한국ITS학회논문지 J. Korea Inst. Intell. Transp. Syst.

텍스트 임베딩을 이용한 자율주행자동차 교통사고 분석에 관한 연구

Study of Analysis for Autonomous Vehicle Collision Using Text Embedding

박 상 민* · 이 환 필** · 소 재 현*** · 윤 일 수****

* 주저자 : 아주대학교 교통시스템공학과 연구조교수
** 공저자 : 한국도로공사 도로교통연구원 책임연구원
*** 공저자 : 아주대학교 교통시스템공학과 조교수
**** 교신저자 : 아주대학교 교통시스템공학과 교수

Sangmin Park* · Hwanpil Lee** · Jaehyun(Jason) So*** · Ilsoo Yun****

* Dept. of Transportation System Engineering, Univ. of Ajou
** Division of Transportation Research, Korea Expressway Corporation Research Institute
*** Dept. of Transportation System Engineering, Univ. of Ajou
***** Dept. of Transportation System Engineering, Univ. of Ajou
† Corresponding author : Ilsoo Yun, ilsooyun@ajou.ac.kr

Vol.20 No.1(2021)
February, 2021
pp.160~173

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2021.20.1.160

Received 15 November 2020
Revised 28 November 2020
Accepted 18 February 2021

요 약

최근 전 세계적으로 자율주행자동차 개발을 위한 연구가 증가하고 있으며, 자율주행자동차의 실도로 도입이 증가되고 있는 추세이다. 하지만, 자율주행자동차의 교통사고 발생으로 인해 자율주행자동차 안전성에 대한 관심이 높아지고 있다. 또한, 자율주행자동차 교통사고에 대한 특성 파악 및 분석 방법론 개발의 필요성이 대두되고 있다. 특히 미국 캘리포니아 차량관리국(California Department of Motor Vehicles, DMV)에서는 자율주행자동차의 교통사고 데이터를 수집하여 리포트 형태로 제공하고 있다. 본 연구에서는 DMV에서 제공하는 자율주행자동차 교통사고를 분석하는 방법론을 제시하였다. 또한, 텍스트 임베딩 기법을 이용하여 주요 키워드 및 주요 토픽 도출을 통해 개발된 방법론의 활용도를 검토하였다. 본 연구에서 개발된 방법론은 향후 자율주행자동차 교통사고 데이터가 충분히 수집된다면 자율주행자동차 교통사고 분석 및 자율주행자동차 개발시 활용될 수 있을 것으로 기대된다.

핵심어 : 텍스트 임베딩, 자율주행자동차, 교통사고, 토픽모델링

ABSTRACT

Recently, research on the development of autonomous vehicles has increased worldwide. Moreover, a means to identify and analyze the characteristics of traffic accidents of autonomous vehicles is needed. Accordingly, traffic accident data of autonomous vehicles are being collected in California, USA. This research examined the characteristics of traffic accidents of autonomous vehicles. Primarily, traffic accident data for autonomous vehicles were analyzed, and the text data used text-embedding techniques to derive major keywords and four topics. The methodology of this study is expected to be used in the analysis of traffic accidents in autonomous vehicles.

Key words : Text Embedding, Autonomous Vehicle, Accident, Topic Modelling

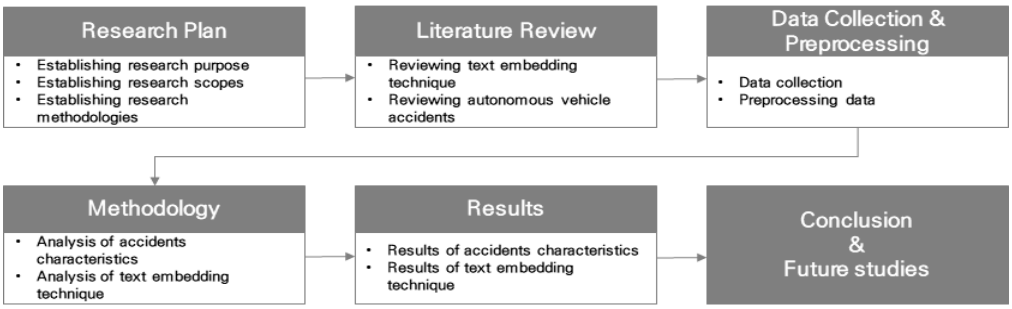
Ⅰ. 서 론

1. 연구의 배경 및 목적

최근 전 세계적으로 자율주행 SAE 레벨 2 수준의 차량이 양산되고 있으며, 레벨 4 수준의 자율주행자동차 상용화를 위해 기술 개발과 연구가 진행되고 있다. 국내에서도 자율주행기술 개발과 이를 지원하기 위한 연구들이 수행 중이다. 하지만, 2016년 구글 자율주행자동차의 잘못된 판단으로 인한 사고 발생과 2016년 테슬라 오토 파일럿의 인지 실패 및 2018년 우버의 교통사고로 인한 사망사건 등으로 인하여 자율주행자동차에 대한 안전성 문제가 대두되고 있다(Park et al., 2018). 또한, 자율주행자동차의 시장 진입률(market penetration rate)이 높아짐에 따라 기존 테스트 베드에서의 실차 실험 및 실도로 주행 실험에서 발견되지 못한 인지 실패 및 판단 오류 등이 실도로에 자율주행자동차 도입시 발생할 수 있다. 이러한 위험요소에 대비하기 위해서는 현재 실도로에서 발생한 자율주행자동차 교통사고를 면밀히 분석하고 이를 관리할 필요가 있다. Park et al.(2019)은 자율주행자동차 실험 시나리오 개발을 위해 일반차량의 교통사고 데이터를 이용하였지만 자율주행자동차의 교통사고 발생원인은 일반차량의 교통사고 원인과 차이가 있을 것으로 판단되며, 일반차량의 교통사고 데이터를 사용했다는 한계가 존재한다. 이에 본 연구에서는 자율주행자동차 교통사고를 분석하기 위해서 미국 캘리포니아 자동차 차량 관리국(California Department of Motor Vehicles, DMV)에서 수집한 자율주행자동차 교통사고 데이터를 분석하고 그 특징을 파악하고자 한다. 또한, 자율주행자동차 교통사고 데이터에 기술된 '사고 상세 설명(accident details description)' 항목을 텍스트 임베딩 기법을 이용하여 분석하고 자율주행자동차 교통사고를 분석하는 방법론을 제시하고자 한다.

