Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

NJIT
New Jersey's Science & Technology University

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

INCIDENT DURATION TIME PREDICTION USING A SUPERVISED TOPIC MODELING METHOD

by
Jihyun Park

Precisely predicting the duration time of an incident is one of the most prominent components to implement proactive management strategies for traffic congestions caused by an incident. This thesis presents a novel method to predict incident duration time in a timely manner by using an emerging supervised topic modeling method. Based on Natural Language Processing (NLP) techniques, this thesis performs semantic text analyses with text-based incident dataset to train the model. The model is trained with actual 1,466 incident records collected by Korea Expressway Corporation from 2016-2019 by applying a Labeled Latent Dirichlet Allocation(L-LDA) approach. For the training, this thesis divides the incident duration times into two groups: shorter than 2-hour and longer than 2-hour, based on the MUTCD incident management guideline. The model is tested with randomly selected incident records that have not been used for the training. The results demonstrate that the overall prediction accuracies are approximately 74% and 82% for the incidents shorter and longer than 2-hour, respectively.

INCIDENT DURATION TIME PREDICTION USING A SUPERVISED TOPIC MODELING METHOD

by
Jihyun Park

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Transportation

John A. Reif, Jr. Department of Civil and Environment Engineering

December 2020

Blank Page

APPROVAL PAGE

INCIDENT DURATION TIME PREDICTION USING A SUPERVISED TOPIC MODELING METHOD

Jihyun Park

Dr. Joyoung Lee, Thesis Advisor Date
Associate Professor of Civil and Environmental Engineering, NJIT

Dr. Grace Wang, Committee Member Date
Professor of Computer Science, NJIT

Dr. Branislav Dimitrijevic, Committee Member Date
Assistant Professor of Civil and Environmental Engineering, NJIT

BIOGRAPHICAL SKETCH

Author:	Jihyun Park
Degree:	Master of Science
Date:	Dec 2020

Undergraduate and Graduate Education:

- Master of Science in Transportation Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2021

- Bachelor of Science in Transportation Engineering,
  Hanyang University, Ansan, South Korea, 2000

Major:	Transportation

Recent Publications and Presentations:

J. Park, J. Lee, and B. Dimitrijevic. Incident Duration Time Prediction Using Supervised Topic Modelling Method. *Transportation Research Record.* Under Review.

J. Park, J. Lee, and B. Dimitrijevic. Incident Duration Time Prediction Using Supervised Topic Modelling Method. *Transportation Research Board Annual Meeting 2021.* Accepted for Presentation.

I. Yang, W. Jeon, J. Lee, and J. Park. (2019) Development of an Integrated Traffic Object Detection Framework for Traffic Data Collection, *The Journal of The Korea Institute of Intelligent Transport System* Vol.18 No.6, pp 191~201

Y. Park, S. Park, S. Lee, J. Park, and O. Kwon. The Impact Evaluation of Speed Reduction Facility in Expressway Using Drone, *Proceeding of 2019 Korean Institute of Intelligent Transport System Conference*, pp 406~411

At the end of a small chapter in my life

with brilliant memories

of my family, friends, and COVID-19

ACKNOWLEDGMENT

First and foremost, praises and thanks to the God, give me such a good opportunity and successful master course.

I would like to thank my advisor, Dr. Joyoung Lee, for his devoted guidance, invaluable advice, and immense knowledge throughout my research. Without his support as a supervisor, this thesis would not have been possible. Also, I express to thank Korea Expressway Corporation for allowing and supporting me to study at NJIT. Specially thanks to the Traffic Management Department and Traffic Information Center colleagues who provide the data used for this thesis. It is a great honor for me to participate in the National Science Foundation project as a researcher.

I would also like to express my sincere appreciation to the rest of my thesis committee, Dr. Grace Wang in the Department of Computer Science, and Dr. Branislav Dimitrijevic in the Department of Civil and Environmental Engineering at NJIT.

Last but not least, I am extremely grateful to my wife Minjung, my son Sugwang, and daughter Sooah for their love, understanding, and continuing support to complete this research course.

TABLE OF CONTENTS

Chapter **Page**

1 INTRODUCTION………….…………………………………………….……….…….. 1

1.1 Background…………………………………………………………….…...... 1

1.2 Problem Statement……………………………………………………………. 4

1.3 Motivation…………………………………………………………………… 6

1.4 Goal and Objectives…………………………………………………………….. 7

1.5 Organization…………………………………………………………….…..... 8

2 LITERATURE REVIEW……………………………………………………………….. 9

2.1 Regression Approach for IDT Prediction…………………………………… 9

2.2 Classification Approach for IDT Prediction…………………………………. 12

3 TOPIC MODELING AND APPLICATION………………………………………. 16

4 METHODOLOGY……………………………………………………………….... 22

4.1 Labeled Latent Dirichlet Allocation (L-LDA) ……………………………… 22

4.2 Dataset………………………………………………………………………...... 22

4.2.1 Dataset Overview…………………………………………………… 22

4.2.2 Dataset Description…………………………………………………. 28

4.2.3 Data Preprocessing………………………………………………….. 31

4.3 L-LDA Based Topic Modeling for Incident Duration Time Prediction………. 33

4.3.1 Document Provision…………………………………………………. 33

4.3.2 Determination of Topic……………………………………………..… 34

4.3.3 Model Calibration…………………………………………………... 35

LIST OF TABLES

TABLE OF CONTENTS (Continued)

Chapter **Page**

5 RESULT…………………………………………………………………...…… 37

5.1 Analysis Result ……………………………………………………………. 37

5.2 Model Comparison ……………………………………………………….. 42

6 CONCLUSIONS AND RECOMMENDATIONS……………………………….. 45

6.1 Conclusions ………………………………………………………………. 45

6.2 Recommendations ………………………………………………………… 46

6.3 Future Research …………………………………………………………… 48

REFERENCES ………………………………………………………………… 50

Table **Page**

4.1 Dataset Description………………………………………………………….......... 29

4.2 Level of Each Experimental Factor for the Calibrations……………………... 36

5.1 Performance Comparison…………………………………………………….. 43

5.2 Performance Comparison of Other Research Result…………………………. 43

LIST OF FIGURES

Figure **Page**

1.1 Cause of congestion in Germany....... 1

1.2 Sources of congestion in the U.S....... 2

1.3 Reasons for congestion of the No 1 Highway in South Korea……………… 2

1.4 Composition of incident duration…………………………………………… 3

1.5 Examples of distribution of incident dataset……………………………... 5

4.1 Graphical model of Latent Dirichlet Allocation (LDA)……………………… 23

4.2 Graphical model of Labeled Latent Dirichlet Allocation (L-LDA)…………... 24

4.3 OneClick system's main interface ……………………….…………….… 26

4.4 Example of real time incident report of OneClick system………………….. 27

4.5 Incident duration time distribution ……...………………………………...... 29

4.6 Relationships between incident duration times and variables ……………… 30

4.7 The sample case of incident report ……………...……………...…………... 33

5.1 Distribution of probabilities for IDT (top: IDT less than 2-hour; bottom: IDT greater than 2-hour) …………………………..……..…………………….. 38

5.2 Cumulative distribution function (top: IDT less than 2-hour; bottom: IDT greater than 2-hour)……………………………………………………………. 39

5.3 Example of match accuracies for each topic (top: IDT less than 2-hour; bottom: IDT greater than 2-hour) ……..………..…………………………… 40

5.4 Overall matching accuracy …………………………………………………. 42

6.1 Example of segmented incident duration time …….……………..…………... 47

LIST OF ACRONYMS

Acronym	Definition
AFT	Accelerated Failure Time
ANN	Artificial Neural Networks
BN	Bayesian Network
BNN	Bayesian Neural Network
BoW	Bag of Word
CART	Classification and Regression Tree
DT	Decision Tree
FHWA	Federal Highway Administration
IDT	Incident Duration Time
ITS	Intelligent Transportation System
KEC	Korea Expressway Corporation
KNN	K-Nearest Neighbor
LAFT	Log-logistic Accelerated Failure Time
LDA	Latent Dirichlet Allocation
L-LDA	Labeled Latent Dirichlet Allocation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NLP	Natural Language Processing
RMSE	Root Mean Square Deviation
SNS	Social Networking Service
STM	Supervised Topic Modeling
SVM	Support Vector Machine
TMC	Traffic Management Center

CHAPTER 1

INTRODUCTION

1.1 Background

Traffic congestion can be classified into recurrent congestion and non-recurrent congestion. Repeatedly occurring in rush hours, the recurrent congestion is primarily caused by the inadequate base capacity of the roadway when demand is at its highest. Non-recurrent congestion is resulted from unexpected incidents such as collision, vehicle break-down, emergency lane closure, fire, etc. [1]. The impact of incident on traffic congestions is significant. Friedrich [2] showed that about 44% of traffic congestions in Germany is caused by incidents (Figure 1.1). He also suggested traffic demand, weather, and traffic incidents as three significant factors of traffic congestion, which affect the reliability of travel time.

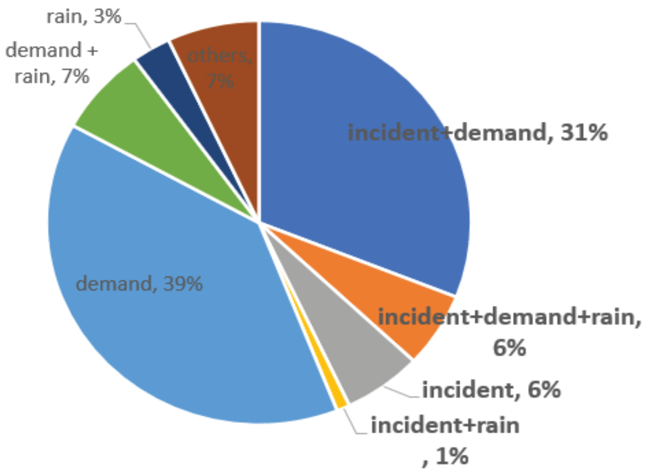


Figure 1.1 Cause of congestion in Germany.

In the United States, incidents account for approximately 25% of total congestions across the country, the second largest portion (Figure 1.2). Furthermore, as incidents create unexpected delays, it is a major source of frustration for road users [3].

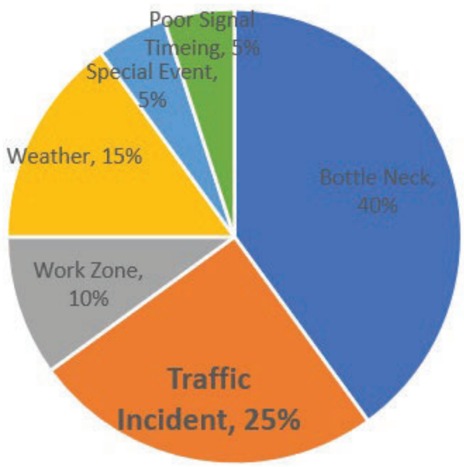


Figure 1.2 Sources of congestion in the U.S.

The Korea Expressway Corp. analyzed 8,306 congestion cases that occurred on Highway No. 1(Kyungbu line) in South Korea (Figure 1.3). It shows that incident is the most influential factor except for traffic volume increasing.

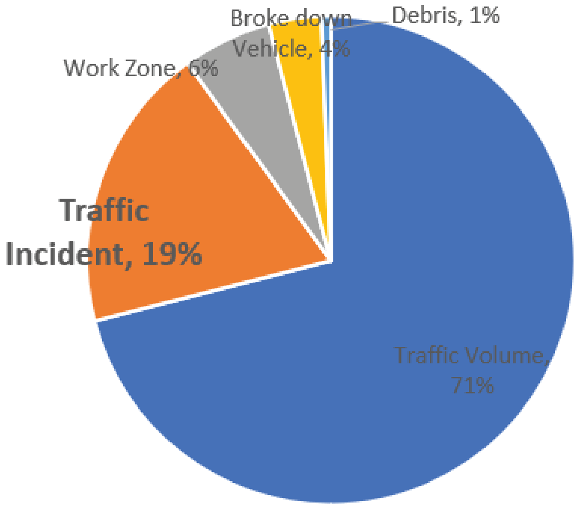


Figure 1.3 Reasons for congestion of the No. 1 highway in South Korea.

One of the most common congestion management strategies dealing with incident in the state-of-the-practice would be to provide road users with detour information to bypass incident scenes. Diverting upstream traffics to alternative routes, the detouring strategy enables not only drivers to take safer routes but also to improve the safety of incident management crews by reducing traffics passing by the scene. Since an incident in nature is unpredictable, irregular, and unrepeated, however, incident-caused congestions often result in undesirable performances of traffic congestion management strategies.

FHWA [3] suggests that transportation agencies need to estimate the precise duration time of an incident to provide accurate information to road users to improve the effectiveness of traffic congestion management. The duration time of an incident consists of the following three elements, as depicted in Figure 1.4 [4]: 1) detection time; 2) response time; and 3) clearance time. It is noted that recovery time in Figure 1 is not included as a part of incident duration time (IDT) by definition. The detection time is the time gap between the incident occurrence and the moment of incident verification by Traffic Management Center (TMC). The response time is from the verification time to the arrival time of incident dispatching units (e.g., SSP, Ambulance) at the incident location. The clearance time is the amount of time spent by the incident dispatching team to clean up the incident on site.

