IATSS Research 46 (2022) 329–344

Contents lists available at ScienceDirect

IATSS Research

Research Article

Road traffic conditions in Kenya: Exploring the policies and traffic cultures from unstructured user-generated data using NLP

Joseph Muguro [a,b,*], Waweru Njeri [b], Kojiro Matsushita [a], Minoru Sasaki [a,**]

[a] Department of Mechanical Engineering, Gifu University, Yanagido 1-1, Gifu, Japan
[b] School of Engineering, Dedan Kimathi University of Technology, 657-10100 Nyeri, Kenya

ARTICLE INFO

Article history:
Received 17 September 2021
Received in revised form 11 January 2022
Accepted 3 March 2022
Available online 8 March 2022

Keywords:
Road traffic accidents (RTA)
Natural language processing (NLP)
Road safety
Twitter
PSVs
Kenya
Matatu
NTSA

ABSTRACT

Road traffic accidents (RTA) are a prevalent cause of fatality with African countries having the highest fatality index (25–34 per quota). The World Health Organization estimates Kenya's fatality rate due to RTA at 28 per quota. From literature, the country's fatality and injuries have increased by 26% and 46.5%, respectively, since the year 2015. The country is faced with incomplete RTA data capturing, hindering effective planning and policy adjustments to curb the menace. In this paper, we scrapped user-generated data (Twitter) and national transport and safety authority's (NTSA) reports to shed light on traffic safety, practices, and cultures in the country. To this end, we gathered 1,000,000 tweets and 8000 speeding entries between 2015 and 2021 and performed natural language processing (NLP) and quantitative study of the data. We applied NLP and n-gram search of keywords to categorize data into 8 topics: traffic, public service vehicle (PSVs), policing, accident, infrastructure, recklessness, robbery, and corruption. From the data, policing, which touches on all police and law-enforcement-related activity was found to be highly correlated with PSVs, recklessness, accidents, traffic congestion, robbery, infrastructure, and corruption with indices of $r(76) = 0.92, 0.91, 0.87, 0.82, 0.81, 0.76$, and 0.70, respectively with $p < 0.001$. The topic modeling confirmed the identified topics to be the latent discussion issues affecting the public. From the study, PSVs, policing and traffic flow were isolated as key issues that ought to be addressed immediately. The research recommended the integration of driver monitoring systems to strengthen policing. The research, which utilized unstructured data, points to the utility of data mining which would greatly benefit traffic research, particularly African-based studies, that suffer from data inadequacy.

1. Introduction

Road traffic accidents (RTA) are a global public health concern. According to the World Health Organization (WHO), 1.35 million people die annually due to RTAs [1]. As per the statistics by WHO and other sources, the fatality rate is highest in African countries with an indexed estimate between 25 and 34 per 1,000,000 population [2–4]. According to the year 2020's RTA trend analysis, Kenya's fatalities and injuries have increased by 26% and 46.5%, respectively, compared to 2015 [5]. Incidences involving vulnerable road users (pedestrians, motorcyclists, cyclists, passengers, and pillion passengers) have reported an increase of over 300% over the same period. RTAs are estimated to cost a country 3–5% of the gross domestic product in terms of Medicare, insurance, and loss of productivity, with 93% of all world accidents occurring in low- and middle-income countries [1,3,6]. The impacts make accidents a social issue worth considering for all stakeholders.

Several interventions and remedies have been investigated and applied in traffic studies with varying successes. Speed regulations are a universal and natural response given the global evidence against excessive speeding [7–9]. Modernizing the road infrastructure with smart streets to cater to various mobility needs has been cited to improve safety [10,11]. Embracing modern technologies is viewed as a way to iteratively monitor and improve the multidisciplinary field of road transport [12,13].

Proactive policing and quality data gathering mechanism is at the center stage of improving road safety. From literature, developing

Abbreviations: NTSA, National Transport and Safety Authority; NPS, National Police Service; PSV, Public Service Vehicle; RTA, Road Traffic Accidents; NLP, Natural Language Processing; ML, Machine Learning; STM, Structural Topic Model; BERT, Bidirectional Encoder Representations from Transformers.

* Correspondence to: Dedan Kimathi University of Technology, Department of Electrical and Electronic Engineering, PRIVATE BAG – 10143 Dedan Kimathi, Nyeri, Kenya.
** Corresponding author at: Graduate School of Engineering, Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan.
E-mail addresses: joseph.muguro@dkut.ac.ke (J. Muguro), sasaki@gifu-u.ac.jp (M. Sasaki).

Peer review under responsibility of International Association of Traffic and Safety Sciences.

https://doi.org/10.1016/j.iatssr.2022.03.003

countries hardly have any traffic data as well as weak institutions that lead to a deteriorating RTA situation [14]. The reported literature focuses on small-scale surveys with hardly any sensor data to validate the findings against except surveys and hospital/police reports [15–18]. Primarily, the data collected from hospitals and police reports only touches on major incidents (accidents) since that is when motorists involve authorities; minor collisions and or traffic flow is hardly documented [19]. As such, alternative data mining methods become necessary to quantify traffic trends and culture. In a previous paper, we looked at the traffic data availed to the public through the police and transport agency [5]. We found the data to be limited in its coverage and grossly underreported.

In a bid to further disintegrate the problem of road traffic and related studies, literature has focused on user-generated data, with Twitter data ranking highest. In this regard, user data is perceived as a complementary source that enriches conventional modes of data gathering like cameras and inductive sensors with the advantage of ubiquity that is, non-geo-limited live data [20]. Using user-generated content, it's possible to get inexpensive and widespread information on both recurrent and non-recurring traffic conditions (e.g., traffic trends and unpredictable incidents such as crashes). Reports indicate an increasing internet penetration in African countries more so with the proliferation of social media. Kenya internet integration is estimated at 85% with 3–9 Million registered Twitter users [21,22]. The data being generated by users would as such be a representative sample of the population.

Previous studies using user-generated data have targeted identification of traffic jams, the occurrence of events like accidents, prediction of travel patterns, sentiment analysis among others [23–27]. The study in [28] suggested a method for generating a machine learning prediction model using geocoded traffic-related data from Twitter to get spatio-temporal traffic congestion information of Mexico city. A paper targeting the generation of real-time localization of accidents in Kenya was conducted in [29]. In the paper, the authors utilized Twitter data from Ma3Route and other machine learning approaches to create crash data and locations. Another research compared the association of traffic volumes and locations between target and ground-truth tweets using variants of Pearson Correlation Coefficients [23].

The way data is treated in user-based studies is either by quantitate/aggregate of tweets, retweets, and mentions [27,30–32]. The other alternative is the use of natural language processing (NLP) which is a subset of machine learning (ML) algorithms. NLP methodologies have been applied in diverse fields in research where the data is text-based [33–38]. Transportation-related studies using NLP and specifically topic extraction have been conducted targeting accidents, traffic flow, logistics among others [39–41]. As computing power increases, ML algorithms and by extension, NLP algorithms will achieve human-like comprehension of data at a massive scale, thus making their integration into society more impactful.

Related research is reported in [39] where the authors utilized NLP-based topical extraction to analyze near-collision events of cyclists data from a crowdsourced platform, BikeMaps.org. Another similar research is a case study done in Washington DC reported in [42]. The study focused on 4-year historic tweets with the keyword “traffic safety” in the country. The corpus of tweets was subjected to sentiment analysis using an extraction–transformation–cleaning procedure. Citizens' opinions were analyzed to evaluate the relevance of reducing deadlines, vehicle security concerns, and the implementation of transportation and official regulations.

1.1. Present study

The present study investigates the policies and traffic-related practices in the country. The main objective is to identify the interlinks between traffic practices and policies in place using user-generated data to derive an overview of traffic conditions in the country. In previous research, we explored fatality and daily incident reports to identify trends and accident fatality in the country [5]. The study mined data from reports and accident statistics compiled and available to the public by the NTSA (https://www.ntsa.go.ke/site/) between year 2015–2020. Some of the challenges identified are scarcity, the extent, and coverage of the data in agreement with other researchers targeting African transportation studies. Additionally, the data does not describe the prevailing conditions of the traffic incidents (e.g., drunk drivers) but rather focuses on the numerical reports (e.g., number of fatalities).

The present study involves data mining from Twitter and national transportation and safety authority (NTSA) data sources. Primarily, we targeted user-generated content (tweets) describing a different aspect of traffic. This way, we hope to derive more generalized behavioral tendencies that characterize the road conditions in the country. To this end, the present study targets to mine information touching on the following broad categorization; accidents, reckless driving, traffic congestion, policing, corruption, robberies, public service vehicles, and road infrastructure.

To achieve this, we employ NLP methodologies to analyze user-generated textual data. The data is aggregated to a representative index or using NLP and ML lexical approach, in this case, topic modeling and sentiment analysis. In summary, the study tries to answer the following questions.

I. Validation of user-generated data with NTSA incident reports.
II. Investigate the correlations between the identified categories.
III. Perform natural language processing on the user-generated data to infer latent interests of road users and the corresponding implications of the identified topics.

The rest of the document is distributed as follows. Section 2 gives an overview and issues faced in the Kenyan roads. Section 3 gives the methodology adopted in the study. Section 4 presents the results. Section 5 & 6 gives the discussion and conclusion of the study, respectively.

2. Kenyan traffic culture and policies

In Kenya, road transport is managed by the ministry of transport, infrastructure housing, urban development, and public works. So far, the government through the ministry has worked towards improving road conditions and safety through policy refinement and infrastructure development. This section explores the prevailing road traffic situations, policies, and challenges and how they contribute to RTAs in the country.

Public transport in the country is driven by privately-owned vehicles, public service vehicles (PSV) popularly known as matatu, operating in the country following licensing through various licensing bodies [43]. At present, a PSV is any vehicle that is licensed to ferry the public on Kenyan roads. This includes buses, mini-buses, vans, and mini-vans, 3-wheel motorcycles, motorcycles, taxis among others. The most popular category of PSV vehicles is the vans and mini-buses that are legalized to carry between 14 and 29 passengers (the usage of the word Matatu conventional points to this category of PSV vehicles). Second in popularity is the motorcycle (Boda-boda) and tricycle taxis (tuk-tuk), with a legalized passenger capacity of 1–3 passengers.

The ministry of transport's sessional paper on integrated national transport policy and other efforts led to the creation of the NTSA which was established in October 2012 to harmonize road transport and improve safety in the country. The body is mandated with licensing, policy formulation, and overseeing road transportation. NTSA has established and implemented several regulatory policies with varying success. Some of the current regulations and policy in place are as follows; All PSVs seats are fitted with seat belts, fitting of PSVs with speed limiters of 80 km/h, painting with a 150-mm width yellow band, and legalized passenger count to distinguish between non-PSVs of the same vehicle make, uniform accompanied by badges of registered driver and conductor, full-time employment of conductor and driver,

requiring PSV owners to belong to a SACCO or registered company with a defined route. The effectiveness and reach of the interventions and regulations established by relevant authorities have been sporadic as reported by various outlets and witnessed in revamping efforts of lax enforcement [43–45].

From the literature review and previous studies, the leading issues affecting road transport in Kenya are listed and explored in the proceeding sections.

i. Public road safety and recklessness as seen in accidents and collision
ii. Policing and impediments of law enforcement through corruption and bribes.
iii. Crimes targeting motorist and other commuter related issues
iv. Road infrastructure, and traffic flow bottlenecks

2.1. Safety, traffic accidents, and recklessness

Traffic accidents and reckless driving are highly correlated with a causal relationship. From literature, RTAs are attributed to either road infrastructure and environment, vehicular problems, or human factors [46]. Over 90% of all accidents are caused by various human factors ranging from psychological to behavioral tendencies. For instance, WHO lists drunk and distracted driving as top of the risky behavior that aggravates accidents. Drunk driving (driving under the influence of alcohol) and any psychoactive substances increase the risk of a crash fivefold compared to non-drunk drivers [1,47]. In contrast, distracted driving increases risk four to fivefold, making it arguably more severe than drunk driving [1]. Researchers have identified over speeding, one of reckless driving behavior, to be correlated with the likelihood and severity of RTAs. A 1% increase in average speed yields a corresponding 3% increase in serious crash risk and a 4% increase in fatal crash risk [1,36]. According to the national highway traffic safety administration, 26% of all traffic fatalities reported in the U.S. were linked to speeding for the year 2018 [47].

Fig. 1 shows incidence counts and estimated impacted users per the involved mode of transport from the 2016 data mining study of traffic accidents in Kenya [49]. From the figure, it is clear that personal vehicles lead to traffic incidents. However, based on the number of users in each car, the impacts of PSV are severest. If we consider a 14-seater minibus, 14 people are endangered every time an accident occurs. Compared with personal cars (maximum occupancy of 5 people), trucks (occupancy of 3 people), and motorcycles (occupancy of 2 people). The traffic accidents situation in the country is dire with outcry from all agencies. Despite this, few research activities and remedial efforts being taken, if any, are minimal in comparison to the problem at hand.