2. 연구의 범위 및 방법

캘리포니아 DMV의 자율주행자동차 교통사고 보고서의 '사고 상세 설명' 데이터를 사용한 본 연구의 특성상 공간적 연구 범위는 미국 캘리포니아이며, 시간적 범위는 2019년 1월부터 2020년 8월까지로 설정하였다. 본 연구에서는 자율주행자동차 교통사고 분석 및 분석방법론 개발을 위해 DMV 자율주행자동차 교통사고 데이터를 수집하고, 통계분석 등을 통해 수집된 자료에서 발생한 자율주행자동차 교통사고 특성을 파악하였다. 또한, 텍스트 임베딩 기법을 이용하여 자율주행자동차 교통사고 '사고 상세 설명' 데이터 에서 키워드를 도출하고, 텍스트 임베딩을 통해 자율주행자동차 교통사고를 군집화 하였다. 또한 자율주행자동차 교통사고에서 발생할 수 있는 주요 상황을 도출하였다. 마지막으로 결론 및 향후 연구 과제를 도출하였다.



<Fig. 1> Research process

Ⅱ. 관련 문헌 고찰

1. 텍스트 마이닝

1) 텍스트 임베딩

텍스트 마이닝(text mining)이란 데이터의 구조가 다양한 비정형 데이터(unstructured data) 중 하나인 텍스트 데이터로부터 의미 있는 정보를 추출하는 방법이다. 특히, 텍스트 임베딩은 텍스트 마이닝의 대표적인 기법으로 인간의 언어로 되어 있는 문자를 컴퓨터의 언어인 숫자로 변환하여 분석하는 방법이다(Chae, 2019). 텍스트 임베딩은 단어나 문장을 벡터로 변환하여 표현하는 방법으로 거대한 단어 집합에서 의미적, 통사적 정보를 추출하는데 효과적인 방법이다(Lai et al., 2016). 텍스트 임베딩은 키워드 분석, 토픽 모델링, 기계번역 등 다양한 방법론이 존재한다. 그 중에서도 토픽 모델링(topic modeling)은 대량의 텍스트를 이용하여 주요 주제를 도출하기 위한 텍스트 임베딩의 주요 기법이다. 특히, 문맥과 관련된 단어들을 이용하여 주요 주제를 도출하는 기법으로 문서들의 군집에 사용된다.

2) 잠재 디리클레 할당

토픽 모델링 기법중 주로 사용되는 잠재 디리클레 할당(latent dirichlet allocation, 이하 LDA)은 다수의 비구조적인 문서에서 어떤 주제들이 존재하는지에 대한 확률 모형이다(Blei et al., 2003). 토픽의 확률 분포와 단어 확률 분포를 추정하기 위한 사전 분포로 디리클레 분포를 사용하여 붙여진 이름이다(Baek, 2018). LDA는 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지를 서술하는 확률적 토픽 모델 기법 중 하나이다(Blei et al., 2003). 다음 식 (1)와 같은 조건부 확률 분포 식(conditional probabilistic distribution)을 갖는다(Blei, 2012).

$$p(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \quad \cdots\cdots (1)$$

$$= (\prod_{n=1}^{K} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}))$$

여기서,

β_k : kth topic in document

θ_d : Topic proportions on dth document

$z_{d,n}$: Topic assignment nth words on dth document

$w_{d,n}$: Observed nth word on dth document

3) LDA를 이용한 연구사례

Cho et al.(2015)은 교통카드 데이터에 대하여 LDA 기법을 적용하여 청주시 버스 승객들의 이동패턴을 분석하였다. 해당 연구에서는 교통카드 데이터에 LDA 분석기법을 적용하고 다차원적 분석을 통해 이동패턴을 추출하였다. 분석결과, 총 5개의 패턴을 도출하였으며 연구결과를 다양한 교통정책의 수립에 활용할 수 있을 것으로 제시 하였다.

Oh et al.(2016) 역시 LDA 모형을 이용하여 도로분야 ITS 정책 이슈를 탐색하는 연구를 수행하였다. 분석결과, 국내 인터넷에 공개된 자료 등을 이용하여 한정된 자원과 시간 내에 빠르게 정책이슈를 발굴할 수 있음을 확인하였다.

Sun and Yin(2017)은 1990년부터 2015년까지 수행된 교통 관련 연구들의 주제와 동향을 분석하기 위해 22개의 교통관련 저널에서 수집한 17,163건의 초록을 수집하여 분석하였다. 분석결과, prediction and forecasting, pedestrian, route choice, vehicle control 등 총 25가지의 토픽이 추출되었으며, 각각의 토픽 관련 연구들이 수행된 국가와 연도별 수행 추세를 도출하였다.

Woo and Lee(2020)는 LDA 모델을 이용하여 국가연구개발사업을 통해 수행되고 있는 ICT(information and communication technology) 분야의 연구과제에 대한 주요 연구 토픽과 동향을 탐색하는 연구를 수행하였다. 연구수행결과, 인공지능, 빅데이터, 사물인터넷(Internet of things)과 같은 토픽이 도출되었으며, 주요 동향으로 초실감 미디어에 관한 연구가 활발하게 진행되고 있는 것을 확인하였다.

Park and Lee(2020)는 LDA 토픽모델링 기법을 활용하여 부산시 민원 빅데이터 분석을 수행하였다. 2015년에서 2017년까지 9,625건의 부산시 전자 민원을 대상으로 20개의 민원토픽을 추출하였으며, 4개의 Hot 민원(버스정차, 택시기사, 칭찬, 민원처리)과 4개의 Cold 민원(CCTV 설치, 버스노선, 공원주차장, 축제 불만)을 도출하였다.

