F1 - score = \frac{2 \times \Pr ecision \times Recall}{\Pr ecision + Recall}$$

*TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative

분류 성능에 대한 검증은 Kohavi[11]의 연구에 따라 실제 정답과 분류기가 예측한 답 간 관계로서 정의할 수 있으며, 구체적인 성능지표는 <Table 1>에 나타낸 바와 같다. 이 경우 예측된 'Positive' 및 'Negative'는 분류기가 예측한 참과 거짓을 각각 나타내며, 'True' 및 'False'는 예측결과와 실제 정답 간 일치여부를 나타낸다.

3.2 데이터 수집 및 분류기 학습

본 연구에서는 대설 피해 관련 언론기사를 수집하여 특정 시점을 기준으로 학습 및 검증 데이터로 절단한 후 분석에 사용하였다. 구체적으로는 2009년 1월부터 2020년 4월까지의 자료를 분류기 학습에 사용하고, 이후 2021년 4월까지의 언론기사는 분류기 검증에 사용하였다. 즉, 표본외 데이터(out-of-sample)를 검증 데이터로 사용하여 완전히 새롭게 수집된 데이터에 대해 피해분야 분류가 가능한지 보수적으로 검증하였다. 최종적으로 학습 데이터 기사 수는 3,144개, 검증 데이터 기사 수는 1,123개로 절단 비율은 7.4 : 2.6이다. <Table 2>는 학습 데이터 셋의 사전확률을 나타낸다. 사전확률을 통해 대설로 인한 주요 피해를 살펴보면 전체 학습 기사 중 33%는 교통사고 피해를 언급하고 있으며, 30%는 도로통제 피해, 그리고 24%가 항공지연취소 피해를 언급하고 있는 것으로 나타났다. 반대로 조

<Table 2> Prior Probabilities of Train Data Set

Prior probability	Class1	Class2	Class3	Class4	Class5	Class6	Class7
	Car accident	Road traffic control	Heavy traffic	Disrupted railroading	Flight delay(cancel)	Harbor control	Public transportation
N. of inclass (%)	1,027 (0.33)	949 (0.30)	595 (0.19)	124 (0.04)	759 (0.24)	436 (0.14)	288 (0.09)
N. of outclass (%)	2,177 (0.67)	2,195 (0.70)	2,549 (0.81)	3,020 (0.96)	2,385 (0.76)	2,708 (0.86)	2,856 (0.91)
Total	3,144						
Prior probability	Class8	Class9	Class10	Class11	Class12	Class13	Class14
	Isolated Town	Closed mountain	Distress situation	Agricultural facilities	General facilities	Falling	Closed school
N. of inclass (%)	234 (0.07)	303 (0.10)	93 (0.03)	400 (0.13)	381 (0.12)	218 (0.07)	342 (0.11)
N. of outclass (%)	2,910 (0.93)	2,841 (0.90)	3,051 (0.97)	2,744 (0.87)	2,763 (0.88)	2,926 (0.93)	2,802 (0.89)
Total	3,144						

<Table 3> Number of Single or Multiple Damage Field News of Train Data Set

Prior probability	Class1	Class2	Class3	Class4	Class5	Class6	Class7
	Car accident	Road traffic control	Heavy traffic	Disrupted railroading	Flight delay(cancel)	Harbor control	Public transportation
N. of single (%)	237 (0.23)	194 (0.20)	81 (0.14)	29 (0.23)	280 (0.37)	28 (0.06)	18 (0.06)
N. of multiple (%)	790 (0.77)	755 (0.80)	514 (0.86)	95 (0.77)	479 (0.63)	408 (0.94)	270 (0.94)
Inclass Total	1,027	949	595	124	759	436	288
Prior probability	Class8	Class9	Class10	Class11	Class12	Class13	Class14
	Isolated Town	Closed mountain	Distress situation	Agricultural facilities	General facilities	Falling	Closed school
N. of single (%)	60 (0.26)	26 (0.09)	41 (0.44)	125 (0.31)	147 (0.39)	26 (0.12)	104 (0.30)
N. of multiple (%)	174 (0.74)	277 (0.91)	52 (0.56)	275 (0.69)	234 (0.61)	192 (0.88)	238 (0.70)
Inclass Total	234	303	93	400	381	218	342

난사고는 약 3%, 철도운행장애가 약 4% 비율로 나타나 비주류 피해분야에 해당하였다.