2.2. Policing and impediments to policing through corruption

At the heart of the success of any policy and regulations is the effectiveness of policing. Road regulation enforcement is a difficult task considering the vast number of motorists contrasted with the low numbers of law enforcement officers. As such, a great deal of cooperation between road users, policymakers, road designers, and law enforcement is needed. The current approach to road traffic and safety places all responsibility on the road user as opposed to an integrated view as witnessed through initiatives like Zusha. Several authors have called out policing as an impediment and a major barrier to safety [49,50,51]. The consensus is that policing activities in the country need reforms.

From news agencies, police conduct random crackdowns on non-conforming motorists, which inconvenience passengers more than solve the problem of motorists flaunting traffic rules. As pointed out by [52], crackdowns are disruptive and interfere with the citizens' socio-economic well-being in the commuters' eyes. PSVs operators often complain that the police harass and/or mistreat them without any justifiable reason other than to coarse a bribe [53]. This maltreatment might indirectly trigger the flaunting of traffic rules which is more rampant among PSVs than other motorists. According to a report by Transparency International, a body that aims at transparent and corruption-free Kenya, 86%of PSV operators indicated preferring to resolve traffic-related matters outside courts owing to collusion at the courts and the judicial systems [45].

Another drawback in policing has to do with the monitoring strategy used in the country, i.e., static checkpoint systems. The drivers already know of the monitored zones and/or communicate with other drivers to notify them of upcoming checkpoints. This way, reckless driving behavior outside of the monitored zone is hard to quantify. As of 2020, the inspector general of police issued directives to regulate checkpoints and roadblocks along highways, with a focus on mobile patrol units [54]. This measure was considered as a means of putting a stop to widespread

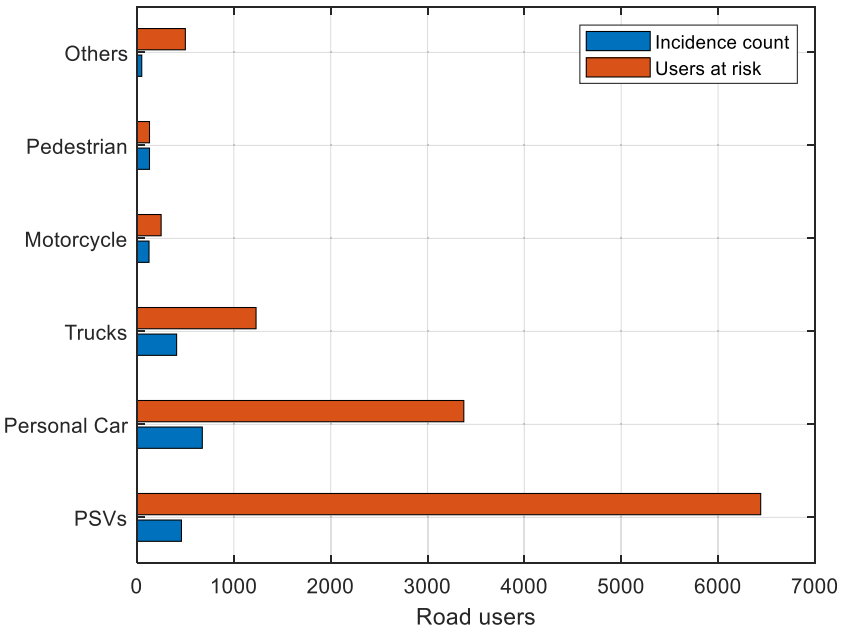

Fig. 1. Accident estimates for involved groups from incidence count in the country.

extortion as reported in the media and by victims of the heinous acts [14,53,51].

2.3. Crimes targeting motorist and road users

In a typical matatu operation, a conductor collects trip-based ticket fees and regulates the collection and dropping off of passengers, sometimes in an illegal manner (non-designated drop areas). The matatu terminals (stage) are often controlled by either privately organized groups or the SACCO management team to beckon passengers, regulate the flow and order of transit. This increases the chaos when boarding a matatu as competing teams will create fracas while beckoning passengers to different vehicles, creating an environment conducive to criminal activities such as pickpocketing. Additionally, muggings and robbery of valuables and car parts (side mirrors and other accessories) have increased particularly during traffic snarls. Vandalism has also been used as a tactic to distract motorists to initiate a full-fledged carjacking or robbery.

Besides insecurity, passenger comfort has been one of the least prioritized aspects of most PSVs, with acts ranging from outright insults to overcrowding passengers with little regard for regulations. Overall, multiple researchers have faulted the mode of public transport in the country concerning safety and conduct [51,56,57].

2.4. Infrastructure and traffic flow

Infrastructure is a major impediment to traffic flow in any country. In Kenya, the conditions are no different. Since 2005, the Kenyan government has actively been improving the road infrastructure. At present, several agencies are mandated with updating and planning of road infrastructure: Kenya national highways authority, the Kenya urban roads authority, and the Kenya rural roads authority. Nairobi-Thika Superhighway, JKIA–Westlands Highway, Eastern, Southern and Western Bypass, and other projects are taking shape to facilitate flow in the city among other places. The construction of such projects principally requires the closure or diversion of currently existing routes which adds to the agony in the form of traffic congestion. As of 2020, the country had a total of 177,800 km of road networks, both classified and unclassified, with 16,902 km of these roads paved [5].

Despite the progress in the road network, the rate of infrastructure development is not fully at par with motorization as witnessed in the congestion and gridlocked nature of traffic in the country. The country employs speed bumps as a deterrent to speeding. Occasionally, the bumps are unregulated, unmarked, and degrade to being a risk hazard [59,60]. The lack of functioning traffic lights, road markings, and signage are also raised as a point of concern. Pavements and designated cycling zones have gained popularity with alternative means of commuting. Particularly in the big cities, pavements and walk paths are usually confiscated by groups of riders and converted to boda-boda terminus if not relegated to kiosks or motor vehicle repair garages.

3. Methodology

Fig. 2 shows the workflow adopted in the present study. Details about each of the steps are described in the sections below.

3.1. Data collection and processing

We mined data from public repositories and transport agency in Kenya to achieve the objectives of the paper. Two sources were considered, the NTSA website and Twitter. NTSA collects, analyzes, and disseminates public traffic data and policies transparently to bring safety sensitization. Data collection involved downloading all entries availed in the website entry by entry. The speeding incident report sample data was published as a word document shown in samples in Table 1. The table lists incidents based on the regional camera. The first column shows the region's name, and the second column shows the province where the station is located and the corresponding number of cases between 1st January to 31st December 2016.

The word files were parsed using Matlab to extract numerical arrays. The total reports available on the website were one hundred and eighty-eight (188) daily reports (out of 365 possible entries in a year). The collected forms featured 43 location-based entries (stations), yielding over 8000 entries. A key point to note is that the data does not provide contextual information per entry but rather incidence count.

From Twitter, we scrapped historical data (tweets) through two libraries to access Twitter API, Tweepy and Twint as commonly utilized in literature [24,27]. The scrapping process was done in compliance with Twitter's Terms of Service. We scrapped tweets that are traffic-related between January 2015 and July 2021, i.e., we searched for #KenyanTraffic and #Matatu as keywords. The leading sources, mentions, and retweets for the keywords were two Twitter handles @Ma3Route and @KenyanTraffic. Ma3Route and KenyanTraffic are mobile/web platforms that crowd-sources for transport data and provide the public with traffic information. Ma3Route data has been at the forefront of generating traffic-related information used in various traffic research in the country [61–64]. As such, we utilized the two Twitter handles and generated ~1 M tweets as JSON (timestamp and tweet) and converted the data to frames using Python®.

3.2. Data cleaning

After extraction, total speeding cases for each entry were recorded and grouped in months. The data is described in Table 2. The collected speeding data had missing data for three months, August, September, and December. April had two entries, while July and February had a complete logging entry. The total entries availed per month are shown as count in the table totaling 188 for the entire year. For April, August, September, and December, we imputed the data with the k-nearest neighbors' regression algorithm with $k = 3$.

Fig. 3 shows the distribution plot of speeding instances recorded in the country. The trend of the data will be compared with other driving behavior data for the same period. There are no extra speeding records availed to the public since then.

From Twitter, a corpus of about 1,000,000 tweets was recorded in the study for data between Jan 1st, 2015, to July 31st, 2021. The data were cleaned to remove English language stop words and common messaging shorthand expressions like “btw”, “u”, “RT”, “lol”, etc. We also removed hyperlinks, punctuation marks, special characters and emojis, numbers, hashtags, and user mentions. We removed commonly appearing names and places like town, road, avenue among others. Since the objective of the study was not to analyze geo-location, we removed visually observable frequent locations and places. Further, we removed entries with less than 3 words from the data. From these processes, we had 770, 482 processed tweets from the initial corpus of 1 M.

We further searched specific phrases (n-gram search with n varying from 1 to 3) to broadly group the processed data into ‘relevant’ or ‘other’ category. The relevant categories were identified from the main issues facing the country as listed in the introduction and distilled into eight topics: traffic, PSV, policing, accident, infrastructure, reckless behavior, robbery, and corruption. The categories, corresponding search phrases, and total counts are listed in Table 3. The count shows the aggregate sum of tweets categorized in that topic. Note that in most cases, multiple topics were registered for the same tweet based on the content. All other tweets that could not be accurately placed, though containing relevant information, were categorized as ‘other topics’ and not considered further.

Table 4 shows time-stamped sample data with tweets, processed tweets, and the categorization of data to the corresponding topic. Topical constitution is indicated with a flag of either 0 or 1 to indicate the presence/absence of mentioned topic. The output of the above processes is captured in Table 4.

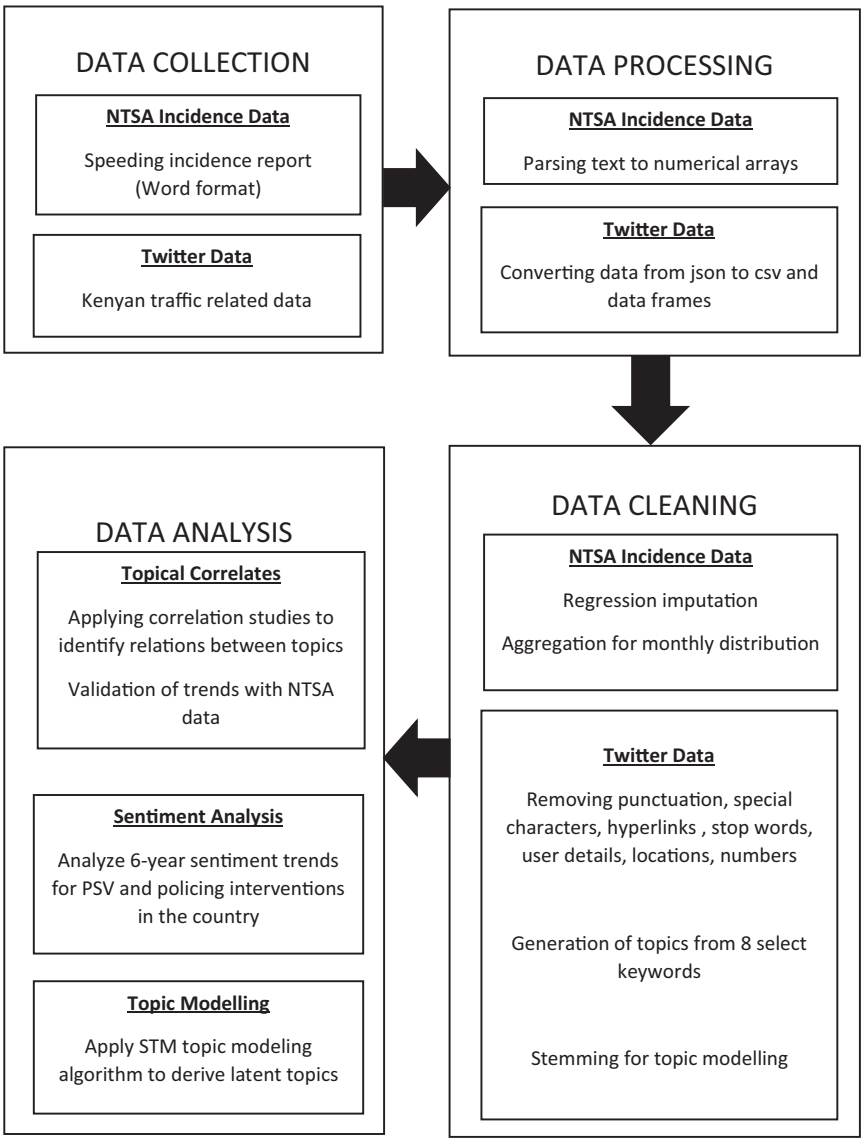

Fig. 2. Flow chart of data collection processing and analysis adopted in the paper.

