

UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

FELIPE CORDEIRO ALVES DIAS

**Caracterização de eventos de exceção e de seus respectivos impactos no
sistema de transporte público por ônibus da cidade de São Paulo**

São Paulo

2017

FELIPE CORDEIRO ALVES DIAS

**Caracterização de eventos de exceção e de seus respectivos impactos no
sistema de transporte público por ônibus da cidade de São Paulo**

Versão original

Texto de Exame de Qualificação apresentado à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo como parte dos requisitos para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Orientador: Prof. Dr. Daniel de Angelis Cordeiro

São Paulo

2017

Resumo

DIAS, Felipe Cordeiro Alves. **Caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo.** 2017. 169 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

A cidade de São Paulo é o município mais populoso do Brasil, caracterizado por uma segregação urbana responsável por inúmeros problemas relacionados a mobilidade urbana. As ações atuais para resolver os problemas de mobilidade urbana têm pouco aprofundamento em questões tecnológicas e melhorias dos sistemas computacionais existentes – como as necessárias ao defasado Sistema Integrado de Monitoramento e Transporte (SIM), utilizado para gestão e monitoramento do transporte público por ônibus de São Paulo. Uma das possíveis melhorias é integrar o SIM às Redes Sociais. Com essa perspectiva de integração, esse trabalho tem como objetivo utilizar tweets e dados do SIM na caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo. Para alcançar tal objetivo, esse trabalho propõe utilizar tweets publicados por instituições governamentais responsáveis por reportar eventos de exceção e dados dos módulos AVL (*Automatic Vehicle Location*) do SIM, responsáveis por rastrear e localizar os ônibus do município. A hipótese é de que é possível identificar e localizar eventos de exceção nos tweets por meio de Processamento de Linguagem Natural e Expressão Regular, e correlacionar esses eventos com os dados históricos do SIM.

Palavras-chaves: Cidades Inteligentes. Transporte Público. Sistemas de Transporte Inteligentes. Eventos de exceção.

Lista de figuras

Figura 1 – Fluxograma do processo do aprendizado supervisionado	33
Figura 2 – Processo de Filtragem	41
Figura 3 – Quantidade de artigos publicados por ano	42
Figura 4 – Porcentagem dos artigos publicados por ano	42
Figura 5 – Nuvem de palavras das <i>keywords</i> dos artigos selecionados	43
Figura 6 – Evidência dos períodos de indisponibilidade de dados AVL referentes a Dezembro de 2017	59
Figura 7 – Arquitetura usada no estudo de caso para visualização e exploração dos dados AVL da SPTrans	65
Figura 8 – Quantidade de dados enviados por dia por ônibus (selecionados aleatoriamente) em janeiro de 2017	66
Figura 9 – Distribuição da quantidade de dados enviados por ônibus (selecionados aleatoriamente) em janeiro de 2017	67
Figura 10 – Localizações enviadas em Janeiro de 2017 de uma linha de ônibus selecionada aleatoriamente	68
Figura 11 – Localizações dos ônibus referente a movimentação de Janeiro de 2017	68
Figura 12 – Metodologia baseada em <i>tweets</i> para encontrar linhas de ônibus impactadas por eventos de exceção na cidade de São Paulo	70
Figura 13 – Histograma da variação dos tamanhos das sentenças dos <i>tweets</i> existentes no <i>Corpus Twitter</i>	76
Figura 14 – Distribuição das classes dos eventos de exceção do <i>Corpus Twitter</i>	77
Figura 15 – Matriz de confusão relacionada a classificação dos <i>tweets</i> em eventos de exceção por meio do algoritmo <i>Multi-layer Perceptron</i>	78
Figura 16 – Endereços mais impactados por eventos de exceção	80
Figura 17 – Distribuição dos eventos de exceção na região central de São Paulo	81
Figura 18 – Dados de geolocalização do código de linha de ônibus “35276” mais afetado por eventos de exceção (ilustração a nível de rua)	82
Figura 19 – Matriz de confusão relacionada a classificação dos <i>tweets</i> em eventos de exceção por meio do algoritmo Regressão Logística	163

Lista de tabelas

Tabela 1 – Descrição e nome dos profiles selecionados do Twitter	16
Tabela 2 – Detalhamento dos arquivos da GTFS	27
Tabela 3 – Quantidades de artigos coletados e fontes de busca	40
Tabela 4 – Intervalo de tempo e número de <i>tweets</i> coletados	57
Tabela 5 – Conjuntos e quantidades de dados especificados em GTFS pela SPTTrans	58
Tabela 6 – Descrição do conjunto de dados AVL	60
Tabela 7 – Meta dados dos dados AVL da SPTTrans	61
Tabela 8 – Métricas das avaliações dos algoritmos utilizados para classificação dos <i>tweets</i> em eventos de exceção	77
Tabela 9 – Quantidade de eventos extraídos por classe	79
Tabela 10 – Linhas de ônibus mais impactadas por eventos de exceção	80
Tabela 11 – Tabela de logradouros com abreviaturas	98
Tabela 12 – Detalhamento dos campos do arquivo <i>agency.txt</i> da GTFS	103
Tabela 13 – Detalhamento dos campos do arquivo <i>stops.txt</i> da GTFS	104
Tabela 14 – Detalhamento dos campos do arquivo <i>routes.txt</i> da GTFS	109
Tabela 15 – Detalhamento dos campos do arquivo <i>trips.txt</i> da GTFS	111
Tabela 16 – Detalhamento dos campos do arquivo <i>stop_times.txt</i> da GTFS	114
Tabela 17 – Detalhamento dos campos do arquivo <i>calendar.txt</i> da GTFS	121
Tabela 18 – Detalhamento dos campos do arquivo <i>calendar_dates.txt</i> da GTFS	124
Tabela 19 – Detalhamento dos campos do arquivo <i>fare_attributes.txt</i> da GTFS	125
Tabela 20 – Detalhamento dos campos do arquivo <i>fare_rules.txt</i> da GTFS . . .	126
Tabela 21 – Detalhamento dos campos do arquivo <i>shapes.txt</i> da GTFS	127
Tabela 22 – Detalhamento dos campos do arquivo <i>frequencies.txt</i> da GTFS . .	128
Tabela 23 – Detalhamento dos campos do arquivo <i>transfer.txt</i> da GTFS	131
Tabela 24 – Detalhamento dos campos do arquivo <i>feed_info.txt</i> da GTFS	133
Tabela 25 – Linhas de ônibus impactadas por eventos de exceção	136

Lista de abreviaturas e siglas

ACM	<i>Association for Computing Machinery</i>
API	<i>Application Programming Interface</i>
APTS	<i>Advanced Public Transportations Systems</i>
ATIS	<i>Advanced Travelers Information Systems</i>
ATMS	<i>Advanced Traffic Management System</i>
AVCS	<i>Advanced Vehicles Control Systems</i>
AVL	<i>Automatic Vehicle Location</i>
CCOI	Centro de Controle Integrado 24 Horas da Cidade de São Paulo
CE	Centro Expandido
CETSP	Companhia de Engenharia de Tráfego de SP
CIMU	Central Integrada de Mobilidade Urbana
CP	Cinturão Periférico
CPTM	Companhia Paulista de Trens Metropolitanos
CRF	<i>Conditional Random Field</i>
CSV	<i>Comma-separated values</i>
CVO	<i>Commercial Vehicles Operation</i>
ETL	<i>Extract, Tranform and Load</i>
GPRS	<i>General Packet Radio Services,</i>
GPS	Global Positioning System
GTFS	<i>General Transit Feed Specification</i>
HDM	<i>Human Driven Method</i>
HP	Hipótese de Pesquisa

HTTP	<i>Hypertext Transfer Protocol</i>
IDF	Inverse Document Frequency
IEEE	<i>Institute of Electrical and Electronics Engineers</i>
ITS	<i>Intelligent Transport System</i>
K-NN	<i>K-Nearest Neighbour</i>
LDA	<i>Latent Dirichlet Allocation</i>
LISA	<i>Local Indicators of Spatial Association</i>
NER	<i>Named Entity Recognition</i>
NLP	<i>Natural Language Processing</i>
NLTK	Natural Language Toolkit
PAC	Programa de Aceleração do Crescimento
PCD	Pessoas com Deficiência
PlanMob/SP	Plano de Mobilidade Urbana de São Paulo
PMESP	Polícia Militar do Estado de São Paulo
PTCS	Sistema de Calibração de Trajetórias Privadas
QP	Questão de Pesquisa
RDBMS	<i>Relational Database Management Systems</i>
RL	Régressão Linear
RTPI	<i>Real Time Passenger Information</i>
SARIMA	<i>Seasonal Autoregressive Integrated Moving Average</i>
SBD	<i>Sentence Boundary Disambiguation</i>
SC	<i>Smart Cities</i>
SIM	Sistema Integrado de Monitoramento e Transporte

SMT	Secretaria Municipal de Transportes
SPCEDEC	Defesa Civil do Estado de São Paulo
SPTrans	São Paulo Transportes
SVM	<i>Support Vector Machine</i>
TDM	<i>Technology Driven Method</i>
TF	<i>Term Frequency</i>
TF-IDF	<i>Term Frequency - Inverse Document Frequency</i>
TIC	Tecnologias da Informação e Comunicação
URL	<i>Uniform Resource Locator</i>
WSD	<i>Word Sense Disambiguation</i>

Sumário

1	Introdução	12
1.1	<i>Motivação</i>	12
1.2	<i>Definição do problema</i>	14
1.3	<i>Objetivos</i>	15
1.4	<i>Hipóteses</i>	16
1.5	<i>Organização do documento</i>	17
2	Fundamentação Teórica	19
2.1	<i>Cidades Inteligentes</i>	19
2.2	<i>Sistemas de Transporte Inteligentes</i>	21
2.3	<i>Conceitos relacionados ao transporte público</i>	22
2.3.1	<i>Acessibilidade</i>	23
2.3.2	<i>Mobilidade</i>	23
2.3.3	<i>Viagem e modais de transporte</i>	24
2.4	<i>General Transit Feed Specification</i>	25
2.5	<i>Redes Sociais</i>	28
2.6	<i>Processamento de Linguagem Natural</i>	28
2.7	<i>Feature Engineering</i>	31
2.8	<i>Algoritmos de Aprendizado de Máquina</i>	32
2.8.1	<i>Algoritmos de aprendizado supervisionado</i>	32
2.8.2	<i>Validação dos modelos de aprendizado supervisionado</i>	34
2.9	<i>Term frequency–Inverse document frequency</i>	34
3	Revisão Sistemática	36
3.1	<i>Planejamento da Revisão Sistemática</i>	36
3.1.1	<i>Justificativa da Revisão Sistemática</i>	37
3.2	<i>Questões de Pesquisa</i>	37
3.3	<i>Coleta de dados</i>	40
3.4	<i>Avaliação de Dados</i>	41
3.5	<i>Análise e Interpretação</i>	43

3.5.1	Tipos de problemas urbanos abordados utilizando o processamento <i>tweets</i> (QP1)	43
3.5.2	Casos de uso relacionados ao transporte público (QP2)	47
3.5.3	Técnicas estatísticas utilizadas no processamento de <i>tweets</i> (QP3)	49
3.5.4	Paradigmas de processamento (QP4)	50
3.5.5	Eventos de exceção relacionados ao transporte público (QP5) .	51
3.5.6	Técnicas de Aprendizado de Máquina utilizadas no processamento de <i>tweets</i> (QP6)	52
3.6	<i>Considerações finais sobre a revisão sistemática</i>	54
4	Construção do conjunto de dados	56
4.0.1	<i>Corpus Twitter</i>	56
4.0.2	<i>Corpus SPTrans</i>	58
5	Exploração e visualização de grandes volumes de dados . . .	62
5.1	<i>Trabalhos relacionados</i>	62
5.2	<i>Druid</i>	63
5.2.1	Real-time nodes	63
5.2.2	Historical nodes	64
5.2.3	Broker nodes	64
5.2.4	Coordinator nodes	64
5.3	<i>Arquitetura utilizada para visualização e exploração dos dados AVL da SPTrans</i>	64
5.4	<i>Estudo de caso com os dados AVL da SPTrans</i>	65
5.5	<i>Consideração sobre a arquitetura utilizada para exploração e visualização dos dados AVL da SPTrans</i>	67
6	Uma metodologia baseada em <i>tweets</i> para encontrar linhas de ônibus impactadas por eventos de exceção na cidade de São Paulo	70
6.1	<i>Pré-processamento</i>	70
6.2	<i>Extração de endereço e geolocalização</i>	72
6.3	<i>Processamento de tweets</i>	73
6.4	<i>Classificação manual do Corpus Twitter</i>	73

6.5	<i>Modelo de classificação de tweets relacionados a eventos de exceção</i>	74
6.6	<i>Encontrando linhas de ônibus afetadas por eventos de exceção</i>	74
6.7	<i>Resultados</i>	75
6.8	<i>Considerações finais sobre a metodologia desenvolvida</i>	81
7	Correlação dos eventos de exceção com os dados AVL da SP-Trans	83
8	Conclusão	84
8.1	<i>Contribuições</i>	84
8.2	<i>Trabalhos publicados</i>	84
8.3	<i>Trabalhos futuros</i>	84
	Referências	87
	APÊNDICES	94
	Apêndice A – Exemplos de tweets	95
	Apêndice B – Logradouros utilizados	98
	Apêndice C – Detalhamento dos campos da GTFS	103
	Apêndice D – Linhas de ônibus impactadas por eventos de exceção	136
	Apêndice E – Matrizes de confusão	163
	Apêndice F – Parametrizações dos algoritmos	169

1 Introdução

Neste capítulo, são apresentadas as seções referentes à motivação da proposta de pesquisa; sobre a definição do problema que pretendemos abordar; a respeito dos objetivos gerais e específicos; sobre as hipóteses a serem verificadas e sobre a organização dos capítulos desse documento.

1.1 Motivação

A cidade de São Paulo é o município mais populoso do Brasil, que passou por um rápido processo de urbanização e tem população atual estimada em 12.106.920 milhões de habitantes (com data de referência em 1º de julho de 2017)¹. Desse total de habitantes, 10% vivem na área do Centro Expandido (CE) e 90% no Cinturão Periférico (CP) (SÁ, T. H. et al., 2017), o que caracteriza uma segregação urbana responsável por inúmeros problemas relacionados a mobilidade urbana.

Um desses problemas é conhecido como o movimento pendular, no qual longas distâncias são percorridas diariamente pelos moradores do CP para acessar os locais de emprego, educação e serviços localizados em maioria no CE. Além disso, o movimento pendular torna o CP uma região dormitória, com parte de seus respectivos moradores dependentes do Sistema de Transporte Público para acessar o CE.

Devido aos problemas de mobilidade urbana existentes no Brasil, como os da cidade de São Paulo, a Lei Federal 12.587/2012², relacionada ao Programa de Aceleração do Crescimento³ (PAC), obrigou os municípios a enviarem seus respectivos planos de mobilidade urbana até o final do ano de 2015, com o objetivo de promover o desenvolvimento sustentável com a mitigação dos custos ambientais e socioeconômicos dos deslocamentos de pessoas. Em resposta a essa lei, o Plano de Mobilidade Urbana de São Paulo (*PlanMob/SP 2015*) foi instituído pelo Decreto

¹ <https://agenciadenoticias.ibge.gov.br/media/com_mediaibge/arquivos/9bc1a0065c49fd6f81dc785b2b8d8c35.xlsx>. Acesso em Outubro, 29 de 2017.

² <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12587.htm>. Acesso em Outubro, 29 de 2017.

³ <<http://www.pac.gov.br>>. Acesso em Outubro, 29 de 2017.

56.834⁴, como instrumento de planejamento e gestão do Sistema Municipal de Mobilidade Urbana para os próximos 15 anos.

No *PlanMob/SP 2015*, a Secretaria Municipal de Transportes (SMT) propõe criar uma central de monitoramento conhecida como Central Integrada de Mobilidade Urbana (CIMU), que tem como objetivo integrar as áreas de trânsito e transporte subordinadas à SMT. Nessa proposta, observam-se os seguintes problemas que poderiam ser resolvidos em paralelo ao desenvolvimento do CIMU: (I) a CIMU não processa conteúdo de Redes Sociais, (II) não aborda melhoria dos sistemas computacionais já existentes e (III) será integrada com o defasado Sistema Integrado de Monitoramento e Transporte (SIM), da São Paulo Transportes (SPTrans), responsável pelo monitoramento da infraestrutura de ônibus.

O SIM utiliza a tecnologia *Automatic Vehicle Location* (AVL) para localizar e rastrear os ônibus, fornecer informações em tempo real aos passageiros (*Real Time Passenger Information* (RTPI)), monitorar 1.353 rotas de ônibus⁵, 10 corredores de ônibus⁶, 28 terminais de ônibus⁷ e 19.933 mil paradas de ônibus⁵ que serviram em 2016 a aproximadamente 8 milhões de passageiros por dia⁸. Apesar da importância do SIM, há inúmeras defasagens tecnológicas (que causam discrepância nas informações recebidas pelos usuários, dentre outros problemas) (CONSULO et al., 2016), que precisariam ser resolvidas antes de integrá-lo ao CIMU.

Sistemas como o SIM são classificados como Sistemas de Transporte Inteligente (ITS — *Intelligent Transport System*), e normalmente estão presentes nas Cidades Inteligentes (SC — *Smart Cities*). Por definição, ITS utilizam Tecnologias da Informação e Comunicação (TIC) para explorar dados capazes de contribuir com a melhoria da segurança, do gerenciamento, eficiência dos transportes e redução do impacto ambiental (ANTTIROIKO, 2013). Com isso, nota-se que ITS são essenciais para os objetivos mencionados na Lei Federal 12.587/2012 e no PlanMob/SP 2015.

No entanto, a lei de mobilidade urbana (12.587/2012) e o *PlanMob/SP 2015* não mencionam explicitamente ITS e TIC. O conteúdo de ambos os documentos tem um viés político-urbano, com pouco aprofundamento em questões tecnológicas e

⁴ <<http://www.prefeitura.sp.gov.br/cidade/secretarias/transportes/planmob>>. Acesso em Outubro, 29 de 2017.

⁵ <<http://www.sptrans.com.br/desenvolvedores>>. Acesso em Outubro, 29 de 2017.

⁶ <<http://www.sptrans.com.br/terminais/corredores.aspx>>. Acesso em Outubro, 29 de 2017.

⁷ <<http://www.sptrans.com.br/terminais>>. Acesso em Outubro, 29 de 2017.

⁸ <<http://www.sptrans.com.br/indicadores>>. Acesso em Outubro, 29 de 2017.

melhorias dos sistemas já existentes. Esse cenário é diferente em alguns países, nos quais existem planejamentos para o transporte e mobilidade urbana que estão explicitamente relacionados ao desenvolvimento e uso de novas tecnologias.

Por exemplo, os EUA têm o plano estratégico para 2015-2019 em ITS, abordando temas como veículos conectados, automação, uso de tecnologias emergentes (para apoiar decisões em tempo real), integração de dados corporativos, interoperabilidade (comunicação entre diferentes sistemas) e entrega acelerada de projetos (United States Department of Transportation, 2017). Já a União Européia e o Japão estão centrados em padronizações de tecnologias em ITS, com o objetivo de serem referências nesse setor (CONSULO et al., 2016).

O contraste entre os dois parágrafos anteriores talvez seja devido ao fato de a legislação brasileira e os planos para mobilidade urbana terem sido estabelecidos como consequência do crescimento urbano acelerado e sem planejamento. Ou seja, como solução paliativa para um problema urbano, o que difere dos planos em ITS mencionados, que têm como foco otimizar o transporte e criar padrões tecnológicos.

Apesar dessas diferenças políticas e sociais, o transporte público pode se beneficiar ao explorar ITS (NELSON; MULLEY, 2013), e ao integrar as Redes Sociais com o planejamento, gestão e as atividades operacionais dos transportes públicos, abordando seus respectivos fatores sócio-técnicos (KUFLIK et al., 2017). Por exemplo, um dos benefícios possíveis é o de se conseguir analisar o impacto dos eventos de exceção na operação do sistema de transporte público por ônibus na cidade de São Paulo, usando dados do SIM (AVL) e de Redes Sociais.

1.2 Definição do problema

Eventos de exceção tais como acidentes, greves, falhas na operação do metrô, manifestações, enchentes, eventos sociais, dentre outras, podem comprometer muitos trechos do sistema de transporte público e, dependendo da proporção do impacto causado pela exceção, inúmeras pessoas podem ser afetadas. Tais eventos de exceção e seus respectivos impactos possuem características que podem ser identificadas visando melhor gestão dessas ocorrências.

Com a identificação dessas características, é possível conhecer previamente quais seriam os impactos decorrentes de um determinado evento de exceção no funcionamento normal do transporte público. Tais características podem ser obtidas analisando o histórico do funcionamento do sistema de transportes, e utilizadas posteriormente em simulações de como o sistema responderia a determinados eventos de exceção.

Os dados históricos existentes para essa análise são os do SIM, obtidos utilizando AVL. No entanto, analisá-los envolve problemas como o (I) grande volume de dados, em virtude da frequência com que são enviados (II) e os referentes ao comprometimento da qualidade dos dados enviados, como consequência dos problemas e limitações do *hardware* responsável pela transmissão; interferências e questões meteorológicas.

O uso de conteúdo de Redes Sociais pode ajudar a abordar os problemas anteriormente mencionados, o qual delimitaria o escopo da análise histórica para a identificação das características dos eventos de exceção e dos seus respectivos impactos. Usar o conteúdo de Redes Sociais envolve alguns desafios como o de (I) identificar eventos de exceção nas publicações, (II) geolocalizá-los, (III) determinar seus *timestamps* (IV) correlacioná-las com a base histórica.

1.3 Objetivos

O objetivo geral desse projeto de pesquisa é a caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo. Visando alcançar esse objetivo, serão coletados tweets das contas oficiais das instituições governamentais responsáveis por reportar eventos de exceção na cidade de São Paulo. Todas as contas selecionadas do Twitter estão listadas na tabela 1. Também, serão utilizados os dados históricos dos módulos AVL do SIM.

Além disso, temos como objetivos específicos:

- Identificar os eventos de exceção, quando existentes, dos tweets coletados.
- Extrair os endereços dos eventos de exceção identificados e geolocalizá-los.

- Construir uma base de dados pública com os dados processados, disponibilizada via API (para consumo e contribuição da comunidade de software), mantendo o modelo de dados consistente. Com isso, a necessidade de entrega dos dados a sociedade, apontada por (KUFLIK et al., 2017), será atendida.
- Criação de plataforma para exploração e visualização dos dados coletados e processados das fontes citadas na tabela 1 e da SPTrans.

Tabela 1 – Descrição e nome dos profiles selecionados do Twitter

Descrição do profile no Twitter	Profile no Twitter
Comando do Corpo de Bombeiros da PMESP ^a	@BombeirosPMESP
Companhia de Engenharia de Tráfego de SP	@CETSP_
Companhia Paulista de Trens Metropolitanos	@CPTM_oficial
Defesa Civil do Estado de São Paulo	@SPCEDEC
Governo do Estado de São Paulo	@governosp
Metrô de São Paulo	@metrosp_oficial
Polícia Civil do Estado de São Paulo	@Policia_Civil
Polícia Militar do Estado de São Paulo	@PMESP
São Paulo Agora — CCOI ^b	@saopaulo_agora
São Paulo Transporte	@sptrans_
São Paulo Turismo	@TurismoSaoPaulo
Secretaria Municipal de Transportes de São Paulo	@smtsp_

^a Polícia Militar do Estado de São Paulo (PMESP).

^b Centro de Controle Integrado 24 Horas da Cidade de São Paulo.

Fonte: Felipe Cordeiro Alves Dias

1.4 Hipóteses

Com base na Revisão Sistemática do Cap. 3, os eventos de exceção presentes nos tweets podem ser caracterizados, não exaustivamente, em:

1. **Acidentes.**

- Acidentes nas estações de transporte (ITOH et al., 2016).
- Incêndio (ITOH et al., 2016).

2. **Espaço-temporais.**

- Dia da semana (CHEN et al., 2016).
- Hora do dia (CHEN et al., 2016).

3. Eventos sociais.

- a) Feiras de rua (CHEN et al., 2016).
- b) Festivais (CHEN et al., 2016), (LECUE et al., 2014).
- c) Jogos esportivos (CHEN et al., 2016), (GAL-TZUR et al., 2014).
- d) Passeatas e maratonas (CHEN et al., 2016), (ITOH et al., 2016).

4. Eventos urbanos.

- a) Relacionados ao tráfego (CHEN et al., 2016); (LECUE et al., 2014).

5. Desastres naturais.

- a) Tempestades (ITOH et al., 2016).
- b) Terremoto (ITOH et al., 2016).
- c) Tufões (ITOH et al., 2016).

6. Metereológicas.

- a) Dia claro, nublado, chuvoso, nevando, com neblina (CHEN et al., 2016).
- b) Temperatura do ar (CHEN et al., 2016).

Dito isso, espera-se que seja possível identificar tais características utilizando Processamento de Linguagem Natural (NLP — *Natural Language Processing*) em conjunto com dicionários auxiliares para o contexto dos eventos de exceção mencionados.

Após a identificação dos eventos de exceção, temos como hipótese que seja possível extrair, com confiabilidade, os endereços dos tweets utilizando a técnica de Expressão Regular. Pois em uma análise preliminar observamos que o conteúdo das contas selecionadas, citadas na tabela 1, utilizam padrões de formatação para os endereços publicados. Com isso, podemos afirmar que esses tweets apresentam a característica de serem semi-estruturados, diferentemente dos tweets não estruturados publicados pelos usuários comuns do Twitter; o que consequentemente simplifica o processamento necessário para geolocalizar os eventos de exceção.

1.5 Organização do documento

Neste documento, é apresentado o Cap. 1 sobre a introdução do trabalho; o Cap. 2 a respeito da fundamentação teórica; Cap. o 3 sobre a revisão sistemática

realizada; o Cap. ?? referente a proposta de pesquisa e o Cap. 8 contendo a conclusão da proposta apresentada.

Atualizar organização do documento

2 Fundamentação Teórica

Neste capítulo, são apresentados fundamentos teóricos sobre os conceitos Cidades Inteligentes; Sistemas de Transporte Inteligentes; relacionados ao transporte público; *General Transit Feed Specification (GTFS)*; Redes Sociais; Processamento de Linguagem Natural; *Feature Engineering* e Aprendizado de Máquina.

2.1 Cidades Inteligentes

Embora não haja consenso, o conceito de Cidades Inteligentes (SC — *Smart Cities*) tem sido definido pela literatura principalmente como cidades sustentáveis e socialmente inclusivas (WANG; SINNOTT; NEPAL, 2016), que utilizam Tecnologias da Informação e Comunicação (TICs) para gerir eficientemente seus respectivos recursos naturais, de energia, transporte, lixo, dentre outros (AHVENNIEMI et al., 2017). As SC podem ter viés tecnológico (*TDM* — *Technology Driven Method*; top-down; de fornecimento), ou, humano (*HDM* — *Human Driven Method*; bottom-up; de demanda) (KUMMITHA; CRUTZEN, 2017).

O aspecto humano das Cidades Inteligentes começou a ser explorado recentemente, após críticas referentes aos poucos indicadores humanos existentes para SC (AHVENNIEMI et al., 2017) (FINGER; RAZAGHI, 2017). A abordagem humana das SC foca questões sociais e qualidade de vida, tais como governança participativa, segurança, cultura, lazer, sustentabilidade, desenvolvimento de capital humano, dentre outras (AHVENNIEMI et al., 2017). Na perspectiva tecnológica de SC, argumenta-se que apenas o uso de TICs seja capaz viabilizar o desenvolvimento de capital humano e de soluções para os problemas da cidade (KUMMITHA; CRUTZEN, 2017).

Independentemente dos vieses humano e tecnológico, a cidade pode ser conceituada como um complexo e dinâmico sistema sócio-técnico. Ou seja, uma cidade (região metropolitana) é composta por sistemas urbanos, com espaços físicos para a vida cotidiana e com sistemas de infraestrutura (para transporte, energia, água e tratamento de água, moradia, telecomunicações e áreas verdes). Os sistemas urbanos por natureza nunca estão em equilíbrio, possuem subsistemas imprevisíveis (FINGER; RAZAGHI, 2017).

