A platform to analyse impacts of the exception events on the bus velocity of São Paulo*

No Author Given

No Institute Given

Abstract This work uses supervised learning to classify tweets in exception events, which are correlated with real data sources (AVL and GTFS) to understand the impacts of these events in the buses velocity of São Paulo.

Keywords: Smart cities · social networks · public transportation.

1 Introduction

In São Paulo 10% of population lives in the Expanded Center area and 90% in the Peripheral Belt [29], which characterizes an urban segregation responsible for numerous problems related to urban mobility. Especially in segregated cities, exception events are capable of generating significant delays or even unavailability of the operation of public transport. Exception events are events that happen sporadically or suddenly such as manifestation, sporting events, floods, tree falls, fires, accidents, etc.

All exception events previously mentioned 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 [17]. In this work we aims to use supervised learning techniques to classify tweets (from selected accounts) in exception events, which are correlated with real data sources (AVL and GTFS) to understand the impacts of these events in the buses velocity of the city of São Paulo.

The platform allows to detect messages that refer to exception events published on social networks and automatically detect which lines will be affected and estimate how the velocity those lines will be affected. To achieve this objective we (I) trained several models to classify exception events reported by the selected profiles, (II) developed a process to addresses extraction and geolocalization based on *tweets*, which are (III) correlated with SPTrans's (responsible for the bus lines of the municipality of São Paulo) GTFS (commonly used to

* This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq, proc. 465446/2014-0, CAPES proc.88887.136422/2017-00, and FAPESP, proc. 2014/50937-1.

describe public transport data) data to find bus lines impacted by exception events in the São Paulo city and with (IV) AVL data to velocity impact characterization. Using this methodology we characterized 60,984 events and found 10,027 exception events that impacted 1,073 bus lines. Besides, we found that social events have an average of 87,04% impact on the average speedy of bus lines affected by a radius of 1,000m; urban events 70,11%; accidents 66,51% and natural disasters 59,77%.

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, [31] used *tweets* to analyze the impact of the terrorist attacks in Paris (2015) on mobility patterns regarding the use of public transport. Similarly, [14] 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, [25] proposed a technique to predict passenger flow in the New York Metro and identify events based on *hashtags*. [4] studied the relationship between traffic events and the demand for bicycles.

In respect to public transport planning and management, [21] presents a platform developed and used by the Bangalore Public Transport Agency, which allows report issues related to public transport, improving the operation planning and the service provided to the population. Analogously, [12] 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, [20] used *tweets* to identify human activities patterns and their respective impacts on the demand for public transport.

In [8] 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 [9,10], 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 [11], 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 method, 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 Basic concepts

3.1 Social Networks

Social Networks (SN) can be defined as networks that have many relationships, with large connected components, clustering coefficients and degree of reciprocity. Such features, e.g, are found on *Facebook*[1]. Another SN is *Twitter*[2], which besides having the social networking features mentioned can also be characterized as an Information Network. In this type of network the dominant interaction is the dissemination of information between relationships, with low reciprocity index [22].

On *Twitter* the information (*tweets*) is published containing a maximum of 280 characters; each publication can receive *retweets* (to be shared by other users), comments (directly in the *tweet* — *replies* — or privately via the message box) and *likes* (indicator of how many users liked the post), in addition to these features, *tweets* may contain mentions to other users (@*profile*) and labels (#*hashtag*) indicating subjects, categories, etc. Due to the characteristics mentioned previously, the *Twitter* has been an important social network for sharing information and everyday events. Such events can be classified as social events, capable of describing from routine events to crisis situations (natural disasters, social mobilizations, among others) [3, 33].

3.2 Smart Cities and Public Transport

The concept of Smart Cities (SC) has been defined mainly as sustainable and socially inclusive cities [30], which use Information and Communication Technologies (ICTs) to efficiently manage natural resources, energy, transportation, waste, etc. [1]. ICTs permeates urban systems and physical spaces, which has been accentuated by the increasing number of sensors and devices connected to the Internet of Things (IoT); voluntary data and existing content on SN about daily events. Such heterogeneous sources generate large amounts of data, used to develop SC services [2, 7].