학습 데이터를 통해 피해분야별 단어의 출현확률을 도출하였다. 앞서 언급한 바와 같이, 하나의 언론기사에서 여러 분야의 피해를 언급할 경우를 고려하여 단일 피해분야에서 출현한 단어를 핵심 단어로 선정하고 확률을 계산하였다. 실제로 <Table 3>의 단일 및 다중 피해분야 빈도수를 살펴보면, 각 피해분야의 절반 이상(56% ~ 94%)이 한 기사에서 다양한 분야 피해를 언급한 경우임을 확인할 수 있다. 이후 단어의 문맥적 의미를 반영할 수 있도록 n개 단어의 연속적 나열인 n-gram을 활용하였으며, 최대 3-gram 수준까지 확장하여 각 n-gram 토큰의 확률을 계산하였다. 이후 일반화를 위하여 전체 출현빈도가 최소 5회 이상, 그리고 특정 피해분야 내 출현비율이 5% 이상인 토큰만을 활용하였다.

<Table 4> Example of Token Frequency and I/O Ratio of Class 'Car Accident'

Token	Inclass freq.	Outclass freq.	I/O ratio
'Collision'	347	12	59.6
'Accident'	857	472	3.74
'Occur'	601	633	1.96
'Collision Accident'	247	11	46.29
'Occur Accident'	376	144	5.38

<Table 4>는 실제 피해분야별 토큰의 확률 예시를 나타낸다. '추돌(Collision)'이라는 단어가 교통사고 피해분야 기사에서 출현할 확률은 약 33.8%(=347건/1,027개 교통사고 기사)인 반면, 타 피해분야에서 출현할 확률은 약 0.6%(=12건/2,117개 타 분야 기사)로 약 59.6배의 I/O Ratio 값을 가진다. '사고(Accident)'의 경우 3.74, '발생(Occur)'의 경우 1.96의 I/O Ratio 값을 가져 상대적으로 작다. 즉 향후 특정 단어가 출현한 기사를 분류할 시 I/O Ratio가 높을수록 해당 피해분야로 분류할 확률이 높아지도록 분류기를 학습하였다.

3.3 분류 결과

분류기 학습 이후, 실제 검증 데이터 셋을 활용하여 재해기상 언론기사의 피해분야를 분류하였다. 검증 데이터 셋의 각 기사를 단어벡터로 변환한 이후, 단어벡터의 요소로서 개별 토큰의 사전 학습된 I/O Ratio를 탐색하여 피해분야별 평균 I/O Ratio를 도출하는 방식으로 분류하였다. <Table 5>는 검증 데이터 셋의 분류 성능지표 평균을 나타내고 있다.

<Table 5> Overall Classification Result of Test Data Set

	Accuracy	Recall	Precision	F1-score
Index	0.85	0.72	0.45	0.50

분류 정확도(accuracy)는 85%로 높고, 또한 재현율(recall)이 전체 평균 72%로 실제 해당 피해분야의 기사가 발생할 경우 누락하지 않고 올바르게 분류할 확률이 높다. 반면 정밀도(precision)와 F1-score의 경우 각각 평균 45%, 50%로 다소 낮게 나타났다. 분류 결과에 대한 보다 구체적인 분석을 위해 각 피해 분야별 상세 분류 성능지표를 <Table 6>에 나타내었다.

피해분야별 분류 정확도는 표본 외 데이터로 검증하였음에도 불구하고 최소 71%(도로통제)에서 최대 96%(학교휴교) 사이의 분포를 보이며 높은 수준의 정확도를 확보하였다. 재현율의 경우 학교휴교(100%), 기타시설물(97%), 교통사고(93%) 등으로 높게 나타났다. 다만 피해분야 간 다소 재현율의 편차가 존재하는데, 특히 재현율이 낮은 피