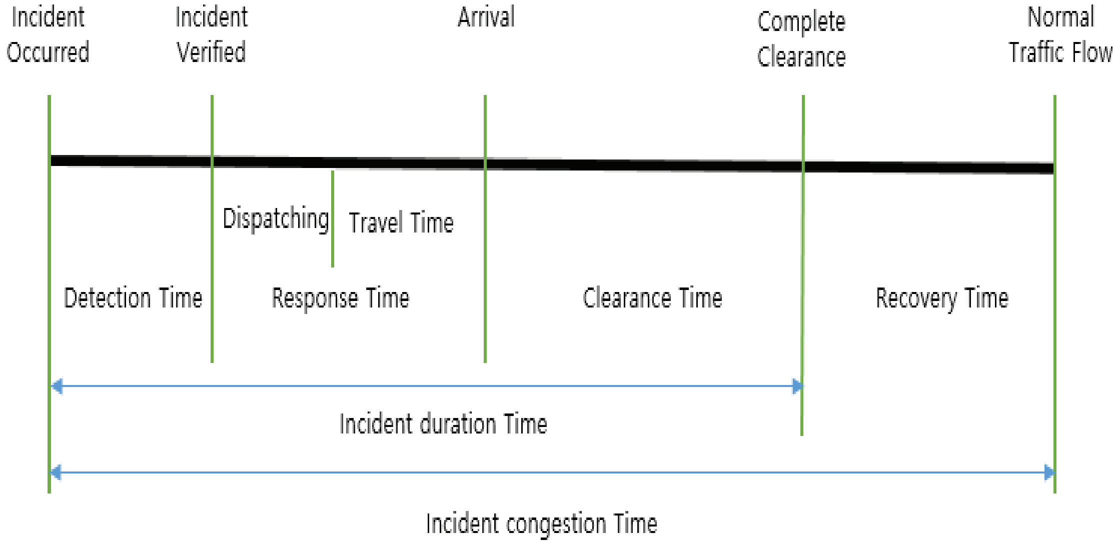


Figure 1.4 Composition of incident duration.

With recent drastic advances in personal mobile device technologies, social networking service (SNS), and traffic surveillance infrastructures, the information of an incident can be reported to TMC in near real-time, which enables TMC operators to rapidly react to the incident. Along with the incident information at which it is reported, if the TMC operators could previse the IDT, their efforts to mitigate the impacts of the incident would become more productive. In that sense, timely and precise prediction of IDT is essential to enable congestion mitigation strategies to be more accurate and reliable.

1.2 Problem Statements

Yet, precisely predicting IDT is a nontrivial task, especially when it is needed to be done instantly. While numerous relevant efforts to forecast IDT have been undertaken for the past decades, only a few of them gained noticeable attentions. Traditionally, IDT prediction models have relied on numerical dataset quantified from incident reports which verbally describe incident situations. In general, an incident report contains information combined by numerous words that vividly explains the incident. However, during the quantifying process to generate numerical dataset, some words in the report can be omitted, adjusted, over- or understated, and/or even garbled to be fitted with a given data format designed for numerical data type. Smith and Smith [5] addressed that the performance of IDT prediction is heavily affected by both the quality (e.g., data structure consistency, information omission) and the quantity (e.g., sample size, sample imbalance) of the data. Unrecorded or omitted information in the incident data are also of critical factor adversely affecting the performance [6]. In addition, inconsistency in data format across agencies who are responsible for collecting, archiving, and managing incident data worsens the performance of incident duration prediction models.

Specifically, duration data has unique features that the longer the duration time, the smaller the number of accident cases. Figure 1.5 shows that the data distribution of the incident duration has a right-skewed shape [5]. Due to the characteristics of left-biased data, the results of many studies show high accuracy in the short duration section where a large number of data exist, but a high error rate in the long duration section where the number of data is small [5, 16, 18, 19, 20].

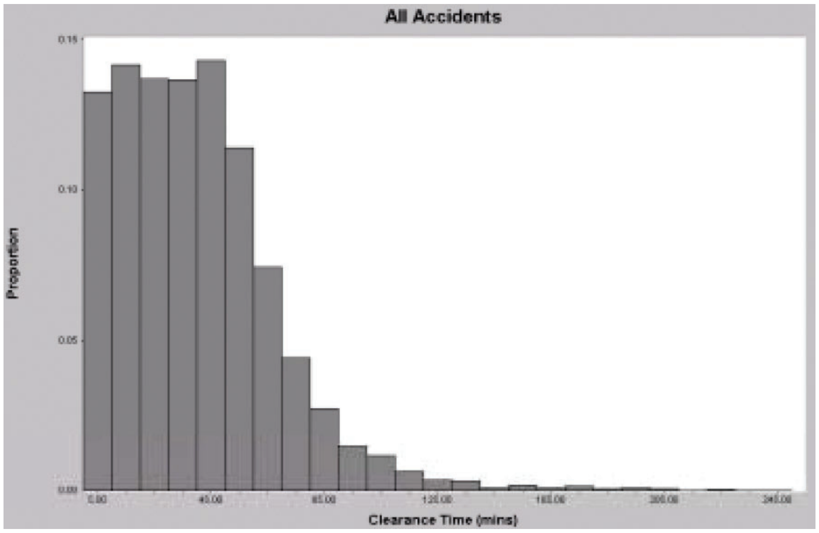


Figure 1.5 Example of distribution of incident dataset.

Source: Smith, K., and Smith, B. 2001. Forecasting the Clearance Time of Freeway Accidents. Center for Transportation Studies, University of Virginia

As an aspect of the practical implementation of the model, duration time should be supported to the operator in TMC who is responsible for impact mitigation in the early stage of incident. To satisfy this assumption, Input data that can collect on time will be utilized from the incident report. Some research suggested that the reaction of incident is a critical contributing factor to estimation, for example, police officer arriving time. [6] Though It is a critical factor to estimate, does not efficient for the operator who decides incident management plan as that information cannot reach in the initial stage of incident.

1.3 Motivation

By using text data, it is possible to overcome the limitations of numerical data. Numerical data has a disadvantage that the case cannot be used for analysis when the missing data occur. Since text analysis focuses on hidden meaning in documents, it is possible to overcome this shortcoming. It is also advantageous for long-duration time analysis, which has small incident cases. Incidents with complex and various factors combined contain more information in the incident report because the text description increases. Therefore, the text analysis method is more effective in terms of the use of traffic incident data.

Recently, machine learning techniques for text analysis has dramatically improved. Text mining, the process of deriving high-quality information from documents, is being used in various fields. Due to the increase in computer and software power, it is possible to analyze a large amount of data in a limited time to extract topics [7]. Recently, NLP-related research efforts using deep learning have been steadily increased [8]. The topic modeling is a part of Natural Language Processing (NLP) techniques which is widely used in computer science nowadays to retrieve topics in the documents. Applying this new technique to incident duration prediction may allow accuracy improvement.

The amount of text data that can be accessed and utilized has increased. In many transportation agencies, accident reports containing a brief summary stored in paper form in the past changed to electric data. Various analyzes and studies are being conducted using text data in these incident reports [35.36.37]. In addition, there is research that conducted data analysis using text information updated in the timestamp [38]. The Korea Expressway Corporation has operated the OneClick system that updates real-time text data of incidents since 2016. The data accumulated over three years is expected to facilitate text analysis.

1.4 Goal and Objectives

The primary goal of this thesis is to develop an accurate incident duration time estimation model that enable TMC's operators to decide to implement traffic detour in an early stage of the incident. The operator’s quick decision is efficient for traffic management by mitigating congestion and preventing secondary accidents. To this end, the following objectives are established:

- To apply suitable machine learning techniques for text analysis to retrieve the hidden meaning of the serious incident report data.
- To develop an accurate estimation model that is helpful for the operator’s decision.
- To develop real-time based model by using initial data which depict incident situation and circumstance.

This thesis presents a novel model for the prediction of incident duration time by employing an emerging supervised topic modeling (STM) method based on cutting-edge Natural Language Processing (NLP) techniques. Dealing with the huddles addressed in problem statements, the developed model performs semantic text analysis using actual incident records to prevent any potential information loss which might provide critical clues for IDT predictions. As an STM method, this thesis adopts Labeled Latent Dirichlet Allocation (L-LDA) [9].

1.5 Organization

This thesis is organized as follows: Chapter 2 covers the literature review with respect to analytical and classification approach for IDT Prediction. Chapter 3 discusses the evaluation of topic modeling and their application with the machine learning approach. The dataset introduction and methodology applying Labelled LDA are included in Chapter 4. Chapter 5 discusses the results in detail obtained from the analysis, and Chapter 6 provides concluding remarks for the future research study.

CHAPTER 2

LITERATURE REVIEW

2.1 Regression Approach for IDT Prediction

In the 1990s, estimating IDT gained an attraction with the introduction of ITS technologies. Khattak et al. [10] stated that the advanced traffic information system has great potentials in automatic incident detection and IDT prediction. The authors presented a duration time prediction method based on a linear regression modeling approach. The variables in the model are as follows; 1) incident characteristics (incident type, vehicle type, injuries and fatalities), 2) response and operational factors (number of rescue vehicles, other agencies assistance, usage of heavy wrecker), 3) environment conditions(weather, visibility), 4) location characteristics, 5) seasonal factors, 6) traffic flow, etc. In order to use reaction variables, this research divided duration time every 5 minutes from 10 to 30 minutes and develop regression models for every truncation. Using the model, the author revealed that the IDT becomes longer if 1) the response time is longer, 2) the incident information is not disseminated, 3) there are severe injuries, 4) trucks with heavy loading are involved, 5) state property is damaged, and 6) the weather is not normal. It is a meaningful trial that the time sequential prediction model was developed considering the factors that were added as time goes by for duration prediction after an incident occurrence.

Peeta et al. [11] developed multiple linear regression models for IDT prediction by using 22,737 incident cases assisted by freeway service patrol in Indiana state from July 1, 1996. to October 28, 1998. The prediction model was designed in part of the development of the driver response model under traffic information. The estimation was conducted for two types of incidents of road debris (1,176 cases) and collisions(835cases). The disablement type which a portion of 91% of the incident are not considered a prediction model. The independent variables are classified into four categories: severity (e.g., number of involved vehicles, fatality), location (e.g., ramp, mainline, merging/diverging), traffic condition level, environmental conditions (illumination, temperature, visibility). They found that the number of vehicles has the most critical impact on IDT, and the severity and environmental conditions also appear statistically significant. However, the overall performance of the models appears inadequate as two models' coefficients of determination (R2) are as low as 0.234 and 0.362, respectively.

Chung et al. [12] applied Log-logistic Accelerated Failure Time (LAFT) model to estimate IDT. LAFT is a kind of method of parametric survival analysis. Survival analysis is a technique that analyzes the occurrence of events by tracking from starting time to end time. It is commonly used in industrial engineering, social science, and medical treatment. In this research, the LAFT was applied to predict IDT with 2,940 incident cases collected by Korea Expressway Corp. in South Korea. The authors discovered that the followings give negative effects to incident duration time; Large truck, trailer, taxies, damage rate over 33%, fatalities, injured person, work zone area, number of vehicles, night time, and so on. Factors such as low damage rate, incident in shoulder location, overturned vehicle, incident reported by freeway service patrol are associated with short duration time. The result shows every variables' estimated time ratio and percentage changes, which can determine positive and negative factors for incident duration time.

Li, R. [13] also applied Accelerated Failure Time (AFT) model using 2,496 incident cases collected in the metropolitan area of Beijing, China, in 2008. In the thesis, duration time is estimated three divided stages: 1) preparation, 2) travel, 3) clearance. The models are developed in each stage to improve the accuracy of prediction by using AFT hazard-based model. The result also suggested critical factors for longer and shorter duration time. For example, in the preparation time stage, while summer and autumn seasons are factors of extending duration time, the winter season impacts reducing duration time. This research revealed many different variables are affecting incident duration time for each stage. However, the prediction model shows low prediction accuracy in extremely short or long incident durations (high MAPE in 1~15min and over 120min section).

Zong et al. [14] developed a hybrid model combining an ordered Probit model and an AFT model to predict IDT by using police-reported 3,914 incident records of Jilin province, China, in 2010. Severity prediction model by ordered Probit model estimates the number of fatalities, injuries, and property damage, and better prediction accuracy compared to the results of Support Vector Machines (SVM). AFT model applied estimation duration time accepts Weibull distribution for value prediction. A two-stage prediction (incident severity and then IDT) model achieved enhanced accuracy. This research reveals a meaningful conclusion that the factors of incident severity significantly affect the incident duration. In other words, accurate prediction of severity gives crucial information for estimation duration time.

Artificial Neural Network (ANN) also appeared as a popular data analysis method to solve complex problems with neurons between input and output data. Many researchers have been trying to apply it to IDT prediction. Guan et al [15] utilized an ANN model with 830 cases in Guangzhou, China. Eight crucial input factors, such as number of trucks, rollover vehicle involvement, fatality damage involvement, etc., are determined by extracting a strong correlation. 660 cases are trained, and 170 cases are used in the validation of the model. Though its performance shows acceptable result as the correlation coefficient of prediction and real data is 0.8535, it is undesirable as its lower value comparing with it of decision tree model and regression model. Two reasons are presented as low correlation coefficient value: small incident cases (830 cases) and hidden input factors not mentioned in the incident report.