The development of SC services has challenges related to connectivity (network infrastructure, interoperability and standards, power consumption and scalability) and related to data (capacity and location of data storage, extraction, processing, analysis, integration and aggregation). Besides, data analysis has issues related to correlation, inference of data from different domains, machine learning, real-time processing, and new-use proposals for data from existing infrastructures [2, 32].

[1] `https://www.facebook.com`. Accessed in December 09, 2018.
[2] `https://twitter.com`. Access in December 09, 2018.

In the public transport context, the GTFS[3] is a specification of a common format (that solves the problem of interoperability and patterns related to public transport data) to exchange static information on public transport. A feed specified in static GTFS consists of text files (which follows certain requirements similar to the CSV format) compressed in ZIP format. In this research we correlate SPTrans's static GTFS and AVL data (i.e. location data related to each bus) with *tweets* from the selected accounts.

3.3 Natural Language Processing

Automatic exception event *tweets* classification involves Natural Language Processing (NLP), which explores how computers can be used to understand and manipulate text or speech in natural language [19]. Before the NLP processing, the *tweets* were preprocessed — removing *URLs*, *datetime*, mentions to other *tweets*, emoticons, punctuations — to remove noise and to reduce the dimension of feature space.

A particular attention was paid to *hashtags*, which are relevant to exception events classification, but adds noise to the address extraction phase. In order to mitigate this problem, *hashtags* are identified and replaced by empty spaces in the address extraction process. Also, it is important to note that *hashtags* are not removed from original *tweets*.

After the preprocessing phase we applied NLP techniques to *tweets*, such as (I) *Tokenization* — process to obtain the words, i.e. tokens (features used to train the classification model), in a *tweet*, removing numbers and characters that do not belong to the alphabet (*TweetTokenizer*[4]); (II) morphological decomposition to get a given word into its inflected form using lemmatization (word lemma identification) or stemming (identification of the root of the word using heuristics to determine the location of its flexion — *RSLPStemmer*[5]); process used to features space reduction, besides of Brazilian Portuguese *stopwords* remotion[6] (common words without meaning) [5, 15, 23, 27, 28].

4 Study

4.1 Classification model

Finding exception events involves the identification of events related to an exception, which is possible through classification. The following classes are often used to classify exception events (that normally occurs in a city) [4, 8, 14, 18]:

[3] Google Transit: `https://developers.google.com/transit`. Accessed in December 11, 2018.

[4] NLTK module used to the tokenization process. `https://www.nltk.org/api/nltk.tokenize`. Accessed in December 09, 2018.

[5] NLTK module used to the stemming process. `https://www.nltk.org/_modules/nltk/stem/rslp`. Accessed in December 09, 2018.

[6] Brazilian Portuguese *stopwords* were obtained from NLTK — `https://www.nltk.org`. Accessed in December 19, 2018.

1. *Accidents*, e.g. accidents occurred at transport stations, fire, collision of vehicles, etc. 2. *Time-space*, e.g. day of the week (mondays, fridays and holidays), time of day (peak times), etc. 3. *Social Events*, e.g. street fairs, festivals, sport games, marches, marathons, etc. 4. *Urban Events*, e.g. related to traffic (deviations), road maintenance, etc. 5. *Natural disasters*, e.g. storms, earthquake, typhoons, etc. 6. *Meteorological*, e.g. clear day, overcast, rainy, snowing, haze, (high and low) temperatures, etc.

Using the found classes, 60,984 *tweets* from selected accounts were manually classified. This labeled data was transformed to a binary representation of features, which was used to train a model to classify *tweets* in exception events. The process of constructing these features is known as *feature engineering*, that is iterative between the phases of feature extraction, feature construction, and feature selection. Before this iteration, the data can be preprocessed using standardization, normalization, noise removal, dimensionality reduction, discretization, expansion, etc; it is important to note that information can be lost when performing these transformations [13].