LDA를 이용한 주요 선행 연구사례를 분석한 결과, 텍스트 문서들을 군집하였으며 토픽과 동향을 탐색하는 연구가 대부분임을 확인할 수 있었으며, 텍스트로 기술된 데이터를 군집화하고 군집된 토픽으로부터 데이터의 주요 주제를 찾는 것이 가능함을 확인하였다.

2. 자율주행자동차 교통사고 분석

1) 관련 연구

Petrović et al.(2020)은 자율주행자동차 교통사고 특성을 분석하기 위해 미국 캘리포니아 주에서 발생한 자율주행자동차 교통사고 자료를 수집하고 분석하였다. 분석을 위해 자율주행 모드인 경우와 일반 모드인 경우를 분리하여 교통사고 특성을 비교하였다. 특히, 교통사고 발생 시의 충돌 유형, 운전자의 조작 및 오류를 중점적으로 분석하였다. 분석결과 자율주행 모드인 경우 후방 추돌 유형의 교통사고가 많이 발생한 것으로 분석되었으며, 보행자 교통사고는 더 적게 발생한 것으로 나타났다.

Favarò et al.(2017)은 미국 캘리포니아 주에서 수집한 자율주행자동차 교통사고 보고서를 이용하여 교통사고의 충돌 유형, 빈도, 원인 등을 분석하였다. 특히, 자율주행 모드 여부, 자율주행 차량의 파손 부위, 사고 원인 등을 종합하고 이를 도식화 하였다. 분석결과, Petrović et al.(2020) 연구의 결론과 유사하게 후방 추돌 사고가 가장 많이 발생한 것으로 분석되었다.

2) DMV Autonomous Vehicle Collision Report

미국 캘리포니아 자동차 차량 관리국에서는 2014년에 자율주행자동차 테스트 프로그램(autonomous vehicle test program)을 수립하고, 제조사의 운전자가 운전석에 앉아 있는 상태로 자율주행자동차 실험을 허가하였으며, 2018년에는 운전자 없는 자율주행자동차 실험을 허가하였다. 이와 함께, 자율주행자동차의 충돌 또는 사고로 인해 재산피해, 신체 상해 또는 사망이 발생하면, 10일 이내에 자율주행자동차의 교통사고 보고서를 의무 제출하도록 하였다. 제출받은 자율주행자동차 충돌 보고서는 웹사이트를 통해 제공하고 있다. DMV의 자율주행자동차 교통사고 보고서에는 사고 시간, 사고 상세 설명, 사고 요인, 사고 대상, 환경 요인, 차량 파손 위치, 날씨, 조도 등 다양한 정보를 제공하고 있다. 특히 Section 5의 사고 상세 설명(accident details description)의 경우 교통사고가 난 상황에 대해 원인차량과 피해차량의 충돌 직전 주행 상황과 충돌 상황, 피해 정도, 경찰 신고여부에 대해 기술하고 있다. 자율주행자동차의 충돌 및 사고 보고서는 총 6개의 section으로 구성되어 있으며, 각 section별 주요 내용은 아래 <Table 1>과 같다.

<Table 1> Autonomous Vehicle Collision Report Contents by Section

Section	Contents
Section 1 – Manufacturer's Information	Manufacturer's name / Business name
Section 2 – Accident Information(Vehicle 1)	Date of accident / Time of accident / Vehicle year / Make / Model Vehicle was (Moving / Stopped in traffic) Involved in the accident (Pedestrian / Bicyclist) Describe vehicle damage Shade in damaged area
Section 3 – Other Party's Information(Vehicle 2)	Vehicle year / Model Vehicle was (Moving / Stopped in traffic) Involved in the accident (Pedestrian / Bicyclist)
Section 4 - In Jury / Death, Property Damage	Injured / Deceased / Driver / Passenger / Bicyclist / Property
Section 5 - Accident Details - Description	Autonomous Mode / Conventional Mode Additional information attached : Weather / Lighting / Roadway surface / Roadway conditions / Movement preceding collision / Type of collision / Other associated factor(s)
Section 6 - Certification	Program director / Authorized representative printed name and title / Signature / Date signed

Ⅲ. 텍스트 임베딩 기법 적용을 통한 자율주행자동차 교통사고 분석 방법론 개발

1. 데이터 수집

본 연구에서는 자율주행자동차 교통사고 분석 방법론을 개발하기 위해 미국 캘리포니아 차량관리국(California Department of Motor Vehicles, DMV)에서 제공하고 있는 자율주행자동차 교통사고 보고서를 수집하였다. DMV의 자율주행자동차 교통사고 보고서에는 사고 시간, 사고 상황, 환경 요인, 차량 파손 위치, 날씨, 조도 등 다양한 정보를 제공하고 있어 자율주행자동차의 교통사고 당시 상황을 분석하는 것이 가능하다. 특히, 비정형 데이터인 텍스트로 기술된 '사고 상세 설명(accident details description)' 항목을 통해 기존 정형 데이터에서 발견하기 어려운 사고 상황을 파악할 수 있는 장점이 있다. 따라서 본 연구에서는 텍스트 임베딩 기법 적용을 통한 자율주행자동차 교통사고 분석 방법론 개발을 위해 2019년 1월부터 2020년 8월까지 발생한 자율주행자동차 교통사고 보고서를 수집하였다. 수집된 각각의 자율주행자동차 교통사고 보고서를 코딩하여 데이터 셋을 구축하였다.

2. 통계 기법을 이용한 데이터 분석

자율주행자동차의 교통사고를 텍스트 임베딩 기법을 이용하여 분석하기에 앞서 사고 발생 현황과 원인데 대한 분석을 위해 데이터 항목별 빈도 분석을 수행하였다. 특히, 자율주행 모드인 경우의 자율주행자동차의 교통사고를 분석하기 위해 차량의 거동 관련 요소, 충돌 및 손상 관련 요소, 도로 환경 요소, 환경 요소, 기타 등으로 구분하였다. 차량의 거동 관련 요소로는 차량들의 충돌 이전의 움직임(movement preceding collision), 객체의 종류(type of objects)가 있으며, 충돌 및 손상 관련 요소로는 차량의 손상된 영역, 충돌 종류(type of collision)로 구성하였다. 또한, 도로 환경 요소는 노면상태(road surface), 도로 상태(roadway condition)이며, 환경 요소로는 날씨(weather), 조도(lighting) 항목을 선정하였다. 다음 <Table 2>는 각 범주별 항목을 나타낸다.