Wu et al. [16] applied Support Vector Machine (SVM) approach, which has strength in analyzing both nonlinear and high-dimensional pattern data to predict IDT by using 1,853 incident cases collected in Utrecht, Netherland. Incidents are divided into three types, Breakdown, Lost-load, and Accident, and prediction models of three cases were established. Basic parameters such as incident type, vehicle type, number of vehicles are used. Specifically, the Yes/No information about the necessity of police, fire brigade, ambulance, tow truck, repair service, road management, fluid to be clean, etc. are included as basic parameters. Bayesian Decision Method-Based Tree Algorithm is compared with the result of the newly predicted value in three divisions of duration time (10~30min, 30~60min, 60~90min). The reliability of predicted duration time less than 60 minutes got high accuracy. However, the result of over 60 minutes is inferior.

2.2 Classification Approach for IDT Prediction

Various classification models based on big data analysis techniques are applied in estimating IDT to improve prediction accuracy. The most common models are: 1) Decision Tree (DT), 2) K-Nearest Neighbor (KNN), 3) Support Vector Machines (SVM), 4) Naive Bayesian Classifiers, and 5) Artificial Neural Networks (ANN) [17].

With the decision tree modeling approach, Chang et al. [18] attempted to improve the prediction accuracy of IDT based on the classification tree analysis. The strength of decision tree model is straightforward for users. The model consists of thirteen predictor variables (notification time, number of vehicles, truck involved, etc.) and three target variables (short, medium, long duration time), and twenty-seven cases are adopted to estimate. The authors employed the number of heavy vehicles and accident types (e.g., overturned, disabled vehicle, etc.) as the primary variables. Although this model produces a reasonable performance (i.e., 75% accuracy) overall, accuracies of mid and long duration appear as low as 17% and 11%, respectively. The researcher added some reasons of inaccuracy such as the imbalance of dataset, other indirect factors (upstream traffic, weather conditions), etc.

He et al. [19] developed a hybrid model integrating the quantile regression model and tree-structure model with 1,245 incident cases that occurred in the Bay area freeways, California. The two-step algorism is suggested. First, incidents are classified by incident type (collision, disabled vehicle, and collision), the presence of injury, number of vehicles, and so on, using the decision tree model. Then quantile regression model is applied to each final node. This model has merits simple interpretation and ease of managing covariates. Comparing to KNN and Classification and Regression Tree (CART), the authors showed the performance of their model with 84% of accuracy in the error range of 15minutes or less. The result also shows that incident characteristics (type, injuries, block lanes, etc.) are critical factors, and location matters such as geometry and jurisdiction also affect duration time.

Yang et al. [20] applied Bayesian Network (BN) model with Florida DOT's 2,629 incident cases collected from January 1, 2005. to October 31, 2006. In this research, duration time categorized three levels as FDOT’s criteria (under 30min as short level, 30min~2hour as median level, over 2hour as long level). Missing data are substituted with the edge value for useful duration prediction, as BM model can accommodate incomplete data and predict. Detailed accuracy is calculated 81.8% in short level, 54.2% median level, 49.8% in long level. Compared with a Logistic model (71.7%) and a decision tree model (72.5%), BN model showed higher overall accuracy (72.6%). BM model has strength by showing straightforward and quickly implementing results suggesting the probability values of every variable, which can classify incidents for different duration classes.

Park et al. [21] developed a hybrid algorithm combining a Bayesian Neural Network (BNN) and a decision tree model using 13,987 Maryland State Highway Administration cases. This algorithm combined the merits of that BNN model can find out critical factors, and the decision tree model is a straightforward method to understand and interpret easily. Duration time was categorized into four groups, under 30min, 30~60min, 60~90min, and over 90min. The result shows that the hybrid model got the lowest error rates (MAE 0.19, MAPE0.27, Proportion of underestimated values 10.2%) compared with Support Vector Machines, Back-propagation Neural Network, and Classification and Regression Tree models. This model could be helpful for operators not only to support higher accuracy than other models but to provide understandable information of duration prediction.

Although a wide variety of modeling approaches have been proposed with various incident cases as reviewed in this section, it is still unsure to determine a model working best. Yu et al. [22] compared the prediction performance of ANN and SVM model. The dataset is composed of 235 cases that took place on Dalian-Shenyang freeway from 2012 to 2014. The MAE, RMSE, MAPE are utilized to compare two model's accuracy, and duration time is categorized into five groups. The authors demonstrated that SVM outperformed ANN in cases of short and medium IDT while ANN appears better for long duration incidents.

Examining the performance of five different modeling approaches for IDT prediction: 1)linear regression, 2) DT, 3) ANN, 4) SVM, and 5) KNN, Valenti et al. [23] showed that the accuracy of each model depends on the IDT. For example, the linear regression approach is the lowest error rate in case of IDT less than 0-30min while ANN shows the best performance for incidents with over 90min of IDT. Similarly, Hamad et al. [24] compared five models: 1) regression tree, 2) SVM, 3) ensemble tree, 4) gaussian process regression, and 5) ANN. The authors discovered that the worst model is a regression tree model, and the SVM model got the lowest error rates. This research also mentioned in the conclusion that there is no best model suitable for all cases.

CHAPTER 3

TOPIC MODELING AND APPLICATION

The topic modeling is a part of Natural Language Processing (NLP) techniques that can identify and classify the subject of a document composed of a combination of words. Each group of words represents the topics that the document stands for. In general, a document deals with multiple topics while multiple documents address one same topic. If documents address the same topic, it is very likely that similar words commonly exist in the documents. The topic modeling technique is a method of deducing topics by semantically clustering words using context-related cues in documents. Several relevant efforts in the diverse fields of engineering have been made to apply the topic modeling methods for medical area [25], analyzing social networks [26], and safety [28, 31, 33, 35, 36, 37, 38].

Term frequency-inverse document frequency (TF-IDF) is a popular method to describe the topics of a document. This method compares essential words in a document and common words in overall documents [27]. This method is developed based on the assumption that the words frequently appeared in a document have a positive relationship with the topic. The frequent words in a document get a weight and the words which appear common across all documents are omitted by giving low weight.

$$x_{ki} = f_{ki} * \log(\frac{N}{n_i}) \tag{3.1}$$

Where,
$f_{ki} =$ frequency of word i in document k
N = the number of documents in the collection

$n_i =$ the number of documents where word i occurs

Goh et al. [28] applied TF-IDF topic modeling approaches to categorize construction accident with 16,323 construction incident reports occurring from 1983 to 2013. The research used ‘title’ and ‘narrative’ in the reports as a dataset for applying TF-IDF text mining techniques. Every incident narrative is tokenized to one word and two or three words combination. Through TF-IDF technology, the words which reflect incident characteristics are remained by removing generally used words in every document. The modified dataset is trained by matching 11 incident cases, and the classification is conducted using the following six machine learning algorithms and compared: 1) SVM, 2) linear regression, 3) random forest, 4) KNN, 5) DT, and 6) NB. The authors discovered that overall SVM outperformed other algorithms. TF-IDF has a critical flaw in that it does not consider the meaning of the word at all [29]. This is because the structure of TF-IDF is designed to focus on the frequency of words.

Deerwester et al. [30] introduced Latent Semantic Indexing (LSI) using Singular Value Decomposition (SVD) for extracting topic words. LSI conceptually assumes that co-occurrence words have the same topics. Based on this assumption, LSI classify the documents based on the frequency of co-occurrence of similar words. Therefore, it is an advanced method than the TF-IDF method, which considers only the frequency of words. SVD is applied in this model which is a calculation method that simplifies numerous computational processes of the input matrix. The method consists of 3 steps: first, creating the input matrix (sentence - term matrix); second, applying the SVD to the matrix; and finally, selecting the sentences for the summary.

Robinson et al. [31] analyzed the narratives of the Aviation Safety Reporting

System (ASRS) using LSI analysis to find significant causes of an aviation accident. A total of 4,497 sentences were trained in labeling primary problems and contributing factors, and a 2,987 data set was used for cross-validation. For topic modeling, 400 words were retained after performing TF-IDF and SVD sequentially, and these words matched 18 cause categories for training. As a result, this research got 44% of the accuracy for primary cause categorization.

Based on the LSI method, Hoffman, T. [32] proposed Probabilistic Latent Semantic Indexing (PLSI). While the LSI method is a model created by a document-word matrix based on the number of occurrences of a particular word in the documents, PLSI has a more comprehensive theoretical foundation accepting the probability of the occurrence of a specific word in a document. accepting the probability means that the document-term matrix is constructed not based on the number of times a specific term appears in the document but based on the probability of a specific word appearing in the document.

Zhao et al. [33] developed a fault diagnosis algorithm for vehicle on-board equipment (VOBE) for the high-speed train using the PLSI method. 1,148 fault cases are collected from 2011 to 2012 in Wuhan-Guangzhou high speed rail in China, in which 848 cases are trained, and 300 cases are used for evaluation. The words in the VOBE reports classified 12 topics using PLIS method and matched fault reasons categorized 2 degrees (7 fault locations and 15 fault reasons). The Bayesian network model is adopted for training as its accuracy is better than K2 algorism, K-Nearest Neighbor, and Back Propagation Artificial Neural Network. The result based on comparing type of input data shows that topic modeling got higher accuracy than the raw feather, containing all the words in the document, and the document frequency, which is one of the simplest word selection methods.

Latent Dirichlet Allocation (LDA) is a frequently used topic modeling algorithm developed by Blei et al. [34]. PLSI has the merit of a probabilistic model. It has, however, inherent problems as follows. First, as the number of words increases, the parameters in the model are growing enormously. Second, while the model's concept is the probability of word's appearance, it is hard to choose in a new document. The LDA algorithm considers the concept of exchangeability among words in a document to reduce computation. The LDA also makes it possible that new terms are automatically identified and ranked by an advanced probabilistic model.

Several efforts have been made to apply LDA in the transportation safety research area using text in the reports of transportation facilities such as airports, railroads, and expressways.

Tanguy et al. [35] applied the LDA model to extract topics from the 167,350 aviation safety reports in the Aviation Safety Reporting System (ASRS). Topics are set 10, 50, 100, 200 for classifying, and labeled an incident reason by experts for each topic. The author found that the number of topics from 50 to 100 have high relevant to experts' interpretation. This research also conducted identifying similar reports and plotting time series. The result could be utilized to find and investigate a series of specific incidents. The researcher suggested that incident signal detection machine learning algorithm using the aforementioned automatic reports classification technique. However, the result of the simulation shows unaffordable accuracy under 50%.

Brown [36] developed the prediction model to estimate the cost of extreme incidents. LDA and Partial Least Squares (PLS) model are used for text mining with eleven years' 42,033 cases in the U.S. The result shows that input data combined with text inform a better understanding of classification. Specifically, it is discovered that the PLS method has a lower root mean squared error (RMSE) than LDA applied by both Random Forest and Gradient Boosting model. It is meaningful that cost estimating is conducted by matching words in LDA topics and words in documents.

Williams and Betak [37] compared LSI and LDA models with 12,447 railroad incident reports obtained from the Federal Railroad Administration’s equipment accident database from 2010 to 2015. The comparison result of identified topics shows that both models have different strengths in each topic. Both techniques have clearly classified as shoving, grade crossing accident, wheels, and hump yards. LSI was effective in identifying accidents related to track maintenance equipment, liquids releasing. On the other hand, LDA was effective in discovering topics having braking accidents, derailment, and hazardous materials.

Notably, Pereira et al. [38] predicted IDT using two year's Singapore incident data. The research focused on real-time estimation, as new information arrived in a traffic information center, and offline analysis. The estimation model is developed based on the LDA algorithm. Five prediction models (Linear Regression, Support Vector Machine, Artificial Neural Network, Decision Tree, and Radial Basis Neural Network) are used to investigate LDA’s performance. As comparison shows median error downed from 16.6 min to 10.5min, the model with text analysis advanced accuracy compared without text analysis.

Many studies focused on the extraction or classification of keywords in the text report using the LDA model. It has an excellent contribution for automation, which reduces a lot of time and effort dealing with it manually in the past. Topic modeling may contribute to increased utilization of text-based report data. Moreover, topic modeling combined prediction models like Brown [36], and Pereira et al. [38] could be an advanced study in terms of expanding the research area and practical utilization.

