International Journal for Traffic and Transport Engineering, 2022, 12(4): 464 - 479
DOI: http://dx.doi.org/10.7708/ijtte2022.12(4).03 *UDC: 656.1.082:004.85(560)*

TRAFFIC EVENT ANALYSIS MODELING WITH MACHINE LEARNING METHODS USING SOCIAL MEDIA DATA

Cihan Çiftçi[1], Halim Kazan[2]
[1,2] *Istanbul University, Faculty of Economics, Business Administration, Beyazıt, Istanbul, Turkey*

Received 24 September 2022; accepted 15 November 2022

Abstract: There is increasing interest in predicting traffic measures by modeling big data-driven complex scenarios with data mining and machine learning methods. In this study, the parameters of the traffic analysis model were created using 35,697 Twitter traffic notifications. The relationships and effects between the parameters of hour, day, month, season, year, lane, accident status, traffic events were revealed. By using the chi-square method, significant relationships were obtained between the day, time, month, season, and lane parameters of the traffic incidents on the D100 highway. Traffic events analysis machine learning methods of the D100 highway, which has a very important place in Istanbul traffic, were carried out. The traffic events prediction accuracy values of the created model were obtained as Naive Bayes 91.2%, Bayes 91.0% and Artificial Neural Network 93.1%. It has been concluded that accident events occur mostly on Fridays, vehicle breakdowns and maintenance-repair works occur mostly on Thursdays, accident events, vehicle breakdowns and maintenance-repair works mostly occur in the right lane. It was concluded that noon times as time, Thursday and Friday as days, January and July as months, winter season as season and right lane as lane are important parameters in terms of ensuring D100 road traffic safety.

Keywords: traffic analysis, data mining, machine learning, chi-square, twitter.

1. Introduction

Big data produced in Intelligent Transportation Systems, which is the future direction of the transportation system, is increasingly becoming the focus of research. Thanks to the big data obtained, it will have an important place in the design and implementation of safer, efficient, and profitable smart transportation systems. Big data obtained through smart cards, GPS, sensors, video detectors and social media in ITS can move ITS to a more efficient point with accurate and effective data analysis (Shi and Abdel-Aty, 2015). Since the amount and complexity of the data obtained in ITS is increasing day by day, the inadequacy of traditional data processing and analysis methods emerges. Big data analysis provides new techniques and methods to ITS in solving this problem. By increasing the processing efficiency of data in ITS with big data analysis, analysis of current and historical data can provide great benefits in terms of real-time traffic flow forecasting, public transport service planning, optimal route planning, traffic accident event and location estimation. He stated that the data collected by electronic sensor technologies, data transmission technologies and smart

[1] Corresponding author: cihan.ciftci@ogr.iu.edu.tr

control technologies, which are among the advanced technology applications within the scope of smart transportation systems, have the characteristics of speed, volume, and diversity, which are the characteristics of big data. It is stated in the literature that smart cities have a rich variety of data sources and that these data sources can be obtained directly from various devices, networks, applications, data providers, information systems such as Intelligent Transportation.

Machine learning is increasingly used in daily life. It is used to increase productivity, detect diseases, image and voice recognition, product recommendations, sentiment analysis, social media, and detect fraud and malicious situations. Intelligent transportation systems have gained an important place in the analysis of traffic forecasts in recent times. It is used to predict the traffic density that may occur and to determine the density and congestion that may occur. By using daily and historical data, important results have been obtained in determining the regions where congestion may occur. Traffic forecasting has become a very important issue recently in order to improve the advance planning and management of users and decision makers in Intelligent Transportation Systems, cope with traffic congestion, and predict near future traffic measures based on current and future traffic data. With the emergence of data and increasing computational resources, it has become a superior approach for traffic forecasting where different modeling approaches are introduced (Vlahogianni *et al.,* 2014). It focused on predicting traffic in a single location using traffic theory models and classical statistical methods as traffic forecasting methods. As a result of the formation of large amounts of traffic data with Intelligent Transportation Systems, it has led to the creation of data-oriented models. As computational capacities and big data processing methods increase, the development of new technologies and techniques for processing bulk data, the interest in Machine Learning methods, which can achieve complex scenarios, has increased intensely, leaving traditional approaches to forecasting continuous traffic measures based on invisible data (Lana *et al.,* 2018).

As the increasing population and the number of vehicles that come along with it cannot be dealt with efficiently, the problem of traffic congestion has gradually increased. Overcoming, predicting, and dealing with direct or indirect traffic problems, which are becoming an increasingly big problem in smart and big cities around the world, has attracted a great deal of attention recently. Reducing traffic congestion, fuel consumption, number of accidents, waiting time, air pollution, as well as carrying out studies that will help road users to make better decisions by informing them about the road situation in advance are among the studies.

In this study, the analysis of vehicle density, traffic events and speed values of the Istanbul D100 highway was carried out. Big data provided by Istanbul Metropolitan Municipality (IMM, 2020) has been processed and organized. The effects and relationships of non-traffic parameters affecting vehicle density and traffic incidents on the D100 highway were examined. In the analysis of traffic incidents, 35,697 accidents that occurred on the D100 highway were digitized as day, month, year, season, damaged, chain, injury, density, left, right, middle lanes. Chi-square tests were used to determine whether there is a statistical relationship between the classification

variables obtained and traffic incidents such as vehicle breakdown, maintenance-repair work, and accident notification. As a result of statistical analysis, after determining the non-traffic parameters and their effects on the traffic of the D100 highway, machine learning methods were used to predict traffic events. With these models, the traffic events that will take place on the D100 highway will be predicted in advance and will make a great contribution to taking precautions.

