A framework to analyse impacts of the exception events on the buses velocity of São Paulo

Author omitted for blind review

No Institute Given

Abstract This work presents a framework to understand the impacts of exception events in the buses velocity of São Paulo, through modules able to identify exception events and integrate them with real heterogeneous data related to public transportation.

Keywords: Smart cities · social networks · public transportation.

1 Introduction

Exception events happens sporadically or suddenly and are capable of generating significant delays or even unavailability of the operation of public transport. This type of events are reported by citizens and authorities in Social Networks, which can be used by Smart City systems. As an example, the public transport can benefit by integrating Social Networks content with the planning, management and operational activities of public transport, addressing their respective sociotechnical factors [13]. In this work we present a new framework validated with real and heterogeneous data sources (*tweets*, AVL—Automatic Vehicle Location and GTFS[1]—Google Transit Feed Specification data) to understand automatically exception events impacts in the São Paulo' buses velocity.

2 Related Work

Several works studies how to use *tweets* processing for analyzing problems related to public transport. These studies can be classified into *event impact analysis*, *planning* and *management of public transport*. For example, [23] used *tweets* to analyze the impact of the terrorist attacks in Paris (2015) on mobility patterns regarding the use of public transport. Similarly, [9] developed a tool based on *tweets* to visualize and explore the decisions of passengers of the Tokyo Metro in abnormal events such as typhoons, fires, earthquakes, etc. In this same context, [19] proposed a technique to predict passenger flow in the New York Metro and identified events based on *hashtags*. [1] studied the relationship between traffic events and the demand for bicycles.

[1] Google Transit: `https://developers.google.com/transit`. Accessed in April 14, 2019.

With respect to public transport planning and management, [16] presents a framework developed and used by the Bangalore Public Transport Agency, which allowed issues related to public transport to be reported, improving the operation planning and the service provided to the population. Analogously, [8] used *tweets* to identify the popularity of points of interest and age distribution in order to determine the best points for bicycle stations and, thus, encourage the use of this mode of transport. Also related to the points of interest, [15] used *tweets* to identify human activities patterns and their respective impacts on the demand for public transport.

In [4] a hierarchical approach was created to classify *tweets* related to transport. They have demonstrated that it is possible to use this information for transportation planning and management purposes. This technique was applied in a case study associated with sporting events in the United Kingdom. The hierarchy is composed of three levels (I) *tweets* classified among those that express the need for transport services, opinions and incidents; (II) identification of the transport category and (III) topics.

Another study that contributes to the planning of public transport is the one carried out in [5,6], in which *tweets* were processed to identify user disposition to trips related to leisure, suggesting to them activities with less time of travel and probability of delays. Another relevant point considered was the level of access to public transport, which, when high, positively impacts people's happiness and correlates with positive feelings, according to the analysis of feelings carried out by [7], using *tweets* published in Greater London.

Neither of the presented works tackle the identification of different types of exception events from *tweets* published by an authority to characterize the velocity impact on buses of São Paulo. In this work we propose a new framework, explained ahead, for deal with this problem. The cited works are connected to our on aspects related to *tweets* processing for analysis of the impact of events on public transport, planning and management.

3 Dataset

Corpus Twitter. *Twitter* was chosen as a data source for the construction of the dataset related to exception events. Twitter contains public data about the daily life of the city, made available in real time by citizens and public agencies. These characteristics make tweets a rich source of data and is used by numerous studies addressing urban and urban mobility problems, as analyzed in Section 2. In this work, the dataset used to identify the exception events is composed by *tweets*—written in Brazilian Portuguese—of the profiles contained in Table 1, which were selected manually according to the agencies responsible for reporting exception events. Such profiles are public in nature, meaning access to tweets does not involve privacy issues.

Corpus SPTrans. The SPTrans (São Paulo Transportation Company)[2] corpus consist of GTFS and geolocation data of all the buses of São Paulo,

[2] `http://www.sptrans.com.br`. Acessed in April 14, 2019.

