Original Article

https://doi.org/10.12985/ksaa.2023.31.3.042
ISSN 1225-9705(print) ISSN 2466-1791(online)

잠재 디리클레 할당(LDA)을 이용한 항공안전 의무보고 토픽 예측 모형

김준환*, 백현진*, 전성진*, 최영재*

Aviation Safety Mandatory Report Topic Prediction Model using Latent Dirichlet Allocation (LDA)

Jun Hwan Kim*, Hyunjin Paek*, Sungjin Jeon*, Young Jae Choi*

ABSTRACT

Not only in aviation industry but also in other industries, safety data plays a key role to improve the level of safety performance. By analyzing safety data such as aviation safety report (text data), hazard can be identified and removed before it leads to a tragic accident. However, pre-processing of raw data (or natural language data) collected from each site should be carried out first to utilize proactive or predictive safety management system. As air traffic volume increases, the amount of data accumulated is also on the rise. Accordingly, there are clear limitation in analyzing data directly by manpower. In this paper, a topic prediction model for aviation safety mandatory report is proposed. In addition, the prediction accuracy of the proposed model was also verified using actual aviation safety mandatory report data. This research model is meaningful in that it not only effectively supports the current aviation safety mandatory report analysis work, but also can be applied to various data produced in the aviation safety field in the future.

Key Words : Aviation Safety Mandatory Report(항공안전 의무보고), Safety Management System (안전관리체계), Aviation Safety Data(항공안전데이터), Latent Dirichlet Allocation(잠재 디리클레 할당), Topic Modeling(토픽 모델링)

Ⅰ. 서 론

항공기 사고를 예방하기 위해서는 사고 발생 이전에 위해요인을 식별하는 것은 무엇보다 중요하다(Paek et al., 2022). 이러한 관점에서 국제민간항공기구(ICAO; International Civil Aviation Organization)는 부속서 13(Aircraft Accident and Incident Investigation)을 통해 회원국들로 하여금 항공안전 위해요인에 관한 정보를 수집하기 위하여 각 국가의 항공안전 의무보고 체계(Mandatory Occurrence/incident Reporting System)를 구축하도록 권고하고 있다(ICAO, 2020). 우리나라는 현행법령에서 정하는 유형의 항공기 사고, 준사고 및 항공안전장애가 발생한 경우, 항공종사자 또는 관계인이 의무보고서를 통해 해당 이벤트와 관련된 사항을 의무적으로 국가에 보고하도록 하고 있다(국토교통부, 2021).

Received: 07. Jul. 2023, Revised: 18. Jul. 2023,
Accepted: 19. Jul. 2023
* 항공안전기술원 항공데이터실 실장
연락저자 E-mail : yj.choi@kiast.or.kr
연락저자 주소 : 서울 강서구 하늘길 38 김포국제공항 국제선청사 313호

항공안전 의무보고에는 발생한 이벤트의 개황과 더불어 대상 항공편, 기종, 기번 및 조치사항 등에 관한 내용이 담겨 있으므로, 이벤트에 직·간접적으로 영향을 미친 요인을 식별하기에 유용한 정보를 제공한다(de Vries, 2020). 또한, 항공당국이 국가의 전반적인 항공안전 추세를 모니터링하거나 특정 이슈에 관한 정책을 마련하는 등 국가의 항공안전을 증진하고 안전 수준을 유지를 위한 활동을 수행하는 데 있어 가장 핵심적인 데이터로 활용된다(Karanikas and Nederend, 2018).

항공안전 의무보고 데이터를 전문적으로 분석하기 위한 방안을 마련하는 것은 국가 항공안전 증진을 위해 수행해야 하는 핵심적인 업무 중 하나이다. 우리나라의 경우, 항공안전법 시행령 제26조(권한 및 업무의 위임·위탁) 10항의 4에 따라 의무보고 대상 항공안전장애에 대한 연구·분석 업무를 전문기관인 항공안전기술원에서 위탁받아 수행하고 있다(국토교통부, 2023).