<Table 2> Autonomous Vehicle Collision Report Contents by Categories

Category	Factor	Contents
Vehicle's Movements	Movement Preceding Collision	Stopped / Proceeding straight / Ran off road / Making right turn / Making left turn / Making U turn / Backing / Slowing & Stopping / Passing other vehicle / Changing lanes / Parking maneuver / Entering traffic / Other unsafe turning / Xing into opposing lane / Parked / Merging / Traveling wrong way
	Type of Objects	Vehicle/ Bicyclist / Truck / Bus / Scooter / Electric Scooter / Skateboard
Collision and Damage	Type of Collision	Head-on / Side swipe / Rear end / Broadside / Hit object / Overturned / Vehicle & Pedestrian
	Damaged Area	Spot of Damaged Area
Road Environments	Roadway Surface	Dry / Wet / Snowy – icy / Slippery (Muddy, Oily)
	Roadway Conditions	Holes, Deep rut / Loose material on roadway / Obstruction on roadway / Construction-repair zone / Reduced roadway width / Flooded / No unusual conditions
Environments	Weather	Clear / Cloudy / Raining / Snowing / Fog(Visibility) / Other / Wind
	Lightning	Daylight / Dusk(Dawn) / Dark(Street light) / Dark(Street light not functioning)

3. 텍스트 임베딩 기법을 이용한 자율주행자동차 교통사고 분석 방법론

1) 텍스트 전처리

텍스트 임베딩 기법을 적용하기 위해서는 텍스트 전처리 과정이 필수적이다. 본 연구에서는 텍스트 전처리를 위해 파이썬 3.6을 이용하였다. 불용어는 데이터 상에서 큰 의미가 없는 단어이며, 텍스트 전처리를 위해 특수 문자 처리, 대문자 처리 및 불용어 제거를 먼저 수행하였다. 불용어 제거는 자연어 처리 라이브러리인 nltk라이브러리에 포함된 영어 불용어 사전과 수집된 데이터로부터 구축한 추가 불용어 사전을 이용하여 제거하였다.

2) 주요 키워드 도출

자율주행자동차 교통사고의 '사고 상세 설명'을 분석하기 위해서 불용어 처리가 완료된 문장에서 주요 키워드를 도출하였다. 주요 키워드를 도출하기 위해서는 토큰화(tokenization) 작업이 필수적이다. 토큰화는 의미를 갖는 단어 단위로 데이터를 구분하는 것을 의미한다. 본 연구에서는 '사고 상세 설명' 데이터들을 토큰화 시킨 후 텍스트들에서 명사와 빈도를 추출하고 이를 바탕으로 주요 키워드를 도출하였다. 주요 키워드와 단어의 빈도를 이용하여 키워드 기반의 워드클라우드(wordcloud)를 구축하고 이를 통해 의미 있는 정보가 추출 가능한지 검토하였다.

3) 최적 토픽 수 결정

본 연구에서는 자율주행자동차 교통사고의 주요 유형을 도출하기 위해 토픽 모델링을 수행하였다. 토픽 모델링을 수행하기 전에 최적 토픽 개수를 결정하는 것이 필수적이다. 본 연구에서는 최적 토픽 개수 결정을 위하여 대표적인 평가 지표인 perplexity 점수를 사용하였다. perplexity 점수는 불순도 정도를 나타내며, 구축한 모형의 정확도를 평가하는 지표이다. 또한, 전체 모형 중 perplexity 점수가 가장 낮은 토픽의 수를 사용하여 토픽 모델링을 수행한다(Ryu, 2019).

4) 토픽 모델링

'사고 상세 설명'을 이용하여 자율주행자동차 교통사고의 주요 유형을 도출하기 위해 토픽 모델링을 수행하였다. 토픽 모델링은 대량의 텍스트로부터 숨겨져 있는 주제 구조를 발견하는 텍스트 임베딩 기법으로 대량의 텍스트들을 군집화 하는 것이 가능하다. 본 연구에서는 LDA 토픽 모델링 알고리즘을 사용하였다. LDA는 각 문서에 어떤 주제들이 존재하는지에 대한 확률 모형이다. LDA를 구현하기 위해 기계학습 및 데이터 분석에 용이한 프로그래밍 언어인 파이썬 3.6과 텍스트 임베딩 라이브러리인 nltk, 토픽 모델링 및 자연어 처리를 위한 오픈 소스 라이브러리인 Gensim을 이용하였다.

5) 자율주행자동차 교통사고 주요 상황 도출

토픽 모델링 수행결과를 바탕으로, 토픽별로 자율주행자동차 교통사고 주요 상황을 도출하였다. 본 연구에서는 토픽 모델링을 통해 도출된 토픽을 바탕으로 각 토픽에 해당하는 교통사고를 매칭하여 분석을 수행하였다. 특히 토픽별 교통사고 중 가장 빈도가 높은 자율주행자동차 교통사고의 주요 상황을 도출하였다.

Ⅳ. 텍스트 임베딩을 이용한 자율주행자동차 교통사고 분석 결과

1. 자율주행자동차 교통사고 분석 결과

수집된 자료를 이용하여 자율주행자동차 교통사고의 통계를 분석하였다. 자율주행자동차 교통사고 데이터에는 자율주행 모드와 일반 모드로 구분되어 있다. 수집된 자율주행자동차 교통사고 데이터는 총 137건이며, 자율주행 모드일 때의 교통사고는 62건이며, 일반모드인 경우는 75건으로 나타났다.