Apesar disso, as TICs permeiam os sistemas urbanos e espaços físicos, o que tem sido acentuado com o crescente número de sensores e dispositivos conectados à Internet (*IoT — Internet of Things*), de dados voluntários enviados por pessoas via dispositivos móveis e, de conteúdo existente em Redes Sociais sobre os acontecimentos da cidade. Tais fontes heterogêneas geram grandes volumes de dados, utilizados para desenvolver serviços de Cidades Inteligentes (FINGER; RAZAGHI, 2017) (ANG et al., 2017).

O desenvolvimento de serviços de SC envolve desafios relacionados a conectividade (infraestrutura de rede, interoperabilidade e padrões, consumo de energia e escalabilidade) e aos dados (capacidade e local de armazenamento, extração, tratamento, processamento, análise, integração e agregação dos dados) (ANG et al., 2017), (XIAO; LIM; PONNAMBALAM, 2017). Além disso, a análise de dados pode tanger problemas referentes a correlação e inferência de dados de diferentes domínios, aprendizado de máquina, processamento em tempo real e propostas de novo uso para dados provenientes de infraestruturas já existentes (ANG et al., 2017).

Por fim, a seguir estão elencadas algumas frentes de estudo e de desenvolvimento de serviços de SC que ilustram iniciativas em Cidades Inteligentes:

- ***Smart buildings*** (TALARI et al., 2017), (MORENO et al., 2017), (ANG et al., 2017), (FINGER; RAZAGHI, 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017).
- ***Smart citizen / community / people*** (TALARI et al., 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).
- ***Smart economy*** (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (XIAO; LIM; PONNAMBALAM, 2017), (AHVENNIEMI et al., 2017).
- ***Smart environment*** (*electricity, waste, water, green space*) (SANTOS et al., 2017), (FINGER; RAZAGHI, 2017), (TALARI et al., 2017), (ANG et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).
- ***Smart governance*** (TALARI et al., 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).

- **Smart living** (*education, health, safety, cultural*) (SANTOS et al., 2017), (TALARI et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (XIAO; LIM; PONNAMBALAM, 2017), (AHVENNIEMI et al., 2017).
- **Smart transportation / mobility** (TALARI et al., 2017), (MORENO et al., 2017), (ANG et al., 2017), (FINGER; RAZAGHI, 2017), (SANTOS et al., 2017), (KUMMITHA; CRUTZEN, 2017), (BARTH et al., 2017), (AHVENNIEMI et al., 2017).

2.2 Sistemas de Transporte Inteligentes

Sistemas de Transporte Inteligentes (ITS — *Intelligent Transportation Systems*) é uma das mais antigas tecnologias presentes em Cidades Inteligentes (MENOUAR et al., 2017), que tem como fim utilizar TICs para resolver problemas relacionados ao transporte, tais como congestionamento, segurança, eficiência e conservação ambiental (FIGUEIREDO et al., 2001).

É importante notar a diferença entre o termo *Intelligent* e *Smart* de *Smart transportation / mobility*, o primeiro, respectivamente, refere-se apenas ao uso de tecnologias, enquanto que o segundo ao uso de TICs para transformar de forma significativa a vida cotidiana das pessoas (ALBINO; BERARDI; DANGELICO, 2015). A seguir, algumas das categorias de ITS estão enumeradas:

1. **Advanced Traffic Management System (ATMS)** — são sistemas utilizados para melhorar a qualidade do serviço de tráfego e redução de atrasos (FIGUEIREDO et al., 2001), por meio de:
 - a) *Collection data team*: equipe de pessoas responsáveis por monitorar e coletar dados das condições de tráfego.
 - b) *Support systems*: conjunto de câmeras, semáforos, sensores, dentre outros dispositivos auxiliares para gerenciar e controlar o tráfego em tempo real.
 - c) *Real time traffic control systems*: sistemas utilizados para com base nos dados coletados controlar acesso a avenidas, semáforos, envio de mensagens para os dispositivos de monitoramento.

2. ***Advanced Travelers Information Systems (ATIS)*** — são sistemas utilizados para fornecer informação em tempo real aos viajantes (FIGUEIREDO et al., 2001).
3. ***Commercial Vehicles Operation (CVO)*** — são sistemas utilizados para a segurança de veículos comerciais e frotas, por meio de tecnologias relacionadas a gerenciamento de tráfego, controle e gerenciamento de veículos e informações aos viajantes (FIGUEIREDO et al., 2001), tais como:
 - a) *Automatic Vehicles Identification.*
 - b) *Automatic Vehicles Classification.*
 - c) *Automatic Vehicles Location.*
 - d) *Pedestrian Movement Detection.*
 - e) *Board Computers.*
 - f) *Real Time Traffic Transmissions.*
4. ***Advanced Public Transportations Systems (APTS)*** — são sistemas que utilizam ATMS e ATIS para melhorar a eficiência e operação do transporte público coletivo (FIGUEIREDO et al., 2001). É importante observar que APTS também podem utilizar CVO.
5. ***Advanced Vehicles Control Systems (AVCS)*** — são sistemas compostos por sensores, computadores e sistemas de controle para auxiliar e alertar motociclistas, com o objetivo de melhorar a segurança e reduzir congestionamentos (FIGUEIREDO et al., 2001).

As categorias mencionadas anteriormente representam parte da primeira geração de tecnologias em ITS, a próxima geração tem como foco veículos autônomos e conectados, capazes de trocarem informações entre si em tempo real para melhorar a segurança dos condutores (MENOUAR et al., 2017).

2.3 Conceitos relacionados ao transporte público

Esta seção define os conceitos relacionados ao transporte público, de acordo com a perspectiva do Plano de Mobilidade Urbana do Município de São Paulo — PlanMob/SP 2015⁴.

2.3.1 Acessibilidade

A acessibilidade pode ser considerada como um atributo do espaço urbano, o qual é diretamente proporcional a abrangência e adequação das infraestruturas de acesso ao espaço urbano. As regiões da cidade têm diferentes padrões de infraestrutura de transporte e deslocamento, portanto, são diferenciadas no aspecto de acessibilidade. Além disso, a acessibilidade atua como instrumento de acesso as oportunidades socioeconômicas da cidade. Observa-se que a acessibilidade não é entendida como um atributo econômico relacionado ao valor das tarifas do transporte, ou, as condições de uso (como o congestionamento viário).

Uma qualidade específica do espaço urbano é a acessibilidade universal, que o caracteriza como acessível a pessoas com deficiência (PCDs). A acessibilidade universal é garantida ao eliminar as barreiras físicas que impedem a participação plena e efetiva das PCDs ao espaço urbano.

2.3.2 Mobilidade

A mobilidade pode ser entendida como um atributo do indivíduo, o qual está relacionado a sua capacidade de se deslocar pelo território da cidade e a sua respectiva renda (dimensão econômica); ou seja, pessoas ou famílias de maior renda tendem a ter maior número de viagens. Além disso, observa-se que a restrição da mobilidade devido a má qualidade das infraestruturas urbanas é considerada como falta de acessibilidade ao espaço e não como perda de mobilidade do indivíduo.

A condição de mobilidade pode ser calculada pelo indicador conhecido como taxa ou índice de mobilidade, determinado pelo quociente entre o total de viagens realizadas e o total da população residente em uma região. Tal indicador pode ser especializado de acordo o tipo de mobilidade, por exemplo, ao considerar apenas as viagens motorizadas, obtém-se o índice de mobilidade motorizada; e ser caracterizado como crescente ou decrescente de acordo com fatores socioeconômicos.

Além da mobilidade como atributo do indivíduo, existe a mobilidade como atributo da cidade, conhecida como mobilidade urbana. A mobilidade urbana consi-

dera um conjunto de fatores de uma aglomeração urbana que tornam a mobilidade mais qualificada e eficiente, tais como:

1. Transporte público coletivo;
2. transporte de alta capacidade;
3. acessibilidade universal nos passeios e edificações;
4. prioridade ao transporte coletivo no sistema viário;
5. terminais de transporte intermodais;
6. rede de transporte coletivo por ônibus (com acessibilidade universal);
7. rede cicloviária;
8. bicicletários e paraciclos;
9. legibilidade dos sistemas de orientação;
10. comunicação eficaz com os usuários;
11. modicidade tarifária;
12. logística eficiente no transporte de carga, dentre outros itens.

2.3.3 Viagem e modais de transporte

O conceito de viagem no setor de transportes é definido como o deslocamento de uma pessoa entre dois pontos de interesse (origem e destino), com um motivo definido e por meio de um modal de transporte. A saber, os modais de transporte considerados no *PlanMob/SP 2015* estão enumerados a seguir:

1. A pé.
 - a) Independentemente do deslocamento percorrido caso o motivo seja escola ou trabalho;
 - b) Superior a 500 metros de deslocamento.
2. Coletivos.
 - a) Metrô;
 - b) ônibus;
 - c) ônibus fretado;
 - d) ônibus escolar e lotação;
 - e) trem.

3. Individuais.

- a) Automóveis (bicicleta, carro particular, caminhão, moto e táxi).

2.4 General Transit Feed Specification

A *GTFS* — *General Transit Feed Specification*¹, como o próprio nome sugere, é uma especificação de um formato comum (o que permite interoperabilidade) para troca de informações estáticas sobre transporte público. Um *feed* especificado na GTFS estática é composto por arquivos de texto (que seguem determinados requisitos semelhantes aos do formato *CSV*¹) compactados no formato *Zip*², e detalhados na tabela 2. Cada arquivo modela diferentes perspectivas do transporte público, tais como paradas, trajetos, viagens e outros dados relativos a horário.

Além da GTFS estática existe a *GTFS-realtime*¹, que é uma extensão da GTFS estática, assim, para usar *feeds* em tempo real é necessário definir os arquivos estáticos da GTFS, que são utilizados na *GTFS-realtime* para obter as informações do sistema de transporte público. A *GTFS-realtime* é utilizada para transmissões em tempo real de três tipos de *feeds*¹, enumerados e detalhados a seguir:

1. Atualizações dos horários de parada.

- a) Descritor de viagem: viagem programada (de acordo ou próxima a uma programação GTFS), adicionada (não programada e adicionada, por exemplo, para atender à demanda ou substituir um veículo quebrado), desprogramada (que está sendo feita e não está associada a uma programação, por exemplo, quando não há uma programação, e os ônibus rodam em um serviço de translado), cancelada (viagem programada, mas removida), substituição (substitui uma parte da programação estática).
- b) Indefinição: especifica o erro esperado no atraso real como um número inteiro, em segundos.

2. Alertas de serviço.

- a) Intervalo de tempo: o alerta será exibido eventualmente, no intervalo de tempo especificado.

¹ <<https://developers.google.com/transit>>. Acesso em Outubro, 29 de 2017.

² <<https://support.pkware.com/display/PKZIP/APPNOTE>>. Acesso em Outubro, 29 de 2017.

- b) Seletor de entidade: agência (afeta toda a rede de transporte público), trajeto (afeta todo o trajeto), tipo de trajeto (afeta qualquer trajeto desse tipo, por exemplo, todos os ônibus), viagem (afeta uma viagem específica) e parada (afeta uma parada específica).
- c) Causa: desconhecida, outra causa (não representada por nenhuma destas opções), problema técnico, greve, manifestação, acidente, feriado, tempo, manutenção, construção, atividade policial, emergência médica.
- d) Efeito: sem serviço, serviço reduzido, atrasos significativos (atrasos não significativos só devem ser fornecidos por Atualizações de viagem), desvio, serviço adicional, serviço modificado, parada deslocada, outro efeito (não representado por qualquer uma dessas opções), efeito desconhecido.

3. Posições de veículos.

- a) Posição: a posição contém os dados de localização na posição do veículo, com os campos obrigatórios latitude e longitude, e com os campos opcionais rumo (direção que o veículo está seguindo), odômetro (distância que o veículo percorreu) e velocidade (velocidade no momento medida pelo veículo, em metros por segundo).
- b) Nível de congestionamento: congestionamento desconhecido, fluxo estável, paradas frequentes, congestionamento e congestionamento grave.
- c) Status de parada do veículo: chegando em (o veículo está prestes a chegar na parada em questão), parado em (o veículo está parado na parada em questão), em direção a (a parada em questão é a próxima parada do veículo — padrão).
- d) Descritor do veículo: id único (sistema de identificação interna do veículo), etiqueta de identificação (visível ao usuário) e placa real do veículo.

No demais, os *feeds* da GTFS-realtime são atualizados frequentemente, serializados em *Protocol Buffers*³ e transmitidos via protocolo HTTP⁴. A estrutura dos dados é definida em um arquivo *gtfs-realtime.proto*¹, usado para gerar o modelo de dados dos *feeds* em diferentes linguagens de programação, tais como *Java*, *C++* ou *Python*.

³ <<https://developers.google.com/protocol-buffers>>. Acesso em Outubro, 29 de 2017.

⁴ <<https://tools.ietf.org/html/rfc2616>>. Acesso em Outubro, 29 de 2017.

Tabela 2 – Detalhamento dos arquivos da GTFS

Nome do arquivo	Condisional	Contéudo ^a
<i>agency.txt</i>	Obrigatório	Contém uma ou mais agências de transporte público como fonte dos dados.
<i>stops.txt</i>	Obrigatório	Contém os locais individuais em que os veículos peggam ou deixam passageiros.
<i>routes.txt</i>	Obrigatório	Contém os trajetos de um grupo de viagens exibidas aos passageiros como um único serviço.
<i>trips.txt</i>	Obrigatório	Contém as viagens de cada trajeto. Uma viagem é uma sequência de duas ou mais paradas que ocorrem em um horário específico.
<i>stop_times.txt</i>	Obrigatório	Contém os horários de partida e chegada dos veículos em paradas específicas em cada viagem.
<i>calendar.txt</i>	Obrigatório	Contém datas para IDs de serviço que usam uma programação semanal. Especificam quando o serviço começa e termina, bem como os dias da semana em que o serviço está disponível.
<i>calendar_dates.txt</i>	Opcional	Contém as exceções para IDs de serviço definidos no arquivo <i>calendar.txt</i> . Se o arquivo <i>calendar_dates.txt</i> inclui todas as datas de serviço, ele pode ser especificado no lugar do <i>calendar.txt</i> .
<i>fare_attributes.txt</i>	Opcional	Contém informações sobre tarifas dos trajetos de uma empresa de transporte público.
<i>fare_rules.txt</i>	Opcional	Contém regras para implementação das informações de tarifa dos trajetos de uma empresa de transporte público.
<i>shapes.txt</i>	Opcional	Contém regras para desenhar linhas em um mapa para representar os trajetos de uma empresa de transporte público.
<i>frequencies.txt</i>	Opcional	Contém os intervalos entre as viagens nos trajetos.
<i>transfers.txt</i>	Opcional	Contém regras para conexões em pontos de baldeação entre os trajetos.
<i>feed_info.txt</i>	Opcional	Contém informações adicionais sobre o <i>feed</i> , incluindo editor, versão e informações sobre validade.

^a Os campos contidos em cada arquivo da especificação GTFS estão descritos no apêndice C, nas tabela 12 - 24.

Fonte: Google Transit (adaptada)¹

2.5 Redes Sociais

As Redes Sociais podem ser definidas como redes que possuem muitos relacionamentos, com grandes componentes conectados, altos coeficientes de agrupamento e grau de reciprocidade. Tais características, por exemplo, podem ser encontradas na rede social *Facebook*⁵. O *Twitter*⁶ além de possuir as características de rede social mencionadas anteriormente, pode ser caracterizado também como uma Rede de Informações. Nesse tipo de rede a interação dominante é a disseminação de informações entre os relacionamentos, com baixo índice de reciprocidade (MYERS et al., 2014).

No *Twitter* as informações (*tweets*) são publicadas contendo no máximo 280 caracteres; cada publicação pode receber *retweets* (ser compartilhada por outros usuários), comentários (diretamente no *tweet* — *replies* — ou de forma privada via caixa de mensagens) e *likes* (indicador de quantos usuários gostaram da publicação). Além dessas funcionalidades, os *tweets* podem conter menções a outros usuários (@*nome do profile*) e rótulos (#*hashtag*) indicando assuntos, categorias, etc.

Devido as características citadas nos parágrafos anteriores, o *Twitter* tem sido uma rede social importante para compartilhamento de informações e acontecimentos do cotidiano. Tais acontecimentos podem ser classificados como eventos sociais, capazes de descrever desde eventos rotineiros (*shows*, jogos esportivos, etc.) a situações de crise (eventos de exceção — desastres naturais, mobilizações sociais, dentre outros) (ZHOU; CHEN, 2014), (ATEFEH; KHREICH, 2015).

2.6 Processamento de Linguagem Natural

O processamento automático de *tweets* envolve o Processamento de Linguagem Natural (NLP — *Natural Language Processing*), que explora como computadores podem ser utilizados para entender e manipular texto ou fala em linguagem natural (LIU; LI; THOMAS, 2017), o que envolve conhecimento interdisciplinar principalmente entre as áreas de ciência da computação, linguística e estatística. A

⁵ <<https://www.facebook.com>>. Acesso em Outubro, 29 de 2017.

⁶ <<https://twitter.com>>. Acesso em Outubro, 29 de 2017.

seguir são detalhados alguns dos problemas relacionadas a NLP, divididos em baixo e alto nível (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011):

1. Baixo nível (problemas comuns a NLP) (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
 - a) ***Sentence Boundary Disambiguation (SBD)***: processamento para identificação do início e fim de uma sentença (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
 - b) ***Tokenization***: processamento realizado para obtenção das palavras (*tokens*) que compõem uma sentença, inclui a remoção de números, pontuações e caracteres que não pertencem ao alfabeto (SETIAWAN; WIDYANTORO; SURENDRO, 2017).
 - c) ***Part-of-speech tagging***: processamento para identificação das classificações gramaticais (verbo, sujeito, adjetivo, etc.) das palavras em uma sentença, considerando seus respectivos significados e contexto no qual estão inseridas (ROY; MAJUMDER; NATH, 2017).
 - d) ***Decomposição morfológica***: processamento para decomposição morfológica de uma determinada palavra para a sua forma inflexionada, usando *lemmatization* (identificação do lema da palavra) ou *stemming* (identificação da raiz da palavra usando heurísticas para determinar a localização de sua respectiva flexão) (SETIAWAN; WIDYANTORO; SURENDRO, 2017), (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011), (KORENIUS et al., 2004).
 - e) ***Shallow parsing (chunking)***: processamento para identificação de segmentos de uma sentença, tais como frases verbais, nominais, etc., com base nos *tokens* que constituem a *part-of-speech* (COLLOBERT et al., 2011), (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
2. Alto nível (aplicação de NLP a problemas específicos, com base nos problemas de baixo nível) (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
 - a) ***Spelling / grammatical error identification and recovery***: processamento iterativo para identificação e correção de erros gramaticais e de digitação. (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

- b) **Named Entity Recognition (NER)**: processamento para identificação e categorização de palavras ou frases específicas (entidades) (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
- c) **Word Sense Disambiguation (WSD)**: processamento para identificação do sentido de uma palavra numa sentença (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
- d) **Negation and uncertainty identification**: processamento para inferir se uma entidade está presente ou não numa sentença, assim como quantificar a quantidade de incerteza da inferência realizada (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
- e) **Extração de relacionamentos**: processamento para identificar relacionamentos entre entidades e eventos (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
- f) **Extração de relacionamento / inferência temporal**: processamento para inferência de expressões e relacionamentos temporais (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).
- g) **Extração de informação**: processamento para extração e transformação para uma forma estruturada de informações específicas a um problema (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Para esta pesquisa, utilizamos o processo de *tokenização TweetTokenizer*⁷ para extrair os *tokens* dos *tweets* (*features* utilizadas para treinar os modelos de classificações) e o processo de *stemming RSLPStemmer*⁸ para redução do espaço de *features*, além da remoção de palavras vazias (*stopwords*^{9,10}) do Português Brasileiro. — palavras comuns do Português Brasileiro.

⁷ <<https://www.nltk.org/api/nltk.tokenize>>. Acessado em 15 de maio de 2018.

⁸ <https://www.nltk.org/_modules/nltk/stem/rslp>. Acessado em 15 de maio de 2018.

⁹ <http://www.nltk.org/howto/portuguese_en>. Acessado em 15 de maio de 2018.

¹⁰ Palavras com alta ou baixa frequência no corpus — comuns ou raras — ou removidas por meio de *feature selection* — <http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>. Acessado em 03 de junho de 2018.

2.7 Feature Engineering

Feature engineering é um processo iterativo que utiliza o conhecimento do domínio dos dados e de suas métricas para criar (*feature construction*), extrair (*feature extraction*) e selecionar *features* (*feature selection*) para serem utilizadas em algoritmos de aprendizado de máquina. Um conjunto de dados pode ser representado por um número fixo de *features* binárias, categóricas ou contínuas. Antes do processo de *feature engineering*, os dados podem ser pré-processados usando técnicas de padronização, normalização, remoção de ruído, redução de dimensionalidade, discretização, expansão, entre outros; é importante notar que informações podem ser perdidas ao realizar essas transformações (GUYON; ELISSEEFF, 2006).

No experimento abordado no Cap.6 usamos uma fase de pré-processamento, explicada na subseção 6.1, e um processo para *feature extraction* (explicado adiante) realizado por meio de uma função que utiliza NLP para preparar os *tweets* coletados para a tarefa de treinamento. As fases de *feature construction* e *feature selection* não são utilizadas pelos experimentos deste trabalho, porém, são mencionadas para um melhor entendimento.

Sendo assim, na fase de *feature construction*, é realizado um processo para descobrir informações ausentes sobre as relações entre as *features* e para aumentar o espaço de *features*, inferindo ou criando novas *features* com o objetivo de melhorar a precisão dos algoritmos de classificação, entender os dados e obter dados ocultos., etc. (MOTODA; LIU, 2002). Neste estágio, de um conjunto de n *features* A_1, A_2, \dots, A_n , é possível construir *features* adicionais $A_{n+1}, A_{n+2}, \dots, A_{n+m}$, por meio de heurísticas, operadores lógicos, algoritmos, etc (MOTODA; LIU, 2002).

Por fim, no processo de extração de *features*, usa uma função de mapeamento para extrair um conjunto mínimo de novas *features* com base nas *features* originais e em métricas de desempenho, diferentemente da análise das relações entre *features* na fase de *feature construction* (MOTODA; LIU, 2002). Assim, com um conjunto inicial de n *features* A_1, A_2, \dots, A_n é possível extrair novas *features* $B_1, B_2, \dots, B_m (m < n)$, $B_i = F_i(A_1, A_2, \dots, A_n)$, onde F_i é a função de mapeamento (MOTODA; LIU, 2002). Analogamente, no processamento de *tweets* realizado no Cap. 6, o espaço de *features* é composto inicialmente por cada palavra extraída do processo de *Tokenization*,

o qual posteriormente é reduzido pelas funções responsáveis pelos processos de *stemming* e remoção de *stopwords*.

2.8 Algoritmos de Aprendizado de Máquina

Os algoritmos de Aprendizado de Máquina podem ser (I) supervisionados, nos quais relações com resultados conhecidos são criadas com base nas características de entrada; (II) não-supervisionado, nos quais são conhecidas as características de entrada, mas não os resultados; (III) semi-supervisionados, nos quais podem ser definidas algumas das relações entre dados de entrada e resultados; (IV) por reforço, nos quais são estabelecidas ações com o foco em maximizar determinado ganho.

No contexto desse trabalho, os dados de entrada são conhecidos e foram classificados manualmente, devido a isso usamos aprendizado de máquina supervisionado para o desenvolvimento do modelo de classificação, abordagem a qual também possui melhor desempenho para a tarefa de classificação textual (DWIVEDI; ARYA, 2016). Com base nisso, realizamos uma revisão não sistemática e, de acordo com a literatura, os seguintes algoritmos são os mais utilizados para aprendizado supervisionado (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007; DWIVEDI; ARYA, 2016; NARAYANAN et al., 2017):

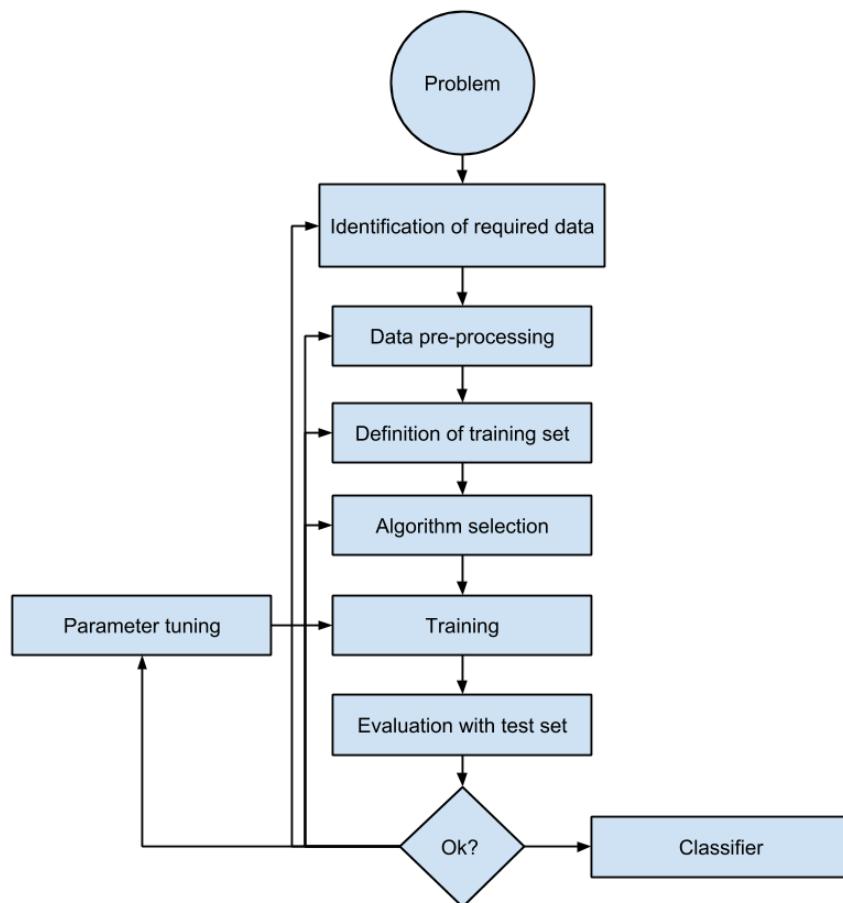
- Árvore de Decisão (*Decision Tree*).
- Floresta Aleatória (*Random Forest*).
- K-ésimo Vizinho mais Próximo (K-NN — *K-Nearest Neighbour*).
- Máquina de Vetores de Suporte (SVM — *Support Vector Machine*).
- *Naive Bayes*.
- Redes Neurais (*Neural Networks*).
- Regressão Logística (*Logistic Regression*).

2.8.1 Algoritmos de aprendizado supervisionado

De acordo com a Fig. 1, a aplicação de algoritmos de aprendizado supervisionado a um problema passa por algumas fases. As primeiras fases se referem aos processos de construção do conjunto de dados (*identification of required data*) e

pré-processamento (*data pre-processing*), descritas respectivamente no Cap. 4 e seção 6.1, as demais fases (*definition of training set* — definição do conjunto de treinamento; *algorithm selection* — seleção do algoritmo; *training* — treinamento; *evaluation with test set* — validação com conjunto de teste; *classifier* — classificador) são explicadas na subseção 6.5. É importante observar que não faz parte do escopo deste trabalho afinar os parâmetros dos algoritmos mencionados na subseção 2.8 (fase *parameter tuning*), devido a isso as parametrizações padrões são utilizadas e descritas no apêndice F.

Figura 1 – Fluxograma do processo do aprendizado supervisionado



Fonte: (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007)

Árvore de Decisão

Floresta Aleatória

K-ésimo Vizinho mais Próximo

Máquina de Vetores de Suporte

Naive Bayes

Redes Neurais

Regressão Logística

Escrever sobre cada algoritmo utilizado

2.8.2 Validação dos modelos de aprendizado supervisionado

The validation of the models to classification tasks can be realized through 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. Following are some of the main metrics utilized to:

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 score = \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

2.9 Term frequency–Inverse document frequency

TF-IDF é um algoritmo de ponderação de variáveis que combina as ponderações *frequência do termo* (TF — Term Frequency) e *inverso da frequência nos*

documentos (IDF — *Inverse Document Frequency*) para calcular os pesos dos termos linguísticos (variáveis) em um determinado corpus. Em outras palavras, o peso da variável é proporcional a frequência com a qual aparece nos documentos, e inversamente proporcional a quantidade de documentos que contém o termo linguístico em questão (WU; YUAN, 2018; YAHAV; SHEHORY; SCHWARTZ, 2018).