항공안전기술원 항공안전데이터분석센터는 해당 업무를 수행하기 위하여 운항·관제·정비·통계 및 데이터 분석 등 유관 분야에 대한 근무 경력과 전문성을 가지고 있는 인력(전문 분석원)을 보유하고 있다. 전문 분석원은 접수된 항공안전 의무보고를 건별로 정독하여 해당 이벤트가 발생하는 데 영향을 준 위해요인을 식별하고, 이벤트 발생유형, 비행 단계, 국가항공안전성과지표 해당 여부 및 이벤트 등급(항공기 사고, 준사고, 항공안전장애)을 구분하는 등 비정형 텍스트 데이터를 정형 데이터로 가공한다.

상기 항공안전 의무보고 데이터 분석 방식은 절대적으로 인력(Manpower)에 의존하는 방식이기 때문에 항공안전 의무보고가 일시에 다량으로 접수되거나, 예상치 못한 인력 손실이 발생하는 경우 정상적인 업무수행이 어려워진다는 한계가 존재한다. 이러한 한계점을 보완하고자, 본 연구에서는 대표적인 텍스트 분석 기법 중 하나인 LDA (Latent Dirichlet Allocation) 토픽 모델링 기법과 실제 항공안전 의무보고 데이터를 기반으로 이벤트 발생유형을 특정할 수 있는 단어 집합(Event Dictionary)을 생성하고, Naive-Bayes 방식을 기반으로 새로 접수되는 항공안전 의무보고서에 대한 이벤트 발생유형을 확률적으로 예측하는 모형을 개발하고, 그 성능을 검증하였다.

II. 본 론

2.1 국내 항공안전 의무보고 시스템

국제민간항공기구는 항공안전 의무보고 시스템이 항공안전 관리 시스템의 일부로 관리되어야 하며, 항공안전 의무보고를 접수, 처리, 평가할 수 있는 체계를 구축하는 것을 권고하였다(ICAO, 2020). 이에 따라 우리나라의 경우, 통합항공안전정보시스템 내에 항공안전 의무보고 분석 시스템을 구축하고 있다. 또한, 항공안전법 제59조(항공안전 의무보고)를 통해 의무보고 대상 이벤트가 발생한 것을 알게 된 항공종사자 또는 관계인이 항공안전법 시행규칙 별지 제65호 서식에 따른 항공안전 의무보고를 수행하도록 규정하고 있다. 해당 양식은 Fig. 1과 같이 보고자 정보, 이벤트 발생개요 및 조치사항 등을 작성할 수 있도록 구성되어 있으며, 의무보고 대상이 되는 이벤트 유형은 항공안전법 시행규칙 별표 20의2 “의무보고 대상 항공안전장애의 범위”에 작성되어 있다(국토교통부, 2023).

항공안전 의무보고는 항공당국이 안전 문제에 대한 추세를 분석하고, 새로운 위협을 식별하며, 관련 정책 및 경감방안을 마련하는 데 있어 유용한 정보를 제공한다(de Vries, 2020). 우리나라에서 항공안전 의무보고를 분석·연구하는 절차는 Fig. 2와 같으며, 의무보고 데이터 분석을 통해 국가의 항공안전을 저해하는 핵심 위해요인을 도출하고 그것의 위험도를 경감시키기 위한 방안을 주기적으로 마련한다.

■ 항공안전법 시행규칙 [별지 제65호서식] <개정 2019. 9. 23.> 통합항공안전정보시스템(https://www.esky.go.kr)에서도 보고할 수 있습니다.

항공안전 의무보고서(Aviation Safety Mandatory Report)