1) 차량의 거동관련 요소 분석 결과

차량의 거동 관련 요소인 충돌 전 차량의 움직임을 차량1(자율주행자동차)과 차량2(일반 차량)으로 구분하여 분석하였다. 분석결과 자율주행자동차가 정지했을 때, 일반 차량의 다양한 움직임으로 인해 발생하는 교통사고의 빈도가 높았다. 그 중에서도 자율주행자동차가 정지한 상황에서 일반 차량이 직진 주행하며 추돌한 사고의 빈도가 높았다. 그리고 자율주행자동차가 직진 주행 중에 후방의 일반 차량도 직진 주행하여 추돌한 사고의 빈도가 높은 것으로 분석되었다. 그 다음으로는 자율주행자동차가 직진 주행 중에 일반 차량의 차로 변경으로 인해 측면을 부딪치는 교통사고의 빈도가 높았다. 다음 <Table 3>은 차량1과 차량2의 충돌 전 움직임을 나타낸다.

<Table 3> Results of Vehicle's Movements Preceding Collision

Vehicle1 (Autonomous Vehicle)	Vehicle2 (Normal Vehicle)	Frequency
Stopped	Proceeding Straight	16
	Making Right Turn	4
	Making Left Turn	2
	Passing Other Vehicle	2
	Changing Lanes	2
	Slowing / Stopping	2
	Entering Traffic	1
	Backing	1
	Xing Into Opposite Lane	1
Slowing / Stopping	Changing Lanes	1
	Other Unsafe Turning	1
	Proceeding Straight	1
Proceeding Straight	Proceeding Straight	8
	Changing Lanes	7
Making Right Turn	Making Right Turn	3
	Entering Traffic	1
	Other	1
Making Left Turn	Making Right Turn	2
	Proceeding Straight	1
	Stopped	1
Changing Lanes	Changing Lanes	1
	Entering Traffic	1
Passing Other Vehicle	Stopped	1
Merging	Proceeding Straight	1
Total		62

다음으로 교통사고에 관여한 객체에 대한 분석결과 차량이 56건이며, 트럭, 버스, 스쿠터, 전기 스쿠터, 스케이트보드, 자전거가 각각 1건씩 나타났다. <Table 4>는 교통사고에 관여한 객체를 나타낸 표이다.

<Table 4> Results of Objects Involved in the Accident

Objects Involved in the Accident	Frequency
Vehicle	56
Scooter	1
Bicyclist	1
Electric Scooter	1
Skateboarders	1
Truck	1
Bus	1

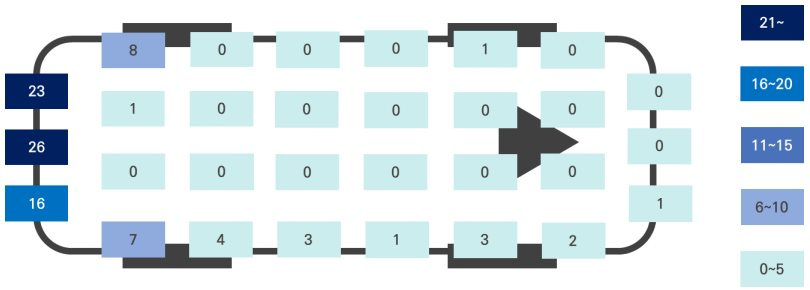
2) 충돌 및 손상 관련 요소 분석 결과

충돌 및 손상 관련 요소로 충돌 종류와 충돌로 인해 손상된 위치를 분석하였다. 분석결과 충돌 종류는 추돌(rear end) 형태의 사고가 49건으로 가장 높았으며, 그 다음으로는 측면 스침(side swipe), 측면 충돌(broad side), 정면 충돌(head-on) 순으로 나타났다. 다음 <Table 5>는 충돌의 종류를 나타낸다.

<Table 5> Results of Type of Collision

Type of Collision	Frequency
Head-on	1
Side Swipe	8
Rear End	49
Broad Side	4
Hit Object	0
Overturned	0
Vehicle/Pedestrian	0

충돌로 인한 손상 위치를 분석결과, 추돌로 인한 후방 범퍼의 손상 빈도가 높았으며, 그 다음으로는 좌측 및 우측 후방 타이어 부근에 손상 빈도가 높았다. 또한, 우측 측면이 좌측 측면에 비해 손상 빈도가 높은 것으로 분석되었다. <Fig. 2>는 손상된 위치를 시각화한 그림이다.



<Fig. 2> Frequency of Damaged Area

3) 도로 및 환경 요소 분석 결과

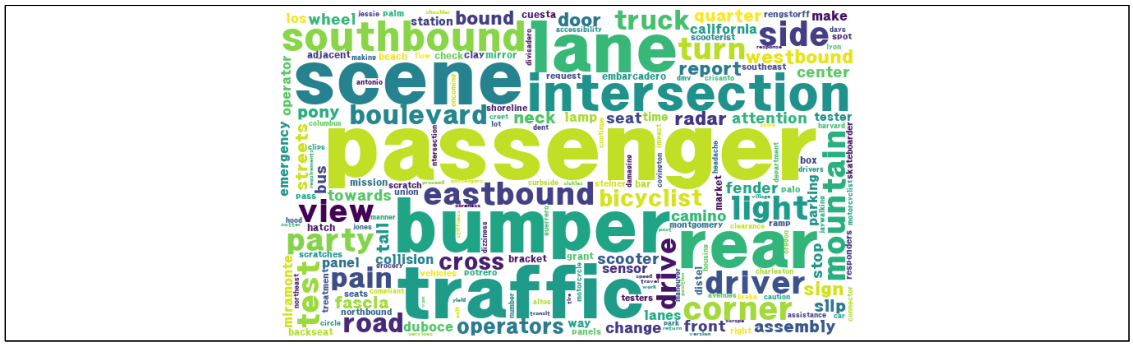
사고 발생시의 도로 환경 요소 분석을 위해 노면 상태와 도로 상태를 분석하였다. 노면 상태는 사고 당시 노면의 상태를 나타내며, 건조(dry)가 58건으로 대부분을 차지했으며, 습윤(wet)은 3건으로 분석되었다. 노면 상태가 기록되지 않은 사고가 1건이 발생하였다. 도로 상태는 사고 당시 도로에서 발생 되었던 도로 차단, 공사, 침수 등의 이벤트를 나타내며, 교통사고 중 1건은 공사구간에서 발생하였으며, 61건의 교통사고는 특별한 이벤트가 없는 일반적인 상황에서 발생하였다.