Dentre as variações de implementação da ponderação $W_{t,d}$ (TF-IDF) existentes, a abordagem tradicional considera uma coleção de termos $t \in T$ que aparecem em um conjunto de N documentos $d \in D$, posto isso, defini-se como o produto entre $tf_{i,j}$ e idf_i — onde $n_{i,j}$ é a frequência do termo t_i no documento d_j , $\sum_k n_{k,j}$ o somatório da frequência de todos os termos do documento d_j e n o número de documentos onde t_i aparece ($n + 1$, caso $n = 0$) — conforme a seguinte equação (WU; YUAN, 2018):

$$\begin{aligned} tf_{i,j} &= \frac{n_{i,j}}{\sum_k n_{k,j}} \\ idf_i &= \log \frac{N}{n + 1} \\ W_{t,d} &= tf_{t,d} * idf_t \end{aligned} \tag{5}$$

No contexto deste trabalho, entendemos documentos como as classes dos eventos de exceção. A *frequência dos termos* (TF — $tf_{t,d}$) é determinada por classe e a *frequência do termo - inverso da frequência nos documentos* (IDF — idf_t) como o inverso dos eventos de exceção, sendo N o tamanho do conjunto dos eventos de exceção, sob o qual df_t é definido. Os eventos de exceção são classificados em suas respectivas classes por meio dos modelos de aprendizado supervisionado, elencados na subseção 2.8.1.

3 Revisão Sistemática

Este capítulo apresenta uma Revisão Sistemática (RS) com o objetivo de encontrar o estado da arte de trabalhos que visam melhorar sistemas de transporte público por meio do processamento de tweets. Além disso, de uma forma mais ampla, busca-se também entender como os tweets têm sido utilizados na caracterização de problemas urbanos. Sendo assim, o capítulo é iniciado com a seção sobre o planejamento da Revisão Sistemática; seguida das questões de pesquisa utilizadas na formulação do problema da RS; do processo de coleta dos estudos primários; da avaliação dos dados coletados; da análise e interpretação dos estudos selecionados, concluindo com as considerações finais.

3.1 Planejamento da Revisão Sistemática

A presente Revisão Sistemática utiliza a metodologia proposta por BIOLCHINI et al. (2005), composta por cinco etapas. A primeira etapa está relacionada à formulação do problema, na qual é levantada uma questão central se referindo ao tipo de evidência que deverá estar contida na revisão. Em seguida, são construídas definições que permitem estabelecer uma distinção entre os estudos relevantes e irrelevantes para o propósito específico do que se está investigando (BIOLCHINI et al., 2005).

A segunda etapa da condução está relacionada à Coleta de Dados, na qual são definidos os procedimentos que serão utilizados para encontrar a evidência relevante que foi definida na etapa anterior. Nesta fase é extremamente importante determinar as fontes que podem fornecer estudos relevantes a serem incluídos na pesquisa (BIOLCHINI et al., 2005).

Na terceira etapa a Avaliação de Dados é definida, na qual são selecionadas as fontes primárias que deverão ser incluídas na revisão. Em seguida, são aplicados os critérios de qualidade para separar estudos que podem ser considerados válidos, e determinadas as diretrizes para o tipo de informação que deve ser extraída dos relatórios de pesquisas primárias (BIOLCHINI et al., 2005).

A quarta etapa da revisão é o processo de Análise e Interpretação, na qual os dados dos estudos primários válidos são sintetizados. E, na quinta etapa são realizados os processos de Conclusão e Apresentação (BIOLCHINI et al., 2005).

3.1.1 Justificativa da Revisão Sistemática

Esta Revisão Sistemática se justifica por não terem sido encontradas revisões sistemáticas com o foco em questões urbanas e de transporte público, abordando unicamente o processamento de *tweets*. Em (CHANIOTAKIS; ANTONIOU; PEREIRA, 2016), por exemplo, foi realizado um mapeamento de forma não sistemática dos trabalhos sobre o uso das mídias sociais em problemas relacionados ao transporte público; (STEIGER; ALBUQUERQUE; ZIPF, 2015), por outro lado, desenvolveram uma revisão sistemática sobre o uso do Twitter para questões espaço-temporais; e (JUNGHERR, 2016) no contexto político.

Devido a isso, a presente revisão sistemática se diferencia por ter como objetivo encontrar o estado da arte de trabalhos que visam melhorar sistemas de transporte público por meio do processamento de *tweets*. Além disso, de uma forma mais ampla, busca-se também entender como os *tweets* têm sido utilizados na caracterização de problemas urbanos.

3.2 Questões de Pesquisa

Nesta seção, são apresentadas as questões de pesquisa utilizadas para a formulação dos problemas abordados por essa Revisão Sistemática. Por meio das quais, busca-se atender os objetivos já mencionados na seção 3.1.1.

1. Quais os tipos de problemas urbanos abordados utilizando processamentos de *tweets*?

O propósito da QP1 é identificar quais são as contribuições do processamento de *tweets* para a mitigação de problemas urbanos. A resposta a essa questão de pesquisa ajudará especialistas das áreas multidisciplinares relacionadas ao Urbanismo (como a de Análise de Redes Sociais e Políticas Públicas) a terem

um panorama de como *tweets* podem ser utilizados para ajudar na solução de problemas urbanos.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP1): alguns dos problemas urbanos abordados estão relacionados ao transporte, mobilidade urbana, turismo e desastres naturais.

2. Como *tweets* têm sido utilizados para abordar problemas relacionados ao transporte público?

O propósito da QP2 é identificar se *tweets* têm sido utilizados para solucionar problemas relacionados ao transporte público. A resposta a essa questão de pesquisa ajudará especialistas das áreas multidisciplinares relacionadas ao Urbanismo (como a de Análise de Redes Sociais e Políticas Públicas) a terem um panorama de como *tweets* podem ser utilizados para ajudar na solução de problemas referentes a mobilidade urbana.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP2): *tweets* têm sido utilizados principalmente para questões relacionadas ao congestionamento, não tendo como foco o transporte público.

3. Quais as técnicas estatísticas utilizadas no processamento de *tweets*?

O propósito da QP3 é identificar quais as técnicas estatísticas utilizadas no processamento de *tweets*, principalmente no que se refere a análise de acurácia de classificação binária. A resposta a essa questão de pesquisa ajudará especialistas a terem um panorama de como garantir a confiabilidade ao utilizar dados oriundos de *tweets*, dentre outros aspectos relacionados a testes estatísticos.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP3): F_1 score é a principal técnica estatística utilizada

para análise de acurácia de classificação binária.

4. Quais os paradigmas de processamento têm sido utilizados ao lidar com *tweets*?

O propósito da QP4 é identificar os paradigmas utilizados para processamento de *tweets*. A resposta a essa questão de pesquisa ajudará especialistas a terem um panorama das técnicas de processamento utilizadas na análise de *tweets*.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP4): o principal paradigma utilizado tem sido o processamento de *tweets* em *batch (offline)*, após um processo de armazenamento. Poucos são os estudos que constroem uma plataforma para processamento de dados em tempo real.

5. Quais são os eventos de exceção relacionados ao transporte público?

O propósito da QP5 é identificar os eventos de exceção relacionados ao transporte público. A resposta a essa questão de pesquisa ajudará especialistas no levantamento de eventos de exceção relacionados ao transporte público, os quais podem ser utilizados em algoritmos de classificação.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP5): há poucos ou nenhum estudo que, ao tratar de problemáticas relacionadas ao transporte público, realizam um levantamento dos eventos de exceção desse contexto.

6. Quais as técnicas de Aprendizado de Máquina utilizadas no processamento de *tweets*?

O propósito da QP6 é identificar as técnicas de Aprendizado de Máquina utilizadas no processamento de *tweets*. A resposta a essa questão de pesquisa ajudará especialistas a terem um panorama das principais técnicas de Aprendi-

zado de Máquina utilizadas no processamento de *tweets*.

Uma análise preliminar dos estudos primários permite elaborar a seguinte Hipótese de Pesquisa (HP6): a técnica *Support Vector Machine* tem sido utilizada na maioria dos estudos que aplicam aos *tweets* algum algoritmo de Aprendizado de Máquina.

3.3 Coleta de dados

Nesta Revisão Sistemática, os artigos foram coletados em quatro fontes de pesquisa, por meio da plataforma de indexação de trabalhos acadêmicos *Google Scholar*¹. Constam na tabela 3 as bases pesquisadas no ano de 2017, quantidades de artigos coletados, descartados no processo de filtragem (Fig. 2, descrito na seção 3.4) e selecionados. Com base na QP1, a seguinte *string* de busca foi construída; restrita aos trabalhos publicados entre 2011 e 2016, escritos no idioma Inglês (devido ao fato das publicações relevantes, na área de Computação, estarem disponíveis nesse idioma):

String de busca: twitter urban planning city (analytics OR patterns OR tweets OR social OR media) AND (public transport)

Palavras-chave: twitter, urban, planning, city, analytics, patterns, tweets, social, media e public transport.

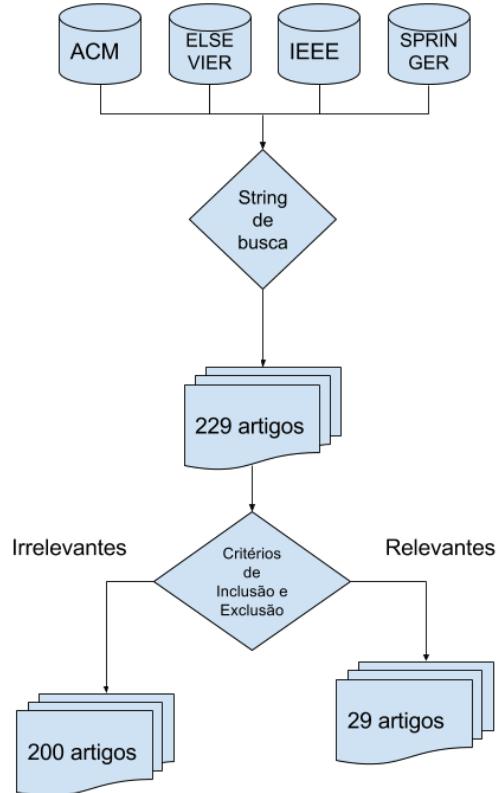
Tabela 3 – Quantidades de artigos coletados e fontes de busca

Fonte	Artigos coletados	Filtragem	Selecionados
ACM	44	34	10
IEEE	82	74	8
Elsevier	81	72	9
Springer	22	20	2
-	229	200	29

Fonte: Felipe Cordeiro Alves Dias

¹ <<https://scholar.google.com>>. Acesso em Outubro, 29 de 2017.

Figura 2 – Processo de Filtragem



Fonte: Felipe Cordeiro Alves Dias, 2017

3.4 Avaliação de Dados

Visando selecionar os artigos relevantes para esta Revisão Sistemática, os seguintes critérios foram utilizados no processo de filtragem:

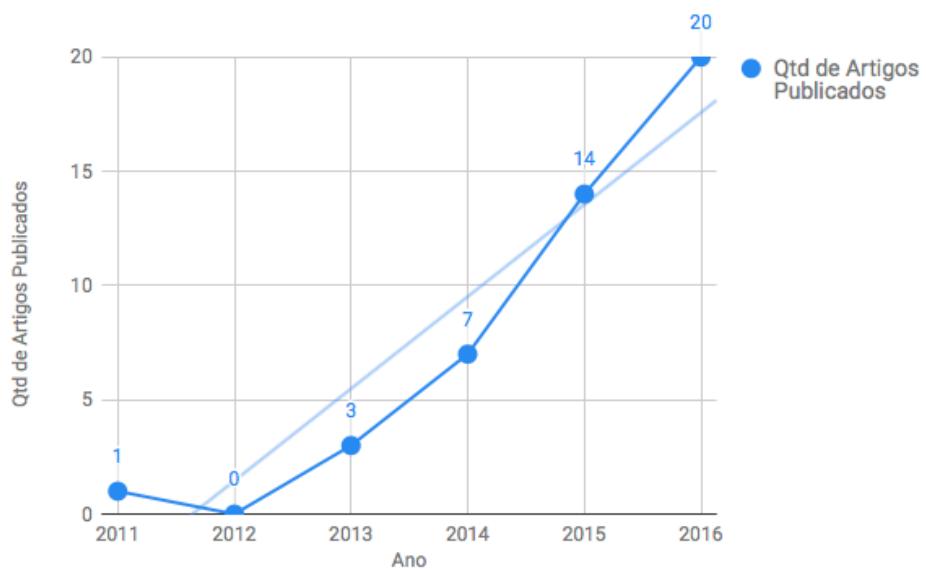
- Trabalho publicado (critério de qualidade).
- Trabalhos que utilizam *tweets* para abordar questões urbanas e de transporte público.
- Trabalhos duplicados.
- Trabalhos que estão fora do escopo da questão de pesquisa.

O processo de condução da Revisão Sistemática foi realizado utilizando os critérios acima mencionados, e está disponível em DIAS (2017) (não incluso neste trabalho com o objetivo de não deixar o texto exaustivo), assim como seu respectivo protocolo (no qual contém o detalhamento dos critérios de inclusão e exclusão,

dentre outros artefatos da condução). Após o processo de condução, alguns dos metadados dos artigos selecionados foram sintetizados.

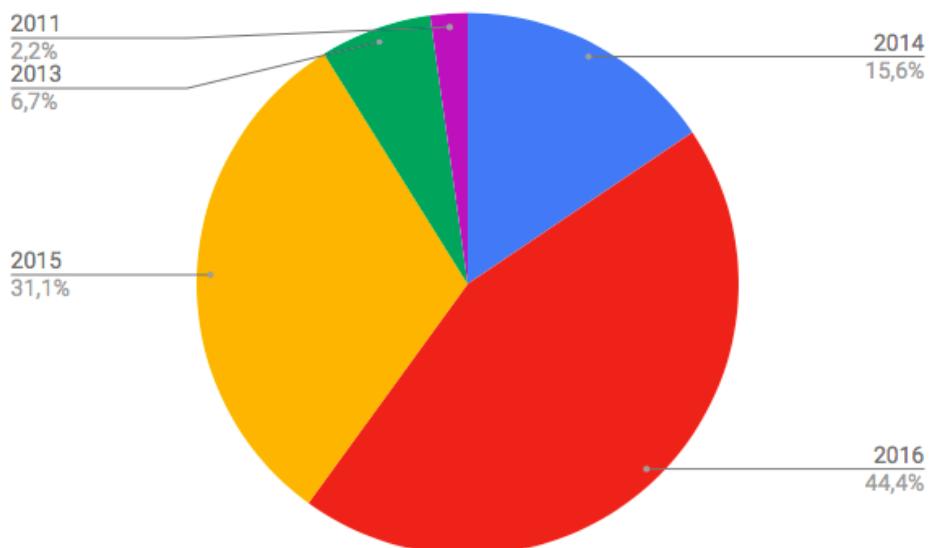
Sendo assim, a Fig. 5 apresenta uma nuvem de tags (gerada com a biblioteca *wordcloud* (Andreas Mueller, 2018)) sintetizando as palavras chaves dos estudos primários selecionados; e a Fig. 3 a quantidade de artigos publicados por ano, sendo possível analisar por meio dela a distribuição dos artigos entre 2011 e 2016, assim como sua respectiva porcentagem, ilustrada na Fig. 4.

Figura 3 – Quantidade de artigos publicados por ano



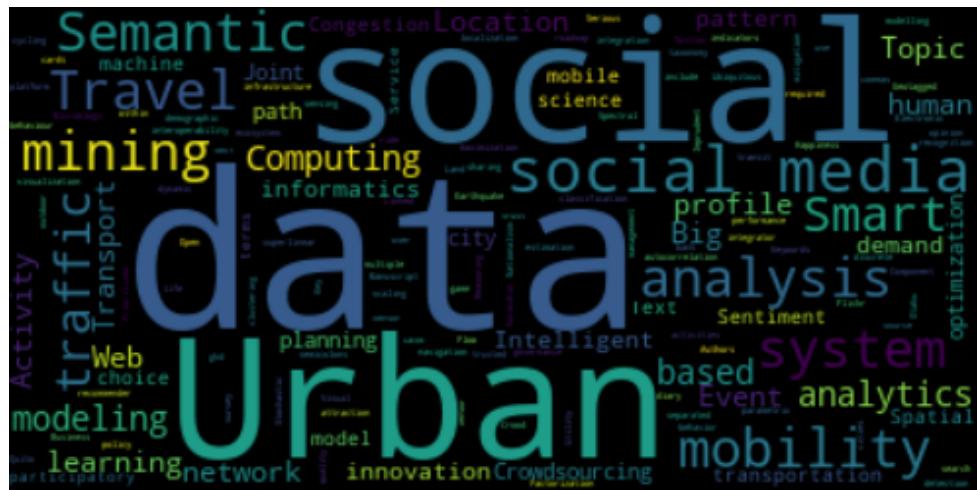
Fonte: Felipe Cordeiro Alves Dias, 2017

Figura 4 – Porcentagem dos artigos publicados por ano



Fonte: Felipe Cordeiro Alves Dias, 2017

Figura 5 – Nuvem de palavras das keywords dos artigos selecionados



Fonte: Felipe Cordeiro Alves Dias, 2017

3.5 Análise e Interpretação

Nesta seção é realizada a análise e interpretação dos estudos primários selecionados pela Revisão Sistemática, sendo as subseções divididas de acordo com as questões de pesquisa.

3.5.1 Tipos de problemas urbanos abordados utilizando o processamento tweets (QP1)

Os tipos de problemas urbanos abordados utilizando o processamento de tweets foram divididos nas seguintes categorias:

1. **e-Participation** (Interação entre cidadãos e órgãos civis) (MUKHERJEE et al., 2015), (SOOMRO; KHAN; HASHAM, 2016).
 2. **Detecção de zoneamento urbano** (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014).
 3. **Identificação de pontos de interesse** (FARSEEV et al., 2015), (GUTEV; NENKO, 2016), (BENDLER et al., 2014), (ABBASI et al., 2015), (GKIOTSALITIS; STATHOPOULOS, 2015), (GKIOTSALITIS; STATHOPOULOS, 2016), (HASAN; UKKUSURI, 2014), (MAGHREBI et al., 2015), (DI LORENZO et al., 2013).
 4. **Mobilidade** (GUTEV; NENKO, 2016), (CHEN et al., 2016), (YOUSSAF et al., 2014).

5. **Padrões demográficos** (FARSEEV et al., 2015), (GUTEV; NENKO, 2016), (STEIGER et al., 2015), (GUO et al., 2016).
6. **Poluição** (ZAGAL; MATA; CLARAMUNT, 2016).
7. **Segurança Pública** (WEN; LIN; PELECHRINIS, 2016), (MATA; CLARAMUNT, 2015).
8. **Turismo** (THOMAZ et al., 2016), (ABBASI et al., 2015), (CHUA et al., 2016), (SOBOLEVSKY et al., 2015).
9. **Tráfego** (ANANTHARAM et al., 2015), (LECUE et al., 2014).

Conforme os estudos primários analisados pela Revisão Sistemática, e enumerados nessa seção, é possível interpretar que *tweets* podem ser utilizados para auxiliar na mitigação de inúmeros problemas urbanos. Apesar disso, (CHANIOTAKIS; ANTONIOU, 2015) observam que os *tweets* contendo informações sobre geolocalização são normalmente publicados em áreas relacionadas ao lazer, além de haver correlação entre regiões urbanas com maior renda *per capita* e o número de *tweets* postados. Tal evidência pode conduzir viés nas análises por representar somente algumas classes econômicas da população.

Considerando a observação anterior, um dos estudos analisados foi o realizado por (ZAGAL; MATA; CLARAMUNT, 2016), na Cidade do México. Nesse estudo, foram mapeados os pontos da cidade referenciados em publicações relacionadas a doenças respiratórias e poluição, orientando tomadas de decisão no aspecto ambiental.

Além disso, há também exemplos de trabalhos relacionados a Segurança Pública, como o estudo de caso realizado por (WEN; LIN; PELECHRINIS, 2016), no qual foi enriquecido um conjunto de dados com *tweets* geolocalizados, visando analisar o impacto dos ataques terroristas (em Paris, em novembro de 2015) nos padrões de atividades urbanas (relacionadas ao uso de transporte público, serviços, realização de compras, e atividade noturna). Em um outro caso de aplicação, estimou-se por meio de *tweets*, a probabilidade de ocorrência de crimes e ameaças nas ruas da Cidade do México, sugerindo rotas seguras aos pedestres (MATA; CLARAMUNT, 2015).

Também, foram encontrados na literatura estudos que utilizaram *tweets* para inferir padrões demográficos. Por exemplo, em (FARSEEV et al., 2015); (GKIOTSA-LITIS; STATHOPOULOS, 2015); (GKIOTSALITIS; STATHOPOULOS, 2016), *tweets*

foram processados para analisar a distribuição etária e de gênero da população, assim como seus respectivos pontos de interesse (HASAN; UKKUSURI, 2014) e (MAGHREBI et al., 2015) (como locais para entretenimento, residência, trabalho, recriação, compras, educação e serviços sociais).

Tais pontos de interesse, podem ser utilizados em problemas relacionados ao transporte público (GUTEV; NENKO, 2016) e também ao Turismo, como no estudo realizado por (ABBASI et al., 2015) para identificar a locomoção de visitantes e residentes em pontos turísticos de Sydney; por (CHUA et al., 2016), ao caracterizar aspectos espaciais, temporais e demográficos, dos turistas da cidade de Cilento, Itália; e por (THOMAZ et al., 2016) na cidade de Curitiba (Brasil), no contexto da Copa do Mundo de 2014.

Nesse mesmo contexto, (GUO et al., 2016) estudaram algumas questões demográficas via análise de sentimento, encontrando correlação positiva entre oportunidades de emprego e sentimentos positivos, e negativa entre felicidade e número de crianças na população da Grande Londres. Outro caso de uso, foi o desenvolvido em (STEIGER et al., 2015), no qual *tweets* foram processados para identificar diferentes tipos de atividades em Londres, correlacionando-as com informações censitárias; e em (SOBOLEVSKY et al., 2015) ao estudar a atratividade da Espanha a turistas.

Um dos problemas relacionados à identificação de pontos de interesse se refere as incertezas espaço-temporais e de determinação de tópicos, o qual foi abordado pelo trabalho realizado por (BENDLER et al., 2014). Nele, os autores contribuíram com uma técnica para minimizar o problema ao processar *tweets*; analisando a causalidade entre o tempo e local das postagens realizadas, reduzindo assim os índices de incerteza, no contexto da cidade de São Francisco, EUA. Outro problema, relaciona-se com a questão da privacidade, pois as localizações dos usuários podem ser inferidas mesmo quando não disponibilizadas. Nesse cenário, (WANG; SINNOTT; NEPAL, 2016) propõem um Sistema de Calibração de Trajetórias Privadas (PTCS), usando os mecanismos de Privacidade Diferencial e de *k-anonymity*, com isso é possível extrair informações sobre trajetórias sem exposição de informações sensíveis, testado na extração de localizações por meio de *tweets*.

Outro contexto na literatura revisada está relacionado ao processamento dos eventos que acontecem na cidade (idealmente em tempo real, como sugerem (SO-

OMRO; KHAN; HASHAM, 2016)). Um dos estudos encontrados sobre esse assunto, foi o realizado por (ANANTHARAM et al., 2015), no qual foi desenvolvida uma técnica para identificar os diferentes tipos de eventos do cotidiano urbano, rotulando-os sequencialmente, por meio da anotação de *tweets* e extração de eventos, considerando aspectos espaciais, temporais e temáticos. Para isso, utilizou conhecimentos de domínio, tais como informações sobre os locais em uma cidade e possíveis termos para os eventos, identificando assim os relacionados ao tráfego da região da Baía de São Francisco, EUA.

Sobre a mesma temática, (DI LORENZO et al., 2013) desenvolveram uma ferramenta inteligente e interativa para exploração visual da dinâmica de eventos sociais ao longo das dimensões espacial, temporal e organizacional. O tráfego também foi objeto de estudo em (CHEN et al., 2016), ao relacionar eventos do trânsito com a demanda por bicicletas; e em (LECUE et al., 2014), ao demonstrar uma plataforma para análise inteligente do tráfego (em tempo real), com base em fontes heterogêneas de dados (incluindo *tweets* de agências oficiais de trânsito).

Em uma abordagem mais genérica, (MUKHERJEE et al., 2015) propuseram uma plataforma para processar (em *near real time*) questões urgentes da cidade, oriundas de diversas fontes (incluindo o *Twitter*), atuando como intermediadora entre cidadãos e agências civis. No que se refere a mobilidade urbana, mas não utilizando informações sobre pontos de interesse, (YOUSAF et al., 2014) inferiram a afinidade entre usuários por meio da análise de *retweets*, possibilitando que rotas de corridas sejam compartilhadas entre pessoas com interesses em comum, tornando a viagem mais agradável.

De forma inusitada, (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014) utilizaram apenas *tweets* geolocalizados para analisar suas respectivas distribuições no espaço urbano, visando identificar a caracterização do uso da terra, considerando os zoneamentos urbanos industriais, residenciais, comerciais e de lazer. O trabalho foi realizado no contexto da cidade de Manhattan (EUA), Londres (Reino Unido) e Madrid (Espanha).

3.5.2 Casos de uso relacionados ao transporte público (QP2)

Nesta seção, são identificados os estudos primários que utilizam processamento de *tweets* tendo como foco a mitigação dos problemas relacionados ao transporte público; enumerados a seguir:

1. Impacto de eventos no transporte público.

- a) Impacto dos ataques terroristas em Paris no uso do transporte público (WEN; LIN; PELECHRINIS, 2016).
- b) Impacto de eventos relacionados ao tráfego na demanda por bicicletas, em Nova Iorque e Washington D.C, EUA (CHEN et al., 2016).
- c) Impacto dos pontos de interesse na demanda por transporte público (MAGHREBI et al., 2015).
- d) Impacto dos eventos anormais nas tomadas de decisão dos passageiros do Metrô de Tokyo (ITOH et al., 2016).
- e) Predição de fluxo de passageiros no Metrô de Nova Iorque (NI; HE; GAO, 2016).

2. Planejamento e gestão do transporte público.

- a) Análise de sentimento relacionada ao acesso ao transporte público (GUO et al., 2016).
- b) Coleta de informações relacionadas ao transporte público (GAL-TZUR et al., 2014).
- c) Identificação de locais para estações de bicicletas, em St. Petersburg, Rússia (GUTEV; NENKO, 2016).
- d) Identificação da disposição dos usuários para realizar viagens de lazer (GKIOTSALITIS; STATHOPOULOS, 2016).
- e) Plataforma para notificação de problemas relacionados ao transporte público de Bangalore, Índia (MUKHERJEE et al., 2015).

Conforme os estudos primários analisados pela Revisão Sistemática, e enumerados nessa seção, é possível interpretar que os estudos estão classificados em análise de impacto de eventos, planejamento e gestão do transporte público. Por exemplo, (WEN; LIN; PELECHRINIS, 2016) utilizaram *tweets* para analisar o im-

pacto dos ataques terroristas em Paris (2015) nos padrões de mobilidade referentes ao uso de transporte público. Semelhantemente, ITOH et al. (2016) desenvolveram uma ferramenta para analisar e explorar visualmente, com base em *tweets*, as tomadas de decisão dos passageiros do Metrô de Tokyo, ante a eventos anormais, tais como Tufões, Incêndios, Terremotos, dentre outros. Nesse mesmo contexto, (NI; HE; GAO, 2016) propuseram uma técnica de predição de fluxo de passageiros no Metrô de Nova Iorque, identificando eventos com base nas *hashtags* dos *tweets*. Enquanto que em (CHEN et al., 2016), analisaram a relação entre eventos do tráfego com a demanda por bicicletas.

No que se refere aos estudos focados no planejamento e gestão do transporte público, (MUKHERJEE et al., 2015) apresentam uma plataforma desenvolvida e utilizada pela Agência de Transporte Público de Bangalore, na Índia, a qual permite que usuários reportem questões relacionadas ao transporte público, possibilitando a melhoria do planejamento de suas respectivas operações, assim como o serviço prestado para a população. Nessa mesma linha de estudo, em (GUTEV; NENKO, 2016), *tweets* são utilizados para identificar a popularidade de determinados locais, pontos de interesse e distribuição etária, com o objetivo de determinar os melhores pontos para estações de bicicletas e incentivar assim o uso desse modal de transporte. Também relacionado aos pontos de interesse, (MAGHREBI et al., 2015) utilizaram *tweets* para identificar padrões das atividades humanas (em diferentes horários do dia) e seus respectivos impactos na demanda por transporte público.