The main reason for LDA's popularity is its extended function. LDA is not only widely used, but it is also extensively modified. Therefore, many researchers changed to supervised learning models such as supervised LDA [39] and DiskLDA [40]. One of the supervised machine learning models for topic modeling is Labeled-LDA, developed by Daniel Ramage et al. [9]

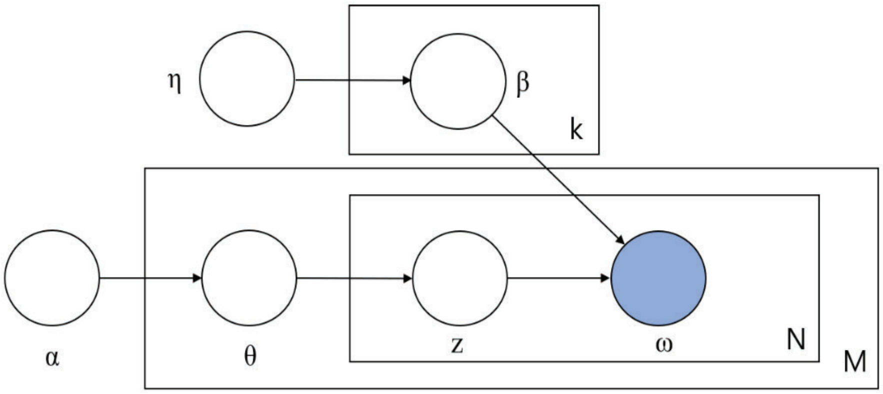
CHAPTER 4

METHODOLOGY

4.1 Labeled Latent Dirichlet Allocation (L-LDA)

As L-LDA is an extended concept of LDA, LDA should be introduced before dealing with L-LDA. LDA assumes that documents contain topics, and every topic contains words, denoted by k and ω, respectively. This assumption is conceived from the process of the document generation. For example, initiating a new document, a writer selects one topic, θ, which is proper for the document. After selecting a topic, the writer chooses one word appropriate for selected topic from multinomial distribution of β. These actions are continuously repeated until the document is completed. In other words, the document is the group of words which is the combination of the probability that a word will exist in a particular topic, and the probability that a particular topic will exist in a document. Repetition for multiple documents in this way enables to find a set of words that parameter θ and β having a multinomial distribution. [34]

Conversely, the LDA model is to infer θ and β from a set of words. For this, the calculation is performed to obtain Z, the number of topics that a selected word belongs to. The characteristics of conjugation between Dirichlet distribution (α, η) and multinomial distribution (θ, β) is used for calculation. Finally, θ and β are derived using the obtained Z. Through this process, the topic in the document can be identified by the distribution of words as shown in Figure 4.1. As of these merits utilizing relationship among topic, document and word, LDA is widely used for subject classification and calculating similarity between documents.



Where,

M: Number of documents

k: Number of topics

N: Number of words in the document

θ: Topic distribution in the document

β: Word distribution for topic k

Z: Number of the topic(topic distribution) which the word belongs

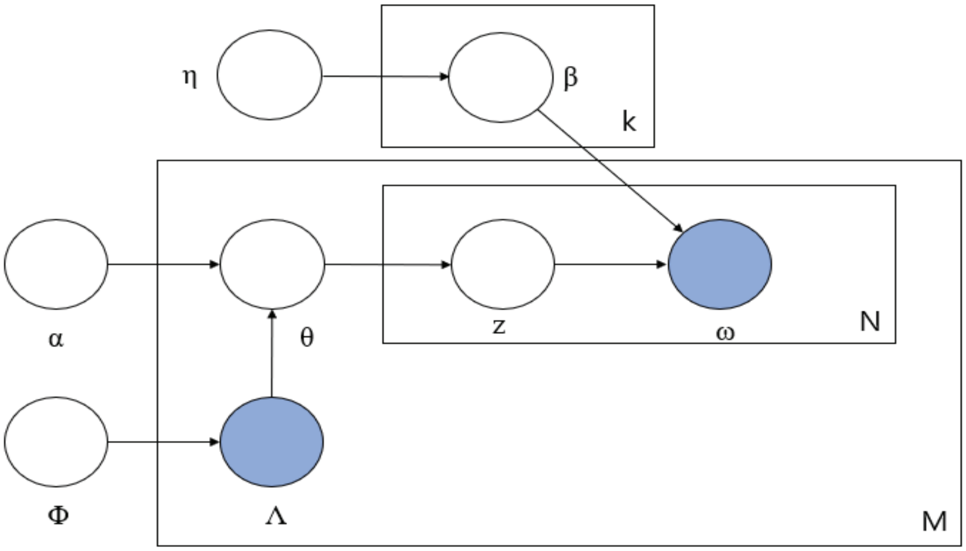
ω: Observed word

α, η: Dirichlet distribution of topic(α) and words(η)

Figure 4.1 Graphical model of Latent Dirichlet Allocation (LDA).

Source : Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 pp. 993-1022

L-LDA is a supervised learning method that utilizes the cause and effect relationships between words, documents, and topics [9]. As L-LDA is an extended concept of LDA, the basic process is the same as LDA. The main difference between the two models is the learning methods. LDA is an unsupervised method of clustering topics using word similarity. This means LDA is clustering model which automatically forms a group of similar words given the number of subjects determined by a user. Unlike LDA, L-LDA starts from known relationships between documents and topics provided by users as a training set. Under the assumption that documents and words become clues that determine the topic, the model proceeds to learn that a set of documents and words labeled with the topics to match it.



Where M: Number of documents

k: Number of topics

N: Number of words in the document

θ: Topic distribution of the document constrained by Λ

β: The word distribution for topic k

Z: The number of the topic to which the word belongs

ω: Observed word

Λ: Topic presence/absence indicator

α, η: Dirichlet distribution of topic(α) and words(η)

Φ: Labeling prior probability

Figure 4.2 Graphical model of Labeled Latent Dirichlet Allocation (L-LDA).

Source: Ramage, D., Hall, D., Nallapati, R., and Manning, C. 2009. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing 248–256

In Figure 4.2, an observed topic presence/absence indicator vector Λ, is added in

LDA model. Comparing to Figure 4.1, the difference is Λ, which restricts acceding all topics to the document by labeling the topics. LDA uses a Dirichlet distribution to model the subject distribution of a document. That is, while the proportion of occurrence of a topic, k, is the same in certain documents, L-LDA serves to limit the topics that will appear in certain documents. Thus, if the Λ of all documents in L-LDA is 1 (there is a possibility that all documents can evenly cover all topics), it will be the same model as the LDA. θ drawn from α and Λ represents the distribution of the topic within the document. Following the Dirichlet distribution, Φ means the distribution of words within the topic.

Using L-LDA for prediction of IDT has several unique benefits compared to other methods reviewed in the section of Literature Review. Firstly, traffic incident reports contain the clue for predicting the IDT. These words describe the incident situation, such as the environment, condition, and severity. The words describing a case creates a cause and effect relationship with the IDT. The supervised learning method is more effective than unsupervised learning, because the dataset is composed of the topics of duration time and related words of those topics. Secondly, the advantage of unsupervised learning is automatic clustering by grouping similar words. However, there are some disadvantages. The users have to find the characteristics of the topics with the results of the clustering. Furthermore, the result may not be matching with the topics that a user thinks before the analysis. L-LDA can solve these shortcomings by initially setting the number of topics that the user wants. Lastly, L-LDA has an advantage in that, when learning is well developed, even if new documents are as input data, it can analyze the material belongs to which topics. The LDA model is also known to be capable of analyzing new documents, but only for small size. [38]

4.2 Dataset

4.2.1 Dataset Overview

The Korea Expressway Corporation (KEC) manages all incidents taking place within its jurisdiction, covering approximately 4,195 km of expressway segments in South Korea. The dataset employed for this study includes a total of 2,220 expressway incident records collected by Korea Expressway Corporation (KEC) from 2016 to 2019. KEC manages incident in four levels, A,B,C, and D, by severity. D level incidents are not considered in this research as the severity is too low to impact IDT.

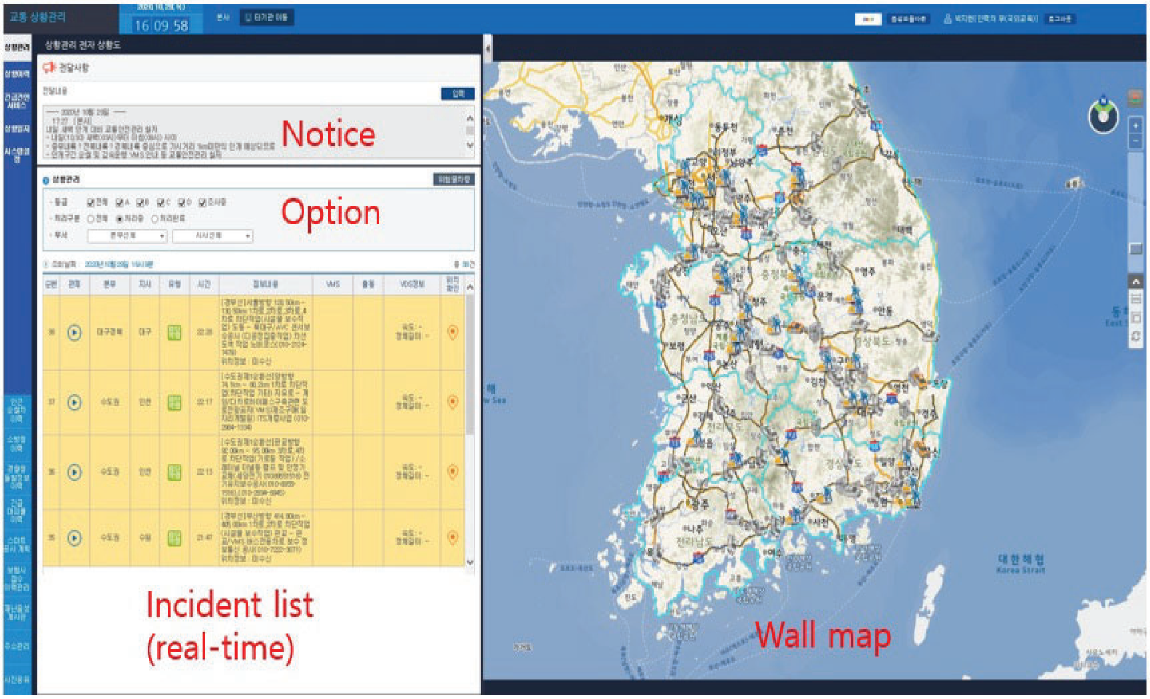
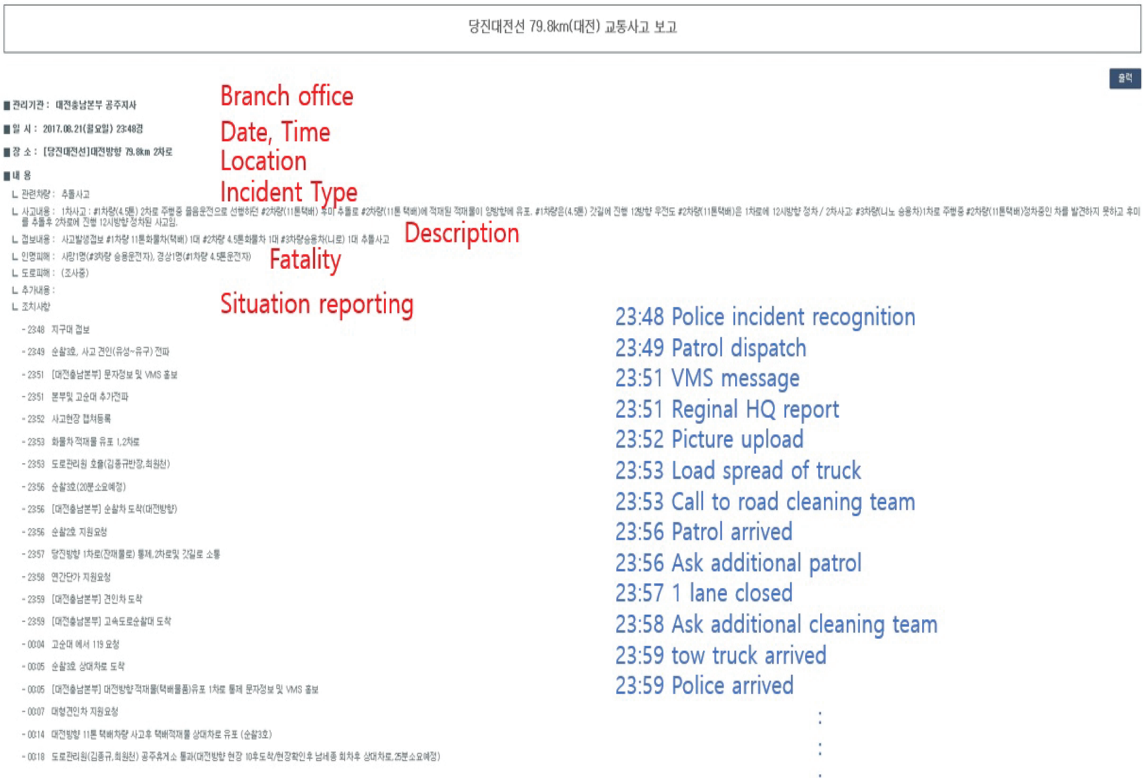


Figure 4.3 OneClick system's main interface.

Every incident text data is archived in OneClick, an incident database system and incident report of the KEC. OneClick is a real-time incident management system developed in 2016. (Figure 4.3) The strength of this system is that the actual field conditions can be shared with all stakeholders in real time. A branch manager manually enters an incident information in real-time based on CCTVs in and around the incident scene and/or instantaneous radio communications with an incident response team. The information is shared in real-time with the local headquarter and TMC manager who is in charge of incident management. Personnel in charge of incident management uses this information to suggest countermeasures, to consider additional support, and to facilitate cooperation with other agencies.