Table 1. TIME INTERVAL AND NUMBER OF TWEETS COLLECTED

***Twitter* profile**	**Total (Ttl.) *tweets***	**Start date**	**End date**
@BombeirosPMESP	6,632	2017-05-21	2017-12-01
*@CETSP_*	5,735	2017-02-20	2017-12-01
@CPTM_oficial	6,301	2017-04-24	2017-12-01
@governosp	6,011	2017-05-10	2017-12-01
@metrosp_oficial	8,621	2017-06-07	2017-12-01
@Policia_Civil	3,417	2015-04-15	2017-11-30
@PMESP	4,365	2016-06-02	2017-11-30
@saopaulo_agora	3,960	2016-11-18	2017-11-30
*@smtsp_*	1,316	2017-04-26	2017-12-01
@SPCEDEC	1,301	2015-06-09	2017-12-01
*@sptrans_*	9,956	2017-06-13	2017-12-01
@TurismoSaoPaulo	3,369	2012-06-12	2017-11-29
Total	60,984	—	—

detailed, respectively, in tables 2 and 3[3], referring to the year of 2017[4]. The original data was converted from *string* to its respective type (*long*, *double*, *int* or *string*), time values were standardized using *POSIX timestamps*, and data referring to latitude and longitude were converted to *legacy coordinate pairs*[5]. In order to enable *geospatial queries*, *geospatial indexes*[5] were created in the *MongoDB* collections containing geolocalized information.

4 Framework to identify impacts on buses speeds in the city of São Paulo

Framework The architecture proposed for the identification framework is depicted in Figure 1. It consists of (1) a *tweets* processing streaming module (which can be replaced by any real time processing software), where the *tweets* are collected, preprocessed and processed; (2) classification models; (3) address extraction and geolocation module and (4, 5 and 6[6]) correlation data modules. These modules are explained in details below.

[3] November missing data: 01/11 — from 12h to 15h. December missing data: 15/12 — from 01h to 09h. We also found inconsistencies in two AVL files of January 11 provided by SPtrans. According to the SPTrans metadata, each file must have 19 fields; however, the file with data from 09h to 10h has 21 fields in line 1,075,548 and the file with data from 10h to 11h has 35 fields in line 60,025. The gaps mentioned before were ignored in processing.

[4] The data was obtained using the Brazilian Law of Access to Information: `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm` (in Portuguese). Accessed in April 14, 2019.

[5] `https://docs.mongodb.com/manual/geospatial-queries`. Accessed in April 14, 2019.

[6] Finding abnormal association rules in the velocity dataset (step 6 in Fig. 1) is still a work in progress.

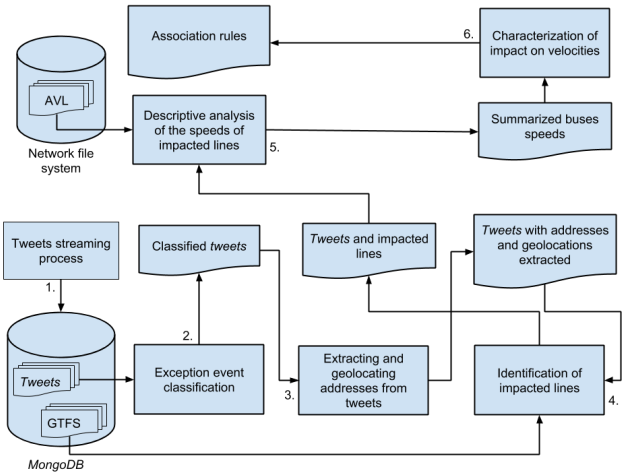
Table 2. Dataset and total records specified in SPTrans' GTFS

Dataset	Ttl. records
agency.txt	1
calendar.txt	6
fare_attributes.txt	6
fare_rules.txt	5,400
frequencies.txt	39,625
routes.txt	291,634
shapes.txt	800,767
stop_times.txt	95,144
stops.txt	19,933
trips.txt	2,273
Total	1,254,779

Table 3. SPTrans'AVL dataset

Month	Ttl. AVL files	Ttl. size (GB)
January	744	102,44
February	672	93,21
March	744	102,64
April	720	97,04
May	744	101,46
June	720	97,13
July	744	104,95
August	744	108,38
September	720	109,89
October	744	110,92
November	717	108,16
December	738	110,89
Total	8,751	1,247,09

Figure 1. Framework Architecture



Natural Language Processing Before the Natural Language Processing (NLP) phase, the *tweets* were preprocessed—removing *URLs*, *datetime*, mentions to other *tweets*, emoticons, punctuations—in order to reduce the dimension of the feature space. A particular attention was paid to *hashtags*, which are relevant to exception events classification, but adds noise to the address extraction phase — all *hashtags* were identified and replaced by empty spaces in the address extraction process, after classification.