2. Literature Review

The fact that transportation data is collected from social media is a new field that has a great importance in overcoming the needs and perspectives of users and its use in transportation planning, management and control is becoming an important traffic data source in the literature. Traffic data sources are collected in six categories (Dabiri and Heaslip, 2019):

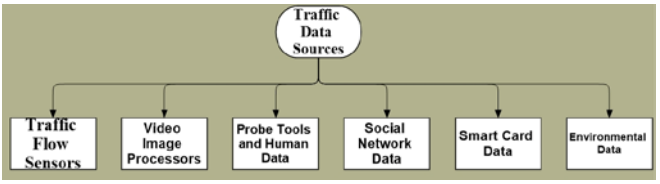

Fig. 1.
Classification of Traffic Data Sources

Depending on the intended use of social media in the field of transport, it is divided into the following categories (Nikolaidou and Papaioannou, 2018):

- incident detection;
- understanding patterns of mobility and activity;
- measuring driver satisfaction;
- information sharing and communication by transport authorities and public institutions;
- conducting transportation research.

They stated that text mining and data mining techniques should be used to extract useful information from social media data in order to better understand, predict and guide human behavior (Abbasi *et al.*, 2015; Maghrebi *et al.,* 2015).

Fu (2015) explored the applicability of social media data to detect traffic incidents. They developed an approach to extract and analyze traffic incident information with data from Twitter. Steuer (2015) concluded that using Twitter data would be a useful source of information for incident detection and event transmission to road users. Daly *et al.* (2013) predicted that social media can be used to visualize current traffic conditions by collecting useful information. They combined dynamic data from social media to update users about traffic accidents and major incidents. Wanichayapong *et al.* (2011) extracted traffic information such as traffic jams and weather events from Twitter, classifying traffic news and associating events with spatial dimensions. As a result of the classification, it was stated that it will help users plan their routes and avoid traffic jams in real time. Abel *et al.* (2012) created a framework to automatically filter, search, and analyze information about emergency events posted on Twitter. Kumar *et al.* (2014) was used to detect bad road conditions

using Twitter data. A sentiment analysis was conducted by collecting tweets about traffic and road hazards. Mai and Hranac (2013) conducted a sensitivity analysis by collecting information about the causes of traffic jams from social media. Social media data was used to determine the activities and spatial mobility of users in transportation research. Hasan *et al.* (2013) examined urban human mobility and activities from location-based data of Twitter users. Chaniotakis *et al.* (2015) explored the potential to use data collected from Twitter for leisure activities demand modelling. Cheng *et al.* (2011) evaluated human mobility models with social media data and analyzed their temporal, spatial and social aspects. Schweitzer (2012) stated that sentiment analysis of Twitter's data can be applied to evaluate opinions about the quality of transportation services, to determine trends in user satisfaction and real-time events. Collins *et al.* (2013) a sensitivity analysis of Twitter data was used to evaluate the satisfaction of users using transit systems. Luong and Houston (2015) examined the attitudes of users from Twitter regarding rail transport services using sentiment analysis.

Gal-Tzur *et al.* (2014) classification of Twitter data into different categories was carried out to identify the need for transport services and to report transport-related events. Hassan and Ukkusuri (2014) proposed a classification-based travel mobility model approach, using social media data and taking advantage of users' location data. Ni *et al.* (2017) social media data was used for metro passenger flow prediction. They developed a prediction model by collecting Twitter data. Cottrill *et al.* (2017) it is aimed to share the data about transportation from the data coming from social media platforms and to ensure the coordination of the public. By analyzing Twitter data, they indicated the potential of timely information sharing to passengers in transportation. Maghrebi *et al.* (2015) using social media data, travel mode selection was made. By analyzing Twitter data and travel mode selection, it was suggested that social media is a potential application area in transportation. Huang *et al.* (2017), analysis and visualization methods of social media data in destination or route selection problems were investigated through spatial and statistical analyzes using Twitter data. Rashidi *et al.* (2017) investigated the size of social media data in the field of transportation. Applications on subjects such as the purpose of travel, mode of transportation, choice of destination are examined. Xu *et al.* (2019) conducted a case study using Twitter data in Toronto, Canada, where there is a large amount of traffic-related data from social media data. By analyzing Twitter data, traffic events classification was carried out based on association rules and machine learning methods. Rahman *et al.* (2019) proposes an analysis of data collected from Twitter account to share real-time traffic information. In order to increase the efficiency of real-time traffic information sharing over social media, a social media-based adaptive real-time traffic feed (SMART-Feed) model has been developed with various measurements. Dabiri and Heaslip (2019) used Twitter data by classifying it to detect traffic events and follow traffic conditions. The model is proposed using convolutional neural network (CNN) and recurrent neural network (RNN). Huang *et al.* (2019), a solution proposal was developed by extracting information about urban traffic from social media data and investigating the potential effect of human activities on daily traffic congestion.