당진대전선 79.8km(대전) 교통사고 보고

■ 관리기관 : 대전충남본부 공주지사
■ 일 시 : 2017.08.21(월요일) 23:48경
■ 장 소 : [당진대전선]대전방향 79.8km 2차로
■ 내 용
ㄴ 관련차량 : 추돌사고
ㄴ 사고내용 : 1차사고 : #1차량(4.5톤) 2차로 주행중 졸음운전으로 선행하던 #2차량(11톤택배) 후미 추돌로 #2차량(11톤 택배)에 적재된 적재물이 양방향에 유포. #1차량은(4.5톤) 갓길에 진행 12방향 우전도 #2차량(11톤택배)은 1차로에 12시방향 정차 / 2차사고: #3차량(니노 승용차)1차로 주행중 #2차량(11톤택배)정차중인 차를 발견하지 못하고 후미를 추돌후 2차로에 진행 12시방향 정차된 사고임.
ㄴ 접보내용 : 사고발생접보 #1차량 11톤화물차(택배) 1대 #2차량 4.5톤화물차 1대 #3차량승용차(니로) 1대 추돌사고
ㄴ 인명피해 : 사망1명(#3차량 승용운전자), 경상1명(#1차량 4.5톤운전자)
ㄴ 도로피해 : (조사중)
ㄴ 추가내용 :
ㄴ 조치사항
- 23:48 지구대 접보
- 23:49 순찰3호, 사고 견인(유성~유구) 전파
- 23:51 [대전충남본부] 문자정보 및 VMS 홍보
- 23:51 본부및 고순대 추가전파
- 23:52 사고현장 캡쳐등록
- 23:53 화물차 적재물 유포 1,2차로
- 23:53 도로관리원 호출(김종규반장,최원천)
- 23:56 순찰3호(20분소요예정)
- 23:56 [대전충남본부] 순찰차 도착(대전방향)
- 23:56 순찰2호 지원요청
- 23:57 당진방향 1차로(잔재물로) 통제,2차로및 갓길로 소통
- 23:58 연간단가 지원요청
- 23:59 [대전충남본부] 견인차 도착
- 23:59 [대전충남본부] 고속도로순찰대 도착
- 00:04 고순대 에서 119 요청
- 00:05 순찰3호 상대차로 도착
- 00:05 [대전충남본부] 대전방향 적재물(택배물품)유포 1차로 통제 문자정보 및 VMS 홍보
- 00:07 대형견인차 지원요청
- 00:14 대전방향 11톤 택배차량 사고후 택배적재물 상대차로 유포 (순찰3호)
- 00:18 도로관리원(김종규,최원천) 공주휴게소 통과(대전방향 현장 10후도착/현장확인후 남세종 회차후 상대차로,25분소요예정)

Figure 4.4 Example of real time incident report of OneClick system.

Figure 4.4 shows an example of an actual incident report used for this study. The incident report contains not only numerical data such as the number of vehicles, fatality but descriptions of the field with the text. In case of an incident, an initial report is created for the incident based on observations by field investigators at the incident scene, such as first responders, police officers, and incident dispatching team. The initial report is then further enhanced with additional information that, if any, was revealed by other information sources (e.g., insurance investigations) after the occurrence of the incident by an incident manager in each branch office. Therefore, the final report contains more information than the initial report. Since the information missing in the initial report depends on the writer's tendency or reporting habits, information describing the situation in the final report was included in the analysis.

For accurate prediction, the data of the following incident were excluded from the dataset for analyzing.

1) Incidents with missing incident duration time: IDT is key data for this analysis.

2) Incidents mismatching OneClick and final report data: inaccurate information was removed as it can distort the analysis results.

3) Incidents with too short description: removed if less than 5 lines in OneClick as it is very likely that operators insincerely wrote the report or inputted it with only summarized text.

4) Incidents without description of the accident situation: since this study predicts IDT based on the incident situation, incidents without information of the situation were excluded. Some managers write the report focusing on the reactions of the response team, not the situation.

As a result, 1,466 incidents were adopted as dataset of this research.

4.2.2 Dataset Description

Table 4.1 and Figure 4.5 show the description of input dataset. The average IDT is 93.2minute, and the median value is 66.0min. The gap between median and mean values means IDT's left-biased characteristics, as also shown in Figure 4.5. Almost half of the incident durations are within 60min (44.7%), and the third quartile of the incident cleared within 2 hours. Due to most of the samples less than 30 minutes excluded from the aforementioned data cleaning process, the number of data in this range is smaller than 30-60 minutes. In practice, as IDT less than 30 minutes does not have a critical impact on traffic, it is reasonable to use this dataset as input data.

Table 4.1 Dataset Description

Mean	Max	Min	25%	50% (Median)	75%	Standard deviation
93.2	893.0	4.0	41.0	66.0	116.5	83.7

Time	30min	60min	90min	120min	150min	180min
Accumulated Percentile	11.6%	44.7%	64.4%	76.5%	83.8%	91.4%

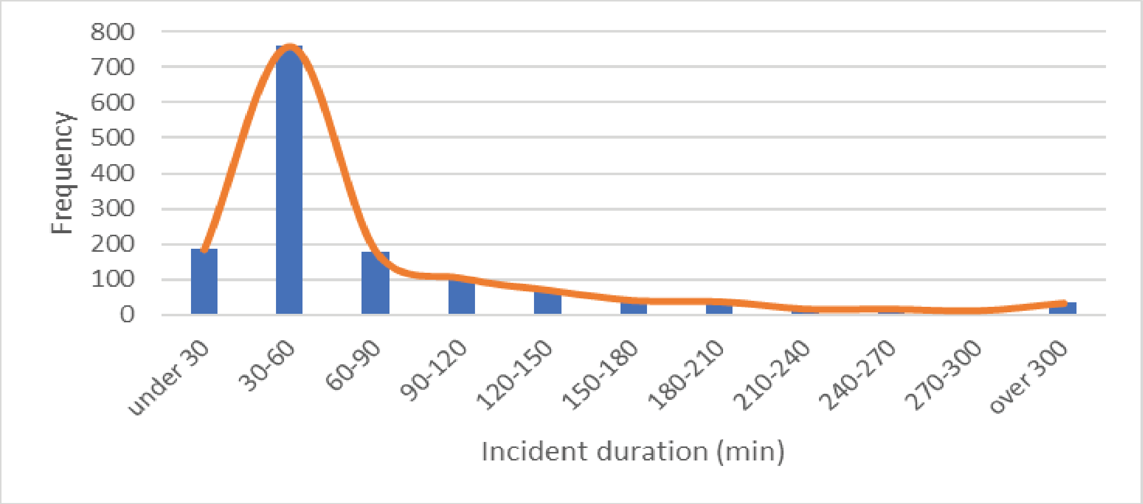


Figure 4.5 Incident duration time distribution.

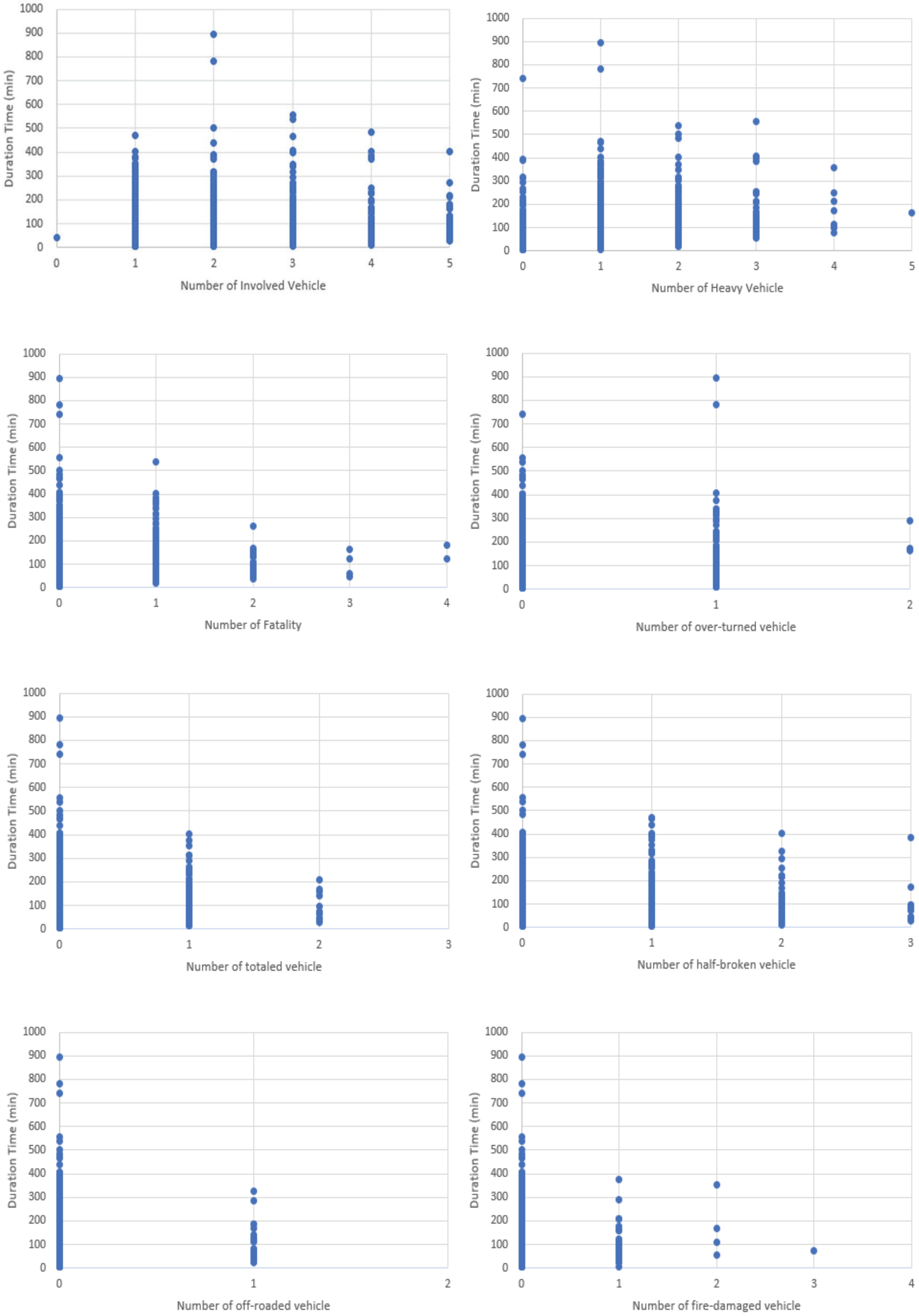


Figure 4.6 Relationships between incident duration times and variables.

Figure 4.6 demonstrates relationships between incident duration times and some selected variables such as the number of involved vehicles, heavy vehicles involved in incident, number of fatalities, etc. In the figure, we can find out variables are widely spread, so cannot easily find correlation with IDT.

4.2.3 Data Preprocessing

The LDA model applies to all languages. Therefore, it is unnecessary to translate the incident records to English. However, it is necessary to clearly state that the entire reporting process is manually conducted, prior to the OneClick system. Thus, it is likely that a report often includes inappropriate information while it omits critical piece of information that could be directly related to the actual incident. To handle this challenge, this thesis performed multi-step data preprocessing to sanitize raw dataset based on Natural Language Processing (NLP) techniques as follows.

- Step 1: **Correction of misspelling and spacing.** The misspelled words have been modified to minimize errors because they can cause a significant bias in topic modeling. Besides, the Korean language, unlike English, has special characteristics that combined two words are used as common words; thus, correcting spaces is also an essential process to transfer exact meaning.

- Step 2: **Matching similar words**. Different sentences and words are used depending on the person although they prepare the same report. If multiple words with the same meaning are used, topic modeling does not recognize a similar meaning.

Therefore, in order to increase the accuracy of the model, words with similar meanings have been unified.

- Step 3: **Bag of Words.** A simple way to analyze a document is to deal with them as a collection of fixed words or vocabulary so-called a 'Bag of Words (BoW)'. The BoW approach was proposed by Zhao et al [27]. This approach is used in case of languages (e.g., Korean, Chinese, and Japanese) that are not fully compatible yet with existing NLP techniques. In reality, not every single word in a KEC incident report is directly related to an incident: e.g., words like the names of first responders, the name of branch office, etc. As the purpose of this study is to predict IDT based on incident-related information, a total of 35 words were selected as BoW to keep all the time as shown below:

"Detach, Chamber, Load, Drop, Fall, Gas, Chemistry, All lane blocking, Cut, Stuck, Congestion, Paint, Green area, Slope, Spread, Chain, Collision, Turned over, Container, Damage, Iron, Leakage, Oil, Fuel, Broken, Outflow, Self-moving, Safety, Rotation, Work, Patrol, Reverse driving, Breaking, Courier, Low-speed"

- Step 4: **Vehicle type unification**. Depending on report generators, the same type of vehicles involved in incidents can be recorded in various ways. For example, a large cargo truck can be written in a report as '15ton cargo truck' or 'Scania truck' or '20ton dump truck'. The vehicle type was expressed in the raw dataset in various forms using the payload capacity, model, and make. Therefore, to maintain consistency, the type of vehicle is unified as shown below:

"Passenger car, Minibus, medium bus, Large bus, Small truck, Medium truck, Heavy cargo truck, Heavy trailer, Large tank lorry"

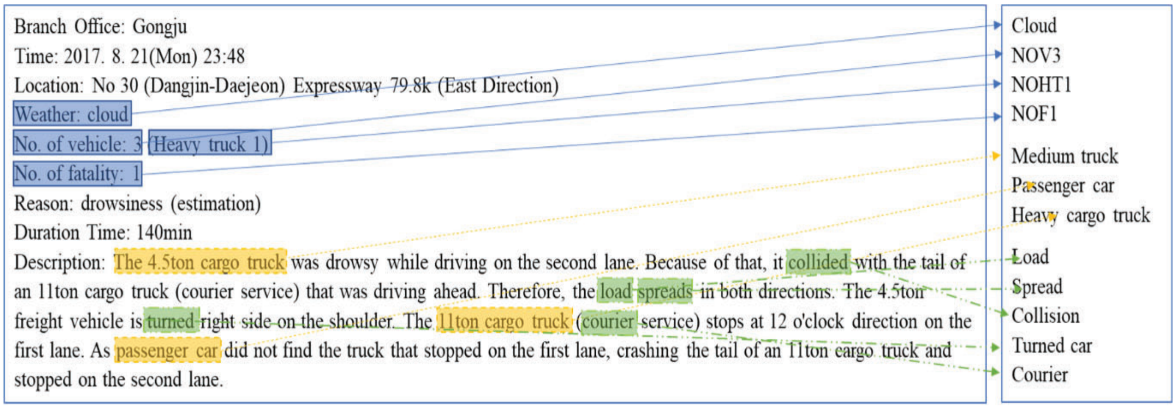
- Step 5: **Integrating numerical data**. The KEC incident report contains common information to supplement the main text-based incident description as shown in Figure 4-6. This type of supplementary information in the raw dataset is given in numerical form. In this study, this numerical information was also utilized by converting it into text form. For example, if the number of fatalities is 2, it was converted into the shape of 'NOF2' and used for text analysis. Below are datasets.