After the preprocessing phase, we applied NLP techniques to *tweets*, such as (I) *Tokenization* — process to obtain the words (tokens) of a sentence (features used to train the classification model), removing numbers and characters

that do not belong to the alphabet (through *TweetTokenizer*[7]), (II) morphological decomposition to get a given word into its inflected form using stemming (through *RSLPStemmer*[8]) and (III) Brazilian Portuguese *stopwords* remotion[9] [2, 11, 17, 21, 22]. These procedures reduced the number of tweets from 60,984 to 59,834, resulting in a corpus with 414,637 words and with a vocabulary of 13,915 words. This corpus was processed to be represented by a bag-of-words, which is composed by feature vectors created using the *Term Frequency - Inverse Document Frequency* (TF-IDF) measure. These feature vectors are used by classification models.

Classification model Finding exception events involves classification task. Based on previous studies on exception events [1,4,9,14] we have chose to use the following exception classes to *tweets*: (I) accidents, (II) social events, (III) urban events, (IV) natural disasters and (V) irrelevant. Using these classes, the *tweets* were manually classified and used to train models to classify them in exception events. In this work, we used supervised learning, since we know how the input data can be classified, being the following algorithms normally applied to classify textual datasets: (Complement and Multinomial) Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Multi-layer Perceptron (MLP), Random Forest (RF) and Support Vector Machine (SVM) [3, 12, 18], more details of the experimental setup in Section 5.

Addresses and geolocation extraction Analyzing the content of *tweets* from the selected accounts, it is possible to observe that the texts published normally follows a given template and, therefore, are actually semi-structured. We used a regular expression to extract addresses from *tweets*: $ER = \{L_1|S_1|L_2|S_2|\ldots|L_n|S_n\}\{[a - z\grave{A} - \ddot{y}\_]+\}$. That expression is divided in two sets, in the first ($\{L_1|S_1|L_2|S_2|\ldots|L_n|S_n\}$), (L—in Portuguese: *logradouro*, meaning public spaces such as avenue, etc.) and (S—public spaces acronyms) are concatenated to specify a filter and identify *strings* initialized with public spaces or its respective acronyms. The second set ($\{[a - z\grave{A} - \ddot{y}\_]+\}$), represents a filter to identify a set of words after L or S, candidates to compose the wanted addresses. These words are treated as *candidates* because it is hard to know in advance how many words after L or S belongs to the address. However, the selected accounts publish *tweets* with visible patterns in the texts, after and before the addresses. As a consequence, a possible method to find the wanted address is the removal of these patterns after and before of the address. After address extraction, we used the Google Maps Geocoding API[10] to geolocate the found

[7] NLTK module used to the tokenization process. `https://www.nltk.org/api/nltk.tokenize`. Accessed in April 14, 2019.

[8] NLTK module used to the stemming process. `https://www.nltk.org/_modules/nltk/stem/rslp`. Accessed in April 14, 2019.

[9] Brazilian Portuguese *stopwords* were obtained from NLTK — `https://www.nltk.org`. Accessed in April 14, 2019.

[10] `https://developers.google.com/maps/documentation/geocoding`. Accessed in April 14, 2019.

address (only 1.5% of *tweets* have geolocation [20]). The HTTP response from this API is processed to get the values from location (which contains latitude and longitude information) and formatted address.

Finding buses lines affected by exception events In order to find the buses lines affected by exception events, it is necessary to match the coordinates of exception events with the existing coordinates in the shapes dataset—a set of latitude and longitude used for drawing buses lines on a map to represent its respective paths—existing in SPTrans' GTFS. All coordinates are stored in MongoDB using legacy pairs and in collections with geospatial indexes. We use the `$near` function provided by MongoDB[11] to find shapes close to the exception event coordinates. The GTFS defines that the *shape_id* (i.e. bus code line) is part of attributes contained in the shape file, which is used as key to integrate with others GTFS files (enriching the information about a bus code line).