As mentioned in Section 3.3, we used a preprocessing phase to feature extraction through a NLP function. The feature construction and selection phases are not used because these processes do not apply to the methodology of this work. After the preprocessing the *tweets* are processed to be represented by a bag-of-words, which contains feature vectors created using the *Term Frequency - Inverse Document Frequency* (TF-IDF) measure. The bag-of-words is randomly partitioned into training (60%) and test (40%) sets, that are inputs to the classification algorithms mentioned in Section 4.2.

4.2 Machine Learning Algorithms

Machine Learning algorithms can be (I) supervised, in which relations with known results are created based on the input characteristics; (II) unsupervised, in which the input characteristics are known, but not the results; (III) semi-supervised, in which some of the relationships between input data and results can be defined.

In this work we used supervised learning, since we know how the input data can be classified, being the following algorithms[7] normally applied to classify textual data sets: (Complement and Multinomial) Naive Bayes, Decision Tree, K-Nearest Neighbors, Logistic Regression, Multi-layer Perceptron, Random Forest and Support Vector Machine [6, 16, 24]. The validation of the models to classification tasks can be realized through 10-fold cross-validation[8] (to validate the generalization of a model) and metrics that has as inputs the number of real positive (P), negative (N) cases in the result of classification, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classifications, such as:

[7] We used the algorithms implemented by Sci-Kit Learn, with the standard hyper-parameters. It is not the focus of this work hyper-parameters tuning.

[8] `https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation`. Accessed in December 26, 2018.

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}; Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}; F_1 score = \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

5 Data set

Corpus Twitter. The Social Network *Twitter* was chosen as data source for the construction of the data set related to the exception events. The choice is due to the fact that each publication is limited in 280 characters, which reduces the complexity of the processing of the published content, and because São Paulo's public agencies use it as an instant channel of communication with its citizens.

The data set used to identify the exception events is composed by *tweets*, written in Brazilian Portuguese, published by the *profiles* cited in Table 1. We chose to use *tweets* from official public service providers to guarantee the reliability of the data analyzed, discarding *retweets* and *replies*. Thus, the data used are related to the unidirectional communication channel (in the context of *e-participation* — interaction between citizens and public authorities). Regarding *profiles* selection, all accounts were manually selected according to the institutions responsible for reporting exception events. Such *profiles* are public in nature, so access to their *tweets* does not involve privacy issues.

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
—	60,984	—	—

Corpus SPTrans. The SPTrans (São Paulo Transportation Company)[9] corpus has data provided by SPTrans specified in GTFS, detailed in Table 2

[9] `http://www.sptrans.com.br`. Acessed in December 11, 2018.

and data of geolocation (movements) of all the buses of São Paulo, referring to the year of 2017 — *obtained by the law on access to information*[10]. In respect to AVL data set, it is important to note inconsistencies in the two AVL files of January 11, according to SPTrans meta data 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, 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*[11]. In order to enable *geospatial queries*, *geospatial indexes*[11] were created in the *MongoDB* collections containing geolocalized information.

Table 2. Data set and total records specified in SPTrans' GTFS

Data set	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
—	1,254,779

Table 3. SPTrans'AVL data set description

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[a]	717	108,16
December[b]	738	110,89
—	8,751	1,247,09

[a] Missing data: 01/11 — from 12h to 15h.
[b] Missing data: 15/12 — from 01h to 09h.

[10] `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm` (in Portuguese). Accessed in December 11, 2018.
[11] `https://docs.mongodb.com/manual/geospatial-queries`. Accessed in December 11, 2018.

6 Analysis

6.1 Addresses and geolocalization 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. So, we used this 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 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[12] to geolocate the found address (only 1.5% of *tweets* have geolocalization [26]). The HTTP response from this API is processed to get the values from location (which contains latitude and longitude information) and formatted address.