Many studies have been carried out in the literature to develop a prediction mechanism that will predict real-time traffic flow and to increase its accuracy, scalability, and applicability. Fancello *et al.* (2018) developed Poisson and Negative Binomial models after applying cluster analysis to accident data to increase road safety. Ma and Yuan (2018) developed a model based on Poisson regression, negative binomial (NB) regression and Zero Inflated Negative Binomial (NINB) regression to reduce the impact of traffic accidents, reduce the number of traffic accidents and increase road safety. The relationship between the number of traffic accidents and factors such as road length and traffic conditions were examined. Gu *et al.* (2018), mutation optimization back propagation neural network prediction model, particle swarm optimization support vector machine model, support vector machine, back propagation neural network, K Nearest Neighbor (K-NN) and Bayesian network methods were used in the prediction model of traffic accident deaths. Contreras *et al.* (2018), Maximum Sensitivity Neural Network method was used to predict traffic accidents. You *et al.* (2017), paired case control method and support vector machines (SVMs) were used to model the risk situation in traffic accidents. The potential impact of weather data on the accident prediction model has been revealed. Taamneh et al. (2017) used Decision Tree (J48), Rule Induction (PART), Naive Bayes and Multilayer Perceptron methods to model the severity of injuries in traffic accidents. Jadaan *et al.* (2014) used the Artificial Neural Network (ANN) method to predict traffic accidents. Ramani and Shanthi (2012), Random Tree, C4.5, J48 and Decision Stump methods, which are Decision Tree algorithms for the classification of road traffic accidents, were applied to the fatal accident database that occurred in Great Britain in 2010. Fu and Zhou (2011) performed traffic accident prediction using artificial neural networks. An artificial neural network model has been developed due to the presence of many non-linear elements such as people, cars, roads, and climate in traffic accidents. Zhang *et al.* (2018), using social media data, used deep learning method to detect traffic accidents. The content of 1 million 300 thousand tweets was examined in detail.

3. Methodology

In recent years, the emergence of big data technology and the successful application of neural network algorithms offer opportunities to conduct data-driven research in transportation. Advances in artificial intelligence give researchers the chance to quickly and deeply learn features from traffic data. Thus, predicting traffic congestions with large and heterogeneous traffic data in both short and long term offers opportunities for improvement using data mining and machine learning approach. In this study, Artificial Neural Networks and Bayesian Networks from Machine Learning methods were used to establish an early warning system to detect the bottleneck situation in advance.

3.1. Problem State of the Research

Traffic congestion is a situation that causes the travel times of vehicles moving to their destinations to increase due to the increase in the number of vehicles in traffic and the decrease in their speed. As the increasing population and the number of vehicles that come along with it cannot be dealt with efficiently, the problem of traffic congestion has gradually increased.

Big data produced in Intelligent Transportation Systems, which is the future direction of the transportation system, is increasingly becoming the focus of research. Thanks to the big data obtained, it will have an important place in the design and implementation of safer, efficient and profitable smart transportation systems.

Overcoming, predicting and dealing with direct or indirect traffic problems, which are becoming an increasingly big problem in smart and big cities around the world, has recently received a great deal of attention.

The emergence of big data technology and the successful application of neural network algorithms offer opportunities to conduct data-based research in transportation.

3.2. Purpose of the Research

The main purpose of our study is to carry out an in-depth analysis of Istanbul traffic, to determine the parameters affecting the traffic and to model and analyze the effects. For this purpose, by using social media and sensor data, the parameters affecting the events in Istanbul traffic were revealed, and a traffic events estimation model was created with machine learning methods, thanks to the parameters obtained. Digitization of traffic incidents was performed using 35,697 traffic incident message data. Relationships of traffic incidents with criteria such as day, time, month, season and lane were determined. The analysis of traffic incidents was determined by the chi-square method. Significant relationships were obtained between these parameters and traffic events. In this study, the relationship of 35,697 traffic incidents that occurred in Istanbul traffic between 2016-2020 with criteria such as hour, day, month, year, season, vehicle density was analyzed using the chi-square method. Prediction modeling was done with machine learning methods according to the obtained parameters.

3.3. Research Data

It was obtained from the data with open access in the Istanbul Metropolitan Municipality Data Portal. Instant notifications of traffic incidents on Twitter are given. Data of 35,697 traffic incidents between 2016-2020 on the D100 highway were obtained. To give an example of the messages received from Twitter regarding traffic incidents: *"D100 15 July Martyrs Bridge Europe-Anatolia Direction, right lane vehicle failure". "Traffic is heavy in the region, D100 Maltepe-Küçükyalı Direction, the traffic in the region is concentrated due to the vehicle malfunction in the right lane". "D100 Maltepe-Küçükyalı Direction, traffic in the region intensified due to vehicle failure in the middle lane". The types of accident events are grouped as follows: Accident Notification, Vehicle Failure, Maintenance-Repair Work. In the classification process, days are defined as Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday. Months: January, February, March, April, May, June, July, August, September, October, November, December. The seasons are defined as: Spring, Summer, Autumn, Winter. The times of accident events are grouped as follows: Morning, Noon, Evening, Nigh. Accident events are grouped as follows: damaged: 1, not damaged: 0, chained: 1, not chained: 0, injured: 1, not injured: 0, intensity: 1, no intensity: 0. The lane where the accident events occurred is grouped as follows: right lane: 1, if not right lane: 0, left lane: 1, if not left lane: 0, middle lane: 1, if not middle lane: 0.*

Table 1

Digitization of Social Media Traffic Messages

Months	Days	Times	Event Type	Damaged	Chaining	Injury	Intensity	Left Lane	Right Lane	Center Lane
January	Monday	Noon	Accident Notification	0	1	0	0	1	0	0
February	Tuesday	Morning	Vehicle Failure	0	0	0	1	0	1	0
March	Wednesday	Noon	Vehicle Failure	0	0	0	1	0	1	0
April	Thursday	Noon	Maintenance-Repair Work	0	0	0	0	0	1	0
May	Friday	Morning	Accident Notification	1	0	0	0	1	0	0
June	Saturday	Evening	Maintenance-Repair Work	0	0	0	0	0	1	0
July	Sunday	Night	Vehicle Failure	0	0	0	1	0	1	0
August	Monday	Evening	Maintenance-Repair Work	0	0	0	0	0	1	0
September	Tuesday	Noon	Accident Notification	1	0	0	0	1	0	0
October	Wednesday	Morning	Vehicle Failure	0	0	0	1	0	1	0
November	Thursday	Night	Accident Notification	0	1	0	0	1	0	0
December	Friday	Night	Vehicle Failure	0	0	0	1	0	1	0