<Table 6> Classification Result of Test Data Set by Classes

Performance index	Class1	Class2	Class3	Class4	Class5	Class6	Class7
	Car accident	Road traffic control	Heavy traffic	Disrupted railroading	Flight delay(cancel)	Harbor control	Public transportation
Accuracy	0.82	0.71	0.81	0.85	0.88	0.82	0.87
Recall	0.93	0.46	0.58	0.56	0.81	0.78	0.22
Precision	0.69	0.81	0.79	0.18	0.80	0.61	0.31
F1-score	0.79	0.58	0.67	0.27	0.80	0.69	0.26
Performance index	Class8	Class9	Class10	Class11	Class12	Class13	Class14
	Isolated Town	Closed mountain	Distress situation	Agricultural facilities	General facilities	Falling	Closed school
Accuracy	0.93	0.79	0.88	0.94	0.85	0.84	0.96
Recall	0.68	0.77	0.83	0.85	0.97	0.68	1.00
Precision	0.14	0.39	0.04	0.35	0.34	0.35	0.45
F1-score	0.24	0.52	0.08	0.50	0.51	0.46	0.62

해분야의 경우 분류에 사용되는 뚜렷한 피해 핵심 단어의 부재에 기인하는 것으로 보인다. 정밀도의 경우 대설재해의 주요 피해분야가 아닌 범주에 대해 낮게 나타나는 경향을 확인할 수 있다. 앞선 <Table 2>의 학습 데이터 셋 구성을 살펴보면, 조난사고(3%), 철도운행장애(4%), 마을고립(7%) 등 전체 검증 데이터 중 5% 내외의 기사만이 피해분야 기사에 해당하였으며, 별도의 표로 기술하지는 않았으나 검증 데이터 셋에서도 해당 비율이 유사하게 나타났다. 즉 전반적으로 피해분야에 해당하는 비율 자체가 낮고, 한 기사에서 여러 분야의 피해를 언급하고 있어 정밀도가 낮게 나타났으나, 향후 학습데이터의 누적 및 I/O Ratio의 개선을 통해 향상될 수 있을 것으로 판단된다.

향후 분류기 개선 및 활용도 제고를 위해 주요 오분류 사례를 정성적으로 분석하였다. 크게 다음의 세 가지 경우로 구분될 수 있는데, 먼저 여러 피해분야에서 핵심단어가 중복되는 경우이다. 예를 들어 교통사고, 조난사고, 낙상사고 피해분야는 '빙판길 미끄러져 사고', '부상자 병원 이송', '구급차 출동', '소방대원 구조' 와 같은 언급이 빈번하다. 따라서 피해 내용과 피해 주체를 구분하고 정밀도를 개선하는 추가적인 방안이 필요할 것으로 보인다. 다음으로 실제 피해 발생이 아닌 사전 예방 및 점검을 강조하는 경우이다. 특히 '~ 피해가 우려', '~는 정상운행', '피해 예방을 위해 ~을 점검'과 같은 경우 오분류 가능성이 높다. 따라서 특정 피해 단어와 예방 관련 단어가 한 문장 내에 동시에 출현할 경우, 이에 대한 페널티(penalty)를 부과하여 단순 예방 차원의 기사를 필터링할 필요가 있다. 마지막으로 단어사전으로 해석되지 않는 문맥적 의미전달의 경우이다. '운항에 차질이 있을 수 있다', '피해 신고는 아직 들어오지 않았다' 등은 n-gram 방법을 적용하여도 문맥상의 의미를 추출하기 어렵다. 이 같은 경우 주요 문장의 패턴을 확보함으로써 개선이 가능할 것으로 보인다. 이외에도 일반적인 경우 학습 데이터의 누적을 통해 분류기가 다양한 사례를 학습함으로써 분류기의 성능은 자연스럽게 개선될 것으로 기대할 수 있다.

4. 결 론

본 연구에서는 재해기상현상에 대한 감시 · 분석 · 예측 기술 지원을 위해 재해기상 관련 언론기사 빅데이터를 활용하여 피해정보를 분석하는 알고리즘을 개발하였다. 구체적으로는 기상영향정보 가공 및 활용을 위한 선택적 처리방안을 제시하였는데, 대설 재해기상에 대하여 언론기사에 언급된 피해분야를 자동으로 추출 · 분류할 수 있는 분류기를 설계하고 검증하였다. 구체적으로는 기계학습 기반 분류기를 활용하였으며, 본 연구목적에 적합하도록 나이브 베이즈 분류기의 세부 분류 알고리즘을 고도화한 분류기를 개발하였다. 특히 언론기사는 하나의 기사에서 여러 분야의 피해를 다룰 수 있음을 고려하여 다중 분류가 가능하도록 설계하였으며, 주류 및 비주류 피해분야 간 표본 크기 차이를 표준화하기 위하여 확률비로서 I/O Ratio를 제시하여 보정하였다. 또한 일반적 분석 단위로서 널리 사용되는 1-gram(uni-gram) 차원이 아닌, n개의 연속적 단어의 묶음으로서 n-gram을 적용하여 개별 단어가 아닌 어절의 의미를 반영하고 분류기의 정확도를 증진하고자 하였다.