3.3. Data analysis

The grouped (topical) data was aggregated monthly/quarterly to obtain a numerical value of mentions and tweets related to a given topic. It was theorized that tweet volume (monthly) reflects incidences, relevance, and importance in the mentioned topic. We performed Pearson correlation tests to identify interlinks between the topics as described in the results section.

In addition, we performed sentiment analysis to identify opinion trends and the general sentiment of the two inseparable topics in Kenyan traffic: PSVs and policing. Sentiment analysis is the process of extracting a writer's opinions and categorizing them into positive, negative, and neutral polarities. There is substantial literature on sentiment analysis utilizing various machine learning algorithms over social media data to forecast epidemics and outbreaks, among other things. The general category for sentiment analysis is machine learning, and lexicon-based. The ML models have demonstrated higher classification performance, but they come with the drawback of requiring a large corpus of labeled data, which is typically limited and expensive to create. As a result, pre-trained models that just require fine-tuning with a smaller dataset have become popular. One such method is Bidirectional Encoder Representations from Transformers (BERT), among other pre-trained language models. In this paper, we utilized twitter-roberta-base-sentiment pre-trained model from [24]. The model is chosen as it's trained on multilingual tweets as opposed to another text-based lexicon. Models pretrained on multiple languages have been shown to perform satisfactorily in cross-lingual transfer tasks compared to other models [65]. In the current application, the data is comprised of English, Swahili, and other local languages. A detailed explanation and other performance-related metrics of the model in use can be found in [66,67].

Topic modeling was used in conjunction with sentiment analysis to create latent themes from the Twitter data corpus. Topic modeling is a commonly used text-mining method in the field of NLP for the discovery of structures (i.e., themes/topics) from a large unstructured text collection by analyzing the common words in the texts. Topic modeling is unsupervised, which means it does not require any prior annotations or tagging of materials. We applied structural topic modeling (STM) to explore topical content and prevalence of the different documents. STM is an extension of the correlated topic model thus befitting the current task which has correlated terms. In addition, STM features high scalability, replicability, and transparent analyses as cited by multiple authors [36,37]. A detailed explanation of the model in use can be found in [30].

Table 1
Sample speed incidence report showing stations and corresponding counties.

30/3/2016

DAILY SPEED CASES WITH EXISTING REGIONAL CAMERAS

S/NO	STATION	REGION	CASES	REMARKS
1.	ROADSAFETY	HQRS	6	
2.	EMBAKASI	NAIROBI	14	
3.	KABETE	NAIROBI	-	
4.	LANGATA	NAIROBI	-	
5.	THIKA	CENTRAL	-	
6.	NYERI	CENTRAL	7	
7.	MAKUYU	CENTRAL	12	
…	…	…	…	
41.	GARISSA	N/EASTERN	-	
42.	ATHI RIVER	EASTERN	-	
43.	MWINGI	EASTERRN	2	
TOTAL	SPEED CASES COUNTRY WIDE		206	

Table 2
Descriptive statistics of the speeding instances.

	January	February	March	April	May	June	July	August	September	October	November	December
Count	23	29	23	2	28	26	31	–	–	14	12	–
Mean	179.4	174.7	186.8	170.5	137.0	123.5	132.5	131.0*	115.0*	118.4	99.5	101.5*
Std	18.3	15.9	13.4	19.1	23.7	19.9	21.4	–	–	22.8	19.6	–
Min	136.0	125.0	165.0	157.0	83.0	88.0	95.0	131.0	115.0	99.0	73.0	101.5
25%	172.5	168.0	176.5	163.8	125.0	110.0	122.0	131.0	115.0	103.8	80.8	101.5
50%	180.0	176.0	187.0	170.5	141.0	122.5	131.0	131.0	115.0	115.0	101.5	101.5
75%	186.0	185.0	194.5	177.3	152.3	139.5	139.5	131.0	115.0	122.5	112.0	101.5
Max	222.0	203.0	215.0	184.0	175.0	161.0	229.0	131.0	115.0	189.0	132.0	101.5

Note: * Estimated using k-nearest neighbor.

4. Results

Fig. 4 shows the quarterly topical aggregation of tweets between the years 2015 and 2021. The topics of interest, in this case, are traffic, accident, insecurity, reckless, corruption, PSV, policing, and infrastructure. Traffic is shown with an asterisk (*) indicating the value is factored by half to fit within the graph. From Fig. 4, traffic information and tweets decreased abruptly around 2020 April which corresponds to the time

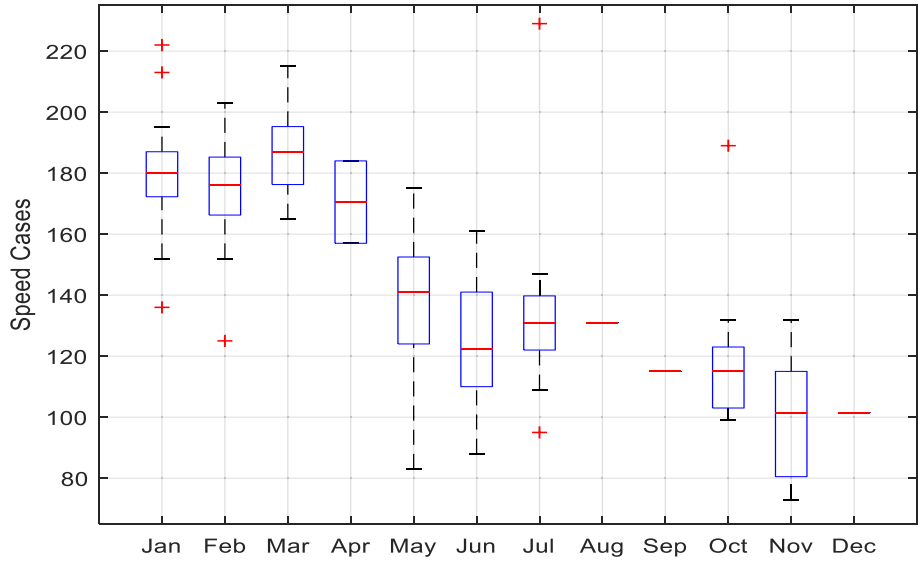

Fig. 3. Monthly distribution of speed cases.

Table 3
Search phrases for categorizing tweets into distinct groups.

Topic	Sample search phrases	Count
Traffic	'traffic jam', 'snarl', 'congestion', 'jam', 'bumper to bumper', 'breakdown', 'clear', 'gridlock', 'standstill', 'moving', 'slowing'	266,988
PSVs	'psv', 'psvs', 'taxi', 'bus', 'bodaboda', 'nduthi', 'motorcycle rider', 'pikipiki', 'tuktuk'	118,518
Policing	'cop', 'cops', 'police', 'karao', 'askari', 'ntsa', 'authorities', 'patrol'	53,285
Accident	'traffic accident', 'collision', 'hit and run', 'injured', 'injuries', 'dead', 'fatal', 'lost life', 'perished', 'head on collision', 'overturned', 'run over', 'crashed', 'knocked down'	52,826
Infrastructure	'road works', 'potholes', 'lane block', 'construction', 'street lights', 'pavements', bumps', 'road markings'	19,031
Recklessness	'overspeed', 'madness', 'mad driver', 'crazy driver', 'carelessness', 'overloading', 'reckless', 'drunkard', 'drunk driving', 'reckless', 'overlapping'	21,009
Robbery	'thieves', 'vandalism', 'vandalize', 'mugging', 'snatch', 'carjacker', 'robber', 'thief', 'thugs', 'pickpocket'	11,930
Corruption	'bribe', 'extortion', 'kitu kidogo', 'hongo'	3737
Others	Any category not fitting the search criterion	332,032

when the country initiated the travel restrictions due to the COVID-19 pandemic. Pre-lockdown between 2019 and 2020, there was an average of 5000 tweets per month compared to 3000 in the aftermath of the COVID-19 and travel restrictions. This may point to the travel and economic disruptions that followed. As of 2021 July, the trend was yet to normalize. The Table 5 below captures the descriptive statistics of the data aggregates for the period.

4.1. Topical correlations

Table 6 shows the Pearson's Correlation Coefficient test with $N = 76$ (monthly data between Jan. 2015 and June 2021). The highest correlation is that of reckless to accident, which is positively correlated, $r(76) = 0.93$, $p < 0.001$. The entire test reported a p-value less than the conventional significance level with at least all tests at $p < 0.001$. From the data, the mode of transport, in this case, PSV is a critical issue in Kenyan traffic with $r(76) = 0.92$, 0.88, 0.84, and 0.80 correlation index with policing, recklessness, accident, and robbery, respectively. In addition, corruption category is principally positively correlated to PSV with $r(76) = 0.68$. Corruption and bribery affect the PSV sector mainly and are similarly highly correlated with policing which are the recipients of bribes with an index of $r(76) = 0.70$ as seen. From the above indices, we identified reckless driving, PSV, accident, and traffic categories as critical issues that need further analysis/attention. Each of these is addressed in the subsequent sections.

4.1.1. Speeding and other reckless driving behavior

From Table 6, accident, policing, traffic, and PSV has $r(76) = 0.93$, 0.91, 0.90, and 0.88 correlation with recklessness, respectively. This is expected as reckless driving behavior impacts all aspects of safe driving. The correlation with policing would indicate an outcry of tweets demanding the arrest of offenders. In relation to PSV, the drivers of these vehicles are well reported in the literature to be notorious for reckless behavior putting the lives of passengers and pedestrians in danger. In most cases, traffic jams and snarl-ups agitate drivers to take reckless behavior like overlapping, tailgating, and other undesirable acts. This is thus captured with high correlation index of $r(76) = 0.90$. The relation between recklessness and accidents is much pronounced with the highest correlation index of $r(76) = 0.93$.

Twitter data attributed to driving behavior was isolated to analyze the relations with speeding data available from NTSA reported in Section 3.2. In addition to the reckless driving behavior explored through topical categories, we generated a subset of data to focus on cases of speed camera warning tweets. Fig. 5 shows the normalized (min-max normalization) plot of reckless behavior data with a highlighted region that represents the available NTSA speeding incidences for comparison. The reckless and accident trends are visually confirmed to be correlated. As mentioned in the introductory section, speed monitoring policies in the country adopt a predictable pattern, and where random measures are employed, the general attitude of motorists is that of penalty avoidance as opposed to compliance as witnessed in warnings. In the figure, the trend in warning tweets agrees with the reckless behavior trend.

The figure shows a discernable declining trend in violations between February 2016 until January 2018 followed by a sharp increase until March 2020. The decline period coincides with collaborative efforts of the national police service (NPS) and NTSA officers in safe road monitoring activity as reported by NPS and other media outlets [68]. The collaborative effort was recalled by the president in January 2018 following a public outcry of worsening accidents [69,70]. The data accurately picks the transition as well as COVID-19-related travel restrictions between March and July 2020 [71,72]. The agreement of user-generated data and policing and country activities gives validity to the acquisition process.

Further validation is derived by comparing NTSA speeding instances with overall reckless data as shown in Fig. 6. The figure shows an extract of reckless driving data that has been for the year 2016 and compared to normalized monthly average NTSA data from Section 3.2. The trends agree with the user-generated data within the duration.

4.1.2. Policing, corruption, and PSV data correlates

From Table 6, policing is positively correlated with highly PSV, robbery, and corruption with $r(76) = 0.92$, 0.80, and 0.70, respectively. Similarly, PSV is positively correlated with accident with $r(76) = 0.84$. From a traffic standpoint, the PSV operators are always in a hurry to beat traffic jams and make as many trips as possible. This inevitably increases their aggressive driving behavior which puts them in crossfire with law enforcers. To get a quick remedy, the PSV operators opt to bribe their way out rather than face the consequences. The issuing of bribes leads to further ruthless disregard for traffic rules. Data relating to policing, corruption, and PSV is isolated and plotted in Fig. 7 for further analysis. From the figure, corruption, which impedes proper policing, reported an all-time high in the final quarter of 2018.

In the figure, PSV, accident, and policing data are reacting to NTSA - NPS collaborative efforts as well as the Covid-19-related travel restriction as identified earlier. The two windows are highlighted in color for ease of tracking. Immediately after the end of the joint efforts (February 2018), an increment of all indices is noted with a peak towards the last quarter of 2018. This is attributed to high traveling routines for the end-year festivities. This is usually the busiest season for PSV operators. Why this peak was not present in other years is unexplainable from this data. Around the festive season, there were major road accidents involving PSVs that saw national and international outcry of the carnage [73].