보고 구분 (Category of Occurrence)	[ ] 항공기사고 (Accident)	[ ] 항공기준사고 (Serious Incident)	[ ] 항공안전장애 (Incident)			
분야 구분 (Fields)	[ ] 항공기운항 (Flight Operation)	[ ] 항공기정비 (Maintenance)	[ ]항공교통관제 (Air Traffic Control)	[ ] 공항항행시설 (Aerodrome and NAVAID)		
발생유형 (Type of Occurrence)						
호출부호 (Call Sign)			등록기호(Registration)			
항공기기종·공항·항행안전시설 명칭 (Type of Aircraft or Name of Aerodrome or NAVAID)						
발생일시 (Date, Time)			발생장소·공항 (Location or Aerodrome)			
발생단계 (Phase of Flight)	[ ] 정지(standing) [ ] 푸시백/견인(push-back/towing) [ ] 유도로이동(taxi) [ ] 이륙(take-off) [ ] 초기 상승 (initial climb) [ ] 순항 (en-route) [ ] 접근(approach)		[ ] 착륙 (landing) [ ] 기동(maneuvering) [ ] 비상강하(emergency descent) [ ] 제어불능상태의 고도강하(uncontrolled descent) [ ] 충돌발생 후(post-impact) [ ] 불분명(unknown)			
비행구간 (Flight Route)				비행고도 (Altitude)		
승객수 (Number of Passengers)		승무원수 (Number of Crew Members)	운항승무원(Flight Crew)			
			객실승무원(Cabin Crew)			
사망자수 (Number of Fatalities)			부상자수 (Number of Injuries)			
기상(Weather)	[ ] VMC		[ ]IMC			
발생 개요(Description of Occurrence)						
사업자의 종류 (Type of Operations)	[ ] 국내 (Domestic Air Carrier)	[ ] 국제 (International Air Carrier)	[ ] 소형 (Small Commercial Air Transport Operator)	[ ] 항공기사용사업 (Aerial Work)	[ ] 기타 (Other)	
보고자의 성명 (Name)			보고자의 연락처 (Telephone)			

「항공안전법」 제59조제1항, 제62조제5항 및 같은 법 시행규칙 제134조제1항에 따라 항공기사고 등을 위와 같이 보고합니다.(In accordance with Paragraph 1, Article 59 and Paragraph 5, Article 62 of the Aviation Safety Act and Paragraph 1, Article 134 of the Ministerial Regulation of Aviation Safety Act, I hereby report the occurrence of mandatory reporting items as described above.)

년 월 일
Date:_____/_____/_____ (YYYY/ MM/ DD)

보고자 (Name) (서명 또는 인) (Signature)

국토교통부장관 또는 지방항공청장 귀하
(Attention : Minister of Ministry of Land, Infrastructure and Transport or Administrator of Regional Aviation Administration)

Fig. 1. Aviation safety mandatory report

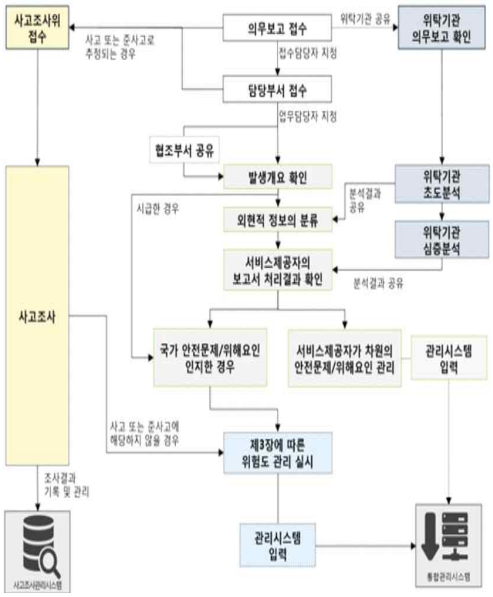

Fig. 2. Aviation safety mandatory report analysis procedure

항공종사자의 노력에도 불구하고, 항공안전 의무보고에는 필수 정보의 누락, 어휘 오류 및 표준 용어 미사용 등 인적 요인에 의한 데이터 결함이 불가피하게 발생하므로, 이를 해석하고 분류하는 데 있어 많은 물적・인적 비용이 소요된다(de Vries, 2020). 또한, 전문 분석원이 접수된 항공안전 의무보고를 수동적으로 분석하는 현행 업무 체계에서 다량의 의무보고가 동시다발적으로 접수되거나, 접수된 의무보고 내 데이터 결함이 존재하는 경우에는 업무 수행의 일관성, 객관성 및 효율성이 저하될 수 있다.

이와 같은 한계점을 보완하고, 국내 통합항공안전정보시스템의 자동화를 이룩하기 위하여 본 연구에서는 토픽 모델링 기법에 기반한 항공안전 의무보고 토픽 예측 모형을 제안하였다.