본 연구에서 개발한 피해분야 분류 알고리즘은 향후 재해기상과 관련한 새로운 기사 발생 시 일정 주기를 거쳐 통합관리시스템에 피해정보를 전달하고, 시스템 사용자의 분류결과 점검을 통해 2차 학습이 가능하다는 장점이 있다. 따라서 전체 학습 알고리즘의 갱신을 통해 자동화 결과의 개선이 가능할 것으로 판단된다. 또한 재해기상 중

대설 재해에 대해 우선적으로 분석을 진행하였으나, 향후 타 재해기상에 대해서도 동일한 방법론을 적용함으로써 연구결과의 확장이 가능하다. 본 연구결과를 통해 재해기상의 피해저감 및 선제적 대응을 위한 피해정보 통합관리 시스템의 구축 기반을 마련하고, 언론기사 빅데이터를 활용하여 실시간에 준하는 기상재해 피해정보 추출 및 관련 의사결정에의 활용 가능성을 확인하였다.

References

[1] Choe, H.S. and Kim, S.J., How to utilize big data in the public sector, *It's Smart Media*, 2013, Vol. 2, No. 3, pp. 18-25.

[2] Choi, S.H., Kim, J.Y., and Lee, J.K., A new solution to disaster management, Big data, *Meteorological Technology & Policy*, 2013, Vol. 6, No. 2, pp. 77-87.

[3] Chung, K.Y., Visions and directions for the Impact-based Forecasting, *Meteorological Technology & Policy*, 2016, Vol. 9, No. 1, pp. 6-22.

[4] Cox, D.R. The regression analysis of binary sequences, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1958, Vol. 20, No. 2, pp. 215-232.

[5] Gupta V. and Lehal G.S., A Survey of Text Mining Techniques and Applications, *Journal of Emerging Technologies in Web Intelligence*, 2009, Vol. 1, pp. 60-76.

[6] Han, S.R., Kang, N.R., and Lee, C.S., Disaster Risk Evaluation for Urban Areas Under Composite Hazard Factors, *Journal of Korean Society of Hazard Mitigation*, 2015, Vol. 15, No. 3, pp. 33-43.

[7] Ihm, H., Jang, K., Lee, K., Jang, G., Seo, M.G., Han, K., and Myaneng, S.H., Multi-source food hazard event extraction for public health, *Proceedings of 2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, February 13-16, Jeju, South Korea, 2017, pp. 414-417.

[8] Jung, J.I., Lee, K.J., and Kim, S.B., Text Mining and Network Analysis of News Articles for Deriving Socio-Economic Damage Types of Heat Wave Events in Korea: 2012~2016 Cases, *Atmosphere*, 2020, Vol. 30, No. 3, pp. 237-248.

[9] Kang, E.S. and Ko, D.S., Automatic Classification Model of Electronic Documents Based on Machine Learning for Job Analysis, *The Journal of Korean Institute of Information Technology*, 2019, Vol. 17, No. 7, pp. 23-29.

[10] Kim, Y.W., Kim, B.H., Ko, G.S., Choi, M.W., Song, H.S., Kim, G.H., Yoo, S.H., Lim, J.T., Bok, K.S., and Yoo, J.S., Design and Implementation of a Flood Disaster Safety System Using Realtime Weather Big Data, *The Journal of the Korea Contents Association*, 2017, Vol. 17, No. 1, pp. 351-362.

[11] Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection, *In International Joint Conference on Artificial Intelligence*, 1995, Vol. 14, No. 2, pp. 1137-1145.

[12] Korea Environment Institute, Building and evaluating climate change adaptation capacity for national risk management: Analysis of direct and indirect effects of heat waves and cold waves based on data, 2019, Research paper.