6.2 Finding bus lines affected by exception events

In order to find the bus lines affected by exception events, it is necessary to correlate the coordinates of exception events with the existing coordinates in the shapes data set — a set of latitude and longitude used for drawing bus lines on a map to represent its respective paths — existing in SPTrans' GTFS. According to Section 5, all coordinates are stored in legacy pairs and in collections with geospatial indexes. Thus, it is possible to use the `$near` function from MongoDB[13] 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 parameter to correlate the bus code line with others GTFS files with details about the bus direction, identification, etc.

6.3 Velocity impact analysis

After we found the bus 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,

[12] `https://developers.google.com/maps/documentation/geocoding`. Accessed in December 20, 2018.

[13] `https://docs.mongodb.com/manual/reference/operator/query/near`. Accessed in December 18, 2018.

24 and 08/31/2017) will be considered in the analysis, this because the days of the week have different patterns of movement (seasonality), e.g. on Fridays many social events occur that usually lead to a more congested traffic. Besides, the months also have different characteristics — holidays at the end of the year, vacations, beginning of school periods, etc. —, as Fig. 3, because of this the selected days are restricted to the month of occurrence of the event.

In another step, 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[14]. 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%.

7 Results

The methodology was applied to the *Corpus Twitter*[15], which contains 60,984 *tweets*. At the end of *tweets* preprocessing and processing, the corpus got 414,637 words, with a vocabulary size of 13,915 words. All *tweets* were manually classified according to identified exception events. This data set is composed of the following labels: Accident, Irrelevant — to non exception events, Natural Disaster, Social Event and Urban Event.

This labeled data set was used to train exception events classification models, based on a *bag-of-words*, described in Section 4.1. According to Table 4, the model using the Multi-layer Perceptron algorithm obtained greater accuracy for the classification task.

Of the 60,984 *tweets* 10,027 were classified into exception events and from that subset we found 7,710 addresses (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 un-

[14] 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*.

[15] Data set publicly available at: `https://github.com/fcas/mobility-analysis/blob/master/datasets/tweets.zip`. Accessed in December 14, 2018.

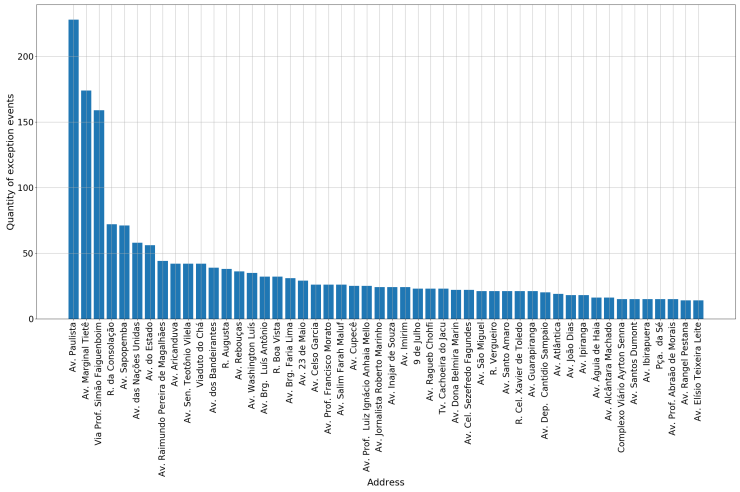
Table 4. Metrics of the evaluations of the algorithms used to classify the *tweets* in exception events

Algorithm	Accuracy	Precision	Recall	*f1-score*
Complement Naive Bayes	0,941	0,949	0,941	0,944
Decision Tree	0,965	0,965	0,965	0,965
K-Nearest Neighbors	0,970	0,971	0,970	0,970
Logistic Regression	0,969	0,968	0,969	0,968
Multi-layer Perceptron	*0,973*	*0,972*	*0,973*	*0,972*
Multinomial Naive Bayes	0,953	0,952	0,953	0,949
Random Forest	0,970	0,970	0,970	0,970
Support Vector Machine	0,833	0,694	0,833	0,757

usual public place name (for example *passageway*, *road complex*, *connection to*). 4. *Tweets* with addresses with concatenated words (for example *avenidapaulista*)