"Number of vehicles, Number of heavy vehicles, Location, Weather, Number of totally broken vehicle Number of half-broken vehicle, Number of fatalities, Number of pedestrian fatality, Number of override vehicle, Number of off-road vehicles, Number of turned vehicle, Number of fire-damaged vehicles"

4.3 L-LDA Based Topic Modeling for Incident Duration Time Prediction

4.3.1 Document Provision

L-LDA training is conducted by modeling the probability of words included in the topic by specifying a topic for each document. Where, the document is the text data of an incident report, and the topic is the IDT. Each incident report going through the data preprocessing steps is used as a document for L-LDA. Similarly, the IDT recorded in each report is used as the topic for the corresponding document. Determining a topic from IDT is discussed in the next section in detail. Figure 4.7 depicts an example of a document used for L-LDA training.



Branch Office: Gongju
Time: 2017. 8. 21(Mon) 23:48
Location: No 30 (Dangjin-Daejeon) Expressway 79.8k (East Direction)
Weather: cloud
No. of vehicle: 3 (Heavy truck 1)
No. of fatality: 1
Reason: drowsiness (estimation)
Duration Time: 140min
Description: The 4.5ton cargo truck was drowsy while driving on the second lane. Because of that, it collided with the tail of an 11ton cargo truck (courier service) that was driving ahead. Therefore, the load spreads in both directions. The 4.5ton freight vehicle is turned right side on the shoulder. The 11ton cargo truck (courier service) stops at 12 o'clock direction on the first lane. As passenger car did not find the truck that stopped on the first lane, crashing the tail of an 11ton cargo truck and stopped on the second lane.

Figure 4.7 The sample case of the incident report.

4.3.2 Determination of Topic

Unlike LDA, L-LDA has to initially determine a topic for each document to perform training as a supervised learning approach. To this end, topics need to be classified based on the range of IDTs in the raw dataset. In this study, the ranges of IDTs are determined by an MUTCD recommendation for a traffic rerouting strategy in traffic incident management.

Depending on the duration time, MUTCD categorizes incidents into three phases: minor, intermediate, and major. [41] A minor incident is defined that can be cleared within 30 minutes. An incident which the duration time is between 30 minutes and 2 hours is grouped into an intermediate incident. The major incident is when the duration time exceeds 2 hours. Given the phases of incidents, MUTCD recommends different mitigation remedies. In case of a minor incident, MUTCD suggests a basic level of traffic management that does not include any queue warning, lane shifting, or detour messages. For an intermediate incident, the MUTCD-suggested mitigation strategies include the deployments of lane shifting signs with tapered lane closure and upstream warning devices to alert road users on the downstream queues. In case a major incident takes place, the provision of detour information to travelers is recommended. Thus, from the perspective of incident management, duration time of 2-hour is a critical boundary to make a decision for implementing the detour strategy. In addition to MUTCD, Alternate Route Handbook (ARH) [42] also presents that several states in the U.S define a major incident that needs to implement detouring when the duration time is over 2-hour. Taking into consideration the criteria for detour appeared in both MUTCD and ARH, this study decided to use binary topics for L-LDA: 1) duration time shorter than 2-hour and 2) duration time longer than 2-hour.

4.3.3 Model Calibration

In order to implement the L-LDA model, it is necessary to determine the initial values of the parameters, α, and η, prior to its implementation. α and η mean Dirichlet distribution of topics and words, respectively. In addition, like other machine learning methods, L-LDA training is implemented based on an iterative approach, which uses a stopping threshold, Λ, to check the convergency of the training. Along with α and η, the training performance of L-LDA model is also affected by the stopping threshold. To achieve the best performance, it is certainly necessary to identify the best combinations of α, η, and Λ for given incident reports that have not been used for training. In this thesis, a full factorial experimental design approach is employed to calibrate the L-LDA model. It is noted that the value of η has been fixed to the tenth of α values to reduce the number of experiments. Table 4.2 summarizes the level of each experimental factor for the calibrations. This study employs a perplexity as a primary measure to determine the best combination of the calibration parameters. Perplexity is a numerical measure to examine the prediction performance of a certain probability model for the observed value; thus, an L-LDA model with low perplexity indicates that it reflects the document well.

It is noted that each experiment was replicated 30 times to reflect random variabilities for the calibration. The calibration was conducted for approximately 62 hours on a computer equipped with a 8th generation Intel Core-i7 CPU, 64GB of RAM and a Nvidia GTX 2080 GPU.

Table 4.2 Level of Each Experimental Factor for the Calibrations

Factor (Parameter)	Level	Cases	Total Experiments
α	0.1 ~ 2.0 at 0.1 interval; 0.01 ~ 0.1 at 0.01interval; and 0.001 ~ 0.01 at 0.002 interval	35	840 (=35X24)
η	Fixed to $\alpha/10$		
Λ	0.1 ~ 0.7 at 0.05 interval; and 0.01 ~ 0.1 at 0.01 interval	24	

CHAPTER 5

RESULTS

5.1 Analysis Results

With the final model selected from the calibration process discussed in the previous section, a set of test documents that are randomly selected from the entire dataset is fed into the selected model to examine the performance of the model. Since L-LDA is based on a stochastic process, results from each test even with the same dataset are different. To handle this, this study performed a total of 30 replications with the same test documents.

Given a document, a trained L-LDA model generates probabilities, or confidence levels, as outputs to indicate how accurately the document fits each topic, including an undefined topic called “common topic”. The common topic is created by the L-LDA model to show a probability that the selected document cannot be classified into any of predefined topics. In general, a training dataset with significant noises, such as typos, unrelated words, broken sentences, would likely increase the probability of common topic. Figure 5.1 shows the probability distribution of IDTs less than 2-hour and larger than 2-hour which are denoted by IDT-Down and IDT-Up for short, respectively. Out of actual 4,131 IDT-Down cases obtained from 30-replication results in the test, Figure 5.1 demonstrates that the developed L-LDA model chose 2,852 cases with 90% or above confidences (i.e., probability) to be IDT-Down. Similarly, 1,299 IDT-Up cases out of 1,776 were predicted by the L-LDA model with at least 90% of confidence.

Figure 5.2 shows the cumulative distribution function of IDT-Down (top figure) and IDT-Up (bottom figure), respectively. Assuming the minimum confidence level for practically acceptable prediction performance to be 0.7 (i.e., 70% confidence), approximately 77% and 81% of IDT-Down and IDT-Up cases are properly predicted by the L-LDA, respectively, as illustrated in Figure 5.2.

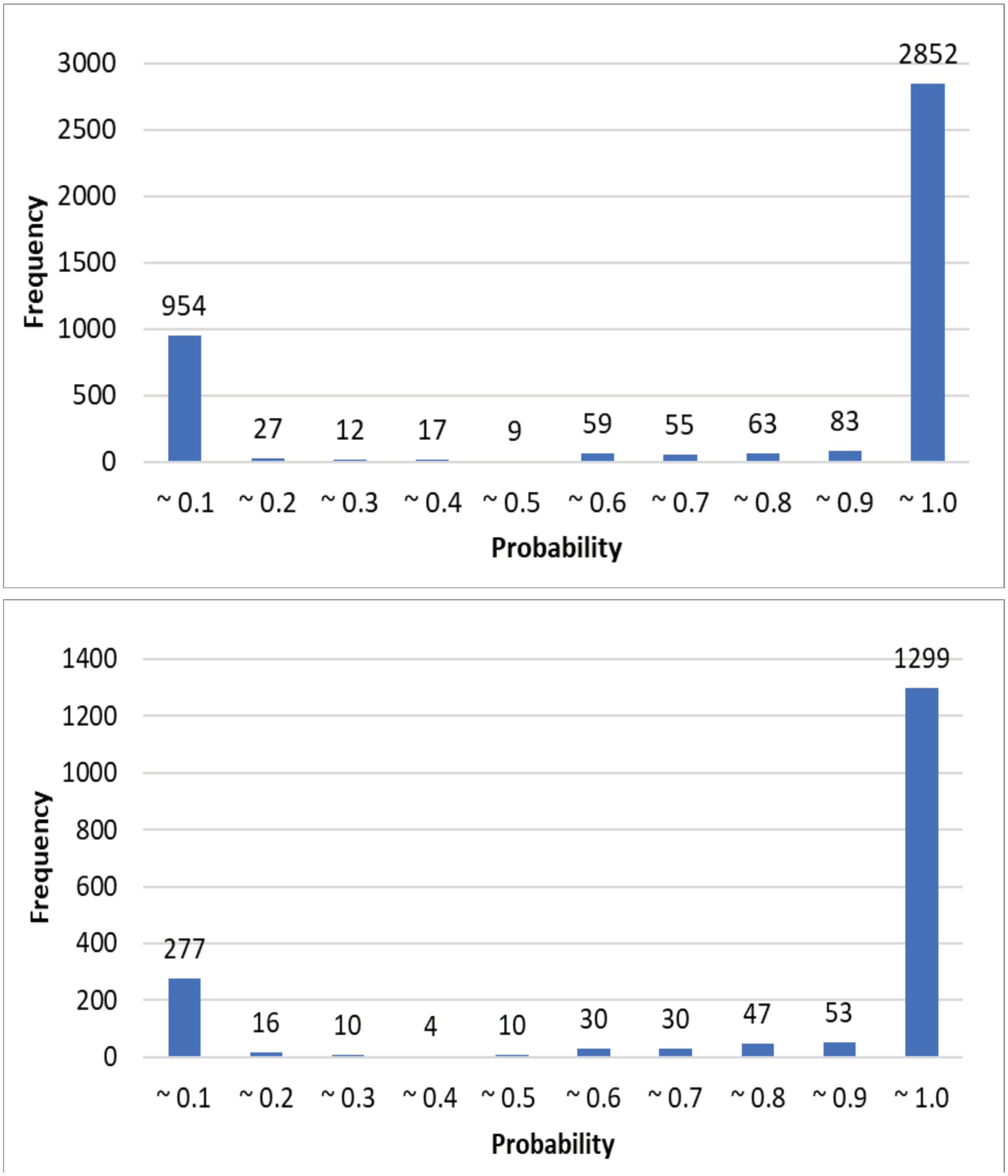


Figure 5.1 Distribution of probabilities for IDT (top: IDT less than 2-hour; bottom: IDT greater than 2-hour).