4.1.3. Traffic flow and infrastructure data correlates

From Table 6, infrastructure is positively correlated to PSVs and accident and traffic with $r(76) = 0.73$, 0.72, and 0.69, respectively. Fig. 8 below captures the trends of data to highlight the effects of infrastructural activities in the target period. There is a direct relation between traffic flow and accidents on the quality of the roads and the data captures the effects as shown. Infrastructure and traffic data do not follow the NTSA-NPS joint policing trend identified earlier. However, there is an abrupt improvement in mid-2016 for both parameters. This is attributed to the commissioning of newly constructed by-pass roads which decongested the city considerably [74,75]. The effect on road commissioning eased traffic congestion considerably.

Table 4
Sample data-frame showing tweets, corresponding processed text data and topical flags.

Datetime	2021-07-07 14:00:47	2021-07-07 13:51:17	2021-07-07 13:45:17	2021-07-07 13:44:47	2021-07-07 13:26:47	2021-07-07 13:09:17
Tweet	08:00 Mombasa rd ni nywee apart from this bus ...	07:51 @KiambuCountyGov @Hon_JamesNyoro Thika ...	07:45 Those people who used to see Uhuru in ca...	07:44 Whats happening on Kiambu Road. We have ...	07:26 Thika Road traffic gives my sleep depriv...	07:09 Cyclist down critically injured two moto...
Processed	clear apart bus stalled inbound exercise pati...	dirty shops empty sex workers workers every	people used see camouflage burials carrying c...	whats happening stuck one spot	traffic gives sleep deprived self extra hour s...	cyclist critically injured motorcycles entangl...
Traffic	1	0	0	0	1	0
PSV	1	0	0	0	0	1
Policing	0	0	0	0	0	0
Accident	0	0	0	0	0	1
Infrastructure	0	0	0	0	0	0
Reckless	0	0	0	0	0	0
Robbery	0	0	0	0	0	0
Corruption	0	0	0	0	0	0
Other	0	1	1	1	0	0

Besides the improvement, several other peaks are highlighted in colored patches that we labeled as El Nino floods, Long and short rains, and construction work. From multiple sources, Kenya experienced El Nino rains and flooding between 2015 and 2016 [76,77]. Normally, the country has two rainy seasons: long rains (April–June) and short rains (October–November) [78]. The rains cause deaths, displacements, floods, and landslides in many parts of East-African countries. The first peak in the data is thought to coincide with the rains in mid-2015. The peaks in 2018 are noted to align perfectly with long and short rains.

Towards the end of 2019, another peak was registered. This is identified as the commencement of Nairobi Expressway construction work. The construction work started mid-2019 but the major disruptions peaked around December 2019 [79,80]. The project is set to end by 2023 but the effects were mitigated by travel restrictions which are seen as a reduction in infrastructure-related tweets by March 2020.

4.2. Topical and sentiment analysis using NLP models

4.2.1. Sentiment analysis

From the previous discourse, traffic flow, PSV, and policing can summarily capture the issues affecting the transport industry in the country. This section looks at the sentiment analysis of these three topics. Fig. 9 below gives a summarized sentiment analysis of 3 selected categories, i.e., PSV, policing, and traffic. The pie charts show the distribution of normalized sentiments for policing, PSV, and traffic. Neutral sentiments are more than polarized tweets, as computed by the BERT model. This was thought to be the result of the difficulty in decoding the mixed usage of language (English, Swahili, Sheng, and local languages) found in the corpus. In this case, we focus on the polarized tweets to identify trends and user-based approvals between 2015 and 2021. The graph on trends is derived from an average of sentiments quarterly.

From the Figure, traffic elicits the highest fluctuation in positive and negative sentiments. The year 2015 had the highest negative sentiments with a drastic improvement in the first quarter of 2016. The positive sentiment was identified in Section 4.1 as the commissioning of roads which greatly eased traffic. As expected, PSV and policing sentiments are closely related (Fig. 9), a relation that has been highlighted in the previous section.

From this data, the NTSA-NPS collaborative efforts are also captured as a steady decrease in negativity between 2016 and 2018 and a corresponding rise thereafter. The other noticeable dip (improvement) in the first quarter of 2020 is related to the travel restrictions imposed by the government to curb the spread of COVID-19 in the country. The sentiment similarly confirms the trends discussed in the previous section.

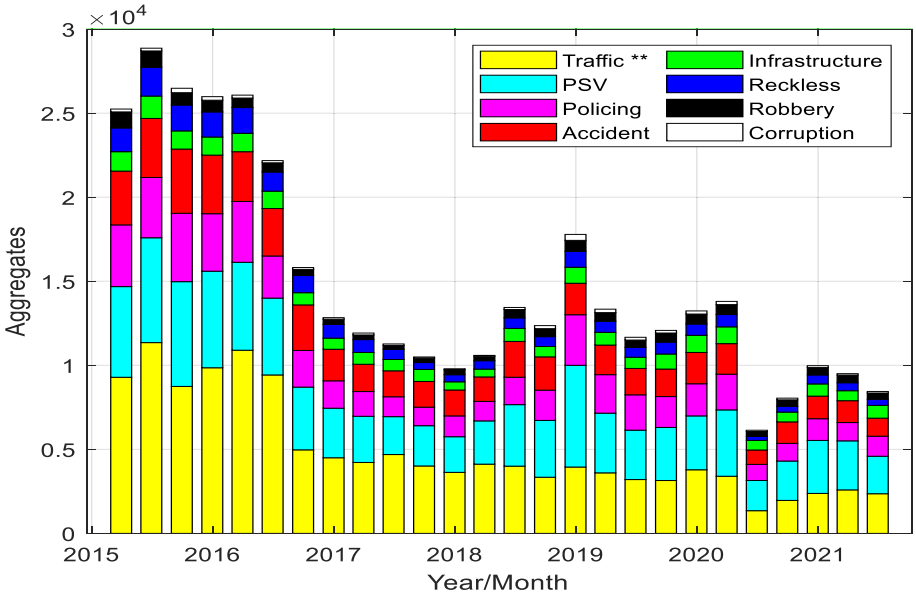

Fig. 4. Quarterly aggregate tweets of generated categories, reported between January 2015 and July 2021.
Quarterly data is plotted for visibility, but the analysis uses monthly data.

Table 5
Descriptive statistics of the monthly aggregates of the data.

	Traffic	PSV	Police	Accident	Infrastructure	Reckless	Robbery	Corruption
Mean	9904.0	3635.1	2042.8	2022.0	812.5	800.6	490.4	142.8
Std	5864.0	1375.3	958.3	812.1	227.4	422.2	197.0	69.5
Min	2699.0	1798.0	958.0	856.0	466.0	230.0	234.0	48.0
25%	6472.5	2615.8	1202.8	1533.0	662.3	507.0	327.5	90.3
50%	7948.0	3182.0	1818.0	1789.0	742.0	665.0	466.5	128.5
75%	9798.5	4414.8	2450.8	2563.5	1014.8	1002.5	586.0	180.0
Max	22,698.0	6237.0	4069.0	3819.0	1331.0	1714.0	971.0	354.0

4.2.2. Topic model

We performed topic modeling using STM with 15 topics as described in the methodology. One of the aims of topic modeling was to validate the general categorization considered in the research, as such, we opted for a higher number of topics that characterizes the corpus as opposed to optimizing the number of topics.

Table 7 shows the top 10 words for each of the topics from the model. Besides the top appearing words, we display FREX (Frequency and Exclusivity) metric which shows frequent and exclusive words (unique) with relation to other topics. Lift and score represent term probability and the terms' logarithmic probability, respectively, within a topic across the entire corpus. Further, topic proportions (in percentage) are shown as a description of the topic prevalence over the entire dataset. The last column infers the topical implications for the clusters (topics) generated by the unsupervised learning model.

Topical interpretation is based on top keywords as listed in Highprob, FREX, Lift and Score. From the Table 7, the model identified 6 variations of traffic flow. The variations are interpreted as follows; clear traffic (topic 3), moving traffic (topic 1), heavy traffic (topic 2), gridlocked traffic jams (topic 8), route suggestion (topic 4), and location-based traffic updates (topic 5). Taken together, this places traffic as the most prevalent issue accounting for 40% of the overall data.

From the model, police-related activity is captured in two topics, topics 9 and 11. Topic 9 captures traffic-related policing from police and NTSA officers. It also features corruption and bribes. Topic 11 can be taken as a treatment of road-related crimes including mugging, robberies, protests among others. This is generally interpreted as a robbery topic in keeping with the topical convention under study. Taken together, policing yields a prevalence of 14.8%.

Infrastructure is captured in topics 12 and 13. Topic 12 isolated issues touching on construction time, potholes, repairs, contractors, etc. Topic 13 identified the issue of parking spaces, county government regulation on parking fees, zones, etc., which is taken as a subset of infrastructural management. These two topics yield a prevalence of 13.9%.

Topic 7 is interpreted as PSVs due to the high probability of terms/keywords like Matatu, Boda-boda, SACCOs, and other PSV related terms. Further, topic 6 captured commuter issues, though largely in Swahili, captured terms like “leo” (today), “watu” (people), “hakuna” (there are no), “gari” (vehicles/means of traveling). This describes the woes of public commute mostly in rush-hours, with crowded passenger picking points, which can be taken as a subset issue in PSV. Taken together, these would yield a prevalence of 13.3%.

Accidents are captured in topic 10 comprising of injuries and fatalities with a prevalence of 7.4%. Reckless driving behavior is captured in topic 14 at 6.6% prevalence with keywords like overlapping, mad driving among others.

From the model, traffic (with variations), infrastructure, policing (including corruption and bribes), PSVs, reckless driving behavior, robberies, and accidents, were identified using STM in agreement with this research's topical categorization. The model with 15 topics had 1 spurious topic which was expected. Unclear/spurious topics have been explained to be inevitable as a cause and the effect of increased topic numbers [34].

Fig. 10 shows the topic correlation as identified by the STM model. In this case, each topic is a node while the connections highlight the co-occurrence of terms with high probability. The size of the node indicates the prevalence. From this, the term correlation identifies accidents (topic 10) uniquely compared to other terms that are highly correlated. The topics can be broadly separated into the related cluster of topics. Traffic-related clusters topics 1, 2, 3, 4, 5, 6, and 8, while other transport issues are clustered as shown in topics 7, 9, 11, 12, 13, and 14 as seen by the network.

From the network, topic 15, which was identified as spurious in Table 7, is correlated with topics 5, 12, and 13, which has to do with location-based traffic flow and construction work. This would suggest that topic 15 is addressing the aftermath of construction works and the mess. The correlation in topics has been discussed in the previous section and confirmed by the model output.

5. Discussion and recommendation

The paper takes a deep look at traffic culture, practices, and implementation of policies related to road transport in Kenya. As pointed out by various authors, missing and incomplete data is a challenge that affects road networks as well as disaster relief planning [14,62,81,56]. In the country, the practices, and policies (enforcement of policing) are explored through user-generated data to identify interlinks and trends. The paper grouped the issues that affected the country into eight topics (traffic flow, PSV, policing, accident, infrastructure, recklessness, robbery, and corruption) and investigated the trends, opinions, and latent expectations of the traffic users.

Table 6
Correlation of topics in the user-generated data.

	Traffic	PSV	Policing	Accident	Infrastructure	Reckless	Robbery	Corrupt
Traffic	1.000							
PSV	0.745	1.000						
Policing	0.822	0.920	1.000					
Accident	0.859	0.840	0.874	1.000				
Infrastructure	0.689	0.732	0.759	0.721	1.000			
Reckless	0.895	0.876	0.910	0.925	0.761	1.000		
Robbery	0.654	0.802	0.810	0.695	0.743	0.738	1.000	
Corrupt	0.291**	0.684	0.696	0.437	0.526	0.488	0.545	1.000

Note: ** ➔ $p = 0.009$.
All other values have $p < 0.001$.