또한, 환경 요소로는 교통사고 발생 당시의 날씨와 조도를 분석하였으며, 날씨에 따른 교통사고 빈도는 맑음은 55건, 흐림은 6건으로 분석되었다. 안개의 경우도 1건으로 분석되었다. 조도의 경우 일광, 새벽, 밤을 나타내며, 밤 시간의 경우 가로등이 있는 경우, 가로등이 없는 경우, 가로등이 있지만 작동하지 않는 경우로 세분화된다. 이 중 교통사고는 일광(daylight)상태에서 41건 발생하였으며, 새벽(dusk-dawn)은 4건, 가로등이 있는 밤 시간에 17건이 발생한 것으로 분석되었다.

2. 텍스트 임베딩을 이용한 자율주행자동차 교통사고 분석 결과

1) 주요 키워드 도출 결과

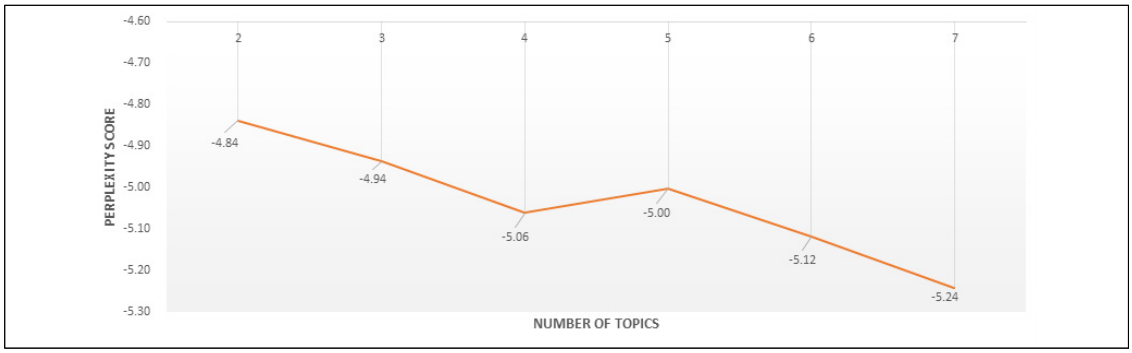
다음 <Fig. 3>과 같이 자율주행자동차 교통사고의 주요 키워드 도출 결과, 승용차(passenger car), 범퍼(bumper), 교통(traffic), 차로(lane), 후방(rear), 교차로(intersection) 등이 주로 도출된 것을 확인하였다. 이를 통해 자율주행자동차 교통사고에 관여하는 주요 객체가 승용차임을 유추할 수 있고, 주요 충돌 부위는 범퍼, 주요 상충은 추돌, 주요 사고 위치는 교차로임을 유추할 수 있다. 이는 앞서 분석한 통계 분석 결과와 유사한 결과를 도출함을 확인하였다.



<Fig. 3> Wordcloud Result of Keywords

2) 최적 토픽 개수 결정

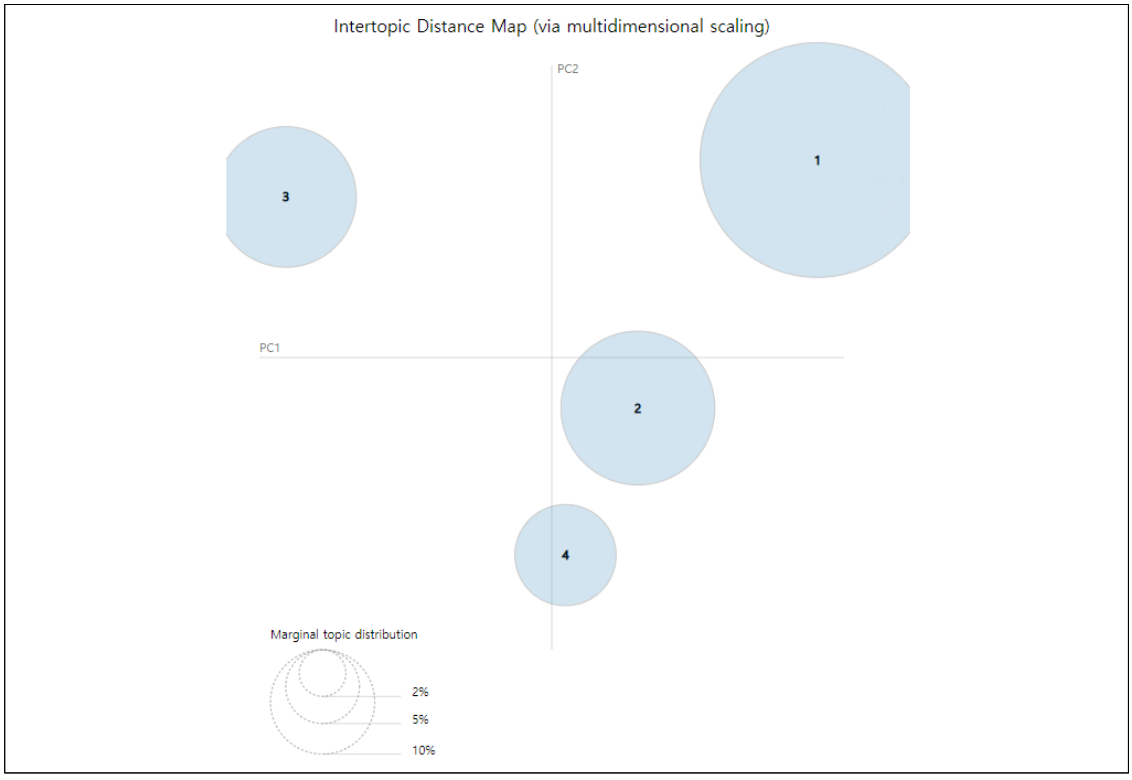
비정형 데이터인 텍스트 데이터로부터 자율주행자동차의 교통사고 유형을 구분하기 위해 본 연구에서는 토픽 모델링인 기법인 LDA 알고리즘을 이용하였다. 토픽 모델링을 수행하기 전 최적 토픽수를 산출하는 것이 필요하며 perplexity 평가 지표를 이용하여 산출하였다. 토픽의 개수를 늘릴수록 perplexity 평가 지표의 값은 감소하는 경향을 보이며, 지표의 값이 작을수록 토픽 모델의 결과가 우수하다고 볼 수 있다. 특히, 본 연구에서 검토한 토픽의 수는 2개에서 7개의 토픽을 검토하였으며, 토픽의 개수가 4개인 경우 -5.06으로 적정한 토픽 수로 분석되었다. 토픽의 개수가 7개 인경우가 perplexity 값이 가장 작으나, 토픽이 분류되지 않고 겹치는 경우가 발생하여 최적 토픽에서 제외하였다. 다음 <Fig. 4>는 perplexity 점수를 토픽별로 나타낸다.