Velocity impact analysis After we found the buses lines impacted by exception events, we select the movement data that will be analyzed, e.g. if the exception event happened on 08/17/2017 (Thursday), every other Thursday in the month of August (3, 10, 24 and 08/31/2017) will be considered in the analysis. Our preliminary analysis of the data showed that each day of the week and month may show different patterns of movements (seasonality)—see Figure 4. For instance, Fridays are well-known for their congested traffics, while December (the beginning of the Brazilian summer and a period of many holidays) is well-known for having a more fluid traffic flow. Because of this seasonality, velocities are analyzed regarding the day of the week in the same month.

We also filtered the data related to the impacted lines within a radius of 100 and 1,000m of the exception event in question, in addition to considering the same time range as the *tweet* time[12]. So, if the *tweet* time is at 5:15 p.m., we considered the AVL data between 5:00 p.m. and 6:00 p.m.

Next, we aggregated the selected data to descriptively analyze the instantaneous speed of each bus line, thereby extracting data on the maximum, minimum, mean, median, variance, standard deviation and percentage of equal and non-zero data. After that, we compared the average speed of the occurrence time range with the average speed of days that do not reference the exception event, for each set of lines affected by the exception event and for each line. Finally, we considered that the line was impacted if the mean of the average speeds of the analyzed days is greater than or equal to the average speed of the day referring to the exception event. Based on this, we assumed that the set of lines has been impacted if the number of impacted lines is greater than or equal to 50%.

[11] `https://docs.mongodb.com/manual/reference/operator/query/near`. Accessed in April 14, 2019.

[12] It is important to note that this work does not consider the exact start and end of the exception events, but a time range of one hour from the time in the *tweet timestamp*.

5 Experimental Analyses

Classification models The *Corpus Twitter*[13] was the labeled dataset used to train exception events classification models, based on a *bag-of-words*, as described in Section 4. The distribution of the examples for each class follows, 7,07% accident, 83,24% irrelevant—to non exception events, 4,52% natural disaster, 2,35% social event and 2,81% urban event, to maintain this distribution we used the stratified 10-fold cross-validation approach[14]. In this technique, for each training and test configuration, the TF-IDF is computed in the training data, then is used to train the classification methods. For the test data, we only transformed the data (using calculated TF-IDF) to be used in the prediction of the classification methods.

In the experiments, we considered the grid search approach, strategy used to choose a set of optimal hyper-parameters for a learning algorithm [10]. Furthermore, given the high dimensionality of the features vectors, we consider also experiments with dimensionality reduction, where we used the Singular Value Decomposition (SVD) technique (more details in: [24]). A cross-validation approach found that 90% of the data variance is explained with 3,000 components. In the Table 4, we present only the best results achieved for each method. For the evaluation of the statistical methods, we considered the accuracy metric and macro approach for precision, recall and f1-score metrics.

Table 4. Metrics of the evaluations of the algorithms used to classify the *tweets* in exception events

Algorithm	**Accuracy**	**Precision**	**Recall**	***f1-score***
Complement Naive Bayes	0,959	0,874	0,853	0,859
Decision Tree	0,966	0,904	0,885	0,894
K-Nearest Neighbors	0,956	0,869	0,856	0,862
Logistic Regression	0,977	0,942	0,924	0,933
Multi-layer Perceptron	0,974	0,92	0,918	0,919
Multinomial Naive Bayes	0,951	0,867	0,812	0,832
Random Forest	0,971	0,926	0,893	0,908
Support Vector Machine	0,978	0,937	0,936	0,936

According to Table 4, the models using the LR, SVM, MLP and RF algorithms obtained satisfactory performance for the classification task. The better results for LR, SVM and MLP models was achieved with SVD technique, in other methods the approach not improved results.

Addresses and geolocation extraction Of the 60,984 *tweets* 10,027 were classified into exception events and from that subset we found 7,710 addresses

[13] Dataset publicly available at: `https://drive.google.com/open?id=1s_HmLHQFFMVrPhYcrAI1XF7rrGbcJohT`. Accessed in April 14, 2019.

[14] `https://scikit-learn.org/0.16/modules/generated/sklearn.cross_validation.StratifiedKFold.html`. Accessed in April 14, 2019.

(which represents 76.89% of the total of *tweets* classified as exception events. The reasons for *tweets* without address extracted are:

1. *Tweets* with only the point of interest, in other words, the address is not explicitly stated. 2. *Tweets* without address information. 3. *Tweets* with unusual public place name (for example *passageway*, *road complex*, *connection to*). 4. *Tweets* with addresses with concatenated words (for example *avenidapaulista*)

Figure 2 shows most affected addresses[15] by exception events and Figure 3 shows the distribution of these events in the central region of São Paulo. It is important to note that the exception events found are concentrated in the addresses and regions where they normally occur in São Paulo, which validates the methodology developed.