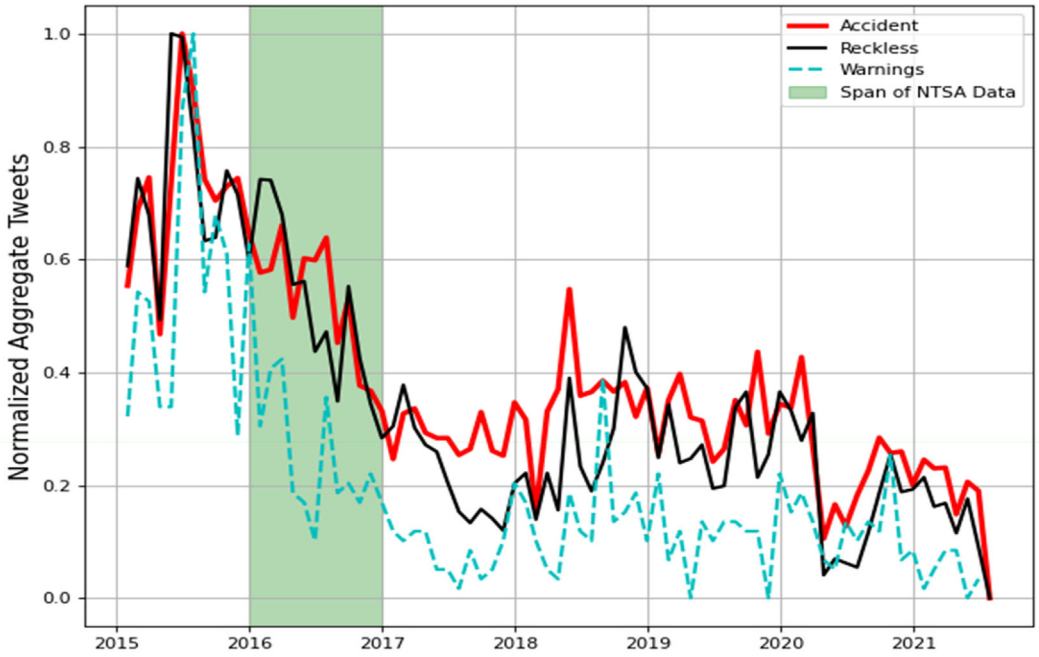

Fig. 5. Trends of reckless driving and impact it has on accidents.

5.1. PSVs, driving behavior and policing practices from user-generated data

From speeding data incidences derived from NTSA, only 30 of the existing 47 counties generated reports in 2016. It is unclear whether the remaining 17 counties had speed tracking cameras or not. It is also unclear why speeding incident reports stopped after 2016. The data derived from NTSA was used to validate the trends of user-generated data. This was done by extracting reckless driving data for the same period and comparing the two. The results (Fig. 6) showed a good agreement between the two datasets. Further, overall improvement in traffic-related incidences was noted between 2016 and 2017, the period which is identified as collaborative efforts between NTSA and NPS in the country.

From user-generated data, accident, policing, traffic, and PSV are positively correlated with an R-value $r(76) = 0.93, 0.91, 0.90$, and 0.88, to recklessness with $p < 0.001$ in all cases, respectively. The interpretation would be that recklessness increases chances of accident occurrence, highly agitates the users who invoke the arrest of the criminal activity, leads to worsening traffic conditions and this behavior is highly observable in PSV operators. Further, Table 6 portrays a gray picture as far as PSVs are concerned with a positive correlation with all the study topics. PSVs are correlated with policing, recklessness, accidents, infrastructure, traffic congestion, robbery, and corruption with indices of 0.93, 0.85, 0.82, 0.81, 0.80, 0.75, and 0.57, respectively. This underpins the unregulated nature of the industry coupled with poor policing

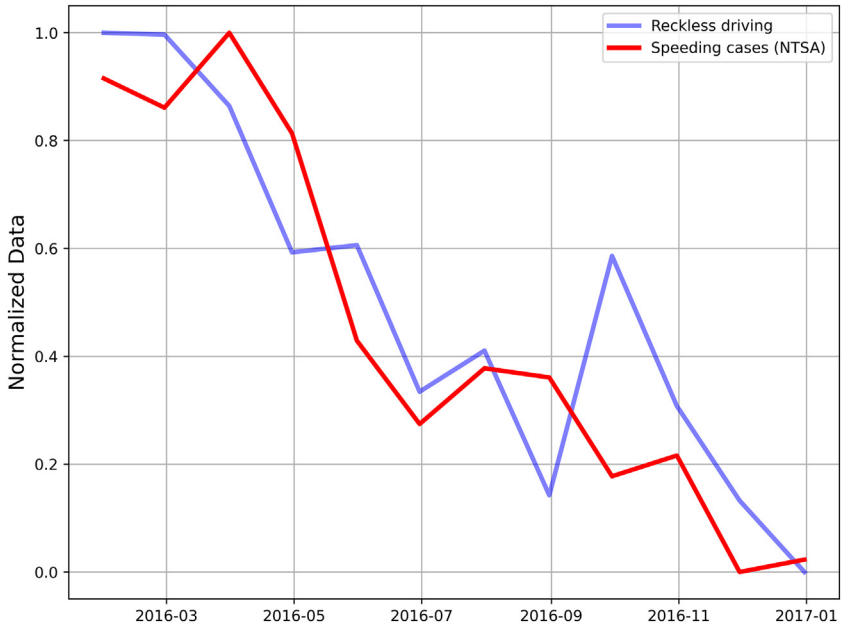

Fig. 6. Normalized reckless behavior and speeding instances as reported in 2016.

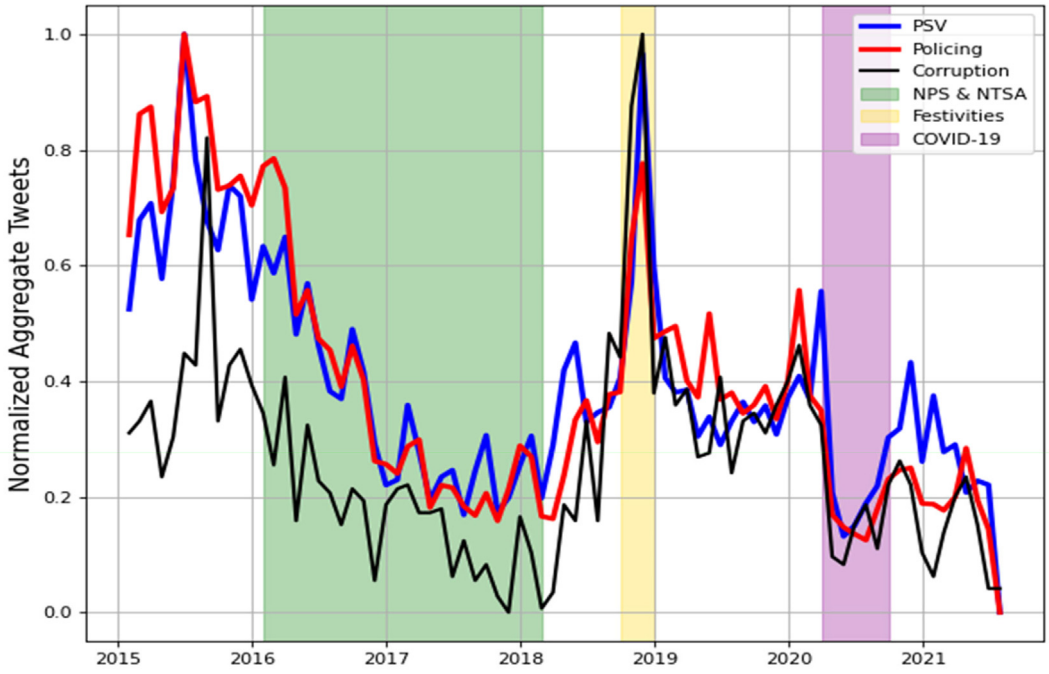

Fig. 7. Effects of policing and events on PSV data.

practices. PSVs driving behavior has been pointed out as below par concerning safety [3,18,19,34]. The results obtained in the study put PSVs in the center stage of the traffic menace in the country ranging from accidents, traffic congestion, reckless driving, robbery as well as corruption.

The police force is spread-thin trying to maintain law and order within the country as well as on the roads with motorists that are apprehensive of policing efforts. The user-generated data captured some attitude of the public with a bias towards penalty-avoidance as opposed to compliance as seen in warning tweets. From Fig. 5, it's a typical occurrence for motorists to warn others, using gestures and/or phones, of forthcoming roadblocks and speed traps and thereby adjust accordingly, which at best elicits zoned compliance with the regulations. On

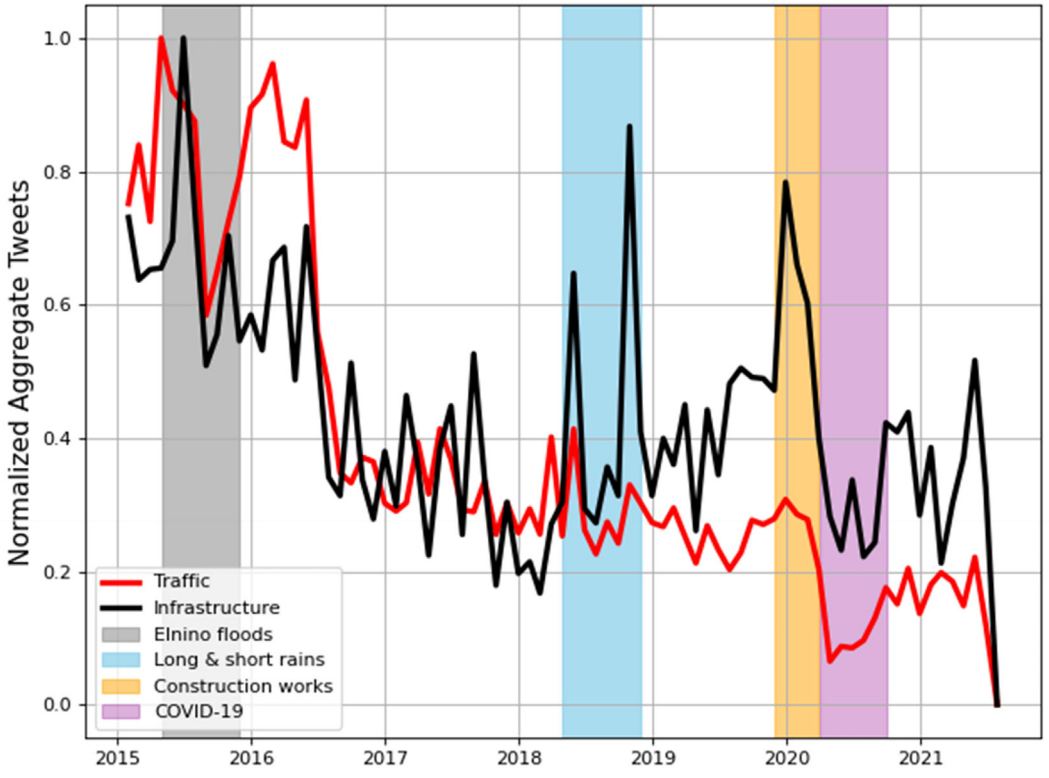

Fig. 8. Effects of infrastructural activities on traffic flow.

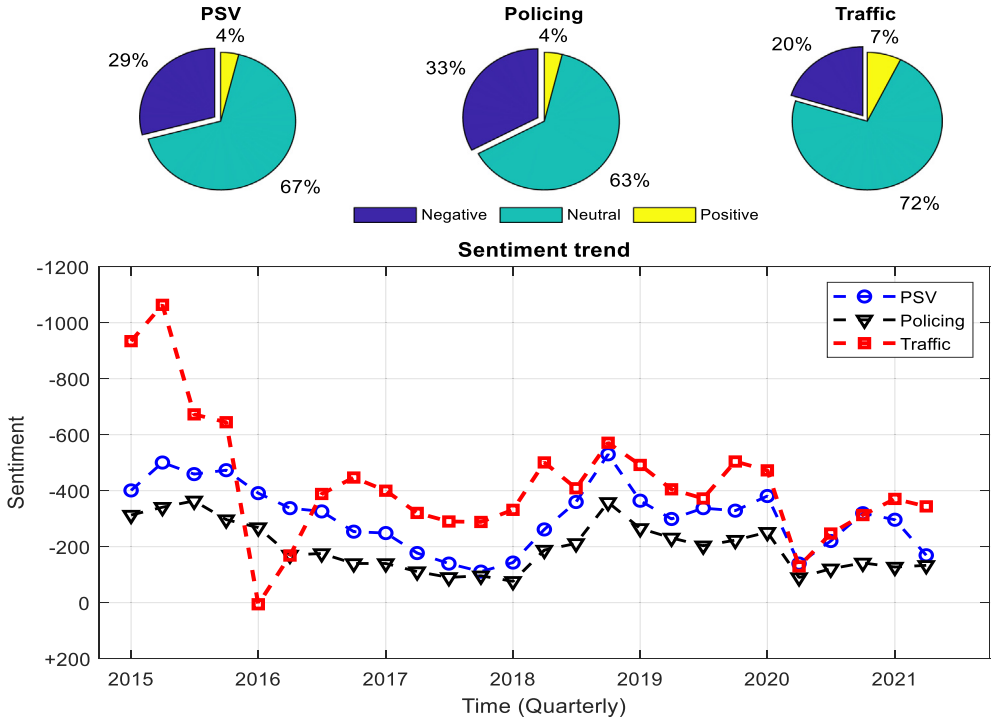

Fig. 9. Sentiment distribution and trends of selected categories.

the other hand, the police force has been implicated negatively in the involvement and soliciting bribes and other corrupt practices. The data indicates a positive correlation between policing and corruption with an all-time high around the last quarter of 2018 (Fig. 7).