3.4. Model of Research

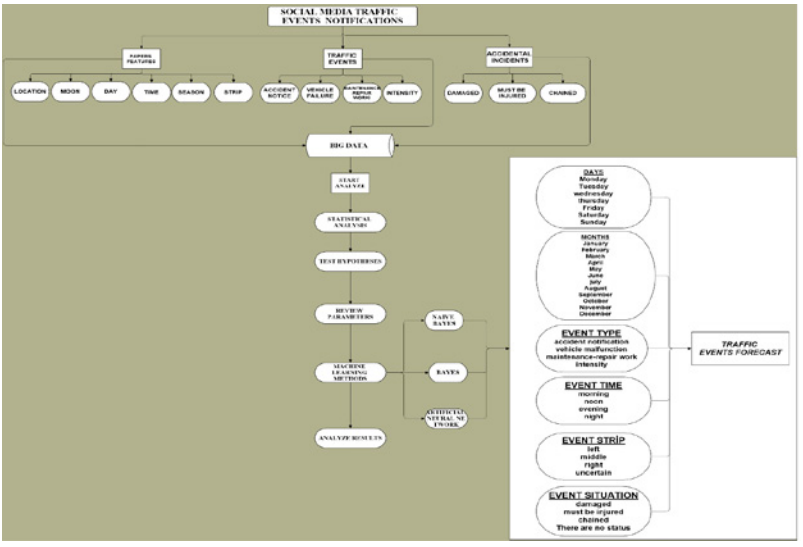

Fig. 2.

Model of the Study

3.5. Research Hypotheses

Istanbul's D100 road accident incidents:

- H^1: There is a significant difference in terms of hours;
- H^2: There is a significant difference in terms of days;
- H^3: There is a significant difference in terms of months;
- H^4: There is a significant difference in terms of stripes;
- H^5: There is a significant difference in terms of seasonal.

4. Data Analysis

As a result of the statistical analyzes carried out, significant relationships were obtained in terms of months, days, seasons and lanes with vehicle breakdown and accident events in 35,697 traffic notification messages that occurred on the D100 highway line. In line with the results obtained, machine learning methods were used in modeling the traffic density situation of the D100 highway line and possible traffic events.

Table 2

Chi-Square Test Results Regarding Traffic Events Occurring According to Hours

Watches	Vehicle Failure	*f*	Maintenance Study	*f*	Accident Notice	*f*	Total	*f*	X^2	*p*
Morning	5,808	31.67%	971	39.98%	4,482	30.02%	11,261	31.55%		
Noon	6,749	36.80%	715	29.44%	5,403	36.19%	12,867	36.05%		
Evening	5,276	28.77%	241	9.92%	4,025	26.96%	9,545	26.74%	1653.957[a]	.000
Night	507	2.76%	502	20.67%	1,018	6.82%	2,030	5.69%		
Total	18,340	51.38%	2,429	6.80%	14,928	41.82%	35,697	100.00%		

According to the results obtained in Table 2, 51.35% of the 35,697 traffic incidents occurring in Istanbul traffic are vehicle breakdowns, 6.80% road maintenance-repair works and 41.82% accident notifications. In addition, 31.55% of the traffic incidents occurred in the morning, 36.05% at noon, 26.74% in the evening and 5.69% at night. The highest vehicle breakdown time was 36.80% at noon, road maintenance-repair work was 39.98% in the morning, and the highest accident reporting time was at noon with 36.19%. According to the results obtained (X^2=*1653.957*[a], *p=0,000<0,005*), it was determined that there was a significant relationship between traffic events and hours, and the H^1 hypothesis was accepted as a result.

Table 3

Chi-Square Test Results for Traffic Events Occurring by Days

Days	Vehicle Failure	*f*	Maintenance Study	*f*	Accident Notice	*f*	*Total*	*f*	X^2	*p*
Monday	3,175	17.31%	295	12.14%	2,284	15.30%	5,754	16.12%	274.379[a]	.000
Tuesday	2,966	16.17%	383	15.77%	2,225	14.90%	5,574	15.61%		
Wednesday	2,956	16.12%	372	15.31%	2,261	15.15%	5,589	15.66%		
Thursday	3,036	16.55%	389	16.01%	2,297	15.39%	5,722	16.03%		
Friday	2,998	16.35%	358	14.74%	2,406	16.12%	5,762	16.14%		
Saturday	1,918	10.46%	325	13.38%	1,833	12.28%	4,076	11.42%		
Sunday	1,291	7.04%	307	12.64%	1,622	10.87%	3,220	9.02%		
Total	18,340	51.38%	2,429	6.80%	14,928	41.82%	35,697	100.00%		

According to the results obtained in Table 3, vehicle breakdown was 16.55% on Thursday, maintenance-repair work was 16.01% on Thursday, and accident notification was 16.12% on Friday. Most of the traffic incidents occurred on Friday with 16.14%. According to the results obtained (X^2=274.379[a], p=0,000<0,005), it was determined that there was a significant relationship between traffic events and days, and the H^2 hypothesis was accepted as a result.