Em (GAL-TZUR et al., 2014), por sua vez, utilizaram uma abordagem hierárquica para classificar *tweets* relacionados ao transporte. Com isso, demonstraram que é possível usar essas informações para fins de planejamento e gerenciamento do transporte. Tal técnica, foi aplicada em um estudo de caso associado a eventos esportivos, ocorridos no Reino Unido. A hierarquia é composta por três níveis, no primeiro, os *tweets* são classificados entre os que expressam a necessidade de serviços de transporte, opiniões e incidentes; o segundo, identifica a categoria do transporte; e último, relaciona *tweets* a tópicos.

Outro estudo que contribui com o planejamento do transporte público, é o realizado em (GKIOTSALITIS; STATHOPOULOS, 2015, 2016), no qual *tweets* foram processados para identificar a disposição dos usuários para realizar viagens relacionadas ao lazer (pontos de interesse), sugerindo a eles atividades com menor

tempo de percurso e probabilidade de atrasos. Além do tempo de percurso, outro ponto relevante considerado foi o de bom nível de acesso ao transporte público, o qual quando existente impacta positivamente na felicidade das pessoas e se correlaciona com sentimentos positivos, segundo a análise de sentimentos realizada por (GUO et al., 2016), utilizando *tweets* publicados na Grande Londres.

3.5.3 Técnicas estatísticas utilizadas no processamento de *tweets* (QP3)

Nesta seção, são apresentadas as técnicas estatísticas utilizadas pelos estudos primários, no processamento de *tweets*, enumeradas a seguir:

1. **Análise de métricas relacionadas a desempenho** (erro de reconstrução relativo, qualidade dos componentes descritivos recuperados e qualidade dos componentes comuns recuperados) (WEN; LIN; PELECHRINIS, 2016).
2. **Cosine similarity** (YOUSAF et al., 2014), (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014).
3. **F_1 score** (ANANTHARAM et al., 2015), (CHEN et al., 2016).
4. **Term frequency-inverse document frequency** (TF-IDF) (MUKHERJEE et al., 2015).
5. **Inverse coefficient of variation** (BENDLER et al., 2014).
6. **Jackknife resampling** (BENDLER et al., 2014).
7. **Local Indicators of Spatial Association** (LISA) (STEIGER et al., 2015).
8. **Local Moran's** (STEIGER et al., 2015).
9. **Maximum likelihood estimation** (MUKHERJEE et al., 2015).
10. **Seasonal Autoregressive Integrated Moving Average** (SARIMA) (NI; HE; GAO, 2016).
11. **Optimization and Prediction with hybrid loss function** (NI; HE; GAO, 2016).

Em (NI; HE; GAO, 2016), os autores utilizaram a técnica *Seasonal Autoregressive Integrated Moving Average* em conjunto com Regressão Linear, propondo uma abordagem baseada em otimização paramétrica e convexa, chamada *Optimization and Prediction with hybrid loss function*, adequada para modelagem utilizando séries

temporais. Com isso, tal técnica foi aplicada na predição de fluxo de passageiros com base em *hashtags* de *tweets*.

Referente aos problemas relacionados a ambiguidade e identificação de contextos, (ANANTHARAM et al., 2015); (CHEN et al., 2016) e (GAL-TZUR et al., 2014) aplicaram a técnica *F₁ score* para analisar a acurácia do processamento de *tweets*. Por outro lado, em (MUKHERJEE et al., 2015), utilizaram a técnica *Maximum likelihood estimation* para determinar a probabilidade de ocorrência de um evento, assim como a confiabilidade da informação.

No que se refere a agrupamento, (YOUSAFA et al., 2014) agruparam usuários de acordo com a *Cosine similarity*, unindo pessoas com interesses em comum nos mesmos grupos. (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014), por outro lado, usou a mesma técnica para agrupar *tweets* de acordo com suas semelhanças quanto aos tipos de zoneamento urbano.

De forma isolada, no trabalho realizado por (MUKHERJEE et al., 2015), utilizaram a técnica TF-IDF na fase de classificação para o definir o *score* de categorias de eventos, escolhendo a mais relevante a ser buscada em um dicionário de categorias. Também isoladamente, (STEIGER et al., 2015) usaram a técnica LISA na identificação de *clusters* espaciais e valores esporádicos espaciais, obtendo assim os locais com atividades sociais. Além disso, também utilizaram a técnica *Local Moran's* para detectar diferentes padrões de atividade de acordo com o espaço geográfico.

Por último, (BENDLER et al., 2014) inovaram ao utilizar a técnica *Jackknife resampling* como inspiração para o desenvolvimento de uma abordagem que visa analisar a estabilidade estatística de um conjunto de categorias. Além disso, usaram também a análise *Inverse Coefficient of variation* para verificar a dispersão negativa da distribuição de um conjunto de variáveis.

3.5.4 Paradigmas de processamento (QP4)

Nesta seção, encontram-se a seguir apenas os paradigmas de processamento extraídos dos estudos primários analisados:

1. ***Batch processing*** (offline) (ANANTHARAM et al., 2015), (WEN; LIN; PELE-CHRINIS, 2016), (FARSEEV et al., 2015), (GUTEV; NENKO, 2016), (MATA;

CLARAMUNT, 2015), (CHEN et al., 2016), (ABBASI et al., 2015), (BENDLER et al., 2014), (BENDLER et al., 2014), (YOUSAF et al., 2014), (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014), (STEIGER et al., 2015), (GAL-TZUR et al., 2014), (GKIOTSALITIS; STATHOPOULOS, 2016), (DI LORENZO et al., 2013), (ITOH et al., 2016), (CHANIOTAKIS; ANTONIOU, 2015).

2. **Near Real Time** (MUKHERJEE et al., 2015).
3. **Real Time** (SOOMRO; KHAN; HASHAM, 2016), (LECUE et al., 2014).

3.5.5 Eventos de exceção relacionados ao transporte público (QP5)

Nesta seção, encontram-se a seguir os eventos de exceção relacionados ao transporte público, extraídos dos estudos primários:

1. **Acidentes.**
 - a) Acidentes nas estações transporte (ITOH et al., 2016).
 - b) Incêndio (ITOH et al., 2016).
2. **Espaço-temporais.**
 - a) Dia da semana (CHEN et al., 2016).
 - b) Hora do dia (CHEN et al., 2016).
3. **Eventos sociais.**
 - a) Feiras de rua (CHEN et al., 2016).
 - b) Festivais (CHEN et al., 2016), (LECUE et al., 2014).
 - c) Jogos esportivos (CHEN et al., 2016), (GAL-TZUR et al., 2014).
 - d) Passeatas e maratonas (CHEN et al., 2016), (ITOH et al., 2016).
4. **Eventos urbanos.**
 - a) Relacionados ao tráfego (CHEN et al., 2016), (LECUE et al., 2014).
5. **Desastres naturais.**
 - a) Tempestades (ITOH et al., 2016).
 - b) Terremoto (ITOH et al., 2016).
 - c) Tufões (ITOH et al., 2016).
6. **Metereológicos.**

- a) Dia claro, nublado, chuvoso, nevando, com neblina (CHEN et al., 2016).
- b) Temperatura do ar (CHEN et al., 2016).

3.5.6 Técnicas de Aprendizado de Máquina utilizadas no processamento de tweets (QP6)

Nesta seção, são apresentadas as técnicas de Aprendizado de Máquina utilizadas para processamento de *tweets*, extraídas dos estudos primários e enumeradas a seguir:

1. ***Bayes classification*** (MATA; CLARAMUNT, 2015).
2. ***C5.0 algorithm*** (ZAGAL; MATA; CLARAMUNT, 2016).
3. ***Conditional Random Field (CRF) with Logistic Regression*** (ANANTHARAM et al., 2015).
4. ***Latent Dirichlet Allocation (LDA)*** (FARSEEV et al., 2015), (ABBASI et al., 2015), (HASAN; UKKUSURI, 2014), (DI LORENZO et al., 2013), (NI; HE; GAO, 2016) .
5. ***Linear Regression*** (GUTEV; NENKO, 2016), (BENDLER et al., 2014), (NI; HE; GAO, 2016), (GUO et al., 2016).
6. ***Monte Carlo simulation*** (CHEN et al., 2016).
7. ***PairFac*** (técnica inovadora que utiliza *Tensor Factorization*) (WEN; LIN; PELECHRINIS, 2016).
8. ***Random Forest classification*** (FARSEEV et al., 2015).
9. ***Support Vector Machine*** (MUKHERJEE et al., 2015), (GAL-TZUR et al., 2014).
10. ***Self-Organizing Maps*** (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014).

No contexto urbano, inúmeros eventos podem acontecer e impactar a população. O trabalho realizado por (WEN; LIN; PELECHRINIS, 2016), desenvolveu uma técnica que utiliza a análise de tensores discriminantes para aprender e de forma automatizada descobrir os impactos de um determinado evento no cotidiano da cidade. Numa abordagem mais simples, (CHEN et al., 2016) utilizou *Monte Carlo simulation* para treinar um modelo para predição de demanda por bicicletas, devido

a dificuldade de encontrar exemplos suficientes para usar outras abordagens de treinamento.

Especificamente sobre as técnicas de classificação, (MUKHERJEE et al., 2015) utilizaram *Support Vector Machine* para classificar os eventos recebidos de diversas fontes. Referente a essa abordagem, (GAL-TZUR et al., 2014) analisaram inúmeras técnicas de Aprendizado de Máquina, obtendo a melhor performance com o SVM, além disso, observaram como principal vantagem a sua capacidade de adaptação ao gênero e tarefas subjacentes.

Apesar disso, (GUO et al., 2016) utilizaram Processamento de Linguagem Natural (baseado em palavras chaves) para rotular sentimentos de *tweets*, devido a facilidade de escalar essa técnica (para processamento de milhões de *tweets*), em comparação a SVM. Outro caso de divergência é o do estudo realizado por (FARSEEV et al., 2015), no qual foi escolhido o modelo de classificação *Random Forest*, devido ao fato de ser mais adequado para classificação em espaço dimensional elevado, em vez das técnicas SVM e *Naive Bayes*, no que se refere a predição de idade e gênero usando *tweets*.

MATA; CLARAMUNT (2015), por sua vez, aplicou a técnica *Bayes Classification* em *tweets*, visando obter probabilidades relacionadas a crimes e ameaças em uma determinada localização. Por outro lado, (ZAGAL; MATA; CLARAMUNT, 2016) usaram o *C5.0 algorithm* devido ao melhor desempenho em relação a *Bayes*, dependendo do tópico que está sendo classificado.

Para anotação de eventos, (ANANTHARAM et al., 2015) treinaram um modelo CRF (usando anotações baseadas em dicionários) para determinar os locais da cidade e os termos relacionados aos eventos expressos em *tweets*. E, isoladamente (FRIAS-MARTINEZ; FRIAS-MARTINEZ, 2014) utilizaram a técnica *Self-Organizing Maps*, tendo como entrada os valores de latitude e longitude de *tweets*. Com isso, construíram um mapa segmentado em áreas urbanas, baseando-se nas regiões com diferentes concentrações de *tweets*.

Em relação a localidades, segundo (FARSEEV et al., 2015), a técnica LDA tem sido muito utilizada para identificação de pontos de interesses mencionados em *tweets*, sendo adequada para grandes bases de dados e agrupamento de *tweets* com tópicos similares, de acordo com (STEIGER et al., 2015). (ABBASI et al., 2015) exemplificou isso ao aplicar LDA para identificação de *tweets* relacionados ao

Turismo; (HASAN; UKKUSURI, 2014), para identificação de padrões de atividades humanas; e (DI LORENZO et al., 2013), para identificação de eventos sociais.

No entanto, (NI; HE; GAO, 2016) em vez de usarem LDA, extraíram hashtags de tweets para um vetor, utilizando-o para medir as atividades sociais e identificar seus respectivos contextos. Segundo (NI; HE; GAO, 2016), isso se justifica devido ao fato de que há uma grande chance do alto volume de tweets não indicar necessariamente eventos e atendimentos a eles. Além disso, afirmam que o método baseado em hashtag é capaz de indicar sobre o que é o evento, mesmo não utilizando o Inglês formal.

Por sua vez, em (GUTEV; NENKO, 2016), os autores utilizaram Regressão Linear (RL) para analisar a demanda por bicicletas de acordo com as localizações extraídas dos tweets. Enquanto que (BENDLER et al., 2014) usaram RL para fornecer evidências de que as categorias dos pontos de interesse se relacionam com as variáveis referentes ao espectro espaço-temporal; e (GUO et al., 2016) para analisar a correlação entre sentimentos positivos com as oportunidades de trabalho, com a quantidade de crianças, e com o acesso a transporte.

3.6 Considerações finais sobre a revisão sistemática

Em uma análise quantitativa dos estudos primários selecionados, podemos concluir que a quantidade de artigos publicados sobre o uso de tweets na caracterização de problemas urbanos e relacionados ao transporte público tem crescido consideravelmente, entre 2011 e 2016. Provavelmente, devido ao fato da popularização das Redes Sociais e grande quantidade de dados disponíveis para processamento.

Tais estudos estão concentrados em maioria na identificação de pontos de interesse, utilizando-os em diferentes contextos, tais como o de turismo, mobilidade. Além disso, abordam também problemas relacionados ao transporte e desastres naturais, confirmando a primeira hipótese (HP1) dessa Revisão Sistemática. As temáticas não abordadas pela HP1 foram as relacionadas a e-Participation, detecção de zoneamento urbano, padrões demográficos e segurança pública, demonstrando a variedade de problemas urbanos explorados com o processamento de tweets.

Referente a segunda hipótese, os estudos exploraram principalmente o impacto de eventos no transporte público, confirmando-a parcialmente. Isso, devido ao fato de um dos trabalhos explorar como os eventos relacionados ao tráfego impactam na demanda por bicicletas; não havendo nenhum outro sobre processamento de *tweets* para mitigação dos problemas envolvendo Tráfego. Outra temática não mencionada pela HP2 e sobre a qual há uma quantidade considerável de estudos, foi a do uso de *tweets* para o planejamento e gerenciamento do transporte público.

Independentemente dos problemas abordados por meio do processamento de *tweets*, dentre as 12 técnicas estatísticas elencadas, F_1 score foi a única referenciada como ferramenta para análise de acurácia de classificação binária, confirmando a terceira hipótese (HP3). Apesar disso, a HP3 não considerou outras técnicas importantes (com propósitos distintos), como a *Linear Regression*, amplamente utilizada nos estudos analisados. Referente as técnicas de Aprendizado de Máquina, a mais utilizada foi a *Latent Dirichlet Allocation* (LDA), seguida da *Support Vector Machine* (SVM), confirmando parcialmente a sexta hipótese (HP6).

Por fim, apenas quatro dos vinte e nove estudos analisados, cerca de 14%, mencionaram *features* relacionadas ao transporte público, confirmando assim a quinta hipótese (HP5). Assim como a quantidade de trabalhos que realizam processamento de *tweets* em tempo real, sendo apenas dois do total analisado, cerca de 6%, que utilizam esse paradigma de processamento, o que confirma a quarta hipótese (HP4). É importante ainda observar que, outros estudos que mencionaram processamento em tempo real, realizaram na verdade coleta de *tweets* em tempo real, para análises a posteriori via processamento em *batch* (offline), categoria na qual a maioria dos estudos foram enquadrados.

4 Construção do conjunto de dados

Nesta seção, são apresentados os conjuntos de dados referentes a proposta de pesquisa.

4.0.1 Corpus Twitter

A Rede Social *Twitter*, foi escolhida como fonte de dados para a construção do conjunto de dados relacionados aos eventos de exceção. Isso devido ao fato de cada publicação ser limitada em 280 caracteres, o que reduz a complexidade de processamento do conteúdo publicado, e devido aos estudos existentes abordando problemas urbanos e de mobilidade urbana, conforme os analisados na revisão sistemática do Cap. 3.

Assim, o conjunto de dados utilizado para a identificação dos eventos de exceção é composto por *tweets*, em português brasileiro, dos *profiles* contidos na tabela 1. É importante observar que, para esse projeto de pesquisa, apenas os *tweets* publicados pelas contas selecionadas serão considerados, descartando os relacionados às interações entre diferentes *profiles* (*retweets* e *replies*). Ou seja, os dados utilizados estão relacionados ao canal unidirecional de comunicação (no contexto de *e-participation*). Com essa restrição, evitamos problemas referentes a confiabilidade dos dados, o que nos permite focarmos na caracterização dos eventos de exceção e de seus respectivos impactos.

Sobre a seleção dos *profiles*, todos foram selecionados manualmente de acordo com os órgãos responsáveis por notificar eventos de exceção. Tais *profiles* são de caráter público, ou seja, o acesso aos *tweets* não envolve questões de privacidade. Apesar do acesso facilitado aos *tweets*, a API do *Twitter* limita a quantidade e frequência de requisições aos *endpoints*. Por exemplo, no protótipo desenvolvido (na linguagem de programação Java), há um artefato que coleta (utilizando o *plugin Twitter4J*¹) os 3.200 *tweets* mais recentes (se disponíveis) de cada conta, através do *endpoint statuses/user_timeline*; o qual permite no máximo 180 requisições, em um intervalo de 15 minutos, com autenticação via conta de usuário².

¹ <twitter4j.org>. Acesso em Outubro, 29 de 2017.

² <<https://dev.twitter.com>>. Acesso em Outubro, 29 de 2017.

Durante a coleta dos *tweets*, eles são mapeados para a seguinte classe do modelo da aplicação: *TweetInfo*, que contém as informações respectivas ao *id*, texto da publicação, *timestamp*, endereço extraído, latitude e longitude. Em seguida, o modelo é persistido no banco de dados não relacional *MongoDB*³ e também no banco de dados de séries temporais *Druid*⁴ para exploração e visualização dos dados, processo explicado na seção 5. Os detalhes sobre o intervalo de tempo e o número de *tweets* coletados constam na tabela 4.

Tabela 4 – Intervalo de tempo e número de *tweets* coletados

Profile no Twitter	# tweets ^a	Timestamp 1 ^b	Timestamp 2 ^c
BombeirosPMESP	5.750	2017-05-21 02:10:39	2017-10-29 23:07:08
CETSP_	5.042	2017-02-20 14:07:04	2017-10-29 21:45:54
CPTM_oficial	5.435	2017-04-24 13:00:17	2017-10-29 10:00:40
governosp	5.450	2017-05-10 17:00:05	2017-10-29 22:00:03
metrosp_oficial	7.296	2017-06-07 17:23:34	2017-10-29 17:48:12
Policia_Civil	3.360	2015-04-15 17:44:44	2017-10-27 10:01:53
PMESP	3.956	2016-06-02 17:21:32	2017-10-29 20:25:37
saopaulo_agora	3.671	2016-11-18 07:36:12	2017-10-29 20:56:28
smtsp_	1.128	2017-04-26 10:44:26	2017-10-29 23:00:11
SPCEDEC	945	2015-06-09 10:50:23	2017-10-29 23:38:36
sptrans_	8.447	2017-06-13 15:19:56	2017-10-29 22:01:44
TurismoSaoPaulo	3.308	2012-06-12 22:00:38	2017-10-27 17:46:59
Total	53.788	-	-

^a Número de *tweets* coletados.

^b *Timestamp* mais antigo.

^c *Timestamp* mais recente.

Fonte: Felipe Cordeiro Alves Dias

Além dos *tweets* coletados, foram extraídos 625 endereços e seus respectivos dados de geolocalização. No entanto, por meio de uma análise manual percebemos dois problemas: (I) alguns endereços não foram extraídos; (II) apesar de o endereço ser extraído corretamente, encontramos geolocalizações fora do estado de São Paulo e do país. Assim, pretendemos melhorar o processo de extração dos endereços dos *tweets* e o restringir a geolocalização para a região de São Paulo.

³ <<https://www.mongodb.com>>. Acesso em Outubro, 29 de 2017.

⁴ <<http://druid.io>>. Acesso em Outubro, 29 de 2017.

4.0.2 Corpus SPTrans

O corpus SPTrans é composto por dados obtidos do SIM, transferidos via AVL, e por dados fornecidos pela SPTrans especificados em GTFS, detalhados na tabela 5. Os dados de ambas as fontes não são triviais de serem processados (grande volume de dados, dados sem tipo explicitamente definido — não tratados, dados separados em lotes de dados — um arquivo para cada hora de movimentação dos ônibus, dados fora do formato convencional — por exemplo, 24h em vez de 0h), devido a isso foram desenvolvidos *scripts* para um processo de ETL (*Extract, Tranform and Load*).

No caso dos dados especificados em GTFS, convertemos os dados originais de *string* para os seus respectivos tipos (*long*, *double*, *int* ou *string*) e padronizamos os valores referentes a hora para *POSIX timestamp*, e os referentes a latitude e longitude para *legacy coordinate pairs*⁵. Além disso, visando viabilizar *geospatial queries*, foram criados *geospatial indexes*⁵ nas *collections* contendo informações geolocalizadas, logo após serem criadas no *MongoDB*. Dessa forma, conseguimos usar *geospatial queries* para identificar as linhas afetadas por um determinado evento de exceção.

Tabela 5 – Conjuntos e quantidades de dados especificados em GTFS pela SPTrans

Conjunto de dados	Quantidade de dados
<i>agency.txt</i>	1
<i>calendar.txt</i>	6
<i>fare_attributes.txt</i>	6
<i>fare_rules.txt</i>	5.400
<i>frequencies.txt</i>	39.625
<i>routes.txt</i>	291.634
<i>shapes.txt</i>	800.767
<i>stop_times.txt</i>	95.134
<i>stops.txt</i>	19.933
<i>trips.txt</i>	2.273
Total	1.254.779

Fonte: Felipe Cordeiro Alves Dias

Os dados AVL utilizados nesta analise são referentes aos movimentos de ônibus ocorridos entre janeiro e dezembro de 2017 (intervalo antes da data em que

⁵ <<https://docs.mongodb.com/manual/geospatial-queries>>. Acesso em Outubro, 29 de 2017.

os dados foram solicitados, por meio da *Lei de Acesso a Informação*⁶). De acordo com a tabela 7, alguns dos dados de movimentação referentes a novembro e dezembro estão ausentes, segundo a SPTrans, essa ausência é justificada devido a períodos de indisponibilidade do sistema de monitoramento.

Os períodos indisponíveis foram identificados por meio de um *script*⁷ desenvolvido por este trabalho, para análises descritivas de grandes volumes de dados AVL, tais como: total de arquivos e espaço em disco, por período. O funcionamento do *script* consiste em gerar os respectivos nomes dos arquivos de movimentação que deveriam existir em determinado período, confrontando-os com os existentes na base obtida, além de sumarizar o espaço em disco e total de arquivos; tais metadados estão especificados na tabela 7.

Figura 6 – Evidência dos períodos de indisponibilidade de dados AVL referentes a Dezembro de 2017

Inicio	15/12/2017 01:30	Fim	15/12/2017 12:00
Causa	Manutenção Preventiva da Claro	Ação	Aberto Chamado
Motivo	Indisponibilidade de Sinal GPRS	Responsável	Claro
Observação	<p>B I $\wedge h2$ $\wedge h3$ $\wedge h4$ \equiv \equiv \equiv Δ \leftarrow</p> <p>Data/Hora de Início da Ocorrência: 15/12/2017 06:00 Data/Hora da previsão de restabelecimento: 15/12/2017 12:00 Região afetada: Clientes com saída pelo túnel SP poderão encontrar dificuldades de navegação na rede de dados.</p>		
<input type="button" value="Gravar"/> <input type="button" value="Excluir"/> <input type="button" value="Cancelar"/>			

Fonte: Resposta ao pedido de acesso a informação referente ao protocolo e-SIC 33310, 2017

⁶ <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acessado em 23 de junho de 2018.

⁷ <https://github.com/fcas/mobility-analysis/blob/master/scripts/data_set_analyser.py>. Acessado em setembro de 2018.

Tabela 6 – Descrição do conjunto de dados AVL

Mês	Intervalo (dias)	Total de arquivos AVL	Espaço em disco (GB)
Janeiro	1 - 31	744	102,44
Fevereiro	1 - 28	672	93,21
Março	1 - 31	744	102,64
Abril	1 - 30	720	97,04
Maio	1 - 31	744	101,46
Junho	1 - 30	720	97,13
Julho	1 - 31	744	104,95
Agosto	1 - 31	744	108,38
Setembro	1 - 30	720	109,89
Outubro	1 - 31	744	110,92
Novembro ^a	1 - 30	717	108,16
Dezembro ^b	1 - 31	738	110,89
—	—	8,751	1,247.09

^a Arquivos indisponíveis em novembro:

- Movto_201711011200_201711011300.zip
- Movto_201711011300_201711011400.zip
- Movto_201711011400_201711011500.zip

^b Arquivos indisponíveis em dezembro, devido a falha na rede de transmissão de dados conforme apresentado no sistema interno de registro de interrupções do sistema, Fig. 6 — resposta oficial da SPTrans, responsável: Albino Silva da Rocha, Chefe de Gabinete da SPTrans:

- Movto_201712150100_201712150200.zip
- Movto_201712150400_201712150500.zip
- Movto_201712150500_201712150600.zip
- Movto_201712150600_201712150700.zip
- Movto_201712150700_201712150800.zip
- Movto_201712150800_201712150900.zip

Tabela 7 – Meta dados dos dados AVL da SPTrans

Nome do campo	Descrição do campo
<i>cd_evento_avl_movto</i>	Código sequencial identificador do evento
<i>cd_linha</i>	Código identificador da linha em operação
<i>dt_movto</i>	Data da gravação em banco de dados do evento gerado no AVL
<i>nr_identificador</i>	Código identificador do AVL
<i>nr_evento_linha</i>	Grupo de indicadores relacionados ao evento
<i>nr_ponto</i>	Código do ponto notável
<i>nr_velocidade</i>	Velocidade instantânea
<i>nr_voltagem</i>	Tensão de alimentação
<i>nr_temperatura_interna</i>	Temperatura do processador
<i>nr_evento_terminal_dado</i>	Código do evento relacionado no terminal de dados
<i>nr_evento_es_1</i>	Grupo de indicadores relacionados ao evento
<i>nr_latitude_grau</i>	Latitude da geolocalização do veículo
<i>nr_longitude_grau</i>	Longitude da geolocalização do veículo
<i>nr_indiceregistro</i>	Índice de geração do evento no AVL
<i>dt_avl</i>	Data da geração do evento no AVL
<i>nr_distancia</i>	Distância em metros do evento com relação ao evento anterior do mesmo AVL
<i>nr_tipo_veiculo_geo</i>	Código para identificação no software de mapeamento
<i>cd_avl_conexao</i>	Código interno utilizado para identificar qual a conexão utilizada para transmissão do evento
<i>cd_prefixo</i>	Prefixo do veículo

5 Exploração e visualização de grandes volumes de dados

Este capítulo tem como objetivo apresentar uma arquitetura para visualizar e explorar grandes volumes de dados, a validação da arquitetura proposta é realizada com o Corpus AVL da SPTrans. Isso, porque além do grande volume esse conjunto de dados possui padrões complexos e demanda um sistema distribuído para serem processados, o qual é apresentado neste trabalho, capaz de suportar atividades analíticas, como visualização e exploração de dados. Tais análises, são importantes para o gerenciamento e planejamento do transporte público. Na seção 5.1 são mencionados alguns trabalhos referentes a visualização de dados, encontrados por meio de uma revisão não sistemática da literatura; na 5.2 é descrita a arquitetura do banco de dados *Druid*, principal componente da arquitetura proposta; na 5.3 a arquitetura em questão para processamento e exploração dos dados AVL; na 5.4 os resultados obtidos no estudo de caso e, por fim, na 5.5 as considerações finais.

5.1 Trabalhos relacionados

Em (CHEN; GUO; WANG, 2015) são mencionados conceitos básicos e fluxos de visualização de dados de tráfego (dos dados brutos, pré-processamento ao mapeamento visual, construído com símbolos visuais), além de uma visão geral das técnicas e métodos de processamento de dados relacionados para processar e descrever propriedades temporais, espaciais, numéricas e categóricas de dados de tráfego.

Analogamente, em (ANDRIENKO et al., 2017) é descrita uma tipologia de dados de tráfego, capaz de abordar suas respectivas propriedades, problemas e transformações relevantes para a análise. Além disso, são apresentadas abordagens analíticas visuais para analisar dados de tráfego de veículos, pedestres, passageiros dentro de sistemas de transporte, etc.

Por fim, no trabalho desenvolvido em (SERAJ; MERATNIA; HAVINGA, 2017) é apresentado um novo algoritmo para mapeamento de medições coletivas para monitorar as infraestruturas de transporte terrestre e, aliviar o impacto de imprecisões do GPS para monitoramento contínuo de infraestruturas de transporte por meio de *smart phones*.