In summary, policing needs urgent reinforcement and adaptation to solve the traffic menace experienced in the country. From the data, the NTSA- NPS collaborative efforts yielded a positive and significant change in the country, contrary to popular belief at the time that led to the disbandment of the efforts. Figs. 5 and 7 confirm that aggregates of recklessness and corruption were at their lowest. To reduce abuse of power through unjust arrests and harassment, modern monitoring methodologies that are pervasive and repeatable should be introduced to assist the police force to perform their duties. Borrowing from literature and practices in other countries, the mode of policing using predetermined checkpoints and roadblocks is ineffective in tracking speedsters and other reckless driving behavior [12,13].

5.2. Traffic flow, accidents, and infrastructure in the country

The user-generated data exposed the relations of traffic flow and infrastructure in the country. From the data, infrastructure is positively correlated to accident and traffic with $r(76) = 0.72$, and 0.69, respectively. Intuitively, a traffic accident is either caused by (road) infrastructural deficiency or will lead to infrastructural damages in its wake. In practice, the greater majority would opt to resolve non-fatal accidents without involving the police which increases non-reported cases of RTA. Under normal circumstances, infrastructure is never considered. This affects public property and infrastructure as reported in the literature [19,85]. Destruction and abuse of road resources by motorists is worth considering. The abuse of road facilities takes various forms like overlapping in footpaths/pavements as well as vehicle conflict with barriers and rail guards upon accident impact.

Traffic congestion is highly correlated with accidents. Congestion increases the tendencies of recklessness and aggravation, which leads to more accidents and further congestion. In cases of reckless behavior and an accident ensued, traffic jams are bound to be more aggravated. As such, urgent methods need to be put in place to address the problem. There is severe resource wastage in the traffic congestion alone with estimates of up to K.Sh. 20 billion (182 million USD) annually. The number of lives lost is similarly unacceptable with estimates of 3000–4000 annual fatalities which multiple sources identify as an underestimation with a factor of 3.5–4.5 [5,29,87].

There are instances of robberies even in an accident. From the data, accidents are positively correlated with robbery with indices of 0.63. The appalling behavior is often reported in news outlets [88,89]. The well-wishers, who reach the accident scene first offer immediate assistance to the motorist involved in an accident but in some instances, the valuable (phones and other items) of the victims go missing. Again, the presence of police and other authorities is usually solicited by users to assist in the developing situation.

The recent trends in traffic congestion have spurred the need for use of alternative modes of commute. In this regard, walking and cycling have been on the increase owing to the flexibility of the method. The new concern is on pedestrian and cyclist safety. Topic modeling identified the safety of the vulnerable road users as a relevant study topic, though not explicitly dealt with in the main discourse. The drive for infrastructure that supports inclusive transport is gaining popularity in Twitter traffic debates.

The user-generated data identified several significant events and practices that have altered traffic flow in the country. Infrastructural development of roads had the highest net effect as seen in the commissioning of the bypass road and commencement of construction of the Nairobi expressway. Intuitively, the closure of certain routes would inevitably increase traffic congestion, particularly in such places where there are no alternative routes. Seasonal rains are also a major cause of disruption in transport in the country as seen in sharp peaks of traffic and infrastructure aggregate tweets in Fig. 8. As the country develops its road network, there should be deliberate efforts to make the roads impervious to floods and other weather fluctuations.

5.3. Policies and practices that has shaped transportation in the country

From topic modeling, the topics identified by unsupervised ML agreed with the broader categorization of issues affecting the transport industry

Table 7
The topic model generated by STM.

Topic	Top 10 Words per category	Topic proportions (%)	Interpretation
1	**Highest Prob:** move, slow, bout, slowli, hill, smooth, approach, place, usual, snail **FREX:** move, slow, slowli, snail, eka, chaka, pace, colt, hill, lion **Lift:** chaka, lion, move, albeit, colt, eka, pace, procession, slow, slowli **Score:** move, chaka, slow, slowli, hill, colt, smooth, eka, snail, pace	6.2	Moving traffic
2	**Highest Prob:** heavi, start, happen, today, flow, well, still, light, much, snarl **FREX:** rain, start, heavi, begin, movement, nightmar, today, flow, monday, graduat **Lift:** abnorm, ceremoni, about, begin, kmtc, messi, unbeliev, downpour, freeli, how **Score:** heavi, start, flow, happen, today, well, snarl, light, rain, still	5.7	Heavy traffic (circumstantial traffic, eg. rains, ceremony)
3	**Highest Prob:** clear, inbound, side, good, morn, drive, look, safe, wrong, gathi **FREX:** morn, clear, stay, lang, flood, side, thank, estat, god, far **Lift:** beauti, bless, crystal, bell, bottleneck, clear, estat, flood, motorcad, okay **Score:** clear, inbound, side, good, morn, look, safe, wrong, thank, drive	5.99	Clear traffic
4	**Highest Prob:** high, use, servic, outbound, back, bad, past, total, pack, expect **FREX:** pack, total, altern, servic, high, expect, super, earli, pole, imeanza **Lift:** altern, kuruka, pack, feeder, sail, allov, total, pole, safaripark, imeanza **Score:** high, use, outbound, servic, kuruka, pack, back, super, bad, gridlock	5.38	Traffic (Route suggestion
5	**Highest Prob:** bumper, citi, around, gaen, hapa, situat, till, current, centr, mpaka **FREX:** bumper, citi, gaen, situat, current, centr, lower, nakumatatut, imeshikana, ridg **Lift:** ast, canadian, doonholm, dunga, easter, khoja, kuanzia, librari, shop, thindigua **Score:** bumper, citi, gaen, around, hapa, situat, imeshikana, centr, mpaka, current	3.53	Location based traffic situation
6	**Highest Prob:** lot, kwa, leo, watu, kama, hapo, street, hakuna, gari, spot **FREX:** lot, kwa, leo, watu, kama, hakuna, gari, sana, ama, kwani **Lift:** acha, alafu, apa, baridi, breakfast, bro, chini, devil, eti, haki **Score:** lot, kwa, leo, watu, kama, gari, hapo, hakuna, street, sana	5.33	Commute issues (no means of transport)
7	**Highest Prob:** matatu, driver, matatus, bus, boda, sacco, passeng, buse, drive, psv **FREX:** matatus, boda, sacco, passeng, buse, psv, reckless, boa, careless, psvs **Lift:** nazigi, sanit, matatus, arrestimatatus, boa, buss, careless, conductor, crew, excess **Score:** matatu, informatatuion, sacco, matatus, bus, driver, boda, passeng, buse, reckless	7.96	PSV(Matatu, buses, boadaboda, saccos)
8	**Highest Prob:** traffic, jam, moment, head, caus, along, crazi, standstil, build, gathi **FREX:** moment, traffic, jam, standstil, crazi, eas, rnd., moder, head, updat **Lift:** standstil, moment, rnd., trafficjam, traffic, moder, jam, cleareeee, traffick, eas **Score:** moment, traffic, jam, standstil, caus, admin, head, crazi, build, moder	13.7	Traffic jams (gridlocks)
9	**Highest Prob:** cop, speed, take, offic, day, ntsa, know, stop, kind, law **FREX:** cop, ntsa, bribe, general, corrupt, speed, fine, limit, court, collect **Lift:** expir, map, random, cop, african, airtim, bail, command, court, fine **Score:** general, cop, ntsa, offic, speed, bribe, take, law, day, know	7.64	Policing (cops, ntsa, corruption)
10	**Highest Prob:** accid, caus, near, involv, along, car, lorri, overturn, dead, hit **FREX:** accid, involv, overturn, dead, knock, trailer, lorri, injur, fatal, crash **Lift:** accid, deadlock, overturn, perish, canter, casualti, dead, deadlin, fatal, injur **Score:** accid, involv, dead, overturn, knock, injur, fatal, grisli, crash, car	7.37	Accidents (causes, nature, fatalities, etc)
11	**Highest Prob:** polic, vehicl, along, block, near, motorist, fire, help, arrest, truck **FREX:** fire, thug, protest, rob, men, post, student, emerg, team, demonstr **Lift:** battl, brigad, burn, cctv, helicopt, mug, post, aid, alleg, arm Score: polic, arrest, fire, block, vehicl, motorist, brigad, thug, near, help	7.2	Robbery(police, arrest, thugs etc.)
12	**Highest Prob:** need, peopl, time, pleas, make, construct, bump, right, see, think **FREX:** construct, problem, pothol, bump, poor, contractor, mark, remov, infrastructur, cycl **Lift:** drain, futur, proper, rubbl, tarmack, vulner, arriveal, bigger, bump, consider **Score:** construct, need, bump, time, peopl, problem, pedestrian, make, pothol, pleas	10.77	Infrastructure (Construction works)
13	**Highest Prob:** park, busi, turn, pay, free, counti, becom, govern, fuel, outsid **FREX:** park, busi, pay, free, counti, becom, whole, space, fee, green **Lift:** decongest, issa, kanjo, occupi, paid, payment, rate, reserv, slot, park **Score:** park, busi, turn, pay, counti, free, fee, govern, becom, charg	3.2	Infrastructure (parking zones, streets, etc.)
14	**Highest Prob:** car, one, get, like, hour, guy, overlap, mad, stuck, home **FREX:** mad, overlap, get, top, power, overlapp, beat, end, got, meet **Lift:** batteri, exid, half, hungri, sat, sick, steer, wheel, android, beat **Score:** car, overlap, powerlast, get, one, hour, like, mad, guy, stuck	6.57	Reckless driving (overlapping, madness, etc.)
15	**Highest Prob:** even, work, come, mess, pass, thing, someth, yet, seem, coz **FREX:** come, better, thing, wonder, sure, work, even, coz, entri, noth **Lift:** confus, wangari, come, entri, former, matatuhai, prof, asleep, maathai, better **Score:** work, even, come, matatuhai, pass, mess, someth, thing, yet, better	3.4	Spurious[a]

[a] This is further interpreted as aftermath of construction work after analysis of STM correlation network.

in the country. Thus, the topical categorization utilized by this inquiry is a fair representation of the latent discussions found in the user-generated dataset. The model indicated traffic and traffic-related topics as a key discourse by the public with a prevalence level of 35%. PSV, with all passenger and pedestrians-related issues, was second in prevalence at 18%, infrastructure at 15%, policing at 14%, and accident at 7% of the investigated data. From the topic modeling, we performed sentiment analysis of three topics: PSV, policing, and traffic flow. Sentiment analysis agreed with events and trends identified in the topical aggregation.

The current inquiry has identified the multifaceted nature of road transport and traffic in the country. The study uncovered several relations that have not been identified earlier in literature coming from unstructured user-generated data. Recklessness, accidents, and PSV interplay in the data were established. Police efforts and impeding corrupt activities were also evident. Similarly, traffic flow and infrastructure interlink were pointed out. Taking the stand that PSV captures all commute-related issues, traffic flow embodies all infrastructural activity, policing handles all criminal activity, traffic flow control, and traffic

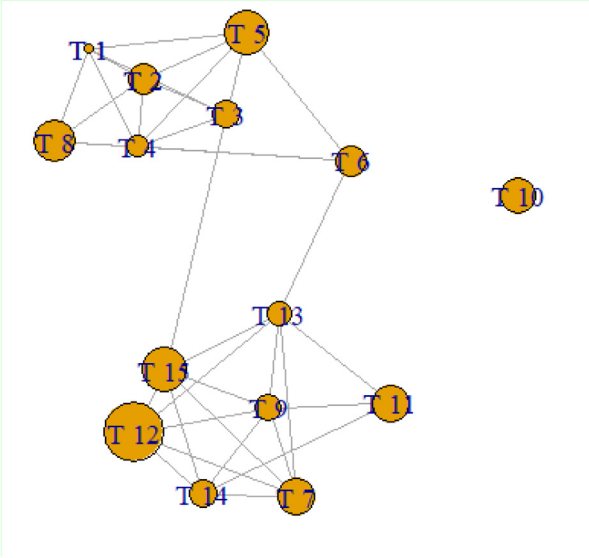

Fig. 10. Network of topic correlation.

offenses; traffic, PSV, and policing thereby becomes a summarization of issues affecting the transport industry in the country.