2.2 토픽 모델링

토픽 모델링은 문서 내 숨겨진 의미 구조(Semantic structure)를 발견하고, 특정 주제에 따른 토픽을 발견하기 위한 하나의 통계 분석 방식이다. 본 논문에서는 항공안전 의무보고의 토픽(이벤트 발생유형)을 식별하기 위하여 잠재 디레클레 할당(Latent Dirichlet Allocation, 이하 LDA) 기법을 사용하였다. LDA 기법은 베이지안(Bayesian) 확률에 기반한 말뭉치(Corpus) 클러스터링 기법이며(Blei et al., 2003; Nam et al., 2018), 특정 문서 내 등장한 단어의 동시 발생 확률(Co-occurrence probability)을 계산하여 해당 문서의 잠재적 주제를 도출하는 비지도 기계 학습 기법(Unsupervised learning technique)이다(Bastani et al., 2019). 특히, LDA 기법은 문서에 대해 특정 토픽별로 유사성을 가지는 단어들의 집합(단어사전)을 만들 수 있다는 점에서 데이터 분류, 감성 분석 연구에 주로 사용되었다(e.g. Bao et al., 2011; Rao et al., 2014; Kozareva, 2015; Kim and Gil, 2019).

본 연구에서는 LDA 기법을 활용하여 항공안전 의무보고 데이터의 토픽 분류를 위한 단어사전을 만들고 이에 기반하여 접수되는 항공안전 의무보고 데이터의 토픽을 자동적으로 분류하기 위한 프로세스를 제안한다.

축적된 의무보고 데이터에 LDA 기법을 적용하여 특정 토픽과 단어들로 구성된 단어사전을 만들고 각 토픽을 이벤트 유형으로 명명하였다. 각 토픽 내 포함된 단어들은 해당 토픽과 높은 연관성을 갖는 단어라고 할 수 있다. 즉, 제안된 방법을 사용하면, 항공안전 의무보고라는 문서에서 특정 이벤트 유형 또는 토픽과 연관성이 높은 단어들을 식별할 수 있으며, 해당 의무보고에 대한 이벤트 발생 유형을 확률적으로(Stochastic) 구분할 수 있다. 이러한 접근 방식은 현재 정성적・수동적으로 수행되고 있는 항공안전 의무보고 이벤트 유형 분류 작업을 자동화 할 수 있을 것으로 기대되며, 데이터 기반의 클러스터링 기법을 적용함으로써 보다 일관성 있는 분석이 가능할 것으로 판단된다.

2.3 방법론

2.3.1 사용 데이터

본 연구에서는 2018년부터 2022년까지 접수된 항공안전 의무보고 데이터를 사용하였으며, 여러 데이터 필드 중에서 이벤트 발생개요와 조치사항 전체를 모델링을 위한 텍스트 데이터로 활용하였다. 발생개요는 항공안전 의무보고 대상 이벤트가 발생한 전반적인 상황을 설명하는 부분으로, 실제 항공안전 의무보고 분석 업무를 수행하는 전문 분석원 또한 해당 내용을 통해 보고된 발생유형이 적합한지 검토하고 있다. 조치사항은 이벤트 발생 이후 보고자 및 관계인이 수행한 사후 조치에 관한 내용으로, 발생한 이벤트에 대한 실제 원인이거나 원인으로 추정되는 부분에 대한 조치 내용을 포함하고 있으므로 이벤트 유형과 관련된 단어를 식별하는 데 유용하다.

2.3.2 LDA 기법 적용 절차

항공안전 의무보고에 LDA 기법을 적용하기 위한 절차는 다음과 같은 4단계로 구성된다.

첫 번째 단계는 데이터 전처리 단계이다. 데이터 전처리는 한글 또는 영문 텍스트 데이터를 형태소 또는 단어(word) 단위로 변환하는 과정으로, ① 특수문자 제거, ② 자음 또는 모음으로만 이루어진 단어 제거(오탈자), ③ 조사 제거 순으로 이루어진다.

두 번째 단계는 불용어(Stopword)를 제거하는 단계로, 접속사 등과 같이 항공안전 의무보고에 자주 등장하지만 실제 의미가 없는 단어들을 삭제하는 단계이다. 본 연구에서는 총 840개의 불용어를 식별 및 제거하였다.