Nenhum dos trabalhos mencionados anteriormente aborda o uso de software livre com suporte a computação distribuída, escalabilidade, tolerância a falhas, processamento em tempo real, baixa latência e visualização de grandes volumes de dados temporais; o que é explorado neste trabalho usando banco de dados *Druid* e o *Apache Superset* para analisar padrões complexos existentes nos dados AVL da SPTTrans.

5.2 *Druid*

O *Druid* é um banco de dados para análises exploratórias em tempo real (latências abaixo da sub-segundos) em grandes conjuntos de dados. A arquitetura distribuída do *Druid* é composta por um *cluster* com diferentes tipos de nós, que operam independentemente uns dos outros e possuem interação mínima entre eles. Existem duas dependências externas: (I) Apache Zookeeper¹, responsável pela coordenação do cluster e (II) um sistema de gerenciamento de banco de dados relacional (RDBMS — *Relational Database Management Systems*), para armazenar parâmetros operacionais adicionais e configurações (YANG et al., 2014).

5.2.1 Real-time nodes

Real-time nodes são responsáveis por ingerir, indexar e consultar fluxos de eventos. Periodicamente, cada nó agenda uma tarefa em segundo plano para procurar todos os índices localmente persistentes, mesclando-os para construir blocos imutáveis de dados com todos os eventos ingeridos em um período de tempo, conhecidos como *segmentos*, os quais podem posteriormente serem carregados para uma camada de *deep storage* (YANG et al., 2014).

Durante os processos mencionados anteriormente não há perda de dados. Além disso, a imutabilidade dos blocos permite a consistência de leitura e um modelo de paralelização simples: *historical nodes* podem simultaneamente examinar e agregar blocos imutáveis de forma não bloqueante (YANG et al., 2014).

¹ <<https://zookeeper.apache.org>>. Acessado em 23 de junho de 2018.

5.2.2 Historical nodes

Os *historical nodes* são responsáveis por carregar, descartar e servir *segmentos* imutáveis por meio de uma arquitetura *shared-nothing* (sem um único ponto de contenção entre os nós) (YANG et al., 2014).

5.2.3 Broker nodes

Os *broker nodes* são responsáveis por receber consultas e mesclar resultados parciais dos *historicals* e *real-time nodes* antes de retornar um resultado final consolidado para o cliente (YANG et al., 2014).

5.2.4 Coordinator nodes

Os *coordinator nodes* são responsáveis pelo gerenciamento e distribuição dos dados nos *historical nodes*, exigindo destes o carregamento, descarte e replicação dos dados (YANG et al., 2014).

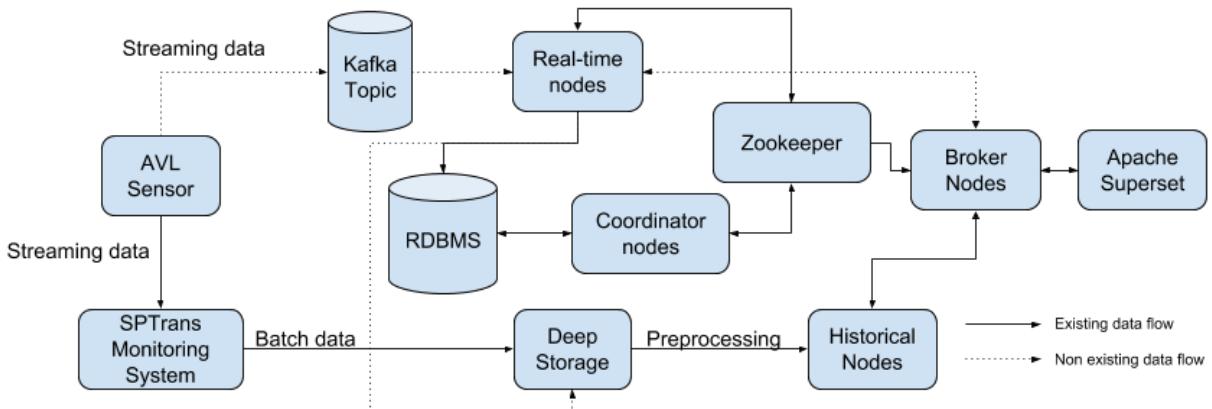
5.3 Arquitetura utilizada para visualização e exploração dos dados AVL da SPTrans

A Fig. 7 mostra a arquitetura utilizada no estudo de caso deste capítulo, composta pelos componentes do *Druid* em conjunto com o módulo *Apache Superset*² — software de código aberto para exploração e análise de dados, nativamente integrado ao *Druid*. Nesta arquitetura, dois fluxos para processamento de dados também são elencados: (I) *batch* e (II) *near real time*.

O fluxo de processamento em *batch* é executado a partir dos dados extraídos do sistema de monitoramento da *SPTrans*, ingeridos nos *historical nodes* (a latência de ingestão máxima medida é 22914.43 eventos / segundo / núcleo, com uma fonte de dados com 30 dimensões e 19 métricas (YANG et al., 2014)) e disponibilizados para o *Apache Superset* por meio dos *broker nodes* (que tem uma latência média de consulta de aproximadamente 550 milissegundos (YANG et al., 2014)). É importante

² <<https://superset.incubator.apache.org>>. Acessado em 23 de junho de 2018.

Figura 7 – Arquitetura usada no estudo de caso para visualização e exploração dos dados AVL da SPTrans



observar que o fluxo de processamento em *batch* é o fluxo de dados implementado neste estudo de caso.

Na arquitetura ilustrada na Fig.7, o fluxo de dados em *streaming* refere-se a uma proposta alvo para a *SPTrans*, a fim de permitir a exploração e visualização dos dados dos ônibus da cidade de São Paulo em *near real time*. Nesta proposta, os tópicos do *Apache Kafka* desempenham o papel de receptores do fluxo de dados, a partir dos quais os dados podem seguir tanto o processamento em *near real time* quanto *batch*.

Ambos os fluxos de dados mencionados anteriormente são válidos, o fluxo *streaming* não exclui a necessidade de um fluxo em *batch*, o qual pode ser usado para análises mais complexas dos dados em questão. Além disso, é importante observar que em ambos os fluxos há um estágio de pré-processamento de dados, para adequar os dados AVL as especificações exigidas para a ingestão no *Druid* (o que adiciona atraso no fluxo de processamento).

5.4 Estudo de caso com os dados AVL da SPTrans

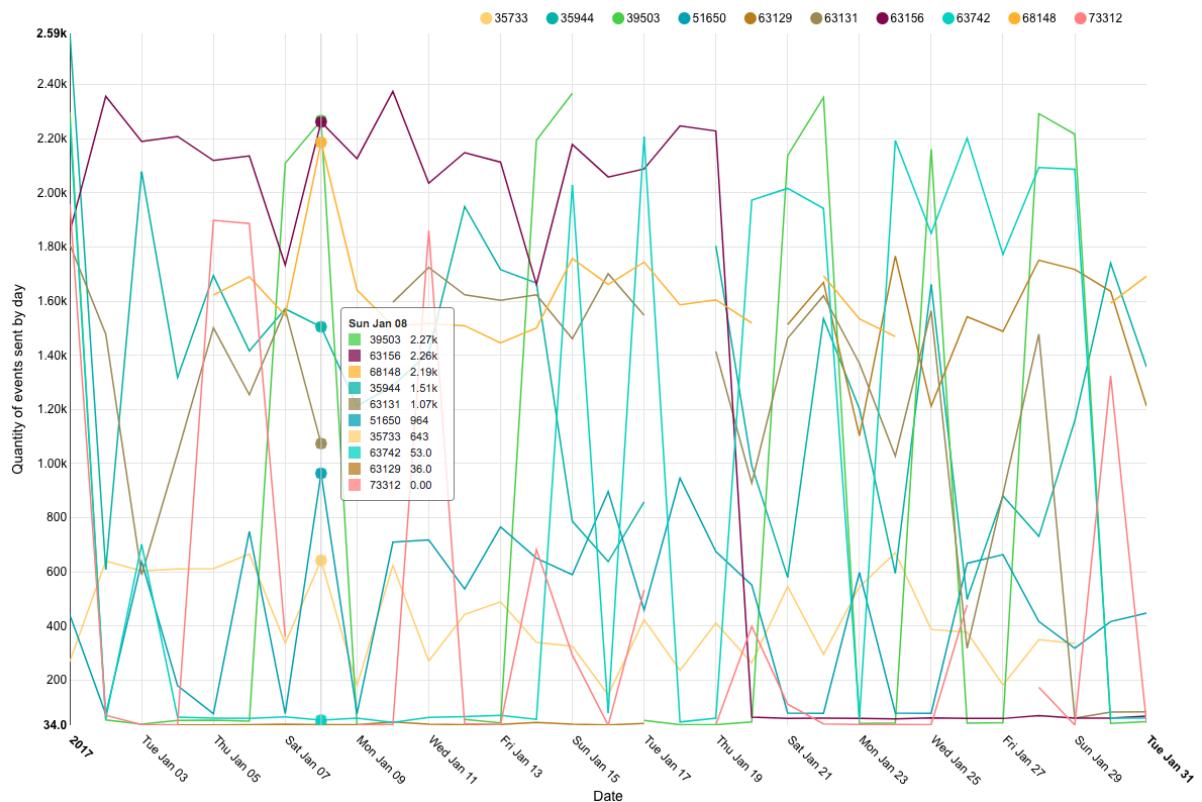
Grandes volumes de dados podem conter padrões complexos e difíceis de serem identificados. Devido a isso, é importante construir visualizações auxiliares para o processo de análise de dados. Com este propósito, usamos o *Apache Superset*³, com suporte nativo ao *Druid*, para exploração e visualização do *corpus* da SPTrans.

³ <<https://superset.incubator.apache.org>>. Acessado em 29 de junho de 2018

As figuras 8, 9, 10 e 11 são exemplos de algumas visualizações construídas a partir dos dados de janeiro das linhas de ônibus selecionadas aleatoriamente.

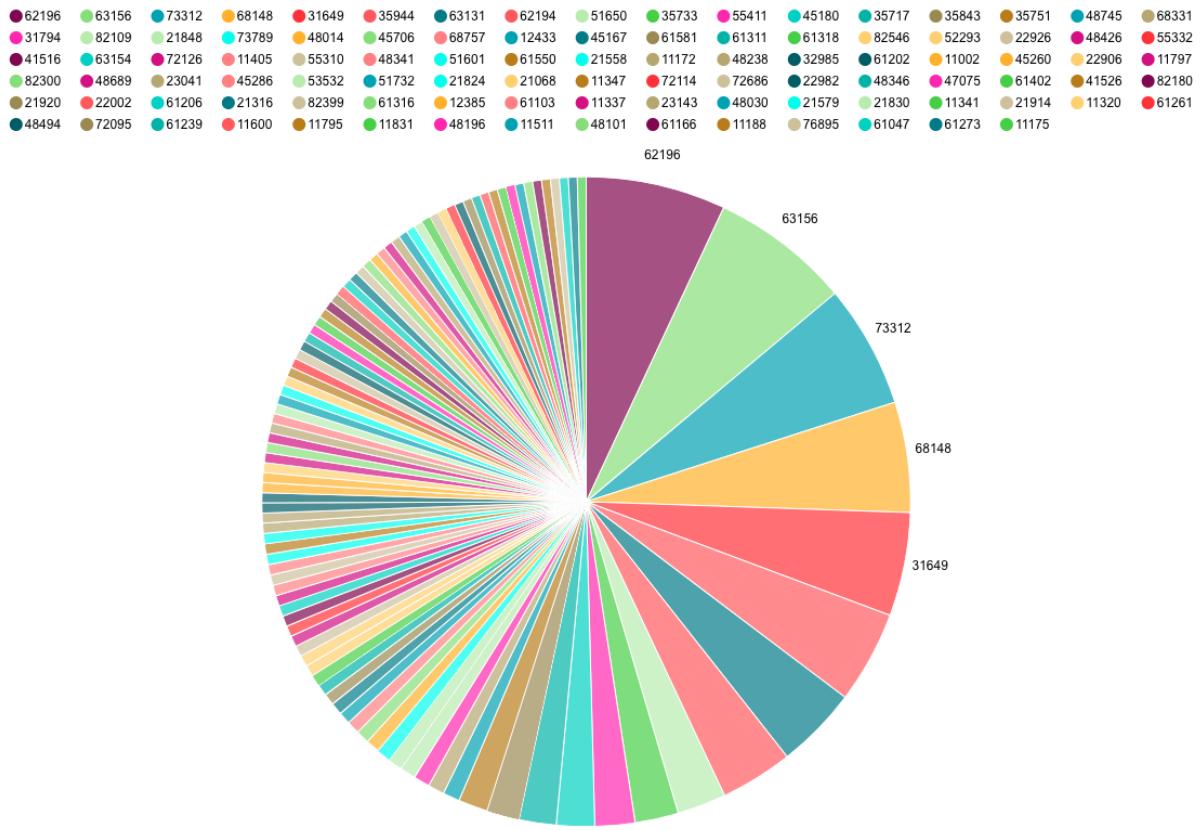
A Fig. 8 ilustra uma série temporal referente à quantidade de dados enviados por ônibus selecionados aleatoriamente, referentes a janeiro de 2017. Com esta visualização é possível observar, por exemplo, a oscilação da quantidade de dados enviados, assim como os picos de maior e menor volume de envio de dados e janelas de tempo com dados ausentes, que podem indicar inúmeros problemas relacionados a essas viagens, como eventos de exceção.

Figura 8 – Quantidade de dados enviados por dia por ônibus (selecionados aleatoriamente) em janeiro de 2017



A Fig. 9, representa a distribuição dos dados enviados em janeiro, a partir de uma amostra aleatória. Nessa figura é possível analisar que a distribuição da quantidade de dados enviados não é normalizada, ou seja, existem ônibus que normalmente enviam mais dados do que os demais. Há muitas razões possíveis para isso, por exemplo: viagens de ônibus mais longas que outras, regiões com diferenças climáticas; módulos AVL desatualizados; maior quantidade de ônibus em uma determinada linha, etc.

Figura 9 – Distribuição da quantidade de dados enviados por ônibus (selecionados aleatoriamente) em janeiro de 2017



Finalmente, os mapas exibidos pelas figuras 11 e 10 ajudam a identificar a localização a partir da qual os dados estão sendo enviados, permitindo visualizar possíveis pontos de falhas durante a transmissão desses dados. O primeiro mapa, respectivamente, refere-se à rota de uma única linha de ônibus e o segundo de todas as rotas; em ambos os casos, referentes aos dados de janeiro. Além disso, na figura 11, é possível observar a segregação urbana da cidade, devido ao fato de algumas regiões terem uma maior densidade de dados enviados, o que também indica regiões de maior tráfego, nas quais eventos de exceção teriam maior impacto.

5.5 Consideração sobre a arquitetura utilizada para exploração e visualização dos dados AVL da SPTrans

Este capítulo apresentou um estudo de caso relacionado à visualização de grandes conjuntos de dados, utilizando dados dos ônibus da cidade de São Paulo. Também, mostramos que é possível encontrar padrões complexos e incomuns e pos-

Figura 10 – Localizações enviadas em Janeiro de 2017 de uma linha de ônibus selecionada aleatoriamente

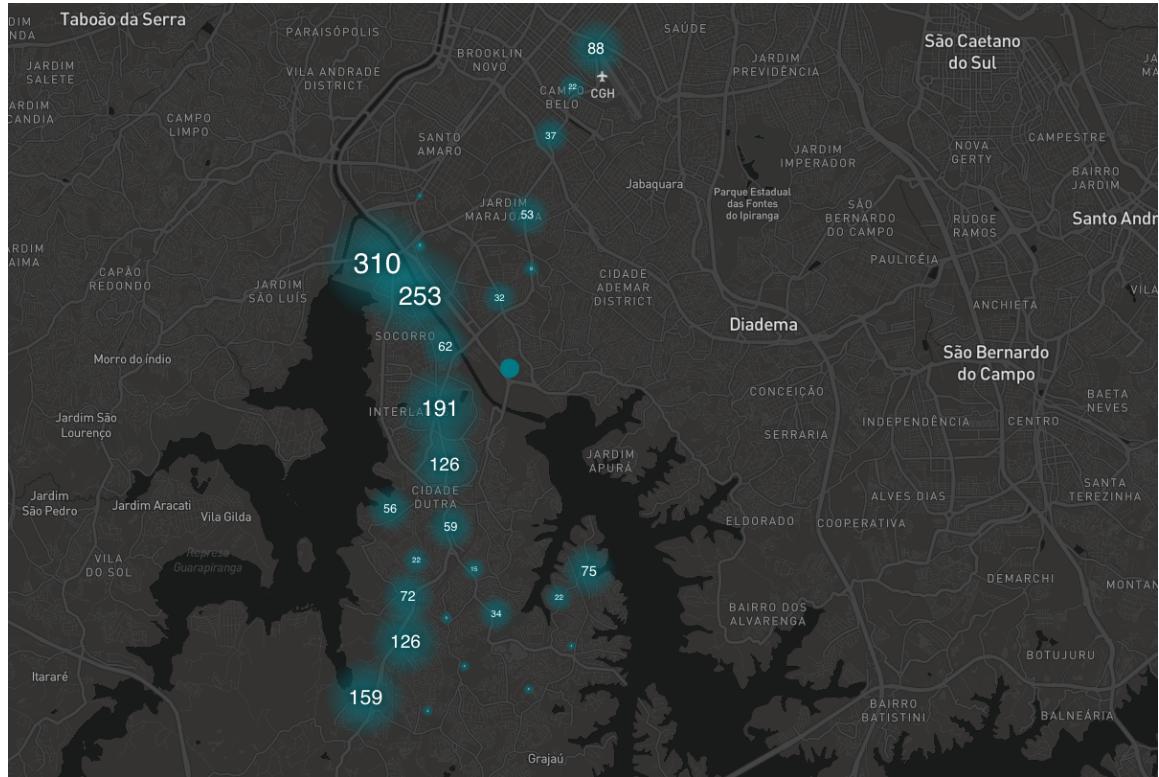
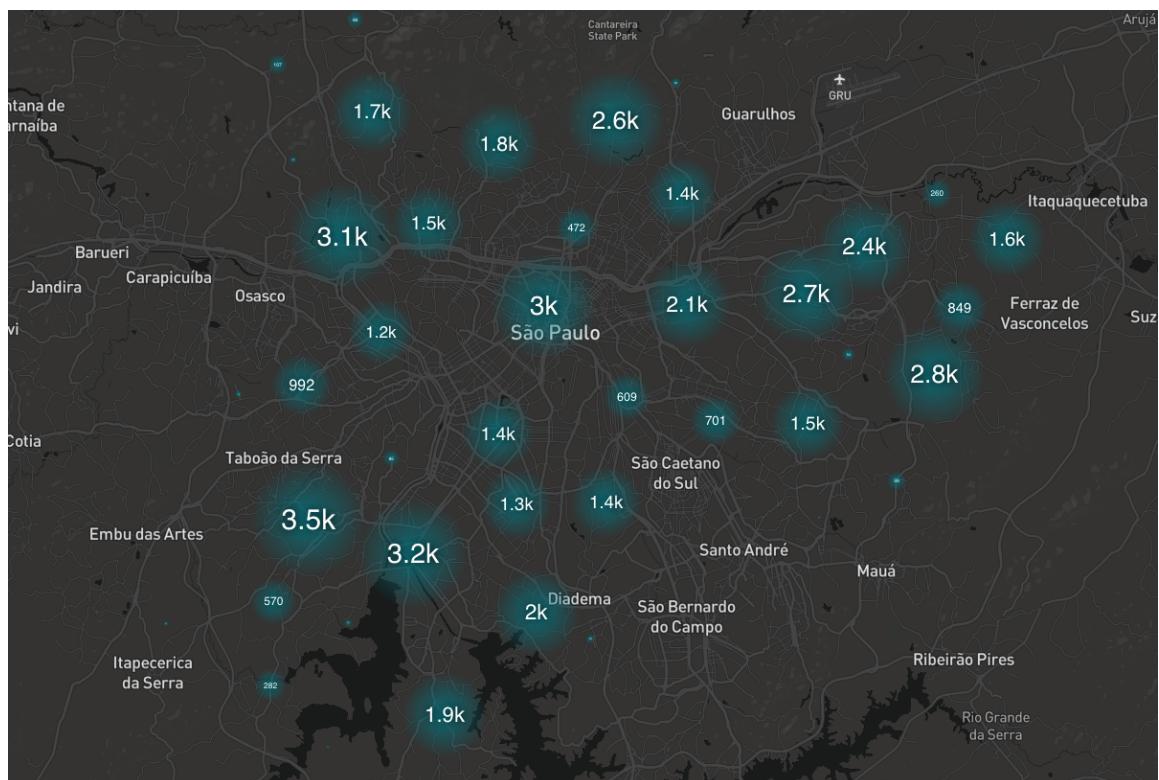


Figura 11 – Localizações dos ônibus referente a movimentação de Janeiro de 2017



síveis eventos de exceção em grandes conjuntos de dados por meio da visualização. O *Druid* e o *Apache Superset* demonstraram suporte a agregação, exploração e

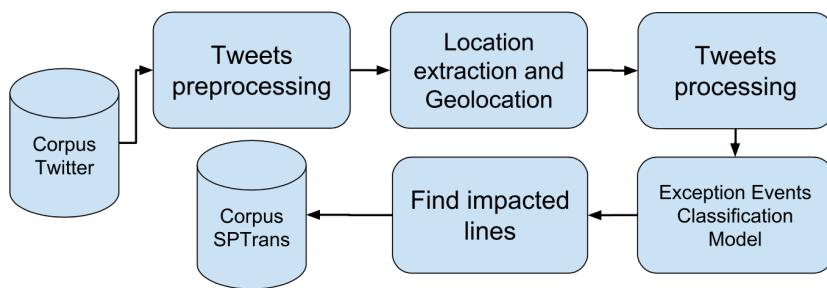
visualização de grandes conjuntos de dados. Como trabalho futuro, pretendemos implementar o fluxo de dados mencionado na Fig. 7, em um cenário de exploração e visualização de dados *near real time*.

6 Uma metodologia baseada em tweets para encontrar linhas de ônibus impactadas por eventos de exceção na cidade de São Paulo

Nesta seção é apresentada uma metodologia baseada em tweets para identificar linhas de ônibus impactadas por eventos de exceção. De acordo com a Fig. 12, a metodologia, explicada em detalhes na seções seguintes, é composta por:

1. Uma base de dados de tweets — *Corpus Twitter*.
2. Pré-processamento dos tweets existentes no conjunto de dados.
3. Extração de localização e geolocalização.
4. Processamento dos tweets.
5. Criação de um modelo de classificação de tweets em classes de eventos de exceção.
6. Identificação das linhas impactadas — por meio de consultas a base GTFS existente no *Corpus SPTrans* — a partir de um raio de cada evento de exceção.

Figura 12 – Metodologia baseada em tweets para encontrar linhas de ônibus impactadas por eventos de exceção na cidade de São Paulo



6.1 Pré-processamento

Numa pré-análise do *Corpus Twitter*, podemos afirmar que os tweets publicados pelos *profiles* selecionados evitam o uso de gírias, abreviações, erros de digitação; conforme consta nos tweets de exemplo contidos no trecho de código em json, no apêndice A. Isso diferencia tais tweets dos tweets publicados por usuários comuns do *Twitter*, que contém erros gramaticais, de sintaxe e que normalmente dependem de análise contextual para que possam ser interpretados.

Apesar disso, com base na literatura analisada ((STEIGER et al., 2015), (MIDDLETON; MIDDLETON; MODAFFERI, 2014), (KOBANI; SCHÜTZE; BURKOVSKI,

2010), (SETIAWAN; WIDYANTORO; SURENDRO, 2017), (ZAGAL; MATA; CLARAMUNT, 2016)), as seguintes etapas de pré-processamento são necessárias para remoção de ruído e redução da dimensão do espaço de *features* e foram realizadas para o *Corpus Twitter*:

- *Case folding*: processamento de normalização de todas as letras do texto (de a-z) para minúsculas.
- Remoção *URLs* e menções a outros *tweets*.
- Remoção de acentos, *emoticons* e pontuações substituídas por espaços vazios.
- *Stemming* — realizado neste trabalho na fase de processamento mencionada na subseção 6.3, com o objetivo de não afetar o processo de extração de endereços.

Além disso, é importante observar que (I) as informações referentes a data e hora mencionadas no conteúdo dos *tweets* (*stopwords* específicas do domínio) são removidas do texto original. As informações de data e hora consideradas para os eventos de exceção são as contidas nos meta dados dos *tweets*, posto que ao analisarmos os *tweets* verificamos que as informações de data e hora contidas no texto normalmente são referentes a eventos futuros, os quais não são considerados por este trabalho; (II) os *retweets* não estão presentes no *Corpus Twitter*; (III) no pré-processamento não há transformação do conteúdo dos *tweets*, embora trabalhos como os relacionados a identificação de sentimentos usem esse meio para transformar *emoticons* nos sentimentos que eles representam (ZAGAL; MATA; CLARAMUNT, 2016); (IV) as *hashtags* não são removidas dos *tweets* originais, pois são importantes para a classificação dos eventos de exceção.

Uma atenção especial foi dada às *hashtags*, que são relevantes para a classificação de eventos de exceção, mas adicionam ruído à fase de extração de endereços. Para mitigar o problema, *hashtags* são identificados e substituídos por espaços vazios no processo de extração de endereço. Além disso, é importante notar que as *hashtags* não são removidas dos *tweets* originais.

6.2 Extração de endereço e geolocalização

Analizando o conteúdo dos *tweets* das contas selecionadas, é possível observar que os textos publicados seguem um determinado padrão e, portanto, são na verdade semi-estruturados. Ante a isso, usamos a seguinte expressão regular para extrair os endereços presentes no conteúdo dos *tweets*:

$$ER = \{L_1|S_1|L_2|S_2|\dots|L_n|S_n\}\{[a - z\grave{A} - \ddot{y}_-] +\} \quad (6)$$

A expressão anterior é dividida em dois conjuntos, no primeiro ($\{L_1|S_1|L_2|S_2|\dots|L_n|S_n\}$), (L — logradouros) e (S — acrônimos de espaços públicos) são concatenados para especificar um filtro e identificar sequências inicializadas com espaços públicos ou seus respectivos acrônimos. No segundo conjunto ($\{[a - z\grave{A} - \ddot{y}_-] +\}$), é especificado um filtro para identificar um conjunto de palavras após L ou S, que são candidatas a compor o endereço desejado.

Essas palavras são candidatas porque é difícil saber quantas palavras após L ou S pertencem ao endereço, no entanto, as contas selecionadas publicam padrões visíveis após os endereços. Como consequência, um método possível para encontrar o endereço desejado é a remoção desses padrões após o início do endereço.

Após a extração do endereço, é necessário geolocalizar o endereço encontrado — apenas 1,5 % de *tweets* têm geolocalização (NIU et al., 2016) — o que é possível, por exemplo, usando a API de geocodificação do Google Maps¹. Os parâmetros de URL utilizados neste trabalho para chamar a API mencionada anteriormente são: (I) *address* — o endereço desejado; (II) *bounds* — uma caixa delimitadora para o resultado retornado, a qual é especificada pelas coordenadas de latitude / longitude dos cantos sudoeste e nordeste de São Paulo; (III) *region* — código da região com dois caracteres, por exemplo, *br* para o Brasil e (IV) *token* — *token* usado na autenticação da API.

Em seguida, a resposta HTTP é processada para obter os valores da localização (que contém informações de latitude e longitude) e o *endereço formatado*. É importante observar que os *tokens* do endereço extraído (*endereço não formatado*) são *stopwords* específicas do *corpus* em caso de alta frequência de eventos

¹ <<https://developers.google.com/maps/documentation/geocoding>>. Acessado em 11 de Abril de 2018.

de exceção localizados neste endereço, devido ao fato de que nesse cenário elas são tratados como *features* relevantes para o modelo de classificação. Portanto, os *tokens* dos endereços extraídos são armazenados para serem removidos na fase de processamento dos *tweets*.

6.3 Processamento de tweets

Nesta fase, os *tweets* são preparados para serem usados para treinar um modelo de classificação de eventos de exceção; neste momento, todos os *tweets* já foram pré-processados. Conforme mencionado na seção anterior, nesta fase, os *tokens* dos endereços extraídos armazenados são removidos para redução de ruído e as *stopwords* do Português Brasileiro filtradas² e todos os demais *tokens* processados por um *stemmer* para o Português Brasileiro³ para reduzir a dimensão do espaço de *features*.