Table 4

Chi-Square Test Results for Traffic Events by Month

Months	Vehicle Failure	*f*	Maintenance Study	*f*	Accident Notice	*f*	*Total*	*f*	X^2	*p*
January	2,013	10.98%	246	10.13%	1,605	10.75%	3,864	10.82%	181.168[a]	0,000
February	1,723	9.39%	228	9.39%	1,314	8.80%	3,265	9.15%		
March	1,683	9.18%	220	9.06%	1,428	9.57%	3,331	9.33%		
April	946	5.16%	239	9.84%	961	6.44%	2,146	6.01%		
May	962	5.25%	128	5.27%	909	6.09%	1,999	5.60%		
June	1,729	9.43%	153	6.30%	1,251	8.38%	3,133	8.78%		
July	1,952	10.64%	287	11.82%	1,403	9.40%	3,642	10.20%		
August	1,376	7.50%	176	7.25%	1,178	7.89%	2,730	7.65%		
September	1,535	8.37%	193	7.95%	1,296	8.68%	3,024	8.47%		
October	1,597	8.71%	246	10.13%	1,302	8.72%	3,145	8.81%		
November	1,348	7.35%	149	6.13%	1,197	8.02%	2,694	7.55%		
December	1,476	8.05%	164	6.75%	1,084	7.26%	2,724	7.63%		
Total	18,340	51.38%	2,429	6.80%	14,928	41.82%	35,697	100.00%		

According to the results obtained in Table 4, vehicle breakdown occurred with 10.98% in January, maintenance-repair with 11.82% in July and accident notification with 10.75% in January. Most of the traffic incidents occurred in January with 10.82%. According to the results obtained ($X^2=181.168^a$, $p=0.000<0.005$), it was determined that there was a significant relationship between traffic events and months, and the H^3 hypothesis was accepted as a result.

Table 5

Chi-Square Test Results Regarding Traffic Events Occurring According to Lane

Lane	Vehicle Failure	*f*	Maintenance Study	*f*	Accident Notice	*f*	Total	*f*	X^2	*p*
Left	2,172	12.41%	879	40.01%	4,959	38.05%	8,010	24.47%	3346.568[a]	0,000
Middle	1,235	7.06%	11	0.50%	1,240	9.51%	2,486	7.60%		
Right	14,094	80.53%	1,307	59.49%	6,835	52.44%	22,236	67.93%		
Total	17,501	53.47%	2,197	6.71%	13,034	39.82%	32,732	100.00%		

According to the results obtained in Table 5, vehicle breakdown occurred in the right lane with 80.53%, maintenance-repair work with 59.49% and vehicle accidents with 52.44% in the right lane. Most of the traffic incidents occurred in the right lane with 67.93%. According to the results obtained ($X^2=3346.568^a$, $p=0.000<0.005$), it was determined that there is a significant relationship between traffic events and lane, and H^4 hypothesis was accepted as a result.

Table 6

Chi-Square Test Results Regarding Traffic Events Occurring According to Seasons

Seasons	Vehicle Failure	*f*	Maintenance Study	*f*	Accident Notice	*f*	Total	*f*	X^2	*p*
Spring	3,591	19.58%	587	24.17%	3,298	22.09%	7,476	20.94%	63.542[a]	0,000
Summer	5,057	27.57%	616	25.36%	3,832	25.67%	9,505	26.63%		
Autumn	4,480	24.43%	588	24.21%	3,795	25.42%	8,863	24.83%		
Winter	5,212	28.42%	638	26.27%	4,003	26.82%	9,853	27.60%		
Total	18,340	51.38%	2,429	6.80%	14,928	41.82%	35,697	100.00%		

According to the results obtained in Table 6, 24.42% vehicle breakdown occurred in winter, 26.72% maintenance-repair in winter, and 26.82% winter vehicle accident. Most of the traffic incidents occurred in winter with 27.60%. According to the results obtained ($X^2=63.542^a$, $p=0.000<0.005$), it was determined that there was a significant relationship between traffic events and seasons, and the H^5 hypothesis was accepted as a result.

Classification analyzes were carried out according to days, time, months, lane, season to estimate traffic events that may occur on the D100 highway. Classification analyzes of 35,697 traffic events were performed with Navie Bayes, Bayes and Artificial Neural Networks. According to the results obtained in Table 7, the density estimation accuracy of the created model was obtained as Navie Bayes, 91.2%, Bayes 91.0% and Artificial Neural Networks as 93.1%.

Table 7
Traffic Events Estimation Results with Machine Learning Methods

Model Evaluation Criteria		Models					
		Naive Bayes		Bayes		Artificial Network	
		Cross Validation-10	Percentage Split (70/30)	Cross Validation-10	Percentage Split (70/30)	Cross Validation-10	Percentage Split (70/30)
Traffic Events Estimation Models	Accuracy	0.912	0.911	0.910	0.911	0.931	0.931
	ROC	0.980	0.967	0.967	0.978	0.989	0.989
	MCC	0.864	0.838	0.864	0.860	0.891	0.891
	Rated((Recall)	0.910	0.911	0.912	0.911	0.931	0.931
	Precision	0.920	0.911	0.912	0.916	0.928	0.930
	F-criterion	0.914	0.900	0.900	0.912	0.927	0.925
	Kappa statistics	0.855	0.843	0.846	0.846	0.880	0.881
	mean absolute error	0.058	0.056	0.052	0.059	0.047	0.047
	Root mean squared error	0.180	0.187	0.185	0.181	0.154	0.154
	Relative absolute error	20.10%	18.37%	18.14%	20.32%	16.11%	16.12%
	Root relative squared error	47.34%	49.22%	48.73%	47.63%	40.38%	40.36%