Figure 2. Addresses most impacted by exception events

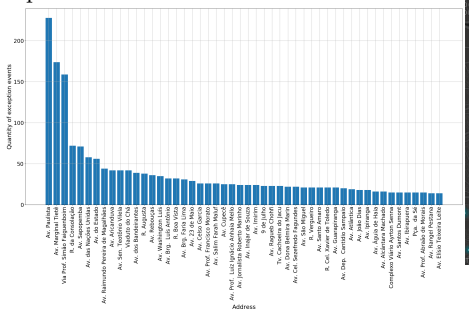
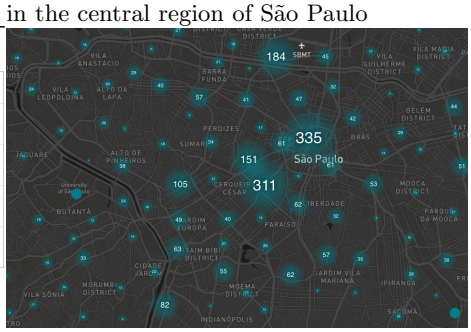


Figure 3. Distribution of exception events in the central region of São Paulo



Source: (Dias et al.) in 2018 IEEE 37th International Symposium on Reliable Distributed Systems Workshops (SRDSW)

Exception event impact analysis We considered that a bus line is affected by an exception event if a coordinate from *shape* is within a radius of 1,000m away from the event. Using this criterion, the total of 1,073 buses lines were affected by exception events during this period, with line "33121 – TERM. PRINC. ISABEL / TERM. STO. AMARO" being the most impacted bus line. This particular line was impacted by 1,623 exception events[16].

Using the methodology described above, we can observe that the social event-related exception events have an average of 87.04% impact on the average speed

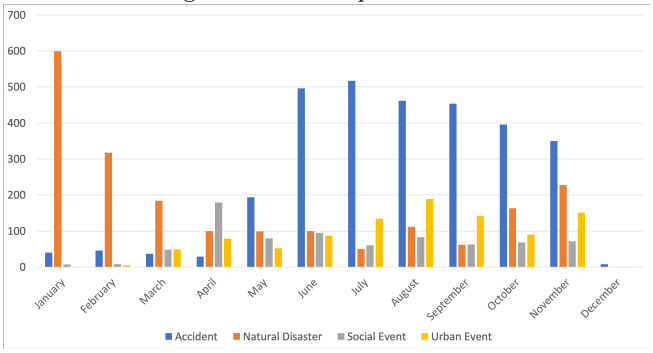
[15] Complete list is available at https://docs.google.com/spreadsheets/d/1gn1cTDifUJEPdgcU67SC45GdYHRKmIHtAfJwRBm088s/edit?usp=sharing. Accessed on April 14, 2019.

[16] The full table is publicly available at `https://docs.google.com/spreadsheets/d/1jIqUuIJg7FhXD5C8MFF8stbvOD3uiUgMfN2bOltT7zE/edit?usp=sharing`. Accessed in April 14, 2019.

Table 5. Percentage of impact on the average speed of the groups of lines affected by exception events at 1,000m and 100m distance respectively, in the months of 2017

Month	***Accident***		***Natural Disaster***		***Social Event***		***Urban Event***	
January	83,33	100	64,23	98,00	100	—	100	—
February	70,58	100	66,25	100	100	100	80	—
March	50,00	—	66,66	100	85,00	100	68,18	100
April	87,50	100	61,11	100	82,75	100	76,92	100
May	65,13	100	58,82	100	93,33	100	50,00	100
June	54,46	100	61,53	100	76,47	100	72,41	100
July	61,48	98,41	66,66	100	69,23	100	58,13	100
August	57,86	87,17	55,35	100	85,54	100	68,10	90,90
September	64,21	100	42,10	100	92,30	100	62,06	100
October	70,49	—	56,81	—	80,00	—	61,11	—
November	66,66	100	57,99	100	92,85	100	74,35	100
December	—	—	—	—	—	—	—	—
Total	66,51	98,39	59,77	99,80	87,04	100	70,11	98,86

Figure 4. Distribution of geolocation exception classes over the months of 2017



in the groups of bus lines affected by a radius of 1,000m and 100% to a radius of 100m, this probably due to the large number of people involved in this type of event, number of avenues with modified or interrupted traffic flow.