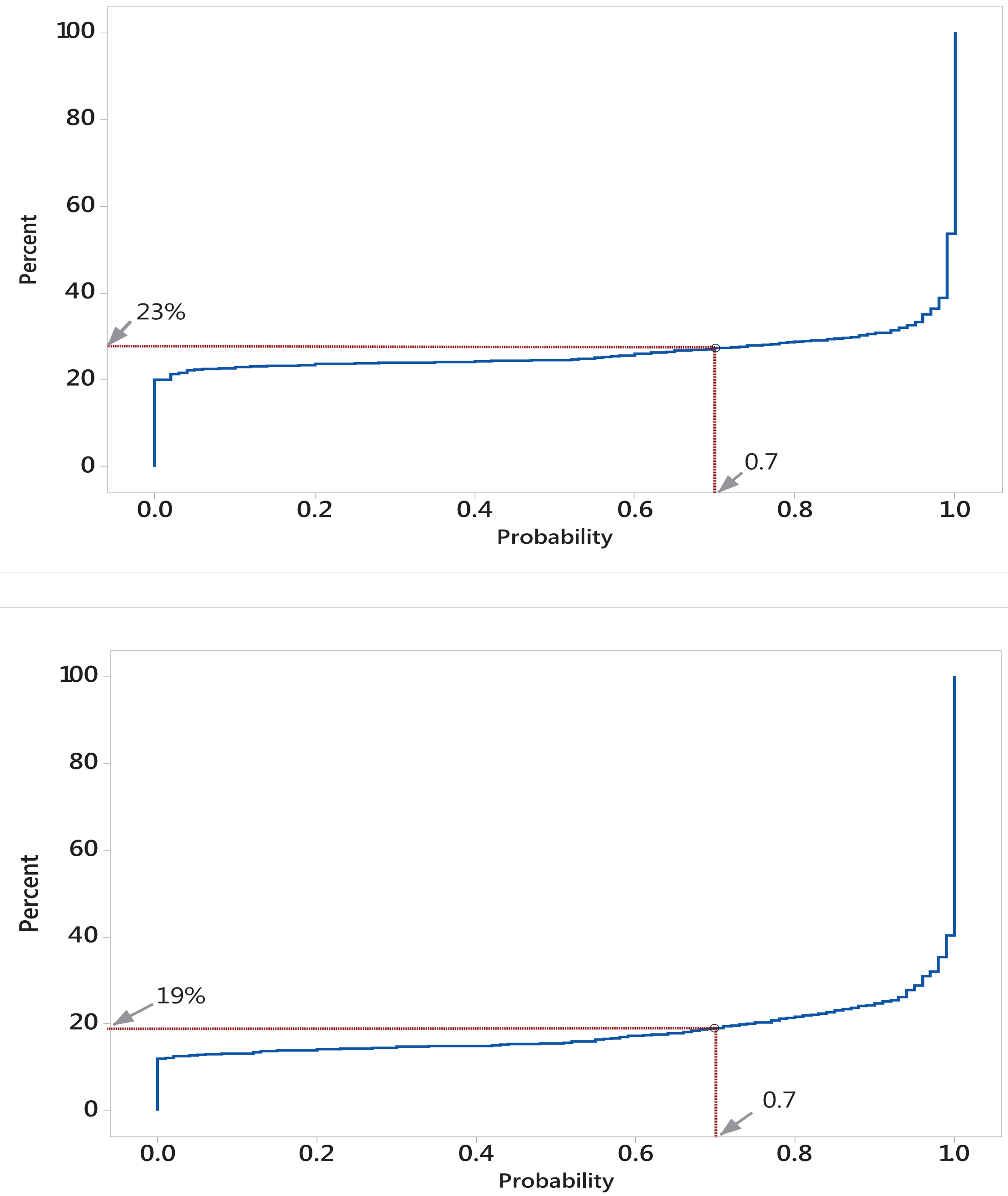


Figure 5.2 Cumulative distribution function (top: IDT less than 2-hour; bottom: IDT greater than 2-hour).

As aforementioned, L-LDA estimates the probabilities of a document to be classified to each predefined topic as well as the common topic. The sum of the probability for each topic is 1.0. As a result, each document has multiple probabilities estimated by L-LDA, as many as the number of predefined topics and the common topic.

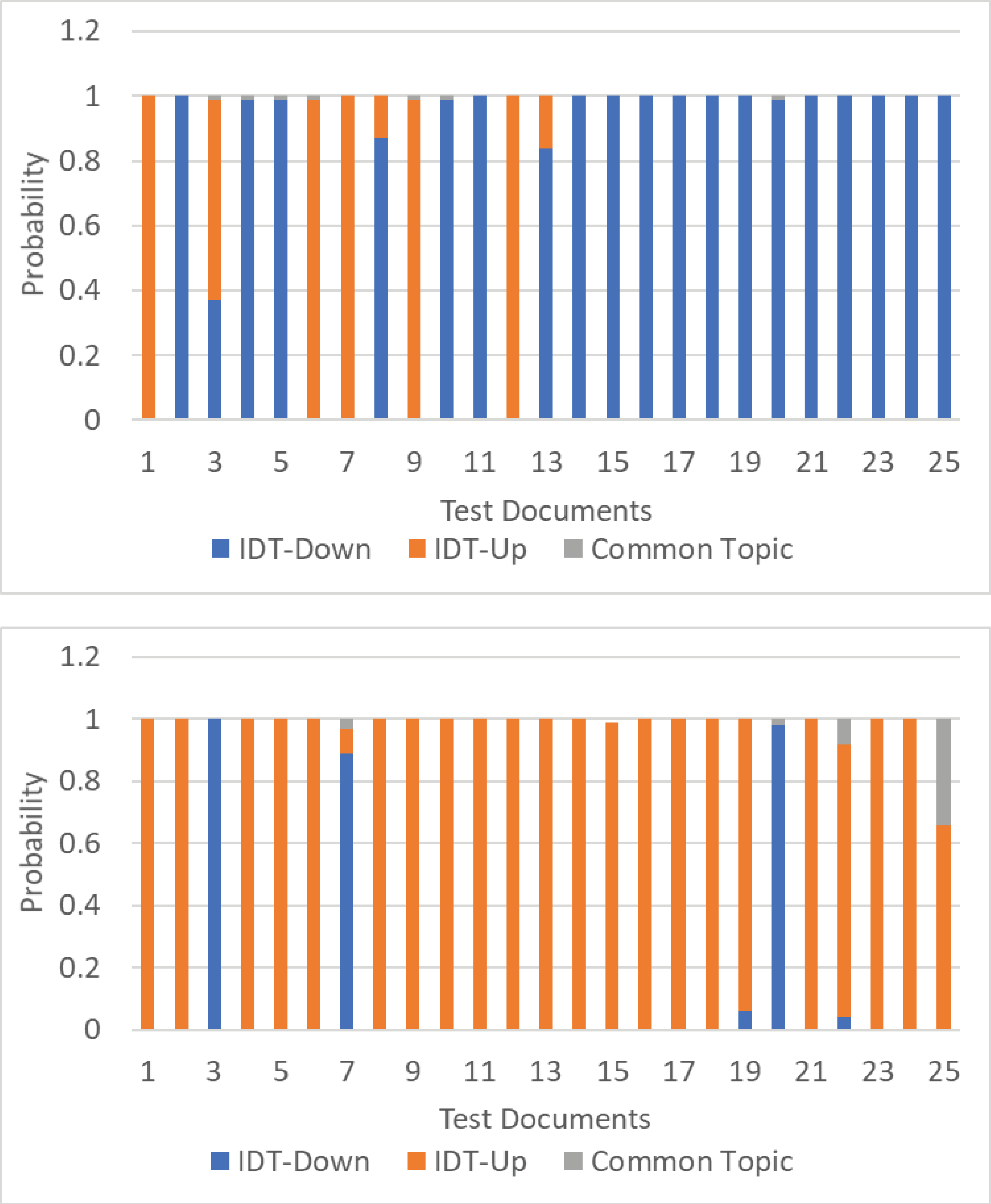


Figure 5.3 Example of match accuracies for each topic (top: IDT less than 2-hour; bottom: IDT greater than 2-hour).

Figure 5.3 shows stacked bar charts as the examples of L-LDA results for each topic: IDT-Down and IDT-Up, respectively, which are randomly excerpted from the test documents to graphically illustrate how the L-LDA results look like. It must be noted that the test documents used in each chart do not necessarily mean the same document. Each of stacked bars in the charts consists of up to three elements in this study, which indicate IDT-Down in blue, IDT-Up in orange, and Common Topic in gray. Thus, a bar filled with a single color means that the document is predicted to be only one topic with 100% confidence. Obviously, if the predicted topic with 100% confidence is identical to the ground truth, then it means a perfect match, such as the document number 2 in the left case; otherwise, it indicates a perfect mismatch as shown in the document number 3 in the right case.

This study examined the matching accuracy of individual test document obtained from all 30 replications to utilize it as an additional performance indicator of the L-LDA-based IDT prediction. Figure 5.4 displays the averages of matching accuracies for IDT-Down and IDT-Up, which look similar to a confusion matrix. As clearly shown in the figure, the overall matching accuracy is 74% and 82% for the IDT-Down and IDT-Up cases, respectively. Blue color in IDT-Down means that 74% of estimates and actual values match (less than 2-hour). Orange color in IDT-Down, means that for 24% of cases, cases of less than 2-hour were incorrectly matched with cases of greater than 2-hour. For IDT-Up cases, Orange color means that 82% of estimates and actual values match (greater than 2-hour). However, Blue color means that 16% of cases are greater than 2-hour is true, but the model estimated as less than 2-hour cases.

It is of interest that the matching accuracy of IDT-Up case is higher than the IDT-Down case by 8%. With in-depth investigations on dataset, it is discovered that the documents dealing with the IDT-Up cases more frequently contain unique and straightforward words that can improve the situational awareness of incident scene.

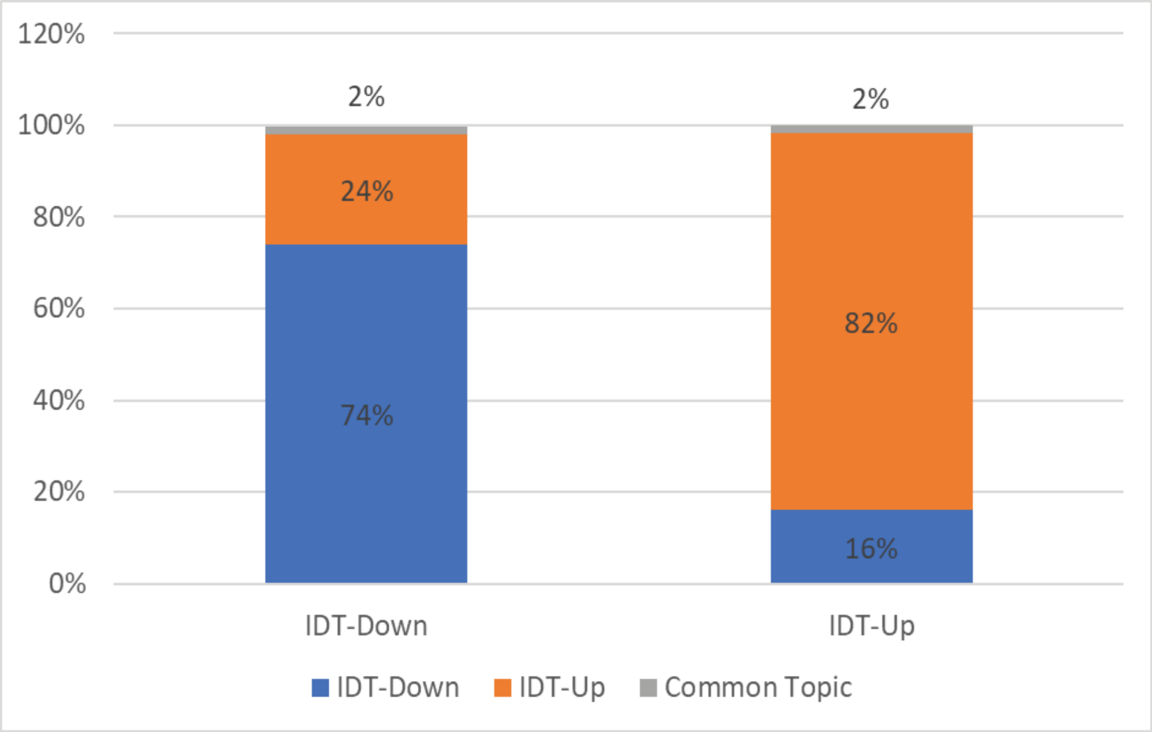


Figure 5.4 Overall matching accuracy.

5.2 Model Comparison

In this thesis, the L-LDA model was compared with three widely used models, SVM, KNN, and the decision tree model. The results are provided in Table 5.1. Though the result shows overall accuracy is low compared to other models, it is discovered that L-LDA shows outperforming prediction accuracy for over 2-hour duration case.

Table 5.1 Performance Comparison

	L-LDA	Decision Tree	KNN	SVM
Total Accuracy	76%	85%	83%	87%
IDT-Down	74%	93%	94%	95%
IDT-Up	82%	57%	45%	60%

This high accuracy is unique even compared to other studies. The result of the study of Yang et al. [20], and Chang et al. [18] are compared as summarized in Table 5.2 for the over 2-hour duration case.

Table 5.2 Performance Comparison of Other Research Result

	L-LDA	BN Model (Yang et al.)	Decision Tree (Chang et al.)
Total Accuracy	76%	72.6%	73.6%
IDT-Down	74%	81.8% (0-30min)	96.7% (0-41min)
		54.2% (30-120min)	17.2% (42-118min)
IDT-Up	82%	49.8%	11.4% (119min~)

From Table 5.2, the findings are as follows. Firstly, the prediction accuracy with small cases is greatly improved. It is shown that the L-LDA model can overcome the limitation of the size of samples that many existing studies have. It is because the hidden meaning was grasped by utilizing the text data analysis. Secondly, the result has great value in terms of the agency's perspectives. Agencies require high accuracy for accidents with long duration times. It is because providing accurate predictions for incidents with high accident severity is more meaningful in practice. Finally, although the overall accuracy is lower than that of the other models, the similar accuracy between the two groups (74%, 82%) means that it is likely to be developed through future model improvement. As mentioned earlier, the accident data has a left shifted shape. Therefore, there is a characteristic that the accuracy of the group with a large amount of data can be increased. Since the L-LDA model overcomes these characteristics, it is expected to develop an improved model in the future.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

This thesis presents a novel method to predict incident duration time using a topic modeling algorithm. As the text data of the incident report has an incomparable amount of information compared with the numerical dataset, the text analysis approach is useful for identifying potential factors affecting the duration times for various types of incidents. However, due to the difficulty of analyzing the text type of data, not enough research efforts have been conducted. Recently, drastic advances in topic modeling and machine learning technologies have made this possible.

The analysis is conducted with Korea Expressway Corporation's 1,466 incident records collected from 2016 to 2019. As the prediction should be done at the beginning of the incident, each incident record is preprocessed for describing only the initial situation of the incident. The analysis was performed using a Labeled LDA model, a supervised topic modeling technique. The result shows that the L-LDA method yields approximately 74% accuracy in less than 2-hour duration cases and 82% in greater than 2-hour duration cases.