세 번째 단계는 LDA 모형의 최적 토픽 개수를 산출하기 위한 단계이다. LDA 모형의 유효성은 토픽 개수에 따라 크게 좌우되기 때문에(Blei et al., 2003), 최적 토픽 개수를 식별하여 모형의 정확도와 유효성을 확보하는 것은 매우 중요하다(Hasan et al., 2021). 본 연구는 최적 토픽 개수를 결정하기 위해 클러스터링 기법을 평가하는 대표적인 기법인 Topic coherence와 Perplexity의 평균값을 사용하였다. Topic coherence는 토픽에 속하는 N개의 단어에 대한 공동 문서 빈도와 문서 빈도 사이의 로그 비율의 합이며(Aletras and Stevenson, 2013), 동일 토픽 내 포함된 단어 간 의미적 유사성을 측정하여 토픽 대표성(Representativeness)을 평가하기 위한 척도로 사용된다(Mimno et al., 2011; Newman et al., 2010; Stevens et al., 2012). 따라서 Topic coherence의 값이 높을수록 모델링 결과가 우수한 것으로 평가할 수 있다. Perplexity는 토픽 모델링의 일반화 정도(Generalization)를 평가하는 척도로, 그 값이 작을수록 개발된 모델의 군집 개수가 더 적절하며, 입력된 데이터에 대한 우도(尤度, Likelihood)가 높다는 의미를 가진다. 본 연구에서 수립한 모형의 경우, 토픽 개수 k를 4부터 100까지 설정하여 각 경우에 대한 Topic coherence와 Perplexity 값을 계산하였으며, 최적 토픽의 개수가 6개인 것을 확인하였다. 또한, 각 토픽에 포함되는 단어의 개수는 단어별 가중치가 1.0% 이상인 단어로 선택하였고, 그 결과 토픽별로 16개의 단어를 포함하는 단어사전이 생성되었다.

마지막 단계는 6개의 토픽을 명명하는 과정이다. 토픽 명명 기준은 항공안전법 시행규칙 별표 20의2에서 규정하는 의무보고 대상 항공안전장애 범위를 활용하였다. 토픽 명명을 위해 분야(항공정비, 항공관제, 항공운항)별 전문가 중 항공안전 의무보고 데이터 분석 경험이 있는 전문가를 대상으로 설문조사를 실시하였다. 본 연구는 선행 연구(Nam and Lee, 2019)에서 제시한 2가지 평가 기준(① 각 토픽에서 등장 확률이 높은 단어가 작은 단어보다 더 큰 설명력을 가짐, ② 여러 토픽에 동시에 존재하는 유사 단어보다는 토픽 간의 차이를 나타낼 수 있는 단어에 초점을 맞출 것)을 응답자에게 충분히 설명한 후 설문을 실시하였다. 또한, 응답자에게 각 토픽에 포함된 단어 중 해당 토픽과 연관성이 가장 높은 단어를 별도로 표시하도록 요청하였다. 각 토픽에 대한 최종 명명은 전체 응답자 중 절반 이상이 동의하며, 가장 응답률이 높은 이벤트 유형(항공안전장애 유형)으로 결정하였다.

2.4 분석 결과

2.4.1 항공안전 의무보고 단어사전

LDA 기법을 적용한 후 생성한 단어사전은 Table 1과 같으며, 단어사전의 토픽은 6개로 구분되고 각 토픽별로 16개의 단어를 포함한다. 단어사전에서 단어별 중요도는 해당 단어의 출현 빈도와 동일 문서 내에서 자주 출현하는 단어 간 동시 출현 가능성에 따라 계산되며, 여러 선행 연구에서 해당 수치가 각 토픽에서 단어가 갖는 중요도로 활용할 수 있음을 입증하였다(e.g. Bi et al., 2019; Nam et al., 2018; Nam and Lee, 2019).

Table 1. LDA results

토픽	단어	중요도
T1	Landing	2.5%
	추정	2.3%
	Hard	2.1%
	Gear	2.0%
	승무원	2.0%
	수신	1.8%
	접지	1.7%
	Well	1.3%
	Wheel	1.3%
	운항	1.3%
	통보	1.2%
	적용	1.2%
	ACC	1.2%
	Hard	1.1%
	FOQA	1.1%
	자료	1.0%