4.1. Discussion of Results

The collection of transportation data from social media is a new field with great importance in overcoming the needs and perspectives of users, and its use in transportation planning, management and control is becoming an important traffic data source in the literature. Useful data can be obtained for the detection of many events (Grant-Muller *et al.,* 2015; Pender *et al.,* 2014). In the literature, Twitter data has been used to detect traffic events, to extract the algorithm of traffic events, and to investigate the causes of traffic jams (Daly *et al.,* 2013; Mai and Hranac, 2013; Fu, 2015; Steuer, 2015; Zhang *et al.,* 2018; Xu *et al.,* 2019; Dabiri and Heaslip, 2019). In this study, traffic analysis modeling was carried out using Twitter traffic notification data. Firstly, Chi-square tests were used to evaluate the traffic incidents that occurred in Istanbul traffic between 2016-2020. Traffic incidents were evaluated in three categories as accident notification, road maintenance-repair works and vehicle breakdown. It has been evaluated whether there is a relationship between parameters such as time, day, month, season and traffic events and whether traffic events are independent of these parameters. According to the results obtained in table 2 (X^2=1653.957^a, p=0.000<0.005), it was determined that there is a significant relationship between traffic events and hours. 51.35% of the 35,697 traffic incidents occurring in the D100 highway line traffic are vehicle malfunctions, 6.80% road maintenance-repair works and 41.82% accident notifications. In addition, 31.55% of the traffic incidents occurred in the morning, 36.05% at noon, 26.74% in the evening and 5.69% at night. The highest vehicle breakdown time was at noon with 36.80%, road maintenance-repair work was in the morning with 39.98%, and the highest accident reporting time was at noon with 36.19%. According to the results obtained in table 3 (X^2=274.379^a, p=0.000<0.005), it was determined that there was a significant relationship between traffic events and days. In the D100 highway traffic, the highest vehicle breakdown was 16.55% on Thursday, maintenance-repair work was 16.01% on Thursday, and accident notification was

16.12% on Friday. Most of the traffic incidents occurred on Friday with 16.14%. According to the results obtained in table 4 ($X^2=181.168^a$, $p=0.000<0.005$), it was determined that there is a significant relationship between traffic events and months. In the D100 highway traffic, the highest number of vehicle breakdowns occurred in January, with 10.98%, in July with 11.82%, with maintenance-repair and in January with 10.75%, with accident reporting. Most of the traffic incidents occurred in January with 10.82%. According to the results obtained in table 5 ($X^2=3346.568^a$, $p=0.000<0.005$), it was determined that there is a significant relationship between traffic incidents and the lane-repair work and vehicle accident occurred in the right lane with 52.44%. Most of the traffic incidents occurred in the right lane with 67.93%. According to the results obtained in table 6 ($X^2=63.542^a$, $p=0.000<0.005$), it was determined that there is a significant relationship between traffic events and seasons. In the D100 highway line traffic, 24.42% vehicle breakdown, 26.72% maintenance-repair in winter and 26.82% winter vehicle accident occurred. Most of the traffic incidents occurred in winter with 27.60%. Machine learning methods in Intelligent Transportation Systems have received increasing attention in recent years. Machine learning methods were also used in the analysis of traffic incidents (Zhang *et al.,* 2018; Fancello *et al.,* 2018; Ma and Yuan, 2018; Özden and Acı, 2018; Gu *et al.,* 2018; Contreras *et al.,* 2018; Alkheder *et al.,* 2017; You *et al.,* 2017; Taamneh *et al.,* 2017; Jadaan *et al.,* 2014; Ramani and Shanthi, 2012; Tayep *et al.,* 2015; Martin *et al.,* 2014; Fu and Zhou, 2011; Ren *et al.,* 2018). In this study, as a result of the statistical results obtained, secondly classification analyzes of traffic events with machine learning methods, according to months, days, time and lane parameters were carried out. As a result of Navie Bayes classification, 91.7% accuracy value, 82.2% sensitivity value, 99.9% conciseness values were obtained. As a result of Bayesian classification, 96.6% accuracy value, 95.0% sensitivity value, 97.9% conciseness values were obtained. As a result of Artificial Neural Network classification, 97.0% accuracy value, 95.7% sensitivity value, 98.1% conciseness values were obtained. In line with these results, important results will be obtained in the prediction of the density that may occur according to the months, days, time, lane and traffic events variables of the D100 highway line.

5. Conclusion

In this study, we performed traffic events modeling by analyzing the big data consisting of traffic notifications, which we obtained thanks to the Twitter social media tool, using machine learning methods. We used non-parametric statistical methods to determine the parameters we used in our model. In our study, we carried out the traffic analysis of the D100 highway, which has a very important place in Istanbul traffic, thanks to the model we built. For traffic analysis, the relationship between traffic events, hour, day, month, season, year, lane, it has any effects or not, whether it creates significant differences on traffic events, were tested with statistical methods. According to the results of the hypotheses, the parameters affecting the traffic of that region were determined. Our model was tested by performing classification and prediction analyzes with Artificial Neural Network, Bayesian Networks, which are machine learning methods. As a result of the analysis of the D100 road safety, a significant relationship was obtained between the

accident events, vehicle breakdowns and maintenance-repair works by months, days, time, season and years. While more vehicle breakdowns and maintenance-repair works occur in the morning hours, accident events occur in the afternoon. In addition, vehicle breakdowns and maintenance-repair works mostly occur on Thursdays, while accident events occur mostly on Fridays. It has been concluded that while the most vehicle breakdowns and accident events occur in January, maintenance-repair works mostly occur in the winter months. Accident events, maintenance-repair works and vehicle breakdown occurred mostly in the right lane. Noon hours as time, Thursday and Friday as days, January and July months as months, Winter season as season and right lane as lane are important parameters for ensuring road traffic safety on the D100 road. Thanks to the D100 Highway Traffic Information Service, month, day, time and lane conditions will be entered and measures will be taken according to the density situations that will occur along the line and the probability of traffic events, and it will make a great contribution to the formation of a more reliable traffic. By integrating this system with navigation programs, informing the users about the density and traffic events of the location before they reach the next location will prevent traffic incidents that may occur and naturally create an obstacle in the occurrence of densities.

Acknowledgement

This study was produced from the doctoral thesis of "Analysis of Istanbul Traffic with Data Mining and Machine Learning: D100 Highway Application" prepared by the first author under the supervision of the second author.