The findings of this study are as follows. Firstly, the topic modeling method has a great potential in incident duration time prediction as it utilizes descriptive information vividly conveying the actual incident scene. This method can contribute to improving the prediction accuracy by complementing the problems of existing numerical type datasets. Secondly, L-LDA is applied for the first time to conduct semantic text analyses of incident records that contain a straightforward cause-and-effect relationship between words used in the records and the actual incident duration times. Unlike unsupervised topic modeling methods, the cause-and-effect relationships observed in the incident records can be used for training, thereby resulting in promising performances. In particular, the high accuracy in greater than 2-hour duration case with less samples proves that the proposed method is appropriate to overcome challenges encountered by the past studies that the accuracy decreases with small number of samples. [20, 24]

6.2 Recommendations

With recent drastic advances in machine learning technologies enabling to handle gigantic amounts of data, autonomous TMC operations become more feasible than ever. Yet rather the quantity of data, an attention should be paid to the quality of data. The followings should be considered to manage data quality.

Firstly, the model's performance relies heavily on the quality of its input dataset; so does the supervised topic modeling method. Since incident reports are compiled by human operators, it is understandable that the reports contain erratic mistakes like typos, information omission. While it would be challenging to achieve in the near future, the performance of the incident duration model could be dramatically improved if the incident reporting procedure becomes more systematic and unified to reduce human errors.

Secondly, the classification of segmented incident duration time could be helpful to predict. As shown in Figure 1.4, HCM divided the duration time into five segments. However, Austroad [43] was subdivided into eight to strengthen TIM as shown in Figure 6.1. If the activities of incident responses can be clarified based on the duration time segments, and each segment has accurate time, the prediction accuracy of IDT could be improved.

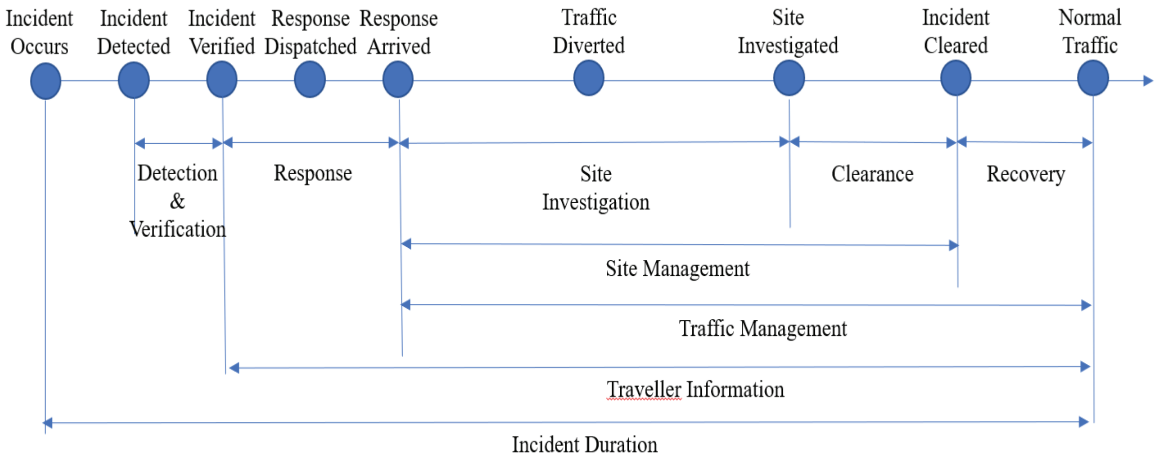


Figure 6.1 Example of segmented incident duration time.

Source : Traffic Incident Management : Best Practices. 2007. Austroads Publication No. AP–R304/07. Austroads Research Report: Sydney, Australia

Finally, from a practical point of view, the texts of the existing incident reports are focusing on the cause and reaction of an incident. It is crucial to establish a business rule for creating an incident report so that the incident report also includes overall situational awareness in and around the incident scene. It is worth noting that data retrieved from the OncClick system used in this study is a good example of real-time text data. Using this, real time duration time estimation is possible. Therefore, the creation of high-quality text data by an operator is a critical factor in the utilization. It is also possible through making clear criteria for creating documents and training operators who make reports.

In summary, agencies need substantial effort to generate and manage incident data for the sake of data quality. In the preprocessing process, it was identified that the value of duration time was recorded incorrectly by the operator due to ambiguous criteria. Because these errors have a huge impact on predictions, agencies should pay great attention to data management. Recently, research for efficient response has attracted a lot of attention. Since the efforts of agencies such as Traffic Incident Management (TIM) lead to a reduction in incident duration time, Data management is more necessary in wide aspect.

6.3 Future Research

As future research, the model may be enhanced by considering the following factors. Firstly, the accuracy could be significantly improved by using the hybrid model which has higher accuracy in less than 2-hour duration case. The L-LDA model showed high accuracy in greater than 2-hour duration case. However, insufficient results showed in case of less than 2-hour comparing greater than 2-hour case. To solve this problem, the L-LDA model combined with other classification models will improve the accuracy. It is also helpful in combining additional data to describe the initial situation, such as images captured at incident scenes. Image data, unlike text data generated by humans, has the advantage of not including errors such as typos. The combination of numerical, text, and image data may improve the accuracy of the analysis.

Secondly, the incident duration time is affected by not only the status in the early stages of the incident but also the response time of every procedure. The accurate prediction will be improved by adding additional texts that reflect incident response. In this study, only the text of the initial situation is considered for research. However, instantaneous response in the early stage of the incident dramatically influences duration time. Therefore, it will also study the adapted method using consideration progress of clearing up incident scene and modifying the prediction time according to the response activity.

Lastly, text data from incident reports can be used to predict the overall severity estimation of incidents. This thesis used the data of incident report to predict only duration time estimation. The scope of prediction can be expanded by considering fatality, hazard spills, etc. If the severity of an incident is predicted early stage, the agile reaction will also possible. This could be a useful study to reduce casualties and reduce duration time.

REFERENCES

1. National Academies of Sciences, Engineering, and Medicine 2014. *Design Guide for Addressing Non-recurrent Congestion*. Washington, DC: The National Academies Press.

2. Friedrich, M., Lohmiller, J. 2012. Factors Influencing the Travel Time Reliability of Motorway Section, Proceedings of the 6th International Symposium Networks for Mobility, Stuttgart.

3. Margiotta, R., Dowling, R., and Paracha, j. 2012. *Analysis, Modeling, and simulation for Traffic Incident Management Applications*. FHWA-HOP-12-045, FHWA. U.S. Department of Transportation, Washington, DC

4. *Transportation Research Board Highway Capacity Manual*. 2010. Washington, D.C.

5. Smith, K., and Smith, B. 2001. Forecasting the Clearance Time of Freeway Accidents. Center for Transportation Studies, University of Virginia

6. Li, R., Pereira, F., and Ben-Akiva, M. 2018. Overview of Traffic Incident Duration Analysis and Prediction. *European Transport Research Review*, pp. 10-22

7 Das, S. Sun, X., and Dutta A. 2016. Text Mining and Topic Modeling on Compendium Papers from Transportation Research Board Annual Meetings. *Transportation Research Record Journal of the Transportation Research Board, pp. 48-56*

8. Young, T., Hazarita, D., Poria, S., and Cambria, E. 2018. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine* Volume: 13, Issue: 3, pp. 55-75

9. Ramage, D., Hall, D., Nallapati, R., and Manning, C. 2009. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing 248–256

10. Khattak, A., Schofer, J., and Wang, M. 1994. *A Simple Time Sequential Procedure for Predicting Freeway Incident Duration.* UCB-ITS-PRR-94-26, California Department of Transportation

11. Peeta. S., Ramos. J., and Gedela. S. 2000, *Providing Real-Time Traffic Advisory and Route Guidance to Manage Borman Incidents On-Line Using the Hoosier Helper Program*. FHWA/IN/JTRP-2000/15, Indiana Department of Transportation and FHWA

12. Chung, Y., Walubita, L., and Choi, K. 2011. Modeling Accident duration and Its

Mitigation Strategies on South Korean Freeway Systems. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2178 pp 49-57

13. Li, R. 2015. Traffic Incident Duration Analysis and Prediction Models Based on the Survival Analysis Approach. *IET Intelligent Transport Systems* Vol. 9, pp. 351–358

14. Zong, F., Zhang, H., Xu, H., Zhu, X., and Wang, L. 2013. Predicting Severity and Duration of Road Traffic Accident. *Mathematical Problems in Engineering*. Volume 2013, Article ID 547904, http://dx.doi.org/10.1155/2013/547904

15. Guan, L., Liu, W., Yin, X., and Zhang, L. 2010. Traffic Incident Duration Prediction Based on Artificial Neural Network. 2010 International Conference on Intelligent Computation Technology and Automation, pp 1076-1079

16. Wu, W., Chen, S., and Zheng, C. 2011. Traffic Incident Duration Prediction Based on Support Vector Regression. Proceedings of the 11th International Conference of Chinese Transportation Professionals, pp 2,412-2,421

17. Gorade, S., Deo, A., and Purohit, P. 2017. A Study of Some Data Mining Classification Techniques. *International Research Journal of Engineering and Technology* Volume: 04 Issue: 04, pp. 3112-3115

18. Chang, H., and Chang, T. 2013. Prediction of Freeway Incident Duration Based on Classification Tree Analysis. *Journal of the Eastern Asia Society for Transportation Studies* vol. 10. pp. 1,964-1,977

19. He, Q., Kamarianakis, Y., Jintanakul, K., and Wynter L. 2011. Incident Duration Prediction with Hybrid Tree-based Quantile Regression. IBM research report

20. Yang, H., Shen, L., Xian, Y., Yao, Z., and Liu, X. 2017. Freeway Incident Duration Prediction Using Bayesian Network. 4th International Conference on Transportation Information and Safety (ICTIS), pp 974-980

21. Park, H., Zhang, X., and Haghani, A. 2013. Interpretation of Bayesian Neural Network for predicting the duration of detected incidents. 92nd annual meeting of the Transportation Research Board, Washington D.C.

22. Yu, B., Wang, Y., Yao, J., and Wang, J. 2016. A Comparison of the Performance of ANN and SVM for the Prediction of Traffic Accident Duration. *Neural Network World* 3, pp 271-287

23. Valenti, G., Lelli, M., and Cucina, D. 2010. A Comparative Study of Models for the Incident Duration Prediction. *European Transport Research Review*. 2, pp. 103–111

24. Hamad, K., Khalil, M., and Alozi, A. 2020. Predicting Freeway Incident Duration Using Machine Learning. *International Journal of Intelligent Transportation Systems Research* 18, pp. 367–380

25. Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* 5:1608

26. Jiang, S., Qian, X., Shen, J., Fu, Y., and Mei, T. 2015. Author topic model-based collaborative filtering for personalized POI recommendations. *IEEE Trans Multimedia* 17(6):907–918

27. Peng, T., Liu, L., and Zuo, W. 2014. PU Text Classification Enhanced by Term Frequency–Inverse Document Frequency-Improved Weighting. *Concurrency and computation practice and experience* pp.728-741

28. Goh, Y., and Ubeynarayana, G. 2017. Construction Accident Narrative Classification: An Evaluation of Text Mining Techniques, *Accident Analysis and Prevention* 108. pp. 122–130

29. Christian, H., Agus, M., and Suhartono, D. 2016. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, Vol. 7, No. 4, pp. 285-294.

30. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*. 41(6):391-407.

31. Robinson, S., Irwin, W., Kelly, T., Wu. X. 2015. Application of Machine Learning to Mapping Primary Causal Factors in Self Reported Safety Narratives. *Safety Science.* 75 pp.118-129

32. Hofmann, T. 1999. Probabilistic Latent Semantic Indexing, Proceedings of the 15th Conference on Uncertainty in AI

33. Zhao, Y., Xu, T., and Wang, H. 2014. Text Mining Based Fault Diagnosis of Vehicle On-board Equipment for High Speed Railway. 2014 IEEE 17th International Conference on Intelligent Transportation Systems pp 900-905

34. Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 pp. 993-1022

35. Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., and Raynal, C. 2016. Natural Language Processing for Aviation Safety Reports: From Classification to Interactive Analysis*.* *Computers in Industry* 78, pp. 80–95

36. Brown, D. 2016 Text Mining the Contributors to Rail Accidents. *IEEE Transactions on Intelligent Transportation Systems*, Volume:17, Issue:2

37. Williams, T., and Betak, J. 2018. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. *Procedia Computer Science* 130 pp.98–102

38. Pereira, F., Rodrigues, F., and Ben-Akiva, M. 2013. Text Analysis in Incident Duration Prediction, *Transportation Research Part C* 37. pp.177-192

39. Blei, D., and McAuliffe, J. 2007. Supervised Topic Models. *NIPS*, volume 21.

40. Lacoste-Julien, S., Sha, F., and Jordan, M.. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. *NIPS*, volume 22.

41. *Manual on uniform traffic control devices for streets and highways.* 2009. FHWA U.S Department of Transportation, Washington, DC

42. *Alternative Route Handbook.* 2006. FHWA-HOP-06-092, FHWA. U.S Department of Transportation, Washington, DC

43. *Traffic Incident Management : Best Practices*. 2007. Austroads Publication No. AP–R304/07. Austroads Research Report: Sydney, Australia