Table 1. Continued

T2	드론	3.3%
	상황	3.0%
	테러	2.0%
	외곽	1.8%
	탐지	1.8%
	이착륙	1.7%
	발행	1.7%
	레이더	1.4%
	중단	1.3%
	착륙	1.3%
	입수	1.2%
	승인	1.1%
	관제탑	1.0%
	재개	1.0%
	신호	1.0%
	지상	1.0%
T3	Level	8.9%
	통합	5.3%
	부식	4.7%
	지침	3.5%
	Doubler	3.4%
	Waste	3.3%
	제거	3.0%
	Corrosion	2.9%
	제작사	2.8%
	Repair	2.7%
	관리	2.6%
	기존	2.6%
	발견	2.6%
	결함	2.5%
	Stringer	2.3%
	Drain	2.2%
T4	접근	5.7%
	TCAS	4.5%
	착륙	3.8%
	관제	2.4%
	지시	2.2%
	기동	2.2%
	강하	2.1%
	고도	1.9%
	ATC	1.9%
	상승	1.8%
	활주로	1.6%
	ILS	1.6%
	RWY	1.5%
	발령	1.5%
	항적	1.4%
	회피	1.4%
T5	부식	7.1%
	수리	5.7%
	정비	3.5%
	STA	2.9%
	SRM	2.6%
	Wing	2.5%
	제작사	2.4%
	소재	1.9%
	지침	1.8%
	부위	1.8%
	발견	1.7%
	초과	1.7%
	결함	1.5%
	허용	1.5%
	교범	1.2%
	등급	1.2%
T6	점검	5.1%
	시스템	4.9%
	대체	4.4%
	이상	3.5%
	연결	2.7%
	프로그램	2.5%
	수리	2.4%
	장비	2.4%
	입력	2.3%
	의무	2.3%
	일반	2.2%
	작업	1.5%
	운항	1.4%
	관제	1.4%
	교체	1.3%
	ATIS	1.0%

2.4.2 토픽 명명 결과

6개 토픽을 적절한 이벤트 유형으로 명명하기 위해 항공안전 의무보고 데이터 분석 경험이 있는 운항, 관제, 정비, 항공안전 분야의 전문가를 구성하여 설문을 진행하였다. 각 분야 평가위원은 토픽 내 단어와 가장 높은 연관성이 있는 유추되는 이벤트 유형을 '항공안전

법 시행규칙 별표 20의2에서 규정하는 의무보고 대상 항공안전장애 범위' 중 1개의 유형으로 지정하였으며, 그 결과는 Table 2와 같다.

2.4.3 항공안전 의무보고 토픽 예측 모형

2018년부터 2022년까지 수집한 항공안전 의무보고 데이터를 기반으로 생성한 LDA 모형(단어사전)을 보면 6개의 토픽($t_1, t_2, t_3, t_4, t_5, t_6$)과 각 토픽($t_x$)에 포함된 16개 단어($N_x = n_x^1, n_x^2, ..., n_x^{16}$) 및 각 단어에 대한 중요도($W_x = w_x^1, w_x^2, ..., w_x^{16}$)가 포함되어 있다. 본 연구에서는 해당 모형을 기반으로 새로운 의무보고 문서의 이벤트 유형(토픽)을 예측하고자 하였으며, 그 절차는 다음과 같다. 첫 번째 과정으로, 새로 접수된 의무보고 문서(d)에 대해 데이터 전처리를 수행한다. LDA 모형 생성을 위한 데이터 가공 방식과 마찬가지로 의무보고 문서(d)에 포함된 특수문자, 접속사, 오탈자 및 불용어 등을 제거한 후, 식 (1)과 같이 접수된 의무보고 문서를 k개의 형태소(m) 단위로 구분한다.

$$d = \{m_1, m_2, ..., m_k\} \tag{1}$$

두 번째 과정으로는, 단어사전에 포함된 토픽별 · 단어별 중요도를 사용하여 해당 의무보고 문서가 각 토픽(t_x)에 포함될 확률(p_x)을 식 (2)와 같은 방법으로 계산한다. 단, 문서(d)내 각 형태소(m)가 특정 토픽에 대한 단어 집합(N_x)에 포함된 경우에만 그에 상응하는 중요도(w_x^t)를 합한다. 또한, 문서(d) 내에서는 동일한 형태소가 중복하여 등장할 수 있으며 이때는 해당 형태소에 상응하는 중요도를 중복하여 합한다. 이는 의무보고 상에 여러 차례 반복하여 등장하는 형태소에 대한 중요성을 고려하기 위함이다.

$$p_x = \sum_{t=1}^{k} w_x^t \tag{2}$$

마지막 세 번째 과정으로, 식 (3)과 같이 접수된 의무보고(d)가 각 토픽에 속할 확률($P_d = \{p_1, p_2, p_3, p_4, p_5, p_6\}$) 중 가장 큰 값을 찾아 해당 의무보고($d$)의 토픽($t_d$)을 예측한다.

$$t_d = t_{\arg max(P_d)} \tag{3}$$

2023년 상반기에 접수된 의무보고 40건을 대상으로 제안된 모델을 적용하여 토픽 예측 정확도를 확인한 결과, 75%(30건)의 의무보고가 실제 전문 분석원이 분석한 결과와 일치하는 것을 확인하였다.