<Fig. 4> Perplexity Score by Number of Topics

3) 토픽 모델링 수행 결과

앞서 도출된 4개의 최적 토픽의 개수를 이용하여 토픽 모델링을 수행하였다. 토픽 모델링 수행 결과 토픽 간 거리가 떨어져 있고, 모든 토픽이 겹치지 않고 독립적으로 분석된 것을 확인하였다. 이는 토픽들 간의 주제가 명확하게 분리되었다고 해석이 가능하다. 다음 <Fig. 5>는 토픽 모델링의 결과를 시각화한 그림이다.



<Fig. 5> Result of Topic Modelling

토픽 모델링을 통해 도출된 4개의 토픽별로 상위 5개의 키워드를 도출하였다. 토픽 1의 경우 승용차, 교통, 교차로, 범퍼, 후방의 단어가 도출되었으며, 토픽 2의 경우 코너, 회전, 후방, 자전거가 도출되었다. 토픽 3의 경우 측면, 문, 거울, 스쿠터, 레이더가 도출되었으며, 토픽 4의 경우 버스, 트럭, 교차로, 테스터가 도출되었다. 다음 <Table 6>은 토픽별 주요 단어를 나타낸다.

<Table 6> 5 Primary Keyword by Topic

Topic	Primary Keyword
1	passenger car, traffic, intersection, bumper, rear
2	corner, turn, rear, bumper, bicyclist
3	side, door, mirror, scooter, radar
4	bus, truck, intersection, tester

4) 주요 상황 도출 결과

각 토픽에 해당하는 교통사고를 분석하여 대표적인 주요 상황을 <Table 7>과 같이 도출하였다. 주요 상황을 도출한 결과, 토픽 1의 경우 주로 자율주행자동차가 주행 중 다른 차량이 후방 범퍼에 충돌하는 상황을 포함하는 토픽으로 분석되었다. 토픽 2의 경우 자율주행자동차가 정지하여 있을 때, 다른 차량이 후방 범퍼에 충돌하는 상황을 주요 상황으로 포함하는 토픽으로 분석되었다. 토픽 3의 경우 자율주행자동차가 정지하여 있을 때, 측면에 충돌하는 접촉사고가 발생하는 상황을 주요 상황으로 포함하는 토픽이었다. 토픽 4의 경우 버스 및 트럭과 자율주행자동차가 충돌한 상황과 기타 사항을 포함하는 토픽으로 분류되었다. 주요 상황 분석결과 수집된 DMV 자율주행자동차 교통사고의 주요 원인은 타 차량에 의한 후방 및 측면 충돌로 분석되었다.

<Table 7> Primary Situation by Topics

Topic	Number of Data by Topic	Primary Situation
1	17	A situation in which another vehicle made a contact with rear bumper while an autonomous vehicle is driving
2	27	A situation in which another vehicle made a contact with rear bumper while an autonomous vehicle is stopped/stopping
3	11	A situation in which another vehicle made a contact with side while an autonomous vehicle is stopped/stopping
4	7	A situation in which truck and bus made a contact with autonomous vehicle

Ⅴ. 결 론

1. 결론

최근 전 세계적으로 자율주행자동차와 관련된 기술 개발과 연구가 증가 되고 있다. 이에 자율주행자동차 교통사고의 특성 파악을 통해 자율주행자동차의 개발을 지원하는 것이 필요하다. 본 연구에서는 미국 캘리포니아 자동차 차량 관리국에서 수집한 자율주행자동차의 교통사고 데이터를 텍스트 임베딩 기법을 이용하여 분석하였다. 우선 수집된 자료의 통계분석을 수행하여 사고 특성을 분석하였다. 또한, '사고 상세 설명' 항목을 텍스트 임베딩 기법을 이용하여 분석하였다. '사고 상세 설명' 데이터의 전처리를 통해 텍스트 임베딩 기법에 적용 가능한 형태로 변환하여 주요한 키워드를 추출하였다. 도출된 키워드로는 승용차(passenger car), 범퍼(bumper), 교통(traffic), 차로(lane), 후방(rear), 교차로(intersection) 등이 높은 빈도로 추출되었으며, 이를 통해 자율주행자동차 교통사고의 주요 객체, 주요 충돌 위치 및 주요 발생 장소 등을 파악하는 것이 가능하였다. 다음으로 텍스트 임베딩 기법 중 많이 사용되는 토픽 모델링 기법을 이용하여 4개의 토픽을 도출하고, 토픽별 주요 상황을 도출하였다. 도출된 주요 상황들은 자율주행자동차가 다른 차량에 의해 추돌 또는 측면 사고를 당하는 상황이 주를 이루고 있음이 발견할 수 있었다. 연구 수행결과 통해 개발된 방법론은 텍스트로 구성된 자율주행자동차 교통사고 분석 및 주요 상황을 도출하는 용도로 사용될 수 있음을 확인하였다. 또한 자율주행자동차의 교통사고를 감소시키기 위해서는 자율주행자동차 전방상황에 대한 고려뿐만 아니라 자율주행자동차 개발시 후방 및 측면에서 발생하는 충돌상황에 대한 고려도 필요할 것으로 판단된다.