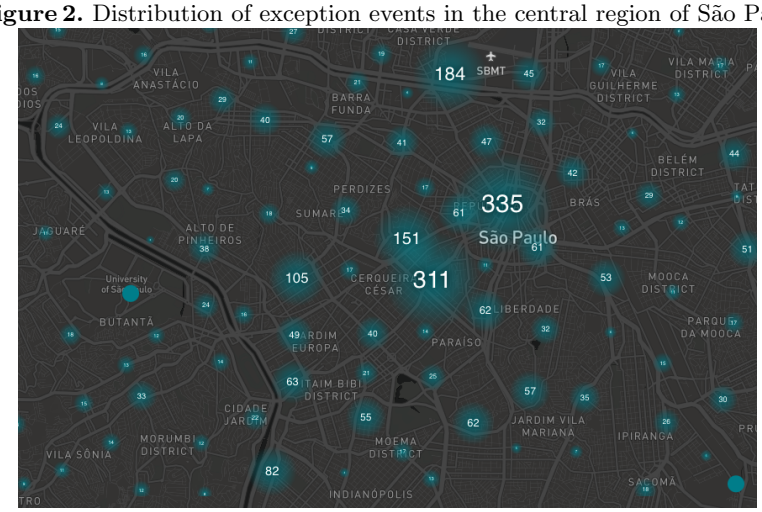
Figure 1 illustrates the addresses[16] most affected by exception events and Figure 2 shows the distribution of these events in the central region of São Paulo. It is important to note that the exception events found are concentrated in the addresses and regions where they normally occur in São Paulo, which validates the methodology developed.

Figure 1. Addresses most impacted by exception events



[16] Complete list is available at https://docs.google.com/spreadsheets/d/1gn1cTDifUJEPdgcU67SC45GdYHRKmIHtAfJwRBm088s/edit?usp=sharing. Accessed on December 20, 2018.

Figure 2. Distribution of exception events in the central region of São Paulo



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 bus lines were affected by exception events during this period, with line "33121" being the most impacted bus line code, according to Table 5. This particular line was impacted by 1,623 exception events.

Table 5. Bus lines most impacted by exception events[a]

Bus code line	Ttl. excepetion events	Bus origin / destination
33121	1623	TERM. PRINC. ISABEL / TERM. STO. AMARO
32826	1502	TERM. PQ. D. PEDRO II / TERM. JOÃO DIAS
32805	1490	TERM. PRINC. ISABEL / CHÁC. SANTANA
34085	1464	TERM. BANDEIRA / JD. VAZ DE LIMA
34233	1418	TERM. BANDEIRA / TERM. VARGINHA
33123	1408	TERM. BANDEIRA / TERM. STO. AMARO
32829	1405	TERM. BANDEIRA / TERM. CAPELINHA
35174	1388	TERM. PQ. D. PEDRO II / TERM. STO. AMARO
32827	1378	TERM. BANDEIRA / TERM. CAPELINHA
33128	1373	TERM. BANDEIRA / SOCORRO