The following are the events, policies, and practices that have shaped the transport industry in the last 5 years.

i. NTSA – NPS collaborative road monitoring efforts (policing)
ii. Floods and heavy rainfall (traffic flow)
iii. Festive seasons (PSV)
iv. Opening (commissioning) of bypass roads (traffic flow)
v. Construction of Nairobi Expressway (traffic flow)
vi. COVID-19 (traffic flow, policies)

Note that the practices and events agree with the triad of summary identified and collaborate with other reports from various sources as found in the discussion. The ramifications of the findings are important concerning concerted efforts of solution refinement. Policing as a topic touch on all aspects of road transport ranging from the safety and rights of the citizens. Condoning and or improper monitoring of unsafe driving, insecurity, and corruption interfere or impede with policing process, denying the citizens of their rights to safety and fair treatment and in most cases leads to lost livelihood through accidents and other injustices. As such, the need for proper policing can never be over-emphasized. Traffic flow, with links to infrastructure and accidents occurrence, will continue bottlenecking the development of the nation if left solely in the hands of law enforcers. Finding a solution that enhances the triad of issues should be embraced by every citizen as an existential problem and treated as such.

5.4. Recommendations

The inquiry has identified three topics of pertinent importance in Kenyan roads: PSVs, policing, and traffic flow. The current policies and interventions are outdated as they focus on the individualistic responsibility of road safety [90]. An alternative would be a system-wide integrated intervention, with a clear-cut fault-finding mechanism and speedy delivery of feedback to failure. From this, the research recommends an integrated driver monitoring targeting mass transport vehicles using state-of-the-art driver-monitoring systems. The use of a monitoring system has been proposed in numerous intelligent transportation systems around the globe [91,92]. In addition, the success of such is attested to by user-based insurance systems that have demonstrated the effectiveness of comprehensive tracking in saving lives and lowering the cost of operating motor vehicles [93,94].

Principally, the PSV industry is unregulated, besides the licensing, and as such, Matatu owners operate with indignity towards the passengers. Additionally, police misconduct in form of bribes is highest in the group. Also, this study and other research confirmed the tendency and prevalence of reckless driving in PSVs. The focus on PSV drivers is guided by the fact public transit is a communal responsibility, which demands stricter regulations and oversight which conforms to international standards [95]. The recommendation would streamline monitoring and regulation of misconduct and reporting process. This would increase transparency in fault-finding as well as improve the desperate state of accident-related data collection in the country.

At the moment, PSVs are required by law to be fitted with (digital) speed governors. Therefore, the inclusion of speed tracking and driver logging system would integrate with the existing and future hardware in the vehicle. We believe that the mode of transport and policing would greatly improve from such a system-wide integration of monitoring using modern technologies.

6. Conclusion

The paper took a deep look at traffic culture, practices, and policies in place concerning road transport in Kenya. The main objective was to identify the interlinks between traffic practices and policies using user-generated data to derive an overview of traffic conditions in the country. Twitter and transport agency data was mined to extract data for the analysis. A corpus of text with approximately 1,000,000 tweets was treated in the study, gathered from Ma3Route and other sources between 2015 January and 2021 July. The data were categorized into distinct topics by searching specific phrases related to the identified issues affecting road transport, i.e., traffic flow, PSV, policing, accident, infrastructure, recklessness, robbery, and corruption. Tweet aggregation and natural language processing methodologies; topic modeling and sentiment analysis, were performed to analyze the data.

From STM, topical categorization utilized by this inquiry was found to be a fair representation of the latent discussions found in the user-generated data. STM model indicated traffic flow and traffic-related topics to have a prevalence of 35%. PSV, with all passenger and pedestrians related issues, had a prevalence of 18%, infrastructure 15%, policing 14% and accident 7% of the investigated data.

In the data, policing and PSVs were found to be correlated with all target study topics. Policing, which touches on all police and law-enforcement-related activity was found to be highly correlated with PSVs, recklessness, accidents, traffic congestion, robbery, infrastructure, and corruption with indices of 0.92, 0. 91, 0.87, 0.82, 0.81, 0.76, and 0.70, respectively. PSV was positively correlated to policing, recklessness, accidents, robbery, traffic flow, infrastructure, and corruption with indices of 0.92, 0.88, 0.84, 0.80, 0.75, 0.73, and 0.68, respectively. The research identified PSV, policing, and traffic flow as a triad that accurately summarizes the issues affecting the transport industry in the country and that need urgent attention.

The results of the user-generated, unstructured data suggested 6 interpretable policy interventions and practices that have shaped the transport industry in the last 5 years. These are NTSA - NPS collaborative road monitoring efforts, floods and heavy rainfalls, an influx in transport during festive seasons, commissioning bypass roads, construction of Nairobi expressway, and COVID-19 travel restrictions. Further, the practices and events agree with the triad of summary identified. These findings are important in relation to concerted efforts in solution refinement. The agreement of this user-generated data with policies and practices is a potential feat that would greatly benefit traffic-related research, particularly African-based studies, that suffer from data inadequacy.

We recommend an integrated driver-monitoring system to ease supplement policing efforts and bring transparency in motorist traffic-related fault-finding. Such a system would integrate properly with the existing technology of speed-governors and further yield traffic-related data that is inefficiently collected at present.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

None.

Acknowledgment

We wish to recognize the tremendous assistance and advice we received from Dr. Justice. O. Odoi., and Mr. Samuel K. Kariuki, ASANTENI SANA!

References

[1] World Health Organization, Global Status Report on Road Safety 2018, 2018.
[2] A.A. Mohammed, K. Ambak, A.M. Mosa, D. Syamsunur, A review of the traffic accidents and related practices worldwide, Open Transp. J. 13 (1) (2019) 65–83, https://doi.org/10.2174/1874447801913010065.
[3] World Health Organization, Road Safety in Ten Countries: Kenya, [Online]. Available: http://data.un.org/CountryProfile.aspx?crName=Kenya 2009 [Accessed: 30-Mar-2020].
[4] World Bank, Mortality caused by road traffic injury (per 100,000 population), World Bank (2019) , [Online]. Available: https://data.worldbank.org/indicator/SH.STA.TRAF.P5 [Accessed: 31-Aug-2021].
[5] J.K. Muguro, M. Sasaki, K. Matsushita, W. Njeri, Trend analysis and fatality causes in Kenyan roads: a review of road traffic accident data between 2015 and 2020, Cogent Eng. 7 (1) (2020) 1797981, https://doi.org/10.1080/23311916.2020.1797981.
[6] M.S. Fraser, B.W. Wachira, A.D. Flaxman, A.Y. Lee, H.C. Duber, Impact of traffic, poverty and facility ownership on travel time to emergency care in Nairobi, Kenya, African J. Emerg. Med. 10 (1) (2020) 40–45, https://doi.org/10.1016/j.afjem.2019.12.003.
[7] E. Petridou, M. Moustaki, Human factors in the causation of road traffic crashes, Eur. J. Epidemiol. 16 (9) (2000) 819–826, https://doi.org/10.1023/A:1007649804201.
[8] X. Kong, S. Das, K. Jha, Y. Zhang, Understanding speeding behavior from naturalistic driving data: applying classification based association rule mining, Accid. Anal. Prev. 144 (2020), 105620 https://doi.org/10.1016/j.aap.2020.105620.
[9] B. Yu, Y. Chen, S. Bao, Quantifying visual road environment to establish a speeding prediction model: an examination using naturalistic driving data, Accid. Anal. Prev. 129 (2019) 289–298, https://doi.org/10.1016/j.aap.2019.05.011.
[10] C. Nicholas Chepchieng, Effects of Road Improvement on Safety: A Case Study of Nairobi-Thika Superhighway, Am. J. Civ. Eng. 3 (6) (2015) 199, https://doi.org/10.11648/j.ajce.20150306.11.
[11] C.B. Casady, Customer-led mobility: a research agenda for mobility-as-a-service (MaaS) enablement, Case Stud. Transp. Policy 8 (4) (2020) 1451–1457, https://doi.org/10.1016/j.cstp.2020.10.009.
[12] D. Prashar, N. Jha, S. Jha, G.P. Joshi, C. Seo, Integrating IoT and Blockchain for Ensuring Road Safety: An Unconventional Approach, Sensors 20 (11) (2020) https://doi.org/10.3390/s20113296.
[13] A.-E.M. Taha, An IoT architecture for assessing road safety in smart cities, Wirel. Commun. Mob. Comput. 2018 (2018) 8214989, https://doi.org/10.1155/2018/8214989.
[14] D. Salon, S. Gulyani, Commuting in urban Kenya: unpacking travel demand in large and small Kenyan cities, Sustain 11 (14) (2019) 1–22, https://doi.org/10.3390/su11143823.
[15] D.M. Matheka, F.A. Omar, C. Kipsaina, J. Witte, Road traffic injuries in Kenya: a survey of commercial motorcycle drivers, Pan Afr. Med. J. 21 (2015) https://doi.org/10.11604/pamj.2015.21.17.5646.
[16] K. Mkutu, T.R. Mkutu, Public health problems associated with 'boda boda' motorcycle taxis in Kenya: The sting of inequality, Aggression and Violent Behavior 47 (2019) 245–252, https://doi.org/10.1016/j.avb.2019.02.009 Elsevier Ltd.
[17] J.G. Myers, et al., Patient characteristics of the Accident and Emergency Department of Kenyatta National Hospital, Nairobi, Kenya: A cross-sectional, prospective analysis, BMJ Open 7 (10) (2017) https://doi.org/10.1136/bmjopen-2016-014974 BMJ Publishing Group.
[18] G.G. Hordofa, S. Assegid, A. Girma, T.D. Weldemarium, Prevalence of fatality and associated factors of road traffic accidents among victims reported to Burayu town police stations, between 2010 and 2015, Ethiopia, J. Transp. Heal. 10 (2018) 186–193, https://doi.org/10.1016/j.jth.2018.06.007.
[19] F. Ye, D. Lord, Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered Probit, and mixed logit, Transp. Res. Rec. 2241 (1) (2011) 51–58, https://doi.org/10.3141/2241-06.
[20] S. Dabiri, K. Heaslip, Twitter-based traffic information system based on vector representations for words, ArXiv (2018) 1–17.
[21] J. Johnson, Number of Internet Users in Selected Countries in Africa as of December 2020, by Country, 2021.
[22] J. Degenhard, Forecast of the Number of Twitter Users in Eastern Africa from 2017 to 2025, 2021.
[23] F. Rebelo, C. Soares, and R. J. F. Rossetti, "TwitterJam: Identification of mobility patterns in urban centers based on tweets," 2015 IEEE 1st Int. Smart Cities Conf. ISC2 2015, pp. 0–5, 2015, https://doi.org/10.1109/ISC2.2015.7366156.
[24] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, P. Chaovalit, Social-based traffic information extraction and classification, 2011 11th Int. Conf. ITS Telecommun. ITST 2011, no. August 2011, pp. 107–112, https://doi.org/10.1109/ITST.2011.6060036.
[25] W. Yao, S. Qian, From twitter to traffic predictor: next-day morning traffic prediction using social media data, Transp. Res. Part C Emerg. Technol. 124 (January) (2021) 102938, https://doi.org/10.1016/j.trc.2020.102938.
[26] N. Chaturvedi, D. Toshniwal, M. Parida, Twitter to transport: geo-spatial sentiment analysis of traffic tweets to discover People's feelings for urban transportation issues, J. East. Asia Soc. Transp. Stud. 13 (2019) 210–220, https://doi.org/10.11175/easts.13.210.
[27] D. Bhowmick, S. Winter, M. Stevenson, Using georeferenced twitter data to estimate pedestrian traffic in an urban road network, Leibniz Int. Proc. Informatics, LIPIcs 177 (1) (2020) 1–15, https://doi.org/10.4230/LIPIcs.GIScience.2021.I.1.
[28] J. Salazar-carrillo, M. Torres-ruiz, C.A. Davis, R. Quintero, M. Moreno-ibarra, G. Guzmán, Traffic congestion analysis based on a web-gis and data mining of traffic events from twitter, Sensors 21 (9) (2021)https://doi.org/10.3390/s21092964.
[29] S. Milusheva, R. Marty, G. Bedoya, E. Resor, S. Williams, A. Legovini, Can crowdsourcing create the missing crash data? COMPASS 2020 - Proc. 2020 3rd ACM SIGCAS Conf. Comput. Sustain. Soc., no. July 2017 2020, pp. 305–306, https://doi.org/10.1145/3378393.3402264.
[30] R. Rahman, K.C. Roy, M. Abdel-Aty, S. Hasan, Sharing real-time traffic information with travelers using twitter: an analysis of effectiveness and information content, Front. Built Environ. 5 (June) (2019) 1–15, https://doi.org/10.3389/fbuil.2019.00083.
[31] D.R. Bild, Y. Liu, R.P. Dick, Z.M. Mao, D.S. Wallach, Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph, ACM Trans. Internet Technol. 15 (1) (2015) https://doi.org/10.1145/2700060.
[32] S. Bashir, et al., Twitter chirps for Syrian people: sentiment analysis of tweets related to Syria chemical attack, Int. J. Disaster Risk Reduct. 62 (2021), 102397 https://doi.org/10.1016/j.ijdrr.2021.102397.
[33] G. Ertek, L. Kailas, Analyzing a decade of wind turbine accident news with topic modeling, Sustainability 13 (22) (2021) https://doi.org/10.3390/su132212757.
[34] N. Hu, T. Zhang, B. Gao, I. Bose, What do hotel customers complain about? Text analysis using structural topic model, Tour. Manag. 72 (March 2018) (2019) 417–426, https://doi.org/10.1016/j.tourman.2019.01.002.
[35] J. Dehler-Holland, M. Okoh, D. Keles, Assessing technology legitimacy with topic models and sentiment analysis – The case of wind power in Germany, Technol. Forecast. Soc. Change (2021) 121354, https://doi.org/10.1016/j.techfore.2021.121354.
[36] K.D. Kuhn, Using structural topic modeling to identify latent topics and trends in aviation incident reports, Transp. Res. Part C Emerg. Technol. 87 (January 2018) (2018) 105–122, https://doi.org/10.1016/j.trc.2017.12.018.
[37] M.E. Roberts, B.M. Stewart, E.M. Airoldi, A model of text for experimentation in the social sciences, J. Am. Stat. Assoc. 111 (515) (2016) 988–1003, https://doi.org/10.1080/01621459.2016.1141684.
[38] M. Chandelier, A. Steuckardt, R. Mathevet, S. Diwersy, O. Gimenez, Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling, Biol. Conserv. 220 (2018) 254–261, https://doi.org/10.1016/j.biocon.2018.01.029.
[39] K.M. Kwayu, V. Kwigizile, K. Lee, J.-S. Oh, T. Nelson, Automatic topics extraction from crowdsourced cyclists near-miss and collision reports using text mining and artificial neural networks, Int. J. Transp. Sci. Technol. (2021) https://doi.org/10.1016/j.ijtst.2021.10.005.
[40] F. Ali, A. Ali, M. Imran, R.A. Naqvi, M.H. Siddiqi, K.-S. Kwak, Traffic accident detection and condition analysis based on social networking data, Accid. Anal. Prev. 151 (2021), 105973 https://doi.org/10.1016/j.aap.2021.105973.
[41] X. Bai, X. Zhang, K.X. Li, Y. Zhou, K.F. Yuen, Research topics and trends in the maritime transport: a structural topic model, Transp. Policy 102 (May 2020) (2021) 11–24, https://doi.org/10.1016/j.tranpol.2020.12.013.
[42] M. Sujon, F. Dai, Social Media Mining for Understanding Traffic Safety Culture in Washington state using twitter data, J. Comput. Civ. Eng. 35 (1) (2021) 04020059, https://doi.org/10.1061/(asce)cp.1943-5487.0000943.
[43] W.V. Mitullah, S.S. Onsate, Formalising the Matatu industry in Kenya: policy twists and turns, Univ. Nairobi, IDS Policy Br. no. 8 (2) (2013) 1–4.
[44] K.R. Hope, Police corruption and the security challenge in Kenya, African Secur. 11 (1) (2018) 84–108, https://doi.org/10.1080/19392206.2017.1419650.
[45] Transparency International, Traffic Legislation Gaps and Drivers of Corruption in Traffic Matters Transparency International Kenya, 2018.
[46] K. Bucsuházy, E. Matuchová, R. Zůvala, P. Moravcová, M. Kostíková, R. Mikulec, Human factors contributing to the road traffic accident occurrence, Transportation Research Procedia 45 (2020) 555–561, https://doi.org/10.1016/j.trpro.2020.03.057.
[47] NHTSA, Research Note: Preview of Motor Vehicle Traffic Fatalities in 2019, 2019.
[49] Resor Elizabeth, Nairobi Accident Map, [Online]. Available: https://nairobiaccidentmap.com/about/ 2016 [Accessed: 26-Mar-2021].
[50] T. Onsarigo, P. Chege, O. Egesah, Government policies, practices and Laws on bribery and how they intersect with Matatu operators narratives in Kisii County, Target J. 1 (2019) 1–7.
[51] N.J. Raynor, T. Mirzoev, Understanding road safety in Kenya: views of matatu drivers, Int. Health 6 (3) (2014) 242–248, https://doi.org/10.1093/inthealth/ihu034.
[52] V. Truelove, J. Freeman, E. Szogi, S. Kaye, J. Davey, K. Armstrong, Beyond the threat of legal sanctions: what deters speeding behaviours? Transp. Res. Part F Traffic Psychol. Behav. 50 (2017) 128–136, https://doi.org/10.1016/j.trf.2017.08.008.