제안된 의무보고 토픽 예측 모형은 접수된 의무보고에 작성된 단어를 기반으로 해당 의무보고가 각 토픽에 해당할 확률을 조건부(Bayes) 확률로 구한 후, 확률이 가장 높은 토픽을 해당 의무보고의 이벤트 유형(토픽)으로 채택한다. 이는 전적으로 데이터에 기반한 방법론으로, LDA 모형 수립 시 사용하는 데이터의 품질과 양에 따라 제안된 모형의 성능이 달라질 수 있다.

또한, 새로 접수된 의무보고가 기존에 접수되었던 의무보고와 전혀 다른 단어를 사용하는 경우에는 모형이 해당 의무보고의 토픽을 부정확하게 분류할 수 있다. 따라서 제안된 모형을 실제 업무에 활용하기 위해서는, 유관 규정에서 사용하는 용어의 변경이나 새롭게 등장하는 단어 및 종사자들이 사용하는 단어의 경향성을 모두 고려할 필요가 있다.

Table 2. Result of topic naming

구분	명명 결과	관련 규정 조항
Topic 1 (t_1)	하드랜딩 (Hard Landing)	항공안전법 시행규칙 [별표 20의2] 2.가.2)
Topic 2 (t_2)	드론위협 (Drone Interference)	항공안전법 시행규칙 [별표 20의2] 7.가.2)
Topic 3 (t_3)	항공기 부식 (Aircraft Corrosion)	항공안전법 시행규칙 [별표 20의2] 5.마
Topic 4 (t_4)	공중충돌경보 (ACAS RA)	항공안전법 시행규칙 [별표 20의2] 1.가
Topic 5 (t_5)	항공기 부식 (Aircraft Corrosion)	항공안전법 시행규칙 [별표 20의2] 5.마
Topic 6 (t_6)	항행안전시설 장애 (Navigation Safety Facilities Malfunction)	항공안전법 시행규칙 [별표 20의2] 6.마

Ⅲ. 결 론

본 연구에서는 국가 항공안전관리의 핵심 업무 중 하나인 항공안전 의무보고 데이터 분석 업무의 효율성을 증진하기 위하여 항공안전 의무보고 토픽 예측 모형을 제안하였다.

이를 위해 대표적 토픽 모델링 기법 중 하나인 LDA 기법을 실제 접수된 항공안전 의무보고 데이터에 적용하여 단어사전을 생성하고, 단어 간 동시 출현 빈도 등을 기반으로 토픽별 단어의 중요도를 산출하였다. 이후, 운항, 정비, 관제 등 각 분야 전문가로 구성된 집단을 구성하고 유관 규정을 참고하여 단어사전의 각 토

픽을 명명하는 작업을 수행하였다. 마지막으로, 생성한 단어사전을 활용하여 접수된 의무보고의 토픽을 예측하는 모형을 제안하였다.

해당 모형의 토픽 예측 정확도는 75%로, 제안된 모형이 실제 업무 수행에 충분히 활용될 가능성이 있음을 확인하였다.

본 연구는 기존 인력 기반의 정성적 방식에 의존하던 항공안전 의무보고 분석 업무체계를 데이터 기반의 자동화 방식으로 전환한다는 점에서 의의가 있다. 기존 정성적 방식의 경우, 항공안전 의무보고 데이터를 분석할 때 분석원의 주관적 판단에 의존한다는 점에서 과학적 근거가 부족하고, 분석원별로 분석 결과의 일관성이 떨어진다는 한계가 존재한다. 또한, 분석원의 인력 변동, 업무량 과중 등 환경적 변화에 따라 업무 처리의 효율성이 감소할 수 있다는 문제점도 존재한다.