6.4 Classificação manual do Corpus Twitter

Encontrar eventos de exceção envolve a identificação de eventos relacionados a uma exceção, o que é possível por meio de classificação de *tweets* (manualmente ou de forma autônoma). De acordo com a revisão sistemática realizado no Capítulo 3, as seguintes classes podem ser usadas para classificar eventos de exceção:

1. Acidentes.

- a) Acidentes nas estações transporte (ITOH et al., 2016).
- b) Incêndio (ITOH et al., 2016).

2. Espaço-temporais.

- a) Dia da semana (CHEN et al., 2016).
- b) Hora do dia (CHEN et al., 2016).

3. Eventos sociais.

- a) Feiras de rua (CHEN et al., 2016).

² Stopwords do Português Brasileiro obtidas da NLTK — <<https://www.nltk.org>>. Acessado em 19 de Abril de 2018.

³ RSLP Stemmer — <http://www.nltk.org/_modules/nltk/stem/rslp.html>. Acessado em 19 de Abril de 2018.

- b) Festivais (CHEN et al., 2016), (LECUE et al., 2014).
- c) Jogos esportivos (CHEN et al., 2016), (GAL-TZUR et al., 2014).
- d) Passeatas e maratonas (CHEN et al., 2016), (ITOH et al., 2016).

4. Eventos urbanos.

- a) Relacionados ao tráfego (CHEN et al., 2016), (LECUE et al., 2014).

5. Desastres naturais.

- a) Tempestades (ITOH et al., 2016).
- b) Terremoto (ITOH et al., 2016).
- c) Tufões (ITOH et al., 2016).

6. Metereológicos.

- a) Dia claro, nublado, chuvoso, nevando, com neblina (CHEN et al., 2016).
- b) Temperatura do ar (CHEN et al., 2016).

Após o estudo do domínio do conhecimento, por meio da revisão sistemática para coletar as classes de exceção, o Corpus Twitter, contendo 60.985, foi classificado manualmente de acordo com suas respectivas classes. Tal conjunto foi usado para treinar o modelo de classificação de tweets em classes de eventos de exceção.

6.5 Modelo de classificação de tweets relacionados a eventos de exceção

O corpus obtido da fase de processamento de tweets é representado por meio de um *bag-of-words*, que contém vetores de *features* criados usando a medida *Term Frequency - Inverse Document Frequency* (TF-IDF). A bag-of-words é particionada aleatoriamente em conjuntos de treinamento (60%) e teste (40%), os quais são entradas para os algoritmos de classificação mencionados na subseção 2.8.1.

6.6 Encontrando linhas de ônibus afetadas por eventos de exceção

Para encontrar as linhas de ônibus afetadas por eventos de exceção, é necessário correlacionar latitude e longitude dos eventos de exceção com as *shapes* da GTFS da SPTrans. Como mencionado anteriormente, as *shapes* da GTFS são um

conjunto de coordenadas de latitude e longitude para desenhar linhas dos ônibus em um mapa, para representar seus respectivos caminhos, por exemplo.

De acordo com a seção 4.0.2, todas as coordenadas são armazenadas em pares no formato *legacy* e em coleções com índices geoespaciais. Assim, é possível usar a função `$near` do MongoDB⁴ para encontrar as *shapes* próximas às coordenadas do evento de exceção. Como consequência da GTFS, o *shape_id* faz parte dos atributos contidos no arquivo de *shapes*, referindo-se a um código de linha de ônibus com o qual é possível procurar mais detalhes sobre o referido código, como direção da linha de ônibus, identificação, paradas, etc.

6.7 Resultados

A metodologia foi aplicada ao Corpus Twitter⁵. No final do pré-processamento e processamento dos *tweets*, o corpus obteve 3.761.226 palavras, com um vocabulário de 33 palavras. O comprimento máximo das sentenças do conjunto de dados é 136, sua respectiva variação é ilustrada pela Figura 13.

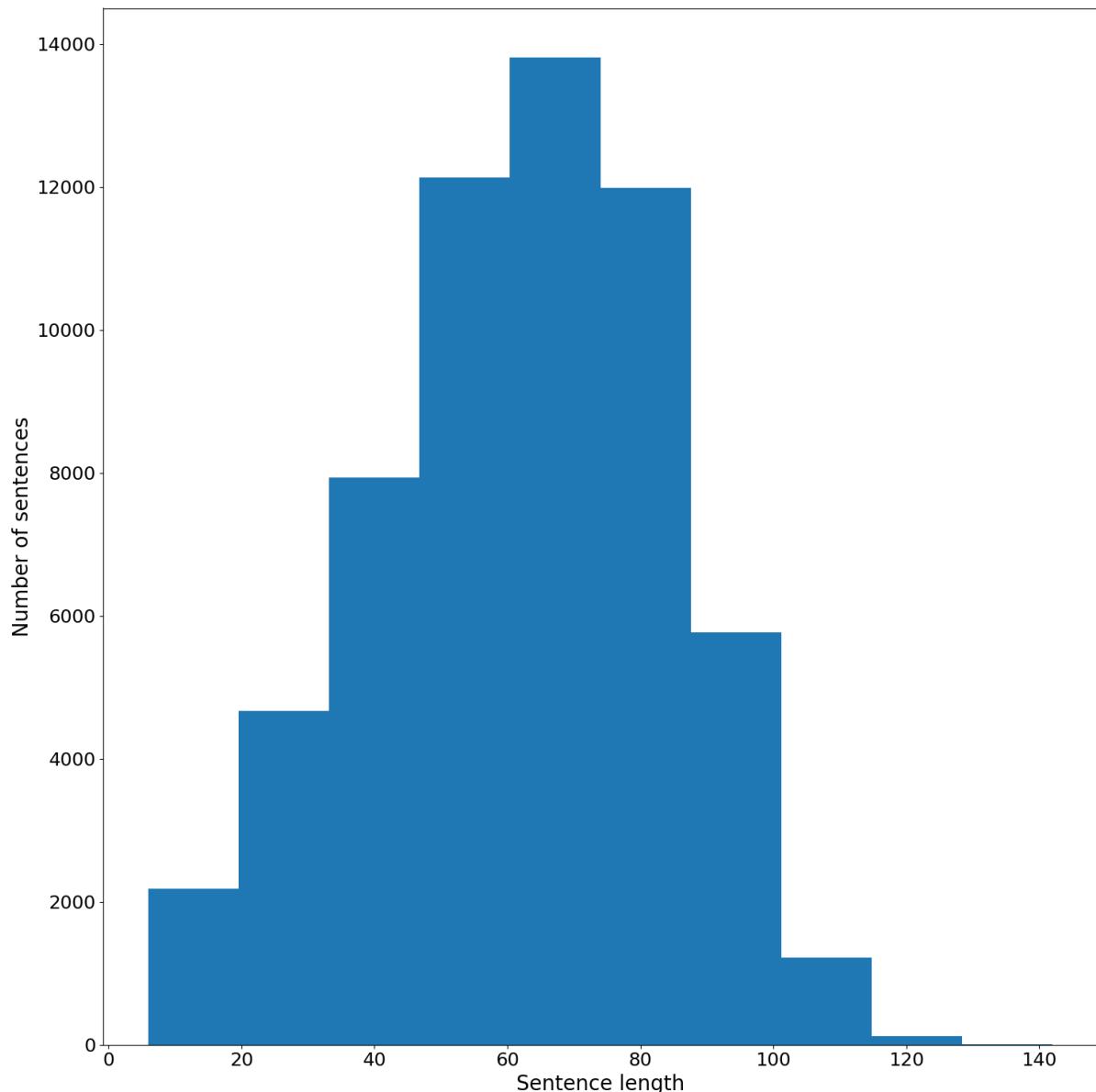
Todos os *tweets* existentes no *Corpus Twitter* foram classificados manualmente de acordo com os eventos de exceção identificados. Este conjunto de dados é composto pelas seguintes classes: Acidente, Irrelevante — quando o *tweet* não é um evento de exceção, Desastre Natural, Evento Social e Evento Urbano. A figura 14 ilustra a distribuição das classes de eventos de exceção em cada conta selecionada.

Esse conjunto de dados rotulado foi usado para treinar modelos de classificação de eventos de exceção, com base em uma *bag-of-words*, descrita na Seção 6.5. De acordo com a Tabela 8, o modelo que usa o algoritmo *Multi-layer Perceptron* para classificação é mais adequado para a tarefa de classificar os *tweets* em eventos de exceção. A matriz de confusão relacionada ao algoritmo de *Multi-layer Perceptron* é ilustrada pela Figura 15, as matrizes de confusão dos demais algoritmos podem ser consultadas no apêndice E.

⁴ <<https://docs.mongodb.com/manual/reference/operator/query/near/>>. Acessado em 18 de Maio de 2018.

⁵ Conjunto de dados disponível em: <<https://drive.google.com/drive/folders/16NIevLsBR0A45UHdPDvv2lZZx6gF4R0p?usp=sharing>>. Acessado em 8 de Setembro de 2018.

Figura 13 – Histograma da variação dos tamanhos das sentenças dos tweets existentes no *Corpus Twitter*



Dos 60.984 tweets 10.027 foram classificados em eventos de exceção e desse subconjunto encontrados 7.674 endereços, de acordo com a Tab. 9 — desconsiderando o tipo de localidade APPROXIMATE — (o que representa 76,53% do total dos tweets eventos de exceção, sem considerar a classe *Irrelevant*). A quantidade de endereços extraídos por classe está descrita na Tab. 9, as razões para tweets sem endereço extraído são:

1. Tweets apenas com o ponto de interesse, ou seja, não consta explicitamente o endereço.

Figura 14 – Distribuição das classes dos eventos de exceção do Corpus Twitter

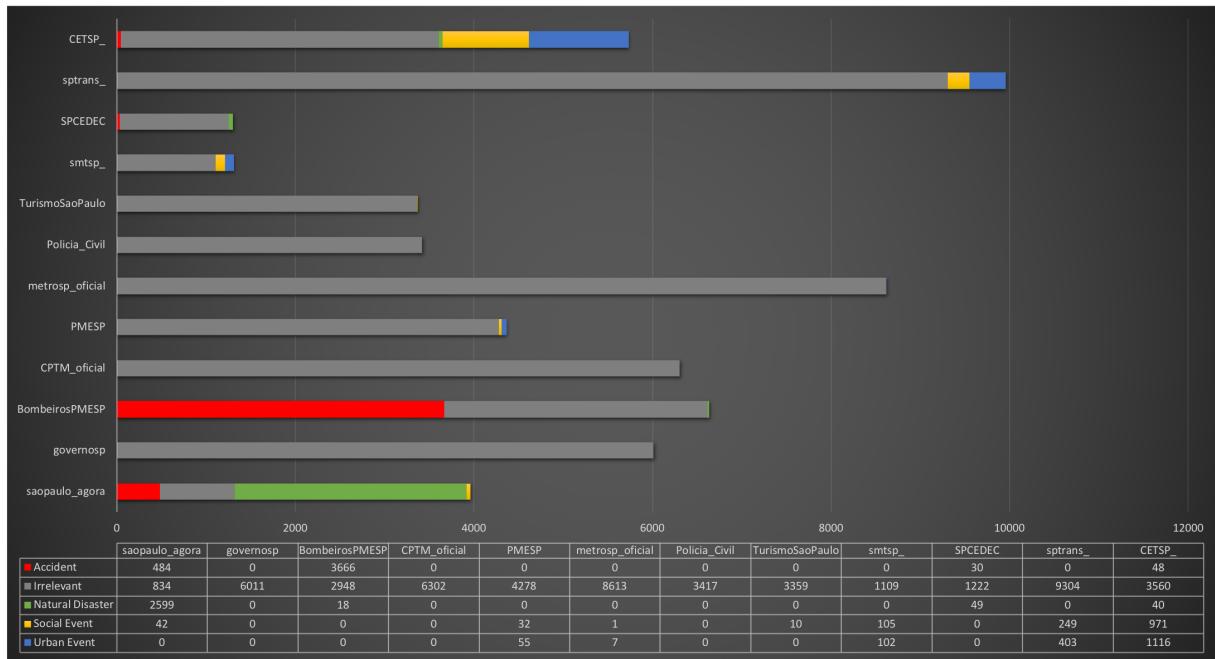


Tabela 8 – Métricas das avaliações dos algoritmos utilizados para classificação dos tweets em eventos de exceção

Algoritmo	Acurácia	Precisão	Revocação	f1-score
Decision Tree	0.966	0.966	0.966	0.966
Gaussian Naive Bayes	0.891	0.919	0.891	0.901
K-Nearest Neighbors	0.970	0.971	0.970	0.970
Logistic Regression	0.970	0.970	0.970	0.969
Multi-layer Perceptron	0.974	0.974	0.974	0.974
Multinomial Naive Bayes	0.954	0.953	0.954	0.951
Random Forest	0.972	0.971	0.972	0.971
Support Vector Machine	0.828	0.686	0.828	0.751

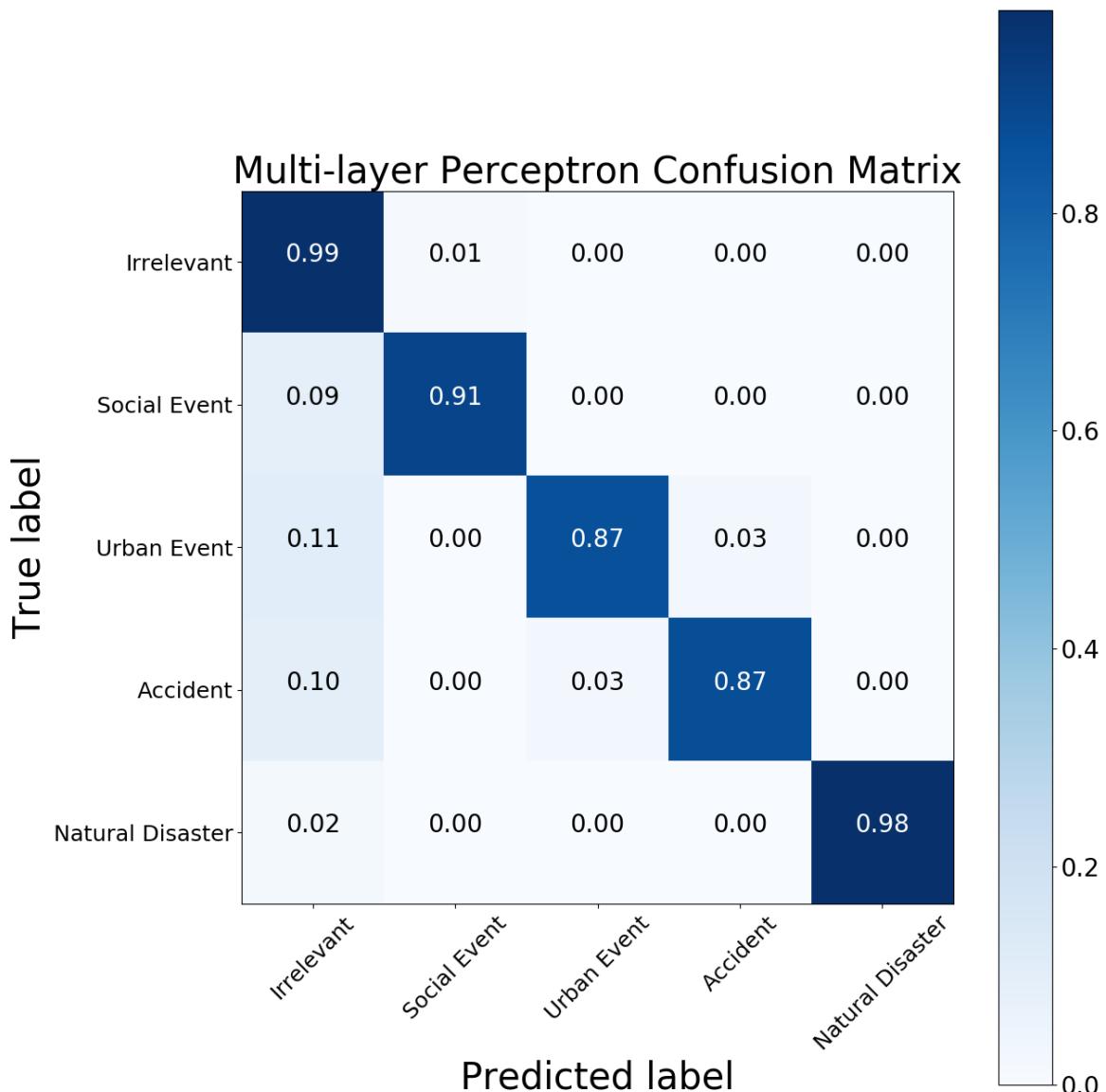
2. Tweets sem informação de endereço.
3. Tweets com nome de logradouro incomum (por exemplo *passagem, complexo viário, ligação sentido*).
4. Tweets com endereços com palavras concatenadas (por exemplo *avenidapaulista*).

Os tipos de localidades⁶ podem ser classificados em:

1. ROOFTOP — Indica que o resultado retornado há informações de localização com precisão a nível do endereço de rua.

⁶ Disponível em <<https://developers.google.com/maps/documentation/geocoding>>. Acessado em 16 de setembro de 2018.

Figura 15 – Matriz de confusão relacionada a classificação dos tweets em eventos de exceção por meio do algoritmo *Multi-layer Perceptron*



2. *RANGE_INTERPOLATED* — Indica que o resultado retornado reflete uma aproximação interpolada entre dois pontos precisos (como interseções). Geralmente, os resultados interpolados são retornados quando os códigos geográficos do *rooftop* não estão disponíveis para um endereço de rua.
3. *GEOMETRIC_CENTER* — Indica que o resultado retornado é o centro geométrico de um resultado.
4. *APPROXIMATE* — Indica que o resultado retornado é aproximado.

Neste estudo de caso, desconsideramos os endereços com classificação *APPROXIMATE*, devido ao fato de poderem comprometer a confiabilidade das análises realizadas.

Tabela 9 – Quantidade de eventos extraídos por classe

Classe	#endereços extraídos ^a	#APP ^b	#GEO ^c	#RANGE ^d	#ROOF ^e
Accident	3.439	7	805	1.130	1.497
Irrelevant	451	13	292	6	140
Natural Disaster	2.464	9	340	719	1.396
Social Event	793	4	761	2	26
Urban Event	1.002	4	942	10	46
-	8.149	37	3.140	1.867	3.105

^a Total de endereços extraídos

^b Total de endereços extraídos com o tipo de localidade *APPROXIMATE*

^c Total de endereços extraídos com o tipo de localidade *GEOMETRIC_CENTER*

^d Total de endereços extraídos com o tipo de localidade *RANGE_INTERPOLATED*

^e Total de endereços extraídos com o tipo de localidade *ROOFTOP*

A Fig. 16 ilustra os endereços⁷ mais afetados por eventos de exceção e a Fig. 17 parte da distribuição desses eventos na região central de São Paulo. É importante ressaltar que os eventos de exceção encontrados estão concentrados em endereços e regiões onde normalmente ocorrem em São Paulo, o que valida a metodologia desenvolvida.

Consideramos que uma linha de ônibus é afetada por uma exceção se uma *shape* em um raio de 100 metros do evento. Utilizando este critério, o total de 799 linhas de ônibus foram afetadas por eventos de exceção durante este período, sendo “35276” o código de linha de ônibus mais impactado (itinerário próximo à concentração de eventos de exceção, conforme Figura 18). Essa linha específica foi impactada por 1.663 eventos de exceção. A Tab. 10 lista as linhas de ônibus que foram impactadas por mais de 1.000⁸ eventos de exceção.

⁷ Lista completa está disponível em <<https://docs.google.com/spreadsheets/d/1gn1cTDifUJEPdgcU67SC45GdYHRKmIHtAfJwRBm088s/edit?usp=sharing>>. Acessado em 09 de setembro de 2018.

⁸ O conteúdo completo da tabela pode ser acessado em:<<https://docs.google.com/spreadsheets/d/1SxnzVeBaSjophbo-slISUcUG0GMD2MYaEM0JAdKiNBs>>. Acessado em 18 de maio de 2018.

Figura 16 – Endereços mais impactados por eventos de exceção

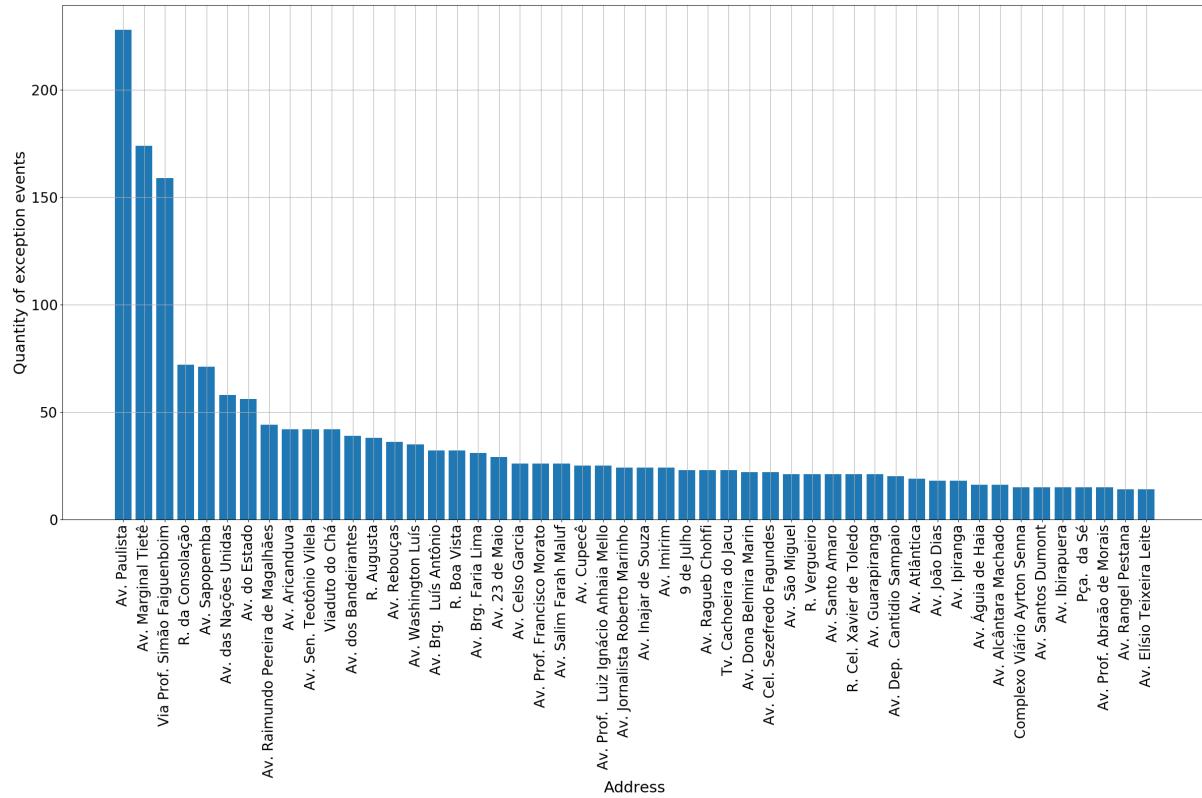
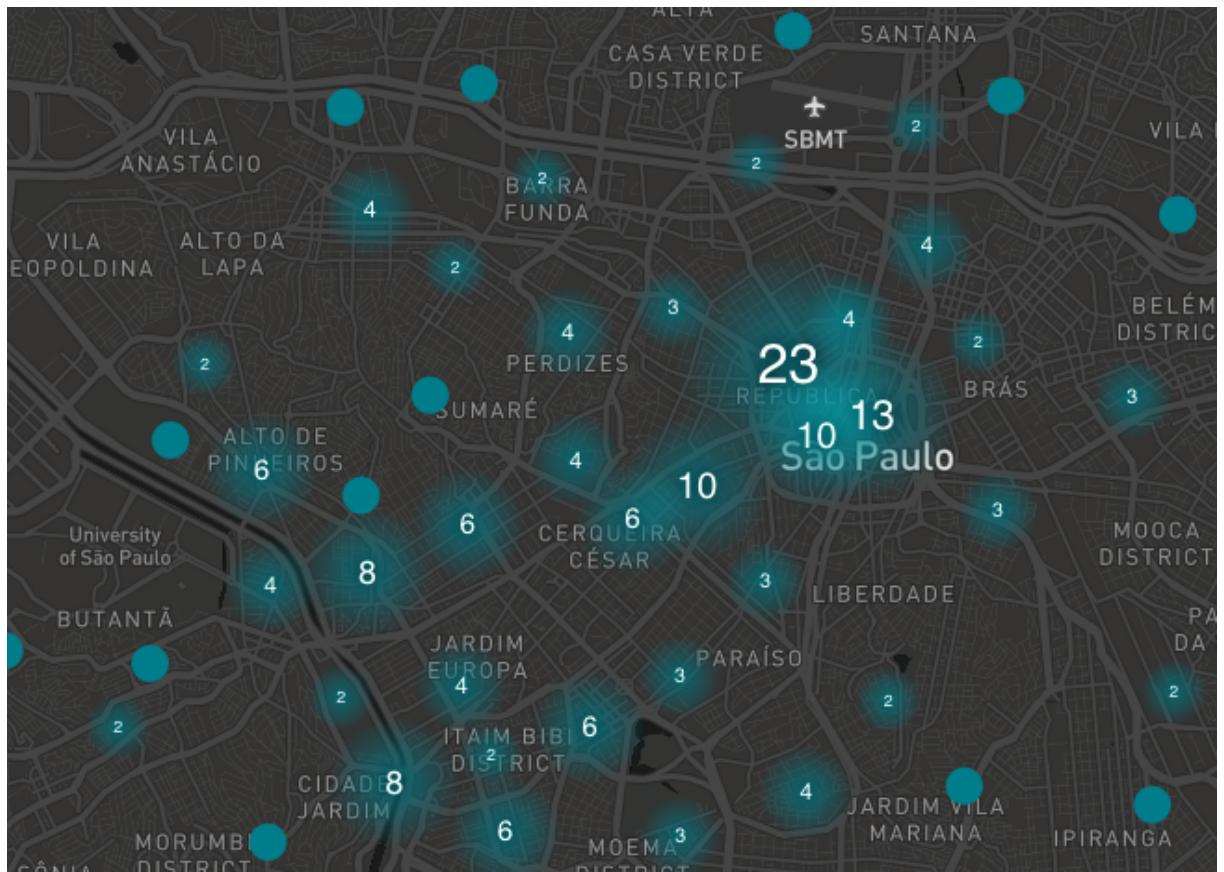


Tabela 10 – Linhas de ônibus mais impactadas por eventos de exceção^a

Código da linha	# eventos de exceção	Letreiro
35276	1,663	PÇA. RAMOS DE AZEVEDO / TERM. CAMPO LIMPO
32953	1,606	TERM. PINHEIROS / TERM. JD. ÂNGELA
32805	1,462	TERM. PRINC. ISABEL / CHÁC. SANTANA
32772	1,423	TERM. PRINC. ISABEL / TERM. STO. AMARO
32846	1,420	METRÔ BRÁS / TERM. GRAJAÚ
32869	1,339	PINHEIROS / GRAJAÚ
34134	1,335	METRÔ ANA ROSA / MORRO GRANDE
33966	1,296	METRÔ VL. MARIANA / TERM. PARELHEIROS
33943	1,295	TERM. STO. AMARO / VARGEM GRANDE
34694	1,245	PARAÍSO / TERM. CAMPO LIMPO
32874	1,210	METRÔ JABAQUARA / PQ. RES. COCAIA
34660	1,184	ACLIMAÇÃO / TERM. CAMPO LIMPO
33272	1,150	PÇA. RAMOS DE AZEVEDO / JD. JOÃO XXIII
34210	1,129	LGO. SÃO FRANCISCO / TERM. VARGINHA
32838	1,105	PÇA. DA SÉ / PQ. RES. COCAIA
32829	1,076	TERM. BANDEIRA / TERM. CAPELINHA
34233	1,061	TERM. BANDEIRA / TERM. VARGINHA
33640	1,056	TERM. CACHOEIRINHA / JD. PRINCESA
33377	1,054	PERDIZES / AEROPORTO
33042	1,022	PÇA. DA SÉ / JD. IV CENTENÁRIO

^a Tabela completa no apêndice D.