References

Abbasi, A.; Rashidi, T. H.; Maghrebi, M.; Waller, S. T. 2015. Utilising location based social media in travel survey methods: Bringing twitter data into the play. In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN 2015 - Held in Conjunction with ACM SIGSPATIAL 2015*, 1-9 https://doi.org/10.1145/2830657.2830660.

Abel, F.; Hauff, C.; Houben, G. J.; Tao, K.; Stronkman, R. 2012. Twitcident: Fighting fire with information from Social Web streams. In *Proceedings of the 21st Annual Conference on World Wide Web Companion WWW'12*, 305–308. https://doi.org/10.1145/2187980.2188035.

Alkheder, S.; Taamneh, M.; Taamneh, S. 2017. Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting*, *36*(1), 100-108.

Chaniotakis, E.; Antoniou, C.; Mitsakis, E. 2015. Data for Leisure Travel Demand from Social Networking Services. In *4th hEART Symposium (European Association for Research in Transportation)*, 1-10.

Cheng, Z.; Caverlee, J.; Lee, K.; Sui, D. Z. 2011. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 81-88. https://doi.org/papers3://publication/uuid/0C46BD5D-4908-4A8A-BD06-5BCB2F1DE282.

Collins, C.; Hasan, S.; Ukkusuri, S. V. 2013. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data, *Journal of Public Transportation* 16(2): 21–45. https://doi.org/10.5038/2375-0901.16.2.2.

Contreras, E.; Torres-Treviño, L.; & Torres, F. 2018. Prediction of car accidents using a maximum sensitivity neural network, *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, *LNICST* 213: 86–95. https://doi.org/10.1007/978-3-319-73323-4_9.

Cottrill, C.; Gault, P.; Yeboah, G.; Nelson, J. D.; Anable, J.; Budd, T. 2017. Tweeting Transit: An examination of social media strategies for transport information management during a large event, *Transportation Research Part C: Emerging Technologies* 77: 421–432. https://doi.org/10.1016/j.trc.2017.02.008.

Dabiri, S.; Heaslip, K. 2019. Developing a Twitter-based traffic event detection model using deep learning architectures, *Expert Systems with Applications* 118: 425–439. https://doi.org/10.1016/j.eswa.2018.10.017.

Daly, E. M.; Lecue, F.; Bicer, V. 2013. Westland row why so slow? Fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the International Conference on Intelligent User Interfaces,* 203–212. https://doi.org/10.1145/2449396.2449423.

Fancello, G.; Soddu, S.; Fadda, P. 2018. An accident prediction model for urban road networks, *Journal of Transportation Safety and Security* 10(4): 387–405. https://doi.org/10.1080/19439962.2016.1268659.

Fu, K. 2015. Social Media Analysis For Traffic Incident Detection And Management. Transportation Research Board 94th Annual Meeting, 1–10.

Fu, H.; Zhou, Y. 2011. The traffic accident prediction based on neural network. In *Proceedings of the 2nd International Conference on Digital Manufacturing and Automation, ICDMA 2011*, 1349–1350. https://doi.org/10.1109/ICDMA.2011.331.

Gal-Tzur, A.; Grant-Muller, S. M.; Kuflik, T.; Minkov, E.; Nocera, S.; Shoor, I. 2014. The potential of social media in delivering transport policy goals, *Transport Policy* 32: 115–123. https://doi.org/10.1016/j.tranpol.2014.01.007.

Grant-Muller, S. M.; Gal-Tzur, A.; Minkov, E.; Nocera, S.; Kuflik, T.; Shoor, I. 2015. Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems*, *9*(4), 407-417.

Gu, X.; Li, T.; Wang, Y.; Zhang, L.; Wang, Y.; Yao, J. 2018. Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization, *Journal of Algorithms and Computational Technology* 12(1): 20–29. https://doi.org/10.1177/1748301817729953.

Hasan, S.; Ukkusuri, S. V. 2014. Urban activity pattern classification using topic models from online geo-location data, *Transportation Research Part C: Emerging Technologies* 44: 363–381. https://doi.org/10.1016/j.trc.2014.04.003.

Hasan, S.; Zhan, X.; Ukkusuri, S. V. 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1-8. https://doi.org/10.1145/2505821.2505823.

Huang, A.; Gallegos, L.; Lerman, K. 2017. Travel analytics: Understanding how destination choice and business clusters are connected based on social media data, *Transportation Research Part C: Emerging Technologies* 77: 245–256. https://doi.org/10.1016/j.trc.2016.12.019.

Huang, W.; Xu, S.; Yan, Y.; Zipf, A. 2019. An exploration of the interaction between urban human activities and daily traffic conditions: A case study of Toronto, Canada, *Cities* 84: 8–22. https://doi.org/10.1016/j.cities.2018.07.001.

IMM Open Data. 2020. Available at: https://data.ibb.gov.tr/en/ (Accessed: 01 January 2020).

Jadaan, K. S.; Al-Fayyad, M.; Gammoh, H. F. 2014. Prediction of Road Traffic Accidents in Jordan using Artificial Neural Network (ANN), *Journal of Traffic and Logistics Engineering* 2(2): 92–94. https://doi.org/10.12720/jtle.2.2.92-94.

Kumar, A.; Jiang, M.; Fang, Y. 2014. Where not to go? Detecting road hazards using Twitter. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 2014*, 1223–1226. https://doi.org/10.1145/2600428.2609550.

Lana, I.; Del Ser, J.; Velez, M.; Vlahogianni, E. I. 2018 Road Traffic Forecasting: Recent Advances and New Challenges, *IEEE Intelligent Transportation Systems Magazine* 10(2): 93–109. https://doi.org/10.1109/MITS.2018.2806634.