2. 향후 연구과제

본 연구는 자율주행자동차 교통사고 데이터와 텍스트 임베딩 기법을 이용하여 자율주행자동차 교통사고 분석 방법론과 자율주행자동차 교통사고를 분석하였으나, 몇 가지 연구의 한계가 존재한다. 우선, 자율주행자동차 교통사고 데이터의 수가 부족하여 텍스트 임베딩 기법에서 추가적인 정보가 도출되지 못한 점이다. 추후 자율주행자동차 교통사고 데이터가 충분히 확보된다면, 빅데이터 기법인 텍스트 임베딩 기법을 통해 추가적인 정보를 도출할 수 있을 것으로 판단되며, 사고의 유형도 다양하게 구분될 수 있을 것으로 판단된다. 이를 위해서는 자율주행자동차가 포함된 교통사고 데이터가 충분히 수집되어야 할 것으로 판단된다.

두 번째로, 수집한 자율주행자동차 교통사고 데이터의 '사고 상세 설명' 항목이 자세하지 않다는 한계가 있다. 텍스트 임베딩 기법을 이용하여 유의한 정보를 추출하기 위해서는 '사고 상세 설명' 부분에 사고의 위치, 객체의 종류, 상충의 종류, 사고 원인 등을 상세히 기록할 필요가 있다. 특히, 자율주행자동차가 원인이 된 경우에는 인지-판단-제어 과정에서 어떤 과정에 문제가 있었는지 추가할 필요가 있다고 판단된다. 추후 이런 추가적인 정보가 반영된다면, 텍스트 임베딩 기법을 이용하여 교통사고 원인 및 주요 상황 등을 추출하는 등 사고 분석에 용이할 것으로 판단된다.

세 번째로, 자율주행자동차 교통사고 데이터에 사고 발생 위치의 기하구조 정보, 교통사고 개요도, 주변 교통 상황 등이 포함되지 않아 교통사고 분석에 용이하지 않은 한계가 있다. 자율주행자동차 교통사고 데이터에 기하구조 정보, 교통사고 개요도, 주변 교통 상황 등이 포함된다면 자율주행자동차 교통사고가 많이 발생하는 기하구조 및 주변 교통상황 도출 등이 가능하여 자율주행자동차 개발 및 인프라 개발 지원 등에 활용될 수 있을 것으로 판단된다.

마지막으로, 자율주행자동차 교통사고 데이터를 이용하여 자율주행자동차 개발 지원에 활용하기 위해서는 기존 사고데이터에 추가적으로 자율주행자동차의 자율주행 수준 및 기하구조 정보, 교통사고 개요도, 주변 교통 상황 정보, 사고 직전 영상 등 다양한 정보를 함께 관리할 필요가 있다. 다양한 정보가 함께 관리된다면, 자율주행자동차 교통사고를 정밀하게 분석할 수 있을 뿐만 아니라 이를 통한 기술 오류 보완과 같은 자율주행자동차 개발 지원 등에 활용할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENTS

본 연구는 국토교통부 도심도로 자율협력주행 안전·인프라 연구 사업의 연구비지원(과제번호 20PQOW-B152473-02)과 2020년도 정부(교육부) 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2020R1I1A1A01072166)의 연구이다.

REFERENCES

Baek S.(2018), *Exploration on utilization of word embedding for topic modeling in Korean data*, Master's Thesis, The Graduate School of Seoul National University.

Blei D. M.(2012), "Probabilistic Topic Models," *Communications of the ACM*, vol. 55, no. 4, pp.77-84.

Blei D. M., Ng A. Y. and Jordan M. I.(2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp.993-1022.

Chae S.(2019), *A Study of Text Embedding for Korean Sentiment Analysis*, Master's Thesis, University of Seoul.

Cho A., Lee K. H. and Cho W. S.(2015), "Latent Mobility Pattern Analysis of Bus Passenger with LDA," *Journal of Korean Data & Information Science Society*, vol. 26, no. 5, pp.1061-1069.

Favarò M. F., Nader N., Eurich O. S., Tripp M. and Varadaraju N.(2017), "Examining accident reports involving autonomous vehicles in California," *PLos ONE*, vol. 12, no. 9, e0184952.

Lai S., Liu K., He S. and Zhao J.(2016), "How to Generate a Good Word Embedding," *IEEE Intelligent Systems*, vol. 31, no. 6, pp.5-14.

Oh C., Lee Y. and Ko M.(2016), "Establishment of ITS Policy Issues Investigation Method in the Road Section applied Text mining," *The Journal of the Korea Institute of Intelligent Transport Systems*, vol. 15, no. 6, pp.10-23.

Park J. and Lee S.(2020), "Big Data Analysis of Busan Civil Affairs Using the LDA Topic Modeling Technique," *Information Policy*, vol. 27, no. 2, pp.66-83.

Park S., Ko H., So J., Wee J. and Yun I.(2018), "Study of Test Scenario for Safety Evaluation of Automated Vehicle(Case of the Community Road in K-City)," *Proceeding of 2018 Korea Institute of Intelligent Transport Systems*, pp.331-334.

Park S., So J., Ko H., Jeong H. and Yun I.(2019), "Development of Safety Evaluation Scenarios for Autonomous Vehicle Tests Using 5-Layer Format(Case of the Community Road)," *The Journal of the Korea Institute of Intelligent Transport Systems*, vol. 18, no. 2, pp.114-128.

Petrović D., Mijailović R. and Pešić D.(2020), "Traffic Accidents with Autonomous Vehicles: Type of Collisions, Manoeuvres and Errors of Conventional Vehicles' Drivers," *Transport Research Procedia*, vol. 45, pp.161-168.

Ryu H.(2019), "Falling Accidents Analysis in Construction Sites by Using Topic Modeling," *Journal of the Korea Convergence Society*, vol. 10, no. 7, pp.175-182.

Sun L. and Yin Y.(2017), "Discovering themes and trends in transportation research using topic modeling," *Transport Research Part C: Emerging*, vol. 77, pp.49-66.

Woo C. W. and Lee J. Y.(2020), "Investigation of Research Topic and Trends of National ICT Research-Development Using the LDA Model," *Journal of the Korea Convergence Society*, vol. 11, no. 7, pp.9-18.