Urban events, in turn, impacts 70.11% at 1,000m and 98,86% at 100m, even though these events are being carried out with alternative routes planning and warn signs on public roads. The third and fourth most affected classes are those of accidents and natural disasters, respectively, 66.51% and 59.77 % at 1,000m and 98.39 % and 99.80 % to 100m, which normally blockages or detours on public roads used by buses.

In addition, January, February and March were the three months most affected by exception events related to natural disasters, a period of high rainfall in São Paulo, where landslides, tree falls and floods usually occurs. In relation to social events, the year 2017 was marked with numerous political manifestations, in this context, May was the most impacted month by this type of exception

event, mainly due to the protests against the former Brazilian president Temer[17]. The events related to accidents usually occur in greater concentration in the periods of holidays and holidays, which can be observed in the months of January and April (single month of 2017 with two long weekends), with a mean impact of 83.33% and 87.50% at the average speeds, respectively. Impacts related to urban events occurs normally during all months, due to which they percentages are uniform.

The months close to 100% of impact at average speeds are justified because of the small volume of events for a given class in a given month, as Figure 4, which also happens for scenarios with geolocated data next to the exception events. Similarly, the months and classes without impact data are months with little data for the analyzed class. The table 5 summarizes the impact on the average speed caused by each type of exception on each month of the year.

6 Conclusions and future works

This work presents a new framework validated with real and heterogeneous data sources to understand automatically exception events impacts in the São Paulo' buses velocity. Using *tweets* from selected public service providers, we achieved satisfactory performance in the experiments to classify *tweets* in exception events. We also showed that it is possible to extract addresses from semi-structured *tweets* using only regular expressions. Classifying these events are the first step to better understand how these exceptional events impacts the velocity of buses, using the proposed framework we found that social events reduces the velocity of 87,04% of a group impacted, urban event 70,11%, accident 66,51% and natural disaster 59,77% from a distance of 1,000m. Although validated using selected Twitter profiles written in Brazilian Portuguese, this method can be generalized for different languages and cities. GTFS is a ubiquitous format for public transport and tools like NLTK supports several languages.

6.1 Future work

There is a need to establish a cooperation between the university and SPTrans for the daily application of the experiments carried out by this work and others related to the analysis of large volumes of public transport data. Another future possibility is to apply the experiments carried out by this work to user publications that represent civil society.

References

1. Chen, L., Zhang, D., Wang, L., Yang, D., Ma, X., Li, S., Wu, Z., Pan, G., Nguyen, T.M.T., Jakubowicz, J.: Dynamic cluster-based over-demand prediction in bike

[17] Anti-Temer protests gather hundreds on av. Paulista and Brasília (in portuguese): `http://folha.com/no1884977`. Accessed on April 14, 2019

sharing systems. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 841–852. UbiComp '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2971648.2971652, http://doi.acm.org/10.1145/2971648.2971652

2. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(Aug), 2493–2537 (2011)
3. Dwivedi, S.K., Arya, C.: Automatic text classification in information retrieval: A survey. In: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. p. 131. ACM, Jeddah, KSA (2016)
4. Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I.: The potential of social media in delivering transport policy goals. Transport Policy **32**, 115 – 123 (2014). https://doi.org/https://doi.org/10.1016/j.tranpol.2014.01.007, http://www.sciencedirect.com/science/article/pii/S0967070X14000225
5. Gkiotsalitis, K., Stathopoulos, A.: Joint leisure travel optimization with user-generated data via perceived utility maximization. Transportation Research Part C: Emerging Technologies **68**, 532 – 548 (2016). https://doi.org/https://doi.org/10.1016/j.trc.2016.05.009, http://www.sciencedirect.com/science/article/pii/S0968090X16300468
6. Gkiotsalitis, K., Stathopoulos, A.: A utility-maximization model for retrieving users willingness to travel for participating in activities from big-data. Transportation Research Part C: Emerging Technologies **58**, 265 – 277 (2015). https://doi.org/https://doi.org/10.1016/j.trc.2014.12.006, http://www.sciencedirect.com/science/article/pii/S0968090X14003568, big Data in Transportation and Traffic Engineering
7. Guo, W., Gupta, N., Pogrebna, G., Jarvis, S.: Understanding happiness in cities using twitter: Jobs, children, and transport. In: 2016 IEEE International Smart Cities Conference (ISC2). pp. 1–7. IEEE, Trento, Italy (Sep 2016). https://doi.org/10.1109/ISC2.2016.7580790
8. Gutev, A., Nenko, A.: Better cycling - better life: Social media based parametric modeling advancing governance of public transportation system in st. petersburg. In: Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia. pp. 242–247. EGOSE '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/3014087.3014123, http://doi.acm.org/10.1145/3014087.3014123
9. Itoh, M., Yokoyama, D., Toyoda, M., Tomita, Y., Kawamura, S., Kitsuregawa, M.: Visual exploration of changes in passenger flows and tweets on mega-city metro network. IEEE Transactions on Big Data **2**(1), 85–99 (March 2016). https://doi.org/10.1109/TBDATA.2016.2546301
10. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, vol. 112. Springer (2013)
11. Korenius, T., Laurikkala, J., Järvelin, K., Juhola, M.: Stemming and lemmatization in the clustering of finnish text documents. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. pp. 625–633. CIKM '04, ACM, New York, NY, USA (2004). https://doi.org/10.1145/1031171.1031285, http://doi.acm.org/10.1145/1031171.1031285
12. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering **160**, 3–24 (2007)

13. Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., Shoor, I.: Automating a framework to extract and analyse transport related social media content: The potential and the challenges. Transportation Research Part C: Emerging Technologies **77**, 275–291 (2017)
14. Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M., Tommasi, P.: Smart traffic analytics in the semantic web with star-city: Scenarios, system and lessons learned in dublin city. Journal of Web Semantics **27-28**, 26 – 33 (2014). https://doi.org/https://doi.org/10.1016/j.websem.2014.07.002, http://www.sciencedirect.com/science/article/pii/S157082681400050X, semantic Web Challenge 2013
15. Maghrebi, M., Abbasi, A., Rashidi, T.H., Waller, S.T.: Complementing travel diary surveys with twitter data: Application of text mining techniques on activity location, type and time. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. pp. 208–213. IEEE, Las Palmas, Spain (Sep 2015). https://doi.org/10.1109/ITSC.2015.43
16. Mukherjee, T., Chander, D., Eswaran, S., Singh, M., Varma, P., Chugh, A., Dasgupta, K.: Janayuja: A people-centric platform to generate reliable and actionable insights for civic agencies. In: Proceedings of the 2015 Annual Symposium on Computing for Development. pp. 137–145. DEV '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2830629.2830642, http://doi.acm.org/10.1145/2830629.2830642
17. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. Journal of the American Medical Informatics Association **18**(5), 544–551 (2011)
18. Narayanan, U., Unnikrishnan, A., Paul, V., Joseph, S.: A survey on various supervised classification algorithms. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). pp. 2118–2124. IEEE, Chennai, India (2017)
19. Ni, M., He, Q., Gao, J.: Forecasting the subway passenger flow under event occurrences with social media. IEEE Transactions on Intelligent Transportation Systems **18**(6), 1623–1632 (June 2017). https://doi.org/10.1109/TITS.2016.2611644
20. Niu, W., Caverlee, J., Lu, H., Kamath, K.: Community-based geospatial tag estimation. In: Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. pp. 279–286. IEEE, Davis, California (2016)
21. Roy, A., Majumder, A.G., Nath, A.: Understanding natural language processing and its primary aspects. International Journal **5**(8) (2017)
22. Setiawan, E.B., Widyantoro, D.H., Surendro, K.: Feature expansion using word embedding for tweet topic classification. In: 2016 10th International Conference on Telecommunication Systems Services and Applications (TSSA). pp. 1–5. IEEE, Denpasar, Indonesia (Oct 2016). https://doi.org/10.1109/TSSA.2016.7871085
23. Wen, X., Lin, Y., Pelechrinis, K.: Pairfac: Event analytics through discriminant tensor factorization. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 519–528. ACM, Indianapolis, Indiana, USA (2016)
24. Zhang, W., Yoshida, T., Tang, X.: A comparative study of tf* idf, lsi and multi-words for text classification. Expert Systems with Applications **38**(3), 2758–2765 (2011)