Luong, T. T. B.; Houston, D. 2015. Public opinions of light rail service in Los Angeles, an analysis using Twitter data. In Proceedings of the *IConference*, 1–4.

Ma, W.; Yuan, Z. 2018. Analysis and Comparison of Traffic Accident Regression Prediction Model. In *3rd International conference on electromechanical control technology and transportation,* 364–369. https://doi.org/10.5220/0006970803640369.

Maghrebi, M.; Abbasi, A.; Rashidi, T. H.; Waller, S. T. 2015.Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. In *Proceedings of the IEEE Conference on Intelligent Transportation Systems, ITSC*, 2015-Octob, 208–213. https://doi.org/10.1109/ITSC.2015.43.

Mai, E.; Hranac, R. 2013. Twitter Interactions as a Data Source for Transportation Incidents. In *Proceedings of the 92nd Annual Meeting Transportation Research Board*. 11 p. Available from Internet: <http://docs.trb.org/prp/13-1636.pdf.

Martín, L.; Baena, L.; Garach, L.; López, G.; de Oña, J. 2014. Using Data Mining Techniques to Road Safety Improvement in Spanish Roads. *Procedia - Social and Behavioral Sciences*, 160, 607–614. https://doi.org/10.1016/j.sbspro.2014.12.174.

Ni, M.; He, Q.; Gao, J. 2017. Forecasting the Subway Passenger Flow under Event Occurrences with Social Media, *IEEE Transactions on Intelligent Transportation Systems* 18(6): 1623–1632. https://doi.org/10.1109/TITS.2016.2611644.

Nikolaidou, A.; Papaioannou, P. 2018. Utilizing Social Media in Transport Planning and Public Transit Quality: Survey of Literature, *Journal of Transportation Engineering, Part A: Systems* 144(4): 04018007. https://doi.org/10.1061/jtepbs.0000128.

Özden, C.; Acı, Ç. 2018. Analysis of injury traffic accidents with machine learning methods: Adana case. *Pamukkale University Journal of Engineering Sciences*, *24*(2), 266-275.

Pender, B.; Currie, G.; Delbosc, A.; Shiwakoti, N. 2014. Social media use during unplanned transit network disruptions: A review of literature. *Transport Reviews*, *34*(4), 501-521.

Rahman, R.; Roy, K. C.; Abdel-Aty, M.; Hasan, S. 2019. Sharing real-time traffic information with travelers using twitter: An analysis of effectiveness and information content *Frontiers in Built Environment* 5(83): 1-15. https://doi.org/10.3389/fbuil.2019.00083.

Ramani, R. G.; Shanthi, S. 2012. Classifier prediction evaluation in modeling road traffic accident data. In *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2012*, 1-4.

Rashidi, T. H.; Abbasi, A.; Maghrebi, M.; Hasan, S.; Waller, T. S. 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges, *Transportation Research Part C: Emerging Technologies* 75: 197–211. https://doi.org/10.1016/j.trc.2016.12.008.

Ren, H.; Song, Y.; Wang, J.; Hu, Y.; Lei, J. 2018. A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2018-November, 3346–3351. https://doi.org/10.1109/ITSC.2018.8569437.

Schweitzer, L. 2012. How are We Doing? Opinion Mining Customer Sentiment in US Transit Agencies and Airlines via Twitter. In *Proceedings of the Transportation Research Board 91st Annual Meeting.* 16 p. Available from Internet: <http://trid.trb.org/view.aspx?id=1129878>.

Shi, Q.; Abdel-Aty, M. 2015. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways, *Transportation Research Part C: Emerging Technologies* 58: 380–394. https://doi.org/10.1016/j.trc.2015.02.022.

Steur, R.J. 2015. Twitter as a spatio-temporal source for incident management. Master's thesis. https://dspace.library.uu.nl/handle/1874/303174.

Taamneh, M.; Alkheder, S; Taamneh, S. 2017. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates, *Journal of Transportation Safety and Security* 9(2): 146–166. https://doi.org/10.1080/19439962.2016.1152338.

Tayeb, A. A. El.; Pareek, V.; & Araar, A. 2015. Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai. *International Journal of Soft Computing and Engineering (IJSCE)*, (4), 2231–2307. Retrieved from http://www.ijsce.org/wp-content/uploads/papers/v5i4/D2679095415.pdf.

Vlahogianni, E. I.; Karlaftis, M. G.; Golias, J. C. 2014. Short-term traffic forecasting: Where we are and where we're going, *Transportation Research Part C: Emerging Technologies* 43: 3–19. https://doi.org/10.1016/j.trc.2014.01.005.

Wanichayapong, N.; Pruthipunyaskul, W.; Pattara-Atikom, W.; Chaovalit, P. 2011. Social-based traffic information extraction and classification. In *Proceedings of the 11th International Conference on ITS Telecommunications, ITST* 2011, 107–112. https://doi.org/10.1109/ITST.2011.6060036.

Xu, S.; Li, S.; Wen, R.; Huang, W. 2019. Traffic event detection using Twitter data based on associated rules, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 4(2/W5): 543–547. https://doi.org/10.5194/isprs-annals-IV-2-W5-543-2019.

You, J.; Wang, J.; Guo, J. 2017. Real-time crash prediction on freeways using data mining and emerging techniques, *Journal of Modern Transportation* 25(2): 116–123. https://doi.org/10.1007/s40534-017-0129-7.

Zhang, Z.; He, Q.; Gao, J.; Ni, M. 2018. A deep learning approach for detecting traffic accidents from social media data, *Transportation Research Part C: Emerging Technologies* 86: 580–596. https://doi.org/10.1016/j.trc.2017.11.027.