Figura 17 – Distribuição dos eventos de exceção na região central de São Paulo

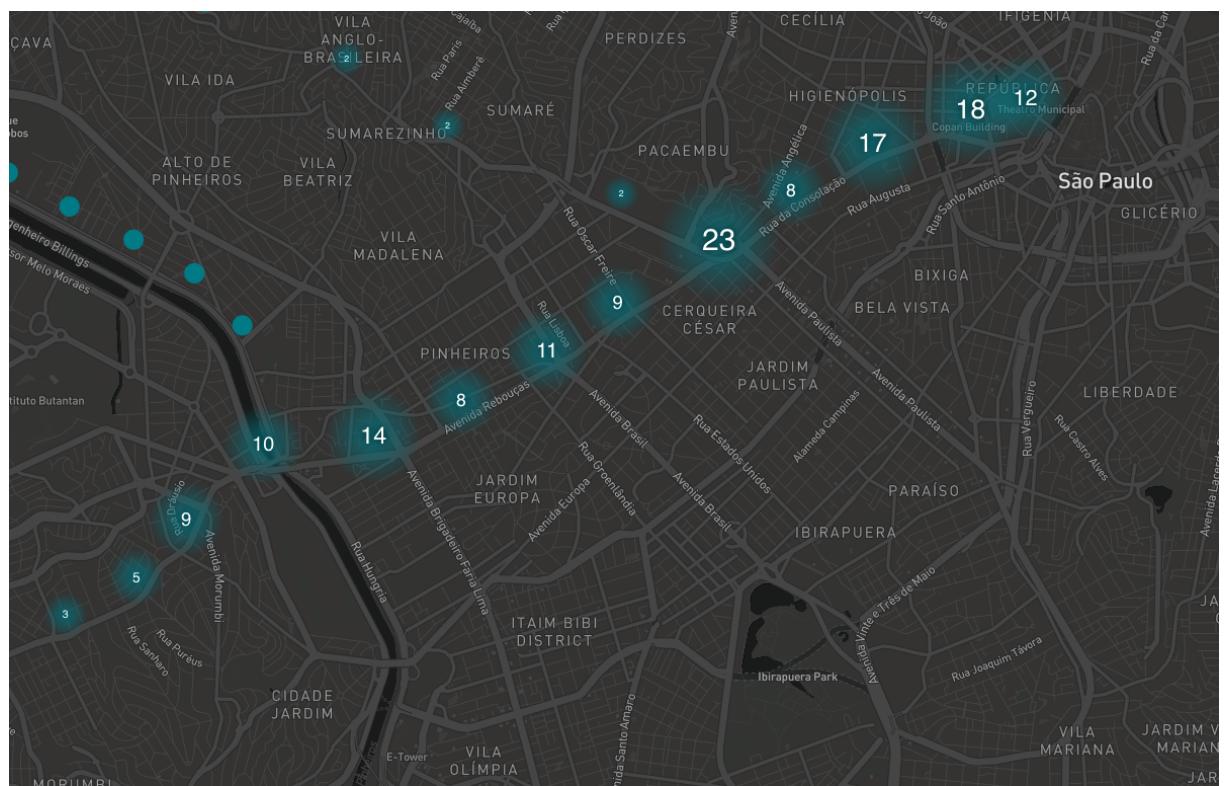


6.8 Considerações finais sobre a metodologia desenvolvida

Este experimento apresenta uma nova metodologia para classificação de eventos de exceção e analisa seus respectivos impactos no sistema de transporte coletivo por ônibus da cidade de São Paulo. Com o conjunto de dados utilizados, descobrimos que o melhor algoritmo para classificar tweets em eventos de exceção foi *Multi-layer Perceptron*. Também, mostramos que é possível extrair endereços de tweets semi-estruturados usando apenas expressões regulares. A classificação desses eventos é o primeiro passo para entender melhor como os eventos de exceção afetam a rede de transporte público.

Embora o método tenha sido validado usando perfis selecionados do Twitter escritos em português do Brasil, o mesmo pode ser generalizado para diferentes idiomas e cidades. A GTFS é um formato ubíquo para o transporte público e ferramentas como a NLTK suporta vários idiomas.

Figura 18 – Dados de geolocalização do código de linha de ônibus “35276” mais afetado por eventos de exceção (ilustração a nível de rua)



7 Correlação dos eventos de exceção com os dados AVL da SPTrans

Escrever Correlação dos eventos de exceção com os dados AVL da SPTrans

Dado que os eventos de exceção podem ser identificados utilizando *tweets* dos *profiles* contidos na tabela 1, há também a possibilidade de caracterizarmos seus respectivos impactos analisando a base histórica dos dados AVL da SPTrans, especificamente os dados referentes a *timestamp*, *latitude*, *longitude*, *bus_id* e *trip_id*. Dito isso, inicialmente pretendemos caracterizar os impactos em:

- Atraso médio induzido nas viagens.
- Ônibus frequentemente afetados por eventos de exceção.
- Ônibus frequentemente afetados por determinado evento de exceção.
- Padrão de ocorrência dos eventos de exceção no espaço-tempo (localizações e *timestamps*).
- Quantidade e viagens afetadas.
- Quantidade e regiões da cidade de São Paulo afetadas.
- Viagens frequentemente afetadas por eventos de exceção.
- Viagens frequentemente afetadas por determinado evento de exceção.

8 Conclusão

Neste capítulo, são apresentadas as contribuições e resultados esperados com o projeto de pesquisa, as limitações a ameaças à validade do estudo.

8.1 Contribuições

A principal contribuição deste projeto é propor uma solução para o problema de caracterização de eventos de exceção e de seus respectivos impactos no sistema de transporte público por ônibus da cidade de São Paulo, por meio de *tweets* e de dados históricos dos módulos AVL do SIM. Além disso, a solução proposta visa disponibilizar os conjuntos de dados que foram construídos e uma plataforma para que esses dados possam ser visualizados e explorados, de forma a contribuir com projetos e pesquisas futuras correlatas.

Em relação a publicações científicas, serão submetidos artigos com os resultados obtidos para veículos de disseminação de conhecimento científico nas áreas de: Análise de Redes Sociais, Sistemas de Transporte Inteligentes, Cidades Inteligentes.

8.2 Trabalhos publicados

Escrever trabalhos publicados

8.3 Trabalhos futuros

Escrever trabalhos futuros

As principais limitações deste projeto estão relacionadas ao processamento de *tweets* em português brasileiro e oriundos das contas selecionadas e referenciadas na tabela 1, o que pode tornar a solução não generalista. Dentre os riscos, apesar das análises preliminares realizadas para extração de endereços dos conteúdos dos *tweets* por meio de Expressão Regular, é possível que sejam encontrados novos desafios que inviabilizem o uso dessa técnica.

Acknowledgment

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.

Notes

Atualizar organização do documento	18
Escrever sobre cada algoritmo utilizado	34
Escrever Correlação dos eventos de exceção com os dados AVL da SPTrans . .	83
Escrever trabalhos publicados	84
Escrever trabalhos futuros	84

Referências

- ABBASI, A. et al. Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play. *Proc. 8th ACM SIGSPATIAL Int. Work. Locat. Soc. Networks - LBSN'15*, p. 1–9, 2015. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2830657.2830660>>. Citado 6 vezes nas páginas 43, 44, 45, 51, 52 e 53.
- AHVENNIEMI, H. et al. What are the differences between sustainable and smart cities? *Cities*, Elsevier B.V., v. 60, p. 234–245, 2017. ISSN 02642751. Disponível em: <<http://dx.doi.org/10.1016/j.cities.2016.09.009>>. Citado 3 vezes nas páginas 19, 20 e 21.
- ALBINO, V.; BERARDI, U.; DANGELICO, R. M. Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, Taylor & Francis, v. 22, n. 1, p. 3–21, 2015. Citado na página 21.
- ANANTHARAM, P. et al. Extracting City Traffic Events from Social Streams. *ACM Trans. Intell. Syst. Technol.*, v. 6, n. 4, p. 1–27, 2015. ISSN 21576904. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2801030.2717317>>. Citado 6 vezes nas páginas 44, 46, 49, 50, 52 e 53.
- Andreas Mueller. 2018. <<https://pypi.python.org/pypi/wordcloud>>. Acesso em Fevereiro, 13 de 2018. Citado na página 42.
- ANDRIENKO, G. et al. Visual analytics of mobility and transportation: State of the art and further research directions. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 18, n. 8, p. 2232–2249, 2017. Citado na página 62.
- ANG, L.-M. et al. Big Sensor Data Systems for Smart Cities. *IEEE Internet Things J.*, v. 4, n. 5, p. 1–1, 2017. ISSN 2327-4662. Disponível em: <<http://ieeexplore.ieee.org/document/7903653/>>. Citado 2 vezes nas páginas 20 e 21.
- ANTTIROIKO, A. V. U-cities reshaping our future: Reflections on ubiquitous infrastructure as an enabler of smart urban development. *AI Soc.*, v. 28, n. 4, p. 491–507, 2013. ISSN 09515666. Citado na página 13.
- ATEFEH, F.; KHREICH, W. A survey of techniques for event detection in twitter. *Computational Intelligence*, Wiley Online Library, v. 31, n. 1, p. 132–164, 2015. Citado na página 28.
- BARTH, J. et al. Informational urbanism . A conceptual framework of smart cities. *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, p. 2814–2823, 2017. Citado 2 vezes nas páginas 20 e 21.
- BENDLER, J. et al. Taming Uncertainty in Big Data. *Bus. Inf. Syst. Eng.*, v. 6, n. 5, p. 279–288, 2014. ISSN 1867-0202. Disponível em: <<http://link.springer.com/10.1007/s12599-014-0342-4>>. Citado 7 vezes nas páginas 43, 45, 49, 50, 51, 52 e 54.
- BIOLOCHINI, J. et al. Techincal report rt-es 679/05: Systematic review in software engineering. *COPPE/UFRJ, 2005*Rio de Janeiro, 2005. Citado 2 vezes nas páginas 36 e 37.

CHANIOTAKIS, E.; ANTONIOU, C. Use of Geotagged Social Media in Urban Settings: Empirical Evidence on Its Potential from Twitter. *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, v. 2015-Octob, n. 1, p. 214–219, 2015. Citado 2 vezes nas páginas 44 e 51.

CHANIOTAKIS, E.; ANTONIOU, C.; PEREIRA, F. Mapping Social media for transportation studies. *IEEE Intell. Syst.*, v. 31, n. 6, p. 64–70, 2016. ISSN 15411672. Citado na página 37.

CHEN, L. et al. Dynamic Cluster-Based Over-Demand Prediction in Bike Sharing Systems. *UBICOMP*, p. 841–852, 2016. Citado 12 vezes nas páginas 16, 17, 43, 46, 47, 48, 49, 50, 51, 52, 73 e 74.

CHEN, W.; GUO, F.; WANG, F.-Y. A survey of traffic data visualization. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 16, n. 6, p. 2970–2984, 2015. Citado na página 62.

CHUA, A. et al. Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tour. Manag.*, Elsevier Ltd, v. 57, p. 295–310, 2016. ISSN 02615177. Disponível em: <<http://dx.doi.org/10.1016/j.tourman.2016.06.013>>. Citado 2 vezes nas páginas 44 e 45.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, v. 12, n. Aug, p. 2493–2537, 2011. Citado na página 29.

CONSULO, M. et al. An evaluation of the proposed its system for the city of são paulo based on the 2015 tender. In: EDP SCIENCES. *MATEC Web of Conferences*. [S.I.], 2016. v. 76, p. 03004. Citado 2 vezes nas páginas 13 e 14.

DI LORENZO, G. et al. EXSED: An intelligent tool for exploration of social events dynamics from augmented trajectories. *Proc. - IEEE Int. Conf. Mob. Data Manag.*, v. 1, p. 323–330, 2013. ISSN 15516245. Citado 5 vezes nas páginas 43, 46, 51, 52 e 54.

DIAS, F. *Repositório contendo os artefatos da Revisão Sistemática*. 2017. Disponível em: <<https://github.com/fcas/dissertacao>>. Citado na página 41.

DWIVEDI, S. K.; ARYA, C. Automatic text classification in information retrieval: A survey. In: ACM. *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. [S.I.], 2016. p. 131. Citado na página 32.

FARSEEV, A. et al. Harvesting Multiple Sources for User Profile Learning. *Proc. 5th ACM Int. Conf. Multimed. Retr. - ICMR '15*, p. 235–242, 2015. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2671188.2749381>>. Citado 5 vezes nas páginas 43, 44, 50, 52 e 53.

FIGUEIREDO, L. et al. Towards the development of intelligent transportation systems. In: IEEE. *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*. [S.I.], 2001. p. 1206–1211. Citado 2 vezes nas páginas 21 e 22.

- FINGER, M.; RAZAGHI, M. Conceptualizing “Smart Cities”. *Informatik-Spektrum*, v. 40, n. 1, p. 6–13, 2017. ISSN 1432122X. Citado 3 vezes nas páginas 19, 20 e 21.
- FRIAS-MARTINEZ, V.; FRIAS-MARTINEZ, E. Spectral clustering for sensing urban land use using Twitter activity. *Eng. Appl. Artif. Intell.*, Elsevier, v. 35, p. 237–245, 2014. ISSN 09521976. Disponível em: <<http://dx.doi.org/10.1016/j.engappai.2014.06.019>>. Citado 7 vezes nas páginas 43, 46, 49, 50, 51, 52 e 53.
- GAL-TZUR, A. et al. The potential of social media in delivering transport policy goals. *Transp. Policy*, v. 32, p. 115–123, 2014. ISSN 0967070X. Citado 8 vezes nas páginas 17, 47, 48, 50, 51, 52, 53 e 74.
- 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.*, Elsevier Ltd, v. 58, p. 265–277, 2015. ISSN 0968090X. Disponível em: <<http://dx.doi.org/10.1016/j.trc.2014.12.006>>. Citado 3 vezes nas páginas 43, 44 e 48.
- GKIOTSALITIS, K.; STATHOPOULOS, A. Joint leisure travel optimization with user-generated data via perceived utility maximization. *Transp. Res. Part C Emerg. Technol.*, Elsevier Ltd, v. 68, p. 532–548, 2016. ISSN 0968090X. Disponível em: <<http://dx.doi.org/10.1016/j.trc.2016.05.009>>. Citado 5 vezes nas páginas 43, 44, 47, 48 e 51.
- GUO, W. et al. Understanding happiness in cities using twitter: Jobs, children, and transport. *IEEE 2nd Int. Smart Cities Conf. Improv. Citizens Qual. Life, ISC2 2016 - Proc.*, 2016. Citado 7 vezes nas páginas 44, 45, 47, 49, 52, 53 e 54.
- 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*, p. 242–247, 2016. Disponível em: <<http://doi.acm.org/10.1145/3014087.3014123>>. Citado 8 vezes nas páginas 43, 44, 45, 47, 48, 50, 52 e 54.
- GUYON, I.; ELISSEEFF, A. An introduction to feature extraction. *Feature extraction*, Springer, p. 1–25, 2006. Citado na página 31.
- HASAN, S.; UKKUSURI, S. V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C Emerg. Technol.*, Elsevier Ltd, v. 44, p. 363–381, 2014. ISSN 0968090X. Disponível em: <<http://dx.doi.org/10.1016/j.trc.2014.04.003>>. Citado 4 vezes nas páginas 43, 45, 52 e 54.
- ITOH, M. et al. Visual Exploration of Changes in Passenger Flows and Tweets on Mega-City Metro Network. *IEEE Trans. Big Data*, v. 2, n. 1, p. 85–99, 2016. ISSN 2332-7790. Disponível em: <<http://ieeexplore.ieee.org/document/7445832/>>. Citado 7 vezes nas páginas 16, 17, 47, 48, 51, 73 e 74.
- JUNGHERR, A. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, Taylor & Francis, v. 13, n. 1, p. 72–91, 2016. Citado na página 37.

- KOBANI, H.; SCHÜTZE, H.; BURKOVSKI, A. Relational feature engineering of natural language processing. *Proc. 19th . . . , n. ii*, p. 1705–1708, 2010. Disponível em: <<http://dl.acm.org/citation.cfm?id=1871709>>. Citado na página 71.
- KORENIUS, T. et al. Stemming and lemmatization in the clustering of finnish text documents. In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2004. (CIKM '04), p. 625–633. ISBN 1-58113-874-1. Disponível em: <<http://doi.acm.org/10.1145/1031171.1031285>>. Citado na página 29.
- KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, v. 160, p. 3–24, 2007. Citado 2 vezes nas páginas 32 e 33.
- KUFLIK, T. et al. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies*, Elsevier, v. 77, p. 275–291, 2017. Citado 2 vezes nas páginas 14 e 16.
- KUMMITHA, R. K. R.; CRUTZEN, N. How do we understand smart cities? An evolutionary perspective. *Cities*, Elsevier, v. 67, n. July 2016, p. 43–52, 2017. ISSN 02642751. Disponível em: <<http://dx.doi.org/10.1016/j.cities.2017.04.010>>. Citado 3 vezes nas páginas 19, 20 e 21.
- LECUE, F. et al. Smart traffic analytics in the semantic web with STAR-CITY: Scenarios, system and lessons learned in Dublin City. *J. Web Semant.*, Elsevier B.V., v. 27, p. 26–33, 2014. ISSN 15708268. Disponível em: <<http://dx.doi.org/10.1016/j.websem.2014.07.002>>. Citado 5 vezes nas páginas 17, 44, 46, 51 e 74.
- 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*. [S.l.: s.n.], 2017. Citado na página 28.
- MAGHREBI, M. et al. 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*, v. 2015-Octob, p. 208–213, 2015. Citado 4 vezes nas páginas 43, 45, 47 e 48.
- MATA, F.; CLARAMUNT, C. A Mobile Trusted Path System Based on Social Network Data. *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, p. 101:1—101:4, 2015. Disponível em: <<http://doi.acm.org/10.1145/2820783.2820799>>. Citado 4 vezes nas páginas 44, 51, 52 e 53.
- MENUAR, H. et al. Uav-enabled intelligent transportation systems for the smart city: Applications and challenges. *IEEE Communications Magazine*, IEEE, v. 55, n. 3, p. 22–28, 2017. Citado 2 vezes nas páginas 21 e 22.
- MIDDLETON, S. E.; MIDDLETON, L.; MODAFFERI, S. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, v. 29, n. 2, p. 9–17, 2014. ISSN 15411672. Citado na página 70.

- MORENO, M. V. et al. Applicability of Big Data Techniques to Smart Cities Deployments. *IEEE Trans. Ind. Informatics*, v. 13, n. 2, p. 800–809, 2017. ISSN 15513203. Citado 2 vezes nas páginas 20 e 21.
- MOTODA, H.; LIU, H. Feature selection, extraction and construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan)* Vol, v. 5, p. 67–72, 2002. Citado na página 31.
- MUKHERJEE, T. et al. Janayuja: A People-centric Platform to Generate Reliable and Actionable Insights for Civic Agencies. *Acm Dev* 2015, p. 137–145, 2015. Citado 9 vezes nas páginas 43, 46, 47, 48, 49, 50, 51, 52 e 53.
- MYERS, S. A. et al. Information network or social network?: the structure of the twitter follow graph. In: ACM. *Proceedings of the 23rd International Conference on World Wide Web*. [S.l.], 2014. p. 493–498. Citado na página 28.
- NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 544–551, 2011. Citado 2 vezes nas páginas 29 e 30.
- NARAYANAN, U. et al. A survey on various supervised classification algorithms. In: IEEE. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. [S.l.], 2017. p. 2118–2124. Citado na página 32.
- NELSON, J. D.; MULLEY, C. The impact of the application of new technology on public transport service provision and the passenger experience: A focus on implementation in Australia. *Res. Transp. Econ.*, Elsevier Ltd, v. 39, n. 1, p. 300–308, 2013. ISSN 07398859. Disponível em: <<http://dx.doi.org/10.1016/j.retrec.2012.06.028>>. Citado na página 14.
- NI, M.; HE, Q.; GAO, J. Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. *IEEE Trans. Intell. Transp. Syst.*, v. 18, n. 6, p. 1623–1632, 2016. ISSN 15249050. Citado 5 vezes nas páginas 47, 48, 49, 52 e 54.
- NIU, W. et al. Community-based geospatial tag estimation. In: IEEE. *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. [S.l.], 2016. p. 279–286. Citado na página 72.
- ROY, A.; MAJUMDER, A. G.; NATH, A. Understanding natural language processing and its primary aspects. *International Journal*, v. 5, n. 8, 2017. Citado na página 29.
- SANTOS, H. et al. Contextual data collection for smart cities. *CoRR*, abs/1704.01802, 2017. Disponível em: <<http://arxiv.org/abs/1704.01802>>. Citado 2 vezes nas páginas 20 e 21.
- SERAJ, F.; MERATNIA, N.; HAVINGA, P. J. An aggregation and visualization technique for crowd-sourced continuous monitoring of transport infrastructures. In: IEEE. *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*. [S.l.], 2017. p. 219–224. Citado na página 62.

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.*, n. 2011, 2017. Citado 2 vezes nas páginas 29 e 71.

SOBOLEVSKY, S. et al. Scaling of City Attractiveness for Foreign Visitors through Big Data of Human Economical and Social Media Activity. *Proc. - 2015 IEEE Int. Congr. Big Data, BigData Congr. 2015*, p. 600–607, 2015. ISSN 2379-7703. Citado 2 vezes nas páginas 44 e 45.

SOOMRO, K.; KHAN, Z.; HASHAM, K. Towards Provisioning of Real-time Smart City Services Using Clouds. *ACM 9th Int. Conf. Util. Cloud Comput. Towar.*, v. 1691, p. 50–59, 2016. ISSN 16130073. Citado 3 vezes nas páginas 43, 46 e 51.

STEIGER, E.; ALBUQUERQUE, J. P.; ZIPF, A. An advanced systematic literature review on spatiotemporal analyses of twitter data. *Transactions in GIS*, Wiley Online Library, v. 19, n. 6, p. 809–834, 2015. Citado na página 37.

STEIGER, E. et al. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Comput. Environ. Urban Syst.*, Elsevier Ltd, v. 54, p. 255–265, 2015. ISSN 01989715. Disponível em: <<http://dx.doi.org/10.1016/j.compenvurbssys.2015.09.007>>. Citado 7 vezes nas páginas 44, 45, 49, 50, 51, 53 e 70.

SÁ, T. H. et al. Health impact modelling of different travel patterns on physical activity, air pollution and road injuries for são paulo, brazil. *Environment International*, v. 108, n. Supplement C, p. 22 – 31, 2017. ISSN 0160-4120. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0160412017305974>>. Citado na página 12.

TALARI, S. et al. A Review of Smart Cities Based on the Internet of Things Concept. *Energies*, v. 10, n. 4, p. 421, 2017. ISSN 1996-1073. Disponível em: <<http://www.mdpi.com/1996-1073/10/4/421>>. Citado 2 vezes nas páginas 20 e 21.

THOMAZ, G. M. et al. Content mining framework in social media: A FIFA world cup 2014 case analysis. *Inf. Manag.*, Elsevier B.V., 2016. ISSN 03787206. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0378720616303354>>. Citado 2 vezes nas páginas 44 e 45.

United States Department of Transportation. *ITS Strategic Plan 2015-2019*. 2017. <<https://www.its.dot.gov/strategicplan.pdf>>. Acesso em Setembro, 17 de 2017. Citado na página 14.

WANG, S.; SINNOTT, R.; NEPAL, S. Privacy-protected social media user trajectories calibration. *Proc. 2016 IEEE 12th Int. Conf. e-Science, e-Science 2016*, p. 293–302, 2016. Citado 2 vezes nas páginas 19 e 45.

WEN, X.; LIN, Y.-R.; PELECHRINIS, K. PairFac: Event Analytics through Discriminant Tensor Factorization. *Cikm*, p. 519–528, 2016. Citado 5 vezes nas páginas 44, 47, 49, 50 e 52.

- WU, H.; YUAN, N. An improved tf-idf algorithm based on word frequency distribution information and category distribution information. In: ACM. *Proceedings of the 3rd International Conference on Intelligent Information Processing*. [S.l.], 2018. p. 211–215. Citado na página 35.
- XIAO, Z.; LIM, H. B.; PONNAMBALAM, L. Participatory Sensing for Smart Cities: A Case Study on Transport Trip Quality Measurement. *IEEE Trans. Ind. Informatics*, v. 13, n. 2, p. 759–770, 2017. ISSN 1551-3203. Citado 2 vezes nas páginas 20 e 21.
- YAHAV, I.; SHEHORY, O.; SCHWARTZ, D. Comments mining with tf-idf: The inherent bias and its removal. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, 2018. Citado na página 35.
- YANG, F. et al. Druid: A real-time analytical data store. In: ACM. *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. [S.l.], 2014. p. 157–168. Citado 2 vezes nas páginas 63 e 64.
- YOUAF, J. et al. Generalized multipath planning model for ride-sharing systems. *Front. Comput. Sci.*, v. 8, n. 1, p. 100–118, 2014. ISSN 20952228. Citado 5 vezes nas páginas 43, 46, 49, 50 e 51.
- ZAGAL, R.; MATA, F.; CLARAMUNT, C. Geographical Knowledge Discovery applied to the Social Perception of Pollution in the City of Mexico. *LBSN*, 2016. Citado 4 vezes nas páginas 44, 52, 53 e 71.
- ZHOU, X.; CHEN, L. Event detection over twitter social media streams. *The VLDB journal*, Springer, v. 23, n. 3, p. 381–400, 2014. Citado na página 28.