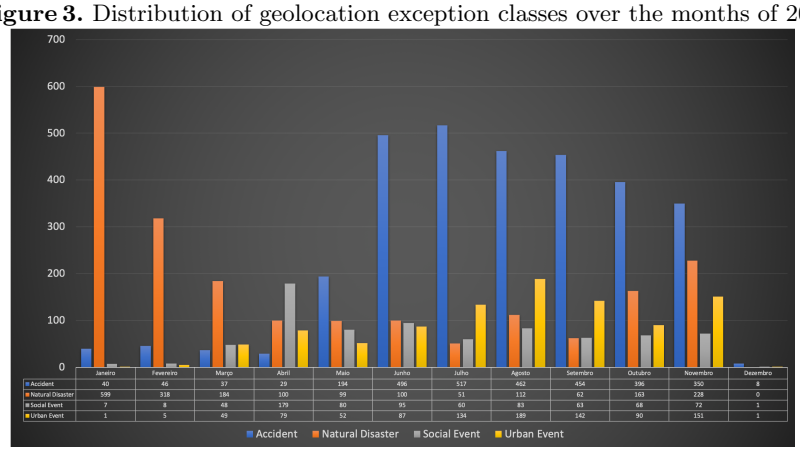
[a] Full table publicly available at `https://docs.google.com/spreadsheets/d/1jIqUuIJg7FhXD5C8MFF8stbvOD3uiUgMfN2bOltT7zE/edit?usp=sharing`. Accessed in December 20, 2018.

Using the methodology described above, we can observe in Tab. 6 that the social event-related exception events have an average of 87,04% impact on the

Table 6. 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	—	—	—	—	—	—	—	—
—	66,51	98,39	59,77	99,80	87,04	100	70,11	98,86

Figure 3. Distribution of geolocation exception classes over the months of 2017



average speed 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 government 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 prolonged holidays), 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 Fig.3, 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.

8 Conclusions and future works

This work presents a new methodology for exception events classification and analyze their respective impacts on velocity of the public transport system by bus of the São Paulo city. Using *tweets* from selected public service providers, we found that Multi-layer Perceptron was the best algorithm for classifying *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 impact the velocity of buses, using the methodology developed 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 language, 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.

Future work

Acknowledgment

This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, FAPESP proc. 14/50937-1, and FAPESP proc. 15/24485-9.

[17] www1.folha.uol.com.br/poder/2017/05/1884977-manifestacao-anti-temer-reune-hundreds-of-people-in-av-paulista.shtml. Accessed on December 2, 2018

References

1. Ahvenniemi, H., Huovila, A., Pinto-Seppä, I., Airaksinen, M.: What are the differences between sustainable and smart cities? Cities **60**, 234–245 (2017). https://doi.org/10.1016/j.cities.2016.09.009, http://dx.doi.org/10.1016/j.cities.2016.09.009
2. Ang, L.M., Seng, K.P., Zungeru, A., Ijemaru, G.: Big Sensor Data Systems for Smart Cities. IEEE Internet Things J. **4**(5), 1–1 (2017). https://doi.org/10.1109/JIOT.2017.2695535, http://ieeexplore.ieee.org/document/7903653/
3. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Computational Intelligence **31**(1), 132–164 (2015)
4. 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 Sharing Systems. UBICOMP pp. 841–852 (2016). https://doi.org/10.1145/2971648.2971652
5. 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)
6. 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 (2016)
7. Finger, M., Razaghi, M.: Conceptualizing "Smart Cities". Informatik-Spektrum **40**(1), 6–13 (2017). https://doi.org/10.1007/s00287-016-1002-5
8. 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. Transp. Policy **32**, 115–123 (2014). https://doi.org/10.1016/j.tranpol.2014.01.007
9. Gkiotsalitis, K., Stathopoulos, A.: Joint leisure travel optimization with user-generated data via perceived utility maximization. Transp. Res. Part C Emerg. Technol. **68**, 532–548 (2016). https://doi.org/10.1016/j.trc.2016.05.009, http://dx.doi.org/10.1016/j.trc.2016.05.009
10. Gkiotsalitis, K., Stathopoulos, A.: A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data. Transp. Res. Part C Emerg. Technol. **58**, 265–277 (2015). https://doi.org/10.1016/j.trc.2014.12.006, http://dx.doi.org/10.1016/j.trc.2014.12.006
11. Guo, W., Gupta, N., Pogrebna, G., Jarvis, S.: Understanding happiness in cities using twitter: Jobs, children, and transport. IEEE 2nd Int. Smart Cities Conf. Improv. Citizens Qual. Life, ISC2 2016 - Proc. (2016). https://doi.org/10.1109/ISC2.2016.7580790
12. Gutev, A., Nenko, A.: Better Cycling - Better Life: Social Media Based Parametric Modeling Advancing Governance of Public Transportation System in St. Petersburg. Proc. Int. Conf. Electron. Gov. Open Soc. Challenges Eurasia pp. 242–247 (2016). https://doi.org/10.1145/3014087.3014123, http://doi.acm.org/10.1145/3014087.3014123
13. Guyon, I., Elisseeff, A.: An introduction to feature extraction. Feature extraction pp. 1–25 (2006)
14. 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 Trans. Big Data **2**(1), 85–99 (2016).