제안된 의무보고 토픽 예측 모형은 단시간 내에 많은 양의 의무보고 데이터 토픽 분류를 빠르게 처리할 수 있고, 관련 지식이 부족한 사람이라도 누구나 쉽게 업무를 수행할 수 있도록 접수된 의무보고의 토픽을 제시해줄 수 있다는 장점이 존재한다. 또한, 실제 과거 데이터에 기반한 객관적 분석결과를 제시하므로 분석결과에 일관성이 존재한다는 강점이 있다.

다만, 모형 개발에 가용할 수 있는 전체 항공안전 의무보고 데이터를 사용했음에도 불구하고 절대적인 데이터 크기(Volume)가 부족할 수 있다는 한계점이 존재한다. 따라서 향후 접수되는 항공안전 데이터를 지속적으로 학습 데이터로 활용하여 모형을 고도화함으로써 산출 결과의 신뢰성을 제고할 필요가 있다.

후 기

본 연구는 국토교통과학기술진흥원의 “빅데이터 기반 항공안전관리 기술개발 및 플랫폼 구축”(20BDAS-B158275-01)의 일환으로 수행되었으며, 지원에 감사드립니다.

References

1. Paek. H., Kim. J. H., Lim. J. J, Jeon. S., and Choi. Y. J., “Quantitative safety risk assessment using aviation safety data”, Journal of the Korean Society for Aviation and Aeronautics, 30(4), 2022, pp.145-158.
2. ICAO, “Annex 13 – Aircraft Accident and Incident Investigation 12th Edition”, 2020.
3. MOLIT, “Aviation Safety Act, Article 59”, 2021.
4. de Vries, V., “Classification of aviation safety reports using machine learning”, 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation, IEEE, Singapore, 2020, pp.1-6.
5. Karanikas, N., Nederend, J., “The controllability classification of safety events and its application to aviation investigation reports”, Safety Science, 108, 2018, pp.89-103.
6. MOLIT, “Aviation Safety Enforcement, Article 26”, 2023.
7. MOLIT, “Aviation Safety Regulation, Enclosure No.65”, 2023.
8. Blei, D. M., Ng, A. Y., and Jordan, M. I., “Latent dirichlet allocation”, Journal of Machine Learning Research, 3, 2003, pp.993-1022.
9. Nam, S., Ha, C., and Lee, H. C., “Redesigning in-flight service with service blueprint based on text analysis”, Sustainability, 10(12), 2018, Online Published.
10. Bastani, K., Namavari, H., and Shaffer, J., “Latent dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints”, Expert Systems with Applications, 127, 2019, pp.256-271.
11. Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y., “Mining social emotions from affective text”, IEEE Transactions on Knowledge and Data Engineering, 24(9), 2011, pp.1658-1670.
12. Rao, Y., Lei, J., Wenyin, L., Li, Q., and Chen, M., “Building emotional dictionary for sentiment analysis of online news”, World Wide Web, 17, 2014, pp.723-742.
13. Kozareva, Z., “Everyone likes shopping! multi-class product categorization for e-commerce” In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1329-1333.
14. Kim, S. W., and Gil, J. M., “Research paper

classification systems based on TF-IDF and LDA schemes. Human-centric", Computing and Information Sciences, 9, 2019, pp.1-21.
15. Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., and Islam, M. J., "Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA)", Proceedings of International Conference on Trends in Computational and Cognitive Engineering, TCCE, Singapore, 2021, pp.341-354.
16. Aletras, N., Stevenson, M., "Evaluating topic coherence using distributional semantics", 10th International Conference on Computational Semantics, IWCS, 2013, pp.13-22.
17. Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A., "Optimizing semantic coherence in topic models", 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Scotland, 2011, pp.262-272.
18. Newman, D., Lau, J. H., Grieser, K., and Baldwin, T., "Automatic evaluation of topic coherence", The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, California, 2010, pp.100-108.
19. Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D., "Exploring topic coherence over many models and many topics", 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, Korea, 2012, pp.952-961.
20. Nam, S., and Lee, H. C., "A text analytics-based importance performance analysis and its application to airline service", Sustainability, 11(21), 2019, Online Published.
21. Bi, J. W., Liu, Y., Fan, Z. P., and Zhang, J., "Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews", Tourism Management, 70, 2019, pp.460-478.