[13] Korea Meteorological Administration, Planning research on integrated data construction methods for forecasting the impact of meteorological disasters, 2016, Researh paper

[14] Kwon, S.H., Anomaly Detection of Big Time Series Data Using Machine Learning, *Journal of Society of Korea Industrial and Systems Engineering*, 2020, Vol. 43, No. 2, pp. 33-38.

[15] Kwon, S.J., Kang, Y.H., Lee, Y.H., Lee, M.H., Park, S.H., and Kang, M.J., Analysis of Disaster Safety Situation Classification Algorithm Based on Natural Language Processing Using 119 Calls Data, *KIPS Transactions on Software and Data Engineering*, 2020, Vol. 9, No. 10, pp. 317-322.

[16] Lee, K.K., Measures to improve reliabiligy of weather forecasts based on big data analysis, *Meteorological Technology & Policy*, 2013, Vol. 6, No. 2, pp. 32-46.

[17] Lee, K.K., Public Satisfaction Analysis of Weather Forecast Service by Using Twitter, *Journal of Society of Korea Industrial and Systems Engineering*, 2018, Vol. 41, No. 2, pp. 9-15.

[18] Lee, K.K. and Kim, T.H., A Business Application of the Business Intelligence and the Big Data Analytics, *Journal of Society of Korea Industrial and Systems Engineering*, 2019, Vol. 42, No. 4, pp. 84-90.

[19] Lee, K.K., Shim, J.K., and Cho, S.J., Estimation of Risk Levels of Impact Forecast for Heavy Snow Event by Using Big Data of Media Articles, *The e-Business Studies*, 2022, Vol. 23, No. 1, pp. 233-245.

[20] National Disaster Management Research Institute, Unstructured big data disaster safety information pattern analysis, 2017, Research paper.

[21] National Institute of Meteorological Sciences, Research

on the use of disaster weather information using Big data, 2019, Research paper

[22] Noh, B.J., Xu, Z.S., Lee, J.U., Chung, Y.W. and Park, D.H., Trend analysis of foot-and-mouth disease using keyword network, in *Proceedings of Conference on Korean Society for Internet Information*, 2016, Vol. 17, No. 1, pp. 217-218.

[23] Oh, H.J., An, S.K., and Kim, Y., Social Issue Risk Type Classification based on Social Bigdata, *The Journal of the Korea Contents Association*, 2016, Vol. 16, No. 8, pp. 1-9.

[24] Palmer, T.N., The economic value of ensemble forecasts as a tool for risk assessment: From days to decades, *Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography*, 2002, Vol. 128, No. 581, pp. 747-774.

[25] Park, J.C., Han K.J., and Chae, Y.R. Correlation Analysis between Livestock Mortality Caused by Heat Wave and News Big Data, *Journal of the Association of Korean Geographers*, 2019, Vol. 8, No. 3, pp. 529-543.

[26] Seo, J.H. and Kim, Y.J., Assessing likelihood of drought impact occurrence in South Korea through Machine Learning, *Proceedings of the Korea Water Resources Association Conference*, 2021, pp.77-77.

[27] Shin, D.H. and Kim, Y.M., The utilization of Big Data's disaster management in Korea, *Journal of the Korea Contents Association*, 2015, Vol. 15, No. 2, pp. 377-392.

[28] Tim K., 80 Percent of Your Data Will Be Unstructured in Five Years, *Data Management Solutions Review*, March 28, 2019. Accessed August 24, 2023. available at: https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/

[29] Vapnik, V., and Chervonenkis, A.Y., A class of algorithms for pattern recognition learning, *Avtomat. i Telemekh*, 1964, Vol. 25, No. 6, pp. 937-945.

[30] WMO, Post-Typhoon Haiyan (Yolanda) Expert Mission to the Philippines, Manila and Tacloban, 7-12 April 2014, Mission Report.

[31] Yeh, S.W., Suggestions for expanding impact forecasting services, *Meteorological Technology & Policy*, 2017, Vol. 10, No. 1, pp. 6-17.

ORCID

Su-Ji Cho | http://orcid.org/0000-0003-1511-5348
Ki-Kwang Lee | http://orcid.org/0000-0003-2291-8376