https://doi.org/10.1109/TBDATA.2016.2546301, http://ieeexplore.ieee.org/document/7445832/

15. 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
16. 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)
17. 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)
18. Lecue, 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. J. Web Semant. **27**, 26–33 (2014). https://doi.org/10.1016/j.websem.2014.07.002, http://dx.doi.org/10.1016/j.websem.2014.07.002
19. Liu, D., Li, Y., Thomas, M.A.: A roadmap for natural language processing research in information systems. In: Proceedings of the 50th Hawaii International Conference on System Sciences (2017)
20. 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. IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC **2015-Octob**, 208–213 (2015). https://doi.org/10.1109/ITSC.2015.43
21. 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. Acm Dev 2015 pp. 137–145 (2015). https://doi.org/10.1145/2830629.2830642
22. Myers, S.A., Sharma, A., Gupta, P., Lin, J.: Information network or social network?: the structure of the twitter follow graph. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 493–498. ACM (2014)
23. 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)
24. 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 (2017)
25. Ni, M., He, Q., Gao, J.: Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. IEEE Trans. Intell. Transp. Syst. **18**(6), 1623–1632 (2016). https://doi.org/10.1109/TITS.2016.2611644
26. 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 (2016)
27. Roy, A., Majumder, A.G., Nath, A.: Understanding natural language processing and its primary aspects. International Journal **5**(8) (2017)

28. Setiawan, E.B., Widyantoro, D.H., Surendro, K.: Feature expansion using word embedding for tweet topic classification. Proceeding 2016 10th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2016 Spec. Issue Radar Technol. (2011) (2017). https://doi.org/10.1109/TSSA.2016.7871085
29. SÁ, T. H., Tainio, M., Goodman, A., Edwards, P., Haines, A., Gouveia, N., Monteiro, C., Woodcock, J.: Health impact modelling of different travel patterns on physical activity, air pollution and road injuries for são paulo, brazil. Environment International **108**(Supplement C), 22 – 31 (2017). https://doi.org/https://doi.org/10.1016/j.envint.2017.07.009, http://www.sciencedirect.com/science/article/pii/S0160412017305974
30. Wang, S., Sinnott, R., Nepal, S.: Privacy-protected social media user trajectories calibration. Proc. 2016 IEEE 12th Int. Conf. e-Science, e-Science 2016 pp. 293–302 (2016). https://doi.org/10.1109/eScience.2016.7870912
31. Wen, X., Lin, Y.R., Pelechrinis, K.: PairFac: Event Analytics through Discriminant Tensor Factorization. Cikm pp. 519–528 (2016). https://doi.org/10.1145/2983323.2983837
32. Xiao, Z., Lim, H.B., Ponnambalam, L.: Participatory Sensing for Smart Cities: A Case Study on Transport Trip Quality Measurement. IEEE Trans. Ind. Informatics **13**(2), 759–770 (2017). https://doi.org/10.1109/TII.2017.2678522
33. Zhou, X., Chen, L.: Event detection over twitter social media streams. The VLDB journal **23**(3), 381–400 (2014)