[53] J. Muchiri, "Matatu Owners Complain about Police Seeking Bribes," Standard Media, Nairobi, 2020.
[54] M. Ongechi, "Static Traffic Officers to Be Removed from Roads, Says IG Mutyambai," Citizen Digital, Nairobi, 2020.
[56] W. Odero, M. Khayesi, P.M. Heda, Road traffic injuries in Kenya: magnitude, causes and status of intervention, Inj. Control. Saf. Promot. 10 (1–2) (2003) 53–61, https://doi.org/10.1076/icsp.10.1.53.14103.
[57] Kenda Mutongi, Matatu, University of Chicago Press, 2017.
[59] A. Walcott-Bryant, et al., Harsh brakes at potholes in Nairobi: Context-based driver behavior in developing cities, IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2016, pp. 675–681, https://doi.org/10.1109/ITSC.2016.7795626.
[60] K. Engineer, Road Carnage: who takes responsibility for wrong design? - Kenya Engineer, [Online]. Available: https://www.kenyaengineer.co.ke/road-carnage-who-takes-responsibility-for-wrong-design/ 2016 [Accessed: 24-Mar-2021].
[61] W. Bank, Kenya - Tweet IDs From Ma3Route 2012–2020, [Online]. Available: https://microdata.worldbank.org/index.php/catalog/3820 2020.
[62] J. Muchiri Njeru, E. Odira Abade, A survey on big data analytics architecture for urban transportation system: a case for Nairobi metropolitan, Int. J. Comput. Appl. 175 (16) (2020) 36–42, https://doi.org/10.5120/ijca2020920665.
[63] S. Milusheva, R. Marty, G. Bedoya, S. Williams, E. Resor, A. Legovini, Applying machine learning and geolocation techniques to social media data (twitter) to develop a resource for urban planning, PLoS One 16 (2) (2021), e0244317 https://doi.org/10.1371/journal.pone.0244317.
[64] D. Santani, et al., CommuniSense: crowdsourcing road hazards in Nairobi, MobileHCI'15, 2015 https://doi.org/10.1145/2785830.2785837.
[65] C.-L. Liu, T.-Y. Hsu, Y.-S. Chuang, H. Lee, What makes multilingual BERT multilingual? ArXiv 2010.10938 (2020).
[66] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, TWEETEVAL: unified benchmark and comparative evaluation for tweet classification, ArXiv:2010.12421 (2020) 1644–1650, https://doi.org/10.18653/v1/2020.findings-emnlp.148.
[67] F. Barbieri, L.E. Anke, J. Camacho-Collados, XLM-T: a multilingual language model toolkit for twitter, ArXiv:2104.12250 no. 2015 (2021).
[68] NPS, NPS to assign 200 police officers to NTSA, *National Police Service (NPS) Media*, 2015 , [Online]. Available: https://www.nationalpolice.go.ke/2015-09-08-17-56-33/news/156-nps-to-assign-200-p0lice-officers-to-ntsa.html.
[69] PDU, President Kenyatta Orders NTSA Off The Roads, Wants Traffic Management Left To Traffic Police, *President Delivery Unit (PDU)*, 2018 , [Online]. Available: https://www.president.go.ke/2018/01/09/president-kenyatta-orders-ntsa-off-the-roads-wants-traffic-management-left-to-traffic-police/.
[70] W. Ndungu, NTSA ordered off roads, now police to enforce traffic rules, Standard Media (2018).
[71] PDU, PRESIDENTIAL ADDRESS ON THE STATE INTERVENTIONS TO CUSHION KENYANS AGAINST ECONOMIC EFFECTS OF COVID-19 PANDEMIC ON 25TH MARCH, 2020, *President Delivery Unit (PDU)*, 2020 , [Online]. Available: https://www.president.go.ke/2020/03/25/presidential-address-on-the-state-interventions-to-cushion-kenyans-against-economic-effects-of-covid-19-pandemic-on-25th-march-2020/.
[72] MoH, President Uhuru lifts movement ban in three counties Nairobi , Monday July 6, 2020, *Ministry of Health (MoH)*, 2020 , [Online]. Available: https://www.health.go.ke/president-uhuru-lifts-movement-ban-in-three-counties-nairobi-monday-july-6-2020/.
[73] S. Stephanie, Busari Farai, At least 50 killed in Kenya bus crash, CNN News (2018).
[74] Xinhua, Chinese-built ring roads reduce traffic jams in Kenya, Xinhua News (2016) , [Online]. Available: http://www.xinhuanet.com/english/africa/2021-07/20/c_1310073053.htm [Accessed: 24-Dec-2021].
[75] O. Mathenge, Presidents Uhuru and Magufuli open southern bypass to decongest city, The Star (2016).
[76] M. Kilavi, et al., Extreme rainfall and flooding over Central Kenya including Nairobi City during the long-rains season 2018: causes, predictability, and potential for early warning and actions, Atmosphere 9 (12) (2018)https://doi.org/10.3390/atmos9120472.
[77] H.W. Njogu, Effects of floods on infrastructure users in Kenya, J. Flood Risk Manag. 14 (4) (2021) 1–10, https://doi.org/10.1111/jfr3.12746.
[78] M. Fortnam, et al., Multiple impact pathways of the 2015-2016 El Niño in coastal Kenya, Ambio 50 (1) (2021) 174–189, https://doi.org/10.1007/s13280-020-01321-z.
[79] P. Njoroge, Letter from Africa: How the Nairobi Expressway is changing Kenya's capital, BBC News (2021) , [Online]. Available: https://www.bbc.com/news/world-africa-55995229 [Accessed: 24-Dec-2021].
[80] A. Denis, Nairobi Expressway project timeline and all you need to know, Construction Review Online (2021) , [Online]. Available: https://constructionreviewonline.com/biggest-projects/nairobi-expressway-project-timeline-and-all-you-need-to-know/.
[81] L. Diaz Olvera, D. Plat, P. Pochet, Looking for the obvious: Motorcycle taxi services in Sub-Saharan African cities, J. Transp. Geogr. (2019) 102476, https://doi.org/10.1016/J.JTRANGEO.2019.102476.
[85] D. Adeloye, et al., The burden of road traffic crashes, injuries and deaths in Africa: a systematic review and meta-analysis, Bull. World Health Organ. 94 (7) (2016) 510–521A, https://doi.org/10.2471/BLT.15.163121.
[87] K.J. Kelly, WHO: Kenya road deaths four times higher than NTSA reported, Daily Nation (2018) , [Online]. Available: https://www.nation.co.ke/news/Kenya-road-deaths-grossly-underreported–WHO/1056-4893792-ve7d07z/index.html.
[88] G. Murage, Residents Loot Flour from Lorry Involved in Accident, The Star, Nairobi, 2020.
[89] I. Otieno, Police Teargas Residents Looting Beer from Overturned Truck, Kenyans, Nairobi, 2020.
[90] D.S. Usami, et al., Defining suitable safe system projects: the experience of the SaferAfrica project in five African countries, IATSS Res. (2021)https://doi.org/10.1016/j.iatssr.2021.08.001.
[91] T. Haramaki, H. Nishino, An edge computer based driver monitoring system for assisting safety driving BT - advances in internet, Data & web Technologies, in: L. Barolli, F. Xhafa, N. Javaid, E. Spaho, V. Kolici (Eds.), Advances in Internet, Data & web Technologies, Springer International Publishing, Cham 2018, pp. 639–650.
[92] D. Grewe, M. Wagner, M. Arumaithurai, I. Psaras, D. Kutscher, Information-centric mobile edge computing for connected vehicle environments: Challenges and research directions, *MECOMM* 2017 - *Proc. 2017 Work. Mob. Edge Commun. Part SIGCOMM 2017* 2017, pp. 7–12, https://doi.org/10.1145/3098208.3098210.
[93] S. Arumugam, R. Bhargavi, A survey on driving behavior analysis in usage based insurance using big data, J. Big Data 6 (1) (2019) Dec, https://doi.org/10.1186/s40537-019-0249-5.
[94] I. Jegham, A. Ben Khalifa, I. Alouani, M.A. Mahjoub, A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD, Signal Process. Image Commun. 88 (Oct. 2020) 115960, https://doi.org/10.1016/j.image.2020.115960.
[95] P.C. Vassallo, José Manuel; Bueno, *Transport Challenges in Latin American Cities: Lessons Learnt from Policy Experiences*, IDB Felipe Herrera Library, Washington, DC, 2019.