Apêndices

Apêndice A – Exemplos de tweets

Exemplos de tweets dos *profiles* selecionados citados na tabela 1

```

1 {
2     "tweet_id" : 895060642952077314,
3     "tweet_account": "BombeirosPMESP",
4     "text" : "19h58 Colisão de Carro x Caminhão, Estrada Sta Isabel,
5         5950 Itaquaquecetuba. 2 Vítimas, 1 Vtr. Aguardando maiores
6         informes"
7 }
8 {
9     "tweet_id" : 894707930217447427,
10    "tweet_account": "CETSP_",
11    "text" : "Referente manifestação Rua Augusta, pista liberada.#ZC"
12 }
13 {
14     "tweet_id" : 894147793060716544,
15     "tweet_account": "CPTM_oficial",
16     "text" : "#L11 Hoje, das 8h à meia-noite, circulação interrompida
17         entre Luz e Brás. P/ seguir viagem, use a L7-Rubi q prestará
18         serviço até a Est. Brás"
19 }
20 {
21     "tweet_id" : 895054721026838530,
22     "tweet_account": "governosp",
23     "text" : "@SANROGE Lamentamos o ocorrido, Rogerio. Estamos
         trabalhando continuamente para melhorar a segurança na região.
         Entre maio e junho, [+][1]"
24 }
25 {
26     "tweet_id" : 895000711284621312,
27     "tweet_account": "metrosps_oficial",

```

```
24     "text" : "08/08/2017 16:16: #metrosp : Linha 5-Lilás: Velocidade  
25       Reduzida. Mais informações em https://t.co/CaeqD26iJR"  
26   }  
27 {  
28     "tweet_id" : 884039273493803008,  
29     "tweet_account": "PMESP",  
30     "text" : "AGORA: Desfile Cívico-Militar de 9 de Julho no Obelisco  
31       - Ibirapuera SP, transmissão ao vivo na página oficial Facebook  
32       da Polícia Militar.",  
33     "dateTime" : "2017-07-09 10:19:22"  
34   }  
35 {  
36     "tweet_id" : 887315002117500932,  
37     "tweet_account": "Policia_Civil",  
38     "text" : "Policia Civil realiza operação para combater a prática  
39       do Jogo conhecido como "Baleia Azul"... https://t.co/kh2HW6UZvT  
40       ",  
41   }  
42 {  
43     "tweet_id" : 895004079910518788,  
44     "tweet_account": "saopaulo_agora",  
45     "text" : "#ItaimPaulista Incêndio na Rua Mateus Barbosa de Resende  
46       n° 235. Defesa Civil Regional acionada para o local. (CCOI) #  
47       spagora"  
48   }  
49 {  
50     "tweet_id" : 894694704989732864,  
51     "tweet_account": "smtpsp_",  
52     "text" : "A @sptrans_ irá modificar 14 linhas na Zona Leste para  
53       obras no Monotrilho Saiba mais: https://t.co/fCA0T7WCSY"  
54   }  
55 {
```

```
48     "tweet_id" : 902953598857949184,  
49     "tweet_account": "SPCEDEC",  
50     "text" : "30-08-2017 - Acidente com produto perigoso em com 36 ,  
51         deixa 21 vítimas feridas e 02 ."  
52     }  
53     {  
54         "tweet_id" : 895065137484320769,  
55         "tweet_account": "sptrans_ ",  
56         "text" : "Obras do Monotrilho desviam itinerários de 14 linhas que  
57             atendem a Av. Sapopemba entre 5 e 11/08, das 23h às 5h: https:  
58             //t.co/jH4LFgrSKZ"  
59     }  
60     {  
61         "tweet_id" : 895042604068458497,  
62         "tweet_account": "TurismoSaoPaulo",  
63         "text" : "Veganos, vegetarianos e simpatizantes: vem aí o Vegan  
64             Club, em 12/08, no Centro de SP! #crueltyfree #veganfood...  
65             https://t.co/7f7ggr4vn4"  
66     }
```

Apêndice B – Logradouros utilizados

Tabela 11 – Tabela de logradouros com abreviaturas

Abreviatura	Logradouro
ACAMP	Acampamento
AC	Acesso
AD	Adro
ERA	Aeroporto
AL	Alameda
AT	Alto
A	Area
AE	Area especial
ART	Arteria
ATL	Atalho
AV	Avenida
AV-CONT	Avenida contorno
BX	Baixa
BLO	Balao
BAL	Balneario
BC	Beco
BELV	Belvedere
BL	Bloco
BSQ	Bosque
BVD	Boulevard
BCO	Buraco
C	Cais
CALC	Calcada
CAM	Caminho
CPO	Campo
CAN	Canal
CHAP	Chacara

Continua na próxima página

Tabela 11 – continuação da página anterior

Abreviatura	Logradouro
CHAP	Chapadao
CIRC	Circular
COL	Colonia
CMP-VR	Complexo viario
COND	Condominio
CJ	Conjunto
COR	Corredor
CRG	Corrego
DSC	Descida
DSV	Desvio
DT	Distrito
EVD	Elevada
ENT-PART	Entrada particular
EQ	Entre quadra
ESC	Escada
ESP	Esplanada
ETC	Estacao
ESTC	Estacionamento
ETD	Estadio
ETN	Estancia
EST	Estrada
EST-MUN	Estrada municipal
FAV	Favela
FAZ	Fazenda
FRA	Feira
FER	Ferrovia
FNT	Fonte
FTE	Forte
GAL	Galeria

Continua na próxima página

Tabela 11 – continuação da página anterior

Abreviatura	Logradouro
GJA	Granja
HAB	Habitacional
IA	Ilha
JD	Jardim
JDE	Jardinete
LD	Ladeira
LG	Lago
LGA	Lagoa
LRG	Largo
LOT	Loteamento
MNA	Marina
MOD	Modulo
TEM	Monte
MRO	Morro
NUC	Nucleo
PDA	Parada
PDO	Paradouro
PAR	Paralela
PRQ	Parque
PSG	Passagem
PSC-SUB	Passagem subterranea
PSA	Passarela
PAS	Passeio
PAT	Patio
PNT	Ponta
PTE	Ponte
PTO	Porto
PC	Praca
PC-ESP	Praça de esportes

Continua na próxima página

Tabela 11 – continuação da página anterior

Abreviatura	Logradouro
PR	Praia
PRL	Prolongamento
Q	Quadra
QTA	Quinta
QTAS	Quinta
RAM	Rama
RMP	Rampa
REC	Recanto
RES	Residencial
RET	Reta
RER	Retiro
RTN	Retorno
ROD-AN	RodoAnel
ROD	Rodovia
RTT	Rotatoria
ROT	Rotula
R	Rua
R-LIG	Rua de ligação
R-PED	Rua de pedestre
SRV	Servidao
ST	Setor
SIT	Sitio
SUB	Subida
TER	Terminal
TV	Travessa
TV-PART	Travessa particular
TRV	Trecho
TRV	Trevo
TCH	Trincheira

Continua na próxima página

Tabela 11 – continuação da página anterior

Abreviatura	Logradouro
TUN	Tunel
UNID	Unidade
VAL	Vala
VLE	Vale
VRTE	Variante
VER	Vereda
V	Via
V-AC	Via de acesso
V-PED	Via de pedestre
V-EVD	Via elevado
V-EXP	Via expressa
VD	Viaduto
VLA	Viela
VL	Vila
ZIG-ZAG	Zigue-zague

Fonte: MS/SAS/DRAC/CGSI - Coordenação Geral dos Sistemas de Informação
 (adaptada)¹

¹ <http://www.pmf.sc.gov.br/arquivos/arquivos/pdf/04_01_2010_10.27.25.2b615e6755138defe1bdb00f1c86031f.PDF>. Acesso em Outubro, 29 de 2017.

Apêndice C – Detalhamento dos campos da GTFS

Tabela 12 – Detalhamento dos campos do arquivo *agency.txt* da GTFS

Nome do campo	Condisional	Descrição
<i>agency_id</i>	Opcional	Identifica uma agência de transporte público. Um <i>feed</i> de transporte público pode representar dados de mais de uma agência. Este campo é opcional para <i>feeds</i> de transporte público que contenham somente dados de uma única agência.
<i>agency_name</i>	Obrigatório	Contém o nome completo da agência de transporte público.
<i>agency_url</i>	Obrigatório	Contém o <i>URL</i> da agência de transporte público.
<i>agency_timezone</i>	Obrigatório	Contém o fuso horário de onde a agência de transporte público está localizada.
<i>agency_lang</i>	Opcional	Contém um código <i>ISO 639-1</i> de duas letras para o idioma principal usado por essa agência de transporte público.
<i>agency_phone</i>	Opcional	Contém um único número de telefone da agência especificada.
<i>agency_fare_url</i>	Opcional	Especifica o <i>URL</i> de uma página da Web que permite que um passageiro compre passagens ou outros instrumentos de tarifas dessa agência <i>on-line</i> .

Fonte: Google Transit (adaptada)¹

¹ <<https://developers.google.com/transit>>. Acesso em Outubro, 29 de 2017.

Tabela 13 – Detalhamento dos campos do arquivo
stops.txt da GTFS

Nome do campo	Condisional	Descrição
<i>stop_id</i>	Obrigatório	Contém um ID que identifica uma parada ou uma estação. Diversos trajetos podem usar a mesma parada.
<i>stop_code</i>	Opcional	Contém um pequeno texto ou um número que identifica a parada para os passageiros. Os códigos das paradas são usados muitas vezes em sistemas de informações sobre transporte público por telefone ou impressos em sinalizações nas paradas para que os passageiros possam obter informações sobre o horário das paradas com mais facilidade ou sobre chegadas de uma parada específica em tempo real. O campo <i>stop_code</i> só deve ser usado para códigos de parada exibidos aos passageiros. Para os códigos internos, use <i>stop_id</i> . Este campo deve ser deixado em branco para as paradas que não têm um código.
<i>stop_name</i>	Obrigatório	Contém o nome de uma parada ou estação. Use um nome compreensível para as pessoas locais e linguagem turística.
<i>stop_desc</i>	Opcional	Contém uma descrição de uma parada. Forneça informações úteis e de qualidade. Não basta repetir o nome da parada.
<i>stop_lat</i>	Obrigatório	Contém a latitude de uma parada ou estação. O valor do campo deve ser uma latitude WGS 84 válida.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>stop_lon</i>	Obrigatório	Contém a longitude de uma parada ou estação. O valor do campo deve ser uma latitude WGS 84 válida entre -180 e 180.
<i>zone_id</i>	Opcional	Define a zona tarifária do ID de uma parada. Os IDs de zonas são obrigatórios para fornecer informações sobre tarifas usando <i>fare_rules.txt</i> . Se esse ID de parada representa uma estação, o ID de zona é ignorado.
<i>stop_url</i>	Opcional	Contém o URL de uma página da Web sobre uma parada específica. Ele deve ser diferente dos campos <i>agency_url</i> e <i>route_url</i> .
<i>location_type</i>	Opcional	Identifica se este ID de parada representa uma parada ou uma estação. Se nenhum tipo de local for especificado ou se o campo <i>location_type</i> estiver em branco, os IDs de parada serão tratados como paradas. As estações podem ter propriedades diferentes das paradas quando são representadas em um mapa ou usadas em planejamento de viagens. O campo de tipo de local pode ter os seguintes valores: 0 ou em branco (para parada) e 1 (estação).

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>parent_station</i>	Opcional	Para paradas que estejam fisicamente localizadas dentro de estações, o campo <i>parent_station</i> identifica a estação associada à parada. Para usar este campo, o arquivo <i>stops.txt</i> também deve conter uma linha em que esse ID de parada tenha o tipo de localização=1.

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>stop_timezone</i>	Opcional	<p>Contém o fuso horário em que a parada ou estação está localizada. Se omitido, assume-se que a parada está localizada no fuso horário especificado por <i>agency_timezone</i> no arquivo <i>agency.txt</i>.</p> <p>Quando uma parada tem uma estação principal, considera-se que a parada esteja no fuso horário especificado pelo valor <i>stop_timezone</i> da estação principal. Se uma parada específica possui um valor <i>parent_station</i>, qualquer valor <i>stop_timezone</i> especificado para essa parada deve ser ignorado. Mesmo que os valores de <i>stop_timezone</i> sejam fornecidos no arquivo <i>stops.txt</i>, os horários em <i>stop_times.txt</i> devem continuar a ser especificados como horários desde a meia-noite no fuso horário especificado por <i>agency_timezone</i> em <i>agency.txt</i>. Isso garante que os valores de tempo em uma viagem sempre aumentam durante uma viagem, independentemente dos fusos horários pelos quais uma viagem passa.</p>

Continua na próxima página

Tabela 13 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>wheelchair_boarding</i>	Opcional	<p>Identifica se é possível o embarque de passageiros em cadeira de rodas na parada ou estação especificada. O campo pode ter os seguintes valores: 0 (ou vazio) - indica que não há informações sobre acessibilidade para a parada; 1 - indica que, pelo menos, alguns veículos nesta parada possibilitam o embarque de passageiros em cadeira de rodas; 2 - o embarque de pessoas em cadeiras de roda não é possível nesta parada. Quando uma parada faz parte de um complexo de estações maiores, como indicado por uma para com um valor <i>parent_station</i>, o campo <i>wheelchair_boarding</i> da parada possui a seguinte semântica adicional: 0 (ou vazio) - a parada herdará o valor para <i>wheelchair_boarding</i> da estação principal, se especificado; 1 - existem vias de acesso na parte externa da estação para a parada/plataforma específica; 2 - não há vias de acesso na parte externa da estação para a parada/plataforma específica</p>

Fonte: Google Transit (adaptada)¹

Tabela 14 – Detalhamento dos campos do arquivo *routes.txt* da GTFS

Nome do campo	Condisional	Descrição
<i>route_id</i>	Obrigatório	Contém um ID que identifica um trajeto.
<i>agency_id</i>	Opcional	Define uma agência para o trajeto especificado. Este valor é indicado no arquivo <i>agency.txt</i> . Campo destinado para quando for fornecido dados para trajetos de mais de uma agência.
<i>route_short_name</i>	Obrigatório	Contém o nome abreviado de um trajeto. Geralmente, será um identificador pequeno e abstrato, como, por exemplo "32", "100X" ou "Verde", que os passageiros usam para identificar um trajeto, mas que não fornece nenhuma identificação de quais lugares são atendidos pelo trajeto. Se o trajeto não tem um nome abreviado, especifique um <i>route_long_name</i> e use uma sequência vazia como o valor deste campo.
<i>route_long_name</i>	Obrigatório	Contém o nome completo de um trajeto. Em geral, esse nome é mais descritivo que <i>route_short_name</i> e incluirá o destino ou a parada do trajeto. Se o trajeto não tem um nome completo, especifique um <i>route_short_name</i> e use uma sequência vazia como o valor deste campo.
<i>route_desc</i>	Opcional	Contém uma descrição de um trajeto. Não basta repetir o nome do trajeto.

Continua na próxima página

Tabela 14 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>route_type</i>	Obrigatório	Descreve o tipo de transporte usado em um trajeto. Os valores válidos deste campo são: 0 - Bonde, ônibus elétrico, veículo leve sobre trilhos; 1 - Metrô, trem subterrâneo; 2 - Via férrea; 3 - Ônibus; 4 - Balsa; 5 - Teleférico; 6 - Gôndola, teleférico suspenso; 7 - Funicular.
<i>route_url</i>	Opcional	Contém o URL de uma página da Web sobre esse trajeto específico. Ele deve ser diferente de <i>agency_url</i> .
<i>route_color</i>	Opcional	Define uma cor que corresponda ao trajeto. A cor deve ser informada como um número hexadecimal de seis caracteres. Se nenhuma cor é especificada, a cor padrão de trajetos é branca (FFFFFF). A diferença de cores entre <i>route_color</i> e <i>route_text_color</i> deve fornecer contraste suficiente quando visualizado em uma tela em preto e branco.
<i>route_text_color</i>	Opcional	Usado para especificar uma cor legível para usar em desenho de texto contra um plano de fundo de <i>route_color</i> .

Fonte: Google Transit (adaptada)¹

Tabela 15 – Detalhamento dos campos do arquivo
trips.txt da GTFS

Nome do campo	Condisional	Descrição
<i>route_id</i>	Obrigatório	Contém um ID que identifica um trajeto. Este valor é indicado no arquivo <i>agency.txt</i> .
<i>service_id</i>	Obrigatório	Contém um ID que identifica um conjunto de datas em que o serviço está disponível para um ou mais trajetos. Este valor é indicado no arquivo <i>calendar.txt</i> ou <i>calendar_dates.txt</i> .
<i>trip_id</i>	Obrigatório	Contém um ID que identifica uma viagem.
<i>trip_headsign</i>	Opcional	Contém o texto que aparece em uma sinalização que identifica o destino da viagem para os passageiros. Use este campo para distinguir diferentes padrões de serviço no mesmo trajeto. Se a placa muda durante uma viagem, você pode substituir o campo <i>trip_headsign</i> , especificando valores para o campo <i>stop_headsign</i> em <i>stop_times.txt</i> .
<i>trip_short_name</i>	Opcional	Contém o texto que aparece em programações e placas de sinalização para identificar a viagem para os passageiros, por exemplo, para identificar números de trens para viagens de trens suburbanos. Se os passageiros não recorrem normalmente aos nomes da viagem, deixe este campo em branco. Um valor de <i>trip_short_name</i> , se possível, deve identificar, com exclusividade, uma viagem em um dia de serviço; ele não deve ser usado para nomes de destino ou designações limitadas/expressas.

Continua na próxima página

Tabela 15 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>direction_id</i>	Opcional	Contém um valor binário que indica a direção de uma viagem. Use este campo para distinguir viagens bidirecionais com o mesmo <i>route_id</i> . Este campo não é usado na criação de trajetos; ele fornece uma maneira de separar viagens por direção durante a publicação de tabelas de horário. Você pode especificar nomes para cada direção com o campo <i>trip_headsign</i> . 0 - viagem em uma única direção (por exemplo, só ida); 1 - viagem na direção oposta (por exemplo, de volta), os campos <i>trip_headsign</i> e <i>direction_id</i> podem ser usados juntos para atribuir um nome a uma viagem em cada direção "1234".
<i>block_id</i>	Opcional	Identifica o quadro a que a viagem pertence. Um bloco consiste em duas ou mais viagens sequenciais feitas usando o mesmo veículo, em que um passageiro pode passar de uma viagem para a próxima permanecendo no veículo. O campo <i>block_id</i> deve ser indicado por duas ou mais viagens no arquivo <i>trips.txt</i> .
<i>shape_id</i>	Opcional	Contém um ID que define a forma da viagem. Este valor é indicado no arquivo <i>shapes.txt</i> . O arquivo <i>shapes.txt</i> permite definir como será traçada uma linha no mapa para representar uma viagem.

Continua na próxima página

Tabela 15 – continuação da página anterior

Nome do campo	Condisional	Descrição
wheelchair_accessible	Opcional	0 (ou vazio) - indica que não há informações sobre acessibilidade para a viagem; 1 - indica que o veículo que está sendo usado nesta viagem específica pode acomodar, pelo menos, um passageiro em cadeira de rodas; 2 - indica que não é possível acomodar passageiros em cadeiras de rodas nesta viagem

Fonte: Google Transit (adaptada)¹

Tabela 16 – Detalhamento dos campos do arquivo

stop_times.txt da GTFS

Nome do campo	Condisional	Descrição
<i>trip_id</i>	Obrigatório	Contém um ID que identifica uma viagem. Este valor é indicado no arquivo <i>trips.txt</i> .

Continua na próxima página

Tabela 16 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>arrival_time</i>	Obrigatório	<p>Especifica o horário de chegada em uma parada específica de uma viagem específica de um trajeto. No caso de horários que ocorram após a meia-noite na data do serviço, digite o horário como um valor maior que 24:00:00 em horário local HH:MM:SS para o dia em que começa a programação da viagem. Se não há horários separados para chegada e partida em uma parada, insira o mesmo valor para <i>arrival_time</i> e <i>departure_time</i>. É necessário especificar os horários de chegada para a primeira e a última paradas de uma viagem.</p> <p>Se essa parada não for programada, use uma sequência vazia para os campos <i>arrival_time</i> e <i>departure_time</i>. As paradas sem horário de chegada são programadas conforme a parada programada anterior mais próxima. Para garantir trajetos precisos, forneça horários de chegada e de partida para todas as paradas programadas.</p> <p>Não intercale as paradas, ou, preencha os horários com espaços. Observação: as viagens que abrangem várias datas terão horários de parada maiores que 24:00:00. Por exemplo, se uma viagem começa às 10:30:00 p.m e termina às 2:15:00 a.m. do dia seguinte, os horários de parada seriam 22:30:00 e 26:15:00. A inclusão desses horários de parada como 22:30:00 e 02:15:00 não produzem os resultados desejados.</p>

Tabela 16 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>departure_time</i>	Obrigatório	<p>Especifica o horário de partida de uma parada específica para uma viagem específica de um trajeto. O horário é medido de "meio-dia menos 12h"(efetivamente meia-noite, exceto para dias do horário de verão), no início da data do serviço. No caso de horários que ocorram após a meia-noite na data do serviço, digite o horário como um valor maior que 24:00:00 em horário local HH:MM:SS para o dia em que começa a programação da viagem. Se não há horários diferentes para a chegada e a saída em uma parada, insira o mesmo valor para <i>arrival_time</i> e <i>departure_time</i>. É necessário especificar os horários de partida da primeira e da última paradas em uma viagem. Se essa parada não for programada, use uma sequência vazia para os campos <i>arrival_time</i> e <i>departure_time</i>. As paradas sem horário de chegada são programadas conforme a parada programada anterior mais próxima. Para garantir trajetos precisos, forneça horários de chegada e de partida para todas as paradas programadas. Não intercale as paradas. Os horários devem ter oito dígitos no formato HH:MM:SS (o formato H:MM:SS também é aceito, se a hora iniciar com 0). Não preencha os horários com espaços.</p>

Continua na próxima página

Tabela 16 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>stop_id</i>	Obrigatório	Contém um ID que identifica uma parada. Diversos trajetos podem usar a mesma parada. O campo <i>stop_id</i> é indicado no arquivo <i>stops.txt</i> . Se <i>location_type</i> é usado no arquivo <i>stops.txt</i> , todas as paradas indicadas em <i>stop_times.txt</i> deverão ter <i>location_type</i> igual a 0. Onde possível, os valores de <i>stop_id</i> devem permanecer consistentes entre as atualizações de feed. Se uma parada não está programada, digite valores em branco para <i>arrival_time</i> e <i>departure_time</i> .
<i>stop_sequence</i>	Obrigatório	Identifica a ordem das paradas de uma viagem específica. Os valores de <i>stop_sequence</i> devem ser números inteiros positivos e devem aumentar ao longo da viagem.
<i>stop_headsign</i>	Opcional	Contém o texto que aparece em uma sinalização que identifica o destino da viagem para os passageiros. Use este campo para substituir o <i>trip_headsign</i> padrão quando as placas mudarem durante as viagens. Se esta placa está associada a uma viagem inteira, use <i>trip_headsign</i> no lugar.

Continua na próxima página

Tabela 16 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>pickup_type</i>	Opcional	Indica se os passageiros são embarcados em uma parada como parte da programação normal ou se não há embarque disponível na parada. Este campo também permite que a agência de transporte público indique se os passageiros devem ligar para a agência ou notificar o motorista para agendar um embarque em uma parada específica. Os valores válidos deste campo são: 0 - Embarque no horário normal; 1 - Sem embarque disponível; 2 - Deve ligar para a agência a fim de agendar o embarque; 3 - Deve combinar com o motorista para agendar o embarque. O valor padrão deste campo é 0.

Continua na próxima página

Tabela 16 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>drop_off_type</i>	Opcional	Indica se há desembarque de passageiros em uma parada, como parte da programação normal ou se não há desembarques na parada. Este campo também permite que a agência de transporte público indique se os passageiros devem ligar para a agência ou notificar o motorista para agendar um desembarque em uma determinada parada. Os valores válidos deste campo são: 0 - Desembarque no horário normal; 1 - Desembarque não disponível; 2 - Deve telefonar para agendar o desembarque; 3 - Deve combinar com o motorista para agendar o desembarque. O valor padrão deste campo é 0.

Continua na próxima página

Tabela 16 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>shape_dist_traveled</i>	Opcional	<p>Quando usado no arquivo <i>stop_times.txt</i>, o campo <i>shape_dist_traveled</i> posiciona uma parada como uma distância a partir do primeiro ponto de forma. O campo <i>shape_dist_traveled</i> representa uma distância real percorrida ao longo do trajeto em unidades como, por exemplo, pés ou quilômetros. Essas informações permitem que o planejador da viagem determine o quanto da forma deve ser desenhado ao exibir parte de uma viagem no mapa. Os valores usados para <i>shape_dist_traveled</i> devem aumentar juntamente com <i>stop_sequence</i>. As unidades usadas para <i>shape_dist_traveled</i> no arquivo <i>stop_times.txt</i> devem corresponder às unidades usadas para este campo no arquivo <i>shapes.txt</i>.</p>

Fonte: Google Transit (adaptada)¹

Tabela 17 – Detalhamento dos campos do arquivo *calendar.txt* da GTFS

Nome do campo	Condisional	Descrição
<i>service_id</i>	Obrigatório	Contém um ID que identifica um conjunto de datas em que o serviço está disponível para um ou mais trajetos. Cada valor de <i>service_id</i> pode aparecer, no máximo, uma vez em um arquivo <i>calendar.txt</i> . Este valor é um conjunto de dados exclusivo. Ele é indicado pelo arquivo <i>trips.txt</i> .
<i>monday</i>	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas as segundas-feiras. O valor 1 indica que o serviço está disponível todas as segundas-feiras durante o período. O período é especificado utilizando-se os campos <i>start_date</i> e <i>end_date</i> . O valor 0 indica que o serviço não está disponível às segundas-feiras no período. Observação: você pode listar exceções para datas específicas, como, por exemplo, feriados, no arquivo <i>calendar_dates.txt</i> .
<i>tuesday</i>	Obrigatório	Contém um valor binário que indica se o serviço é válido para todas as terças-feiras. O valor 1 indica que o serviço está disponível todas as terças-feiras durante o período. O período é especificado utilizando-se os campos <i>start_date</i> e <i>end_date</i> . O valor 0 indica que o serviço não está disponível às terças-feiras no período.

Continua na próxima página

Tabela 17 – continuação da página anterior

Nome do campo	Condisional	Descrição
wednesday	Obrigatório	<p>Contém um valor binário que indica se o serviço é válido para todas as quartas-feiras.</p> <p>O valor 1 indica que o serviço está disponível todas as quartas-feiras durante o período. O período é especificado utilizando-se os campos <i>start_date</i> e <i>end_date</i>. O valor 0 indica que o serviço não está disponível às quartas-feiras no período.</p>
thursday	Obrigatório	<p>Contém um valor binário que indica se o serviço é válido para todas as quintas-feiras.</p> <p>O valor 1 indica que o serviço está disponível todas as quintas-feiras durante o período. O período é especificado utilizando-se os campos <i>start_date</i> e <i>end_date</i>. O valor 0 indica que o serviço não está disponível às quintas-feiras no período.</p>
friday	Obrigatório	<p>Contém um valor binário que indica se o serviço é válido para todas as sextas-feiras.</p> <p>O valor 1 indica que o serviço está disponível todas as sextas-feiras durante o período. O período é especificado utilizando-se os campos <i>start_date</i> e <i>end_date</i>. O valor 0 indica que o serviço não está disponível às sextas-feiras no período.</p>

Continua na próxima página

Tabela 17 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>saturday</i>	Obrigatório	Contém um valor binário que indica se o serviço é válido para todos os sábados. O valor 1 indica que o serviço está disponível todos os sábados durante o período. O período é especificado utilizando-se os campos <i>start_date</i> e <i>end_date</i> . O valor 0 indica que o serviço não está disponível aos sábados no período.
<i>sunday</i>	Obrigatório	Contém um valor binário que indica se o serviço é válido para todos os domingos. O valor 1 indica que o serviço está disponível todos os domingos durante o período. O período é especificado utilizando-se os campos <i>start_date</i> e <i>end_date</i> . O valor 0 indica que o serviço não está disponível aos sábados no período.
<i>start_date</i>	Obrigatório	O campo <i>start_date</i> contém a data de início do serviço. O valor do campo <i>start_date</i> deve estar no formato YYYYMMDD.
<i>end_date</i>	Obrigatório	O campo <i>end_date</i> contém a data final do serviço. Essa data está incluída no intervalo do serviço. O valor do campo <i>end_date</i> deve estar no formato AAAAMMDD.

Fonte: Google Transit (adaptada)¹

Tabela 18 – Detalhamento dos campos do arquivo *calendar_dates.txt* da GTFS

<i>service_id</i>	Obrigatório	Contém um ID que identifica um conjunto de datas em que uma exceção ao serviço está disponível para um ou mais trajetos. Cada par (<i>service_id</i> , <i>date</i>) pode aparecer somente uma vez em <i>calendar_dates.txt</i> . Se um valor de <i>service_id</i> aparece nos arquivos <i>calendar.txt</i> e <i>calendar_dates.txt</i> , as informações contidas em <i>calendar_dates.txt</i> modifica as informações de serviço especificadas em <i>calendar.txt</i> . Este campo é indicado pelo arquivo <i>trips.txt</i> .
<i>date</i>	Obrigatório	Especifica uma data específica em que a disponibilidade do serviço é diferente do normal. Você pode usar o campo <i>exception_type</i> para indicar se o serviço está disponível na data especificada. O valor do campo <i>date</i> deve estar no formato AAAAMMDD.
<i>exception_type</i>	Obrigatório	Indica se o serviço está disponível na data especificada no arquivo <i>date</i> . O valor 1 indica que o serviço foi adicionado para a data especificada. O valor 2 indica que o serviço foi removido para a data especificada.

Fonte: Google Transit (adaptada)¹

Tabela 19 – Detalhamento dos campos do arquivo *fare_attributes.txt* da GTFS

<i>fare_id</i>	Obrigatório	Contém um ID que identifica uma classe de tarifas.
<i>price</i>	Obrigatório	Contém o preço da tarifa, na unidade especificada por <i>currency_type</i> .
<i>currency_type</i>	Obrigatório	Define a moeda usada para pagar a tarifa. Use os códigos de moeda em ordem alfabética ISO 4217.
<i>payment_method</i>	Obrigatório	Indica quando a tarifa deve ser paga. Os valores válidos deste campo são: 0 - A tarifa é paga a bordo; 1 - A tarifa deve ser paga antes do embarque.
<i>transfers</i>	Obrigatório	O campo <i>transfers</i> especifica o número de baldeações permitidas nesta tarifa. Os valores válidos deste campo são: 0 - Não são permitidas baldeações nesta tarifa; 1 - Os passageiros só podem fazer uma baldeação; 2 - Os passageiros podem fazer duas baldeações; (empty) - Se o campo estiver vazio, não há limites para o número de baldeações.
<i>transfer_duration</i>	Opcional	Especifica a duração, em segundos, antes da expiração da baldeação. Quando usado com um valor 0 para <i>transfers</i> , o campo <i>transfer_duration</i> indica por quanto tempo uma passagem é válida para uma tarifa quando as baldeações não são permitidas. A menos que você pretenda usar este campo para indicar a validade da passagem, <i>transfer_duration</i> deve ser omitido ou deve ficar em branco, quando <i>transfers</i> é definido como 0.

Fonte: Google Transit (adaptada)¹

Tabela 20 – Detalhamento dos campos do arquivo *fare_rules.txt* da GTFS

<i>fare_id</i>	Obrigatório	Contém um ID que identifica uma classe de tarifas. Este valor é indicado no arquivo <i>fare_attributes.txt</i> .
<i>route_id</i>	Opcional	Associa o ID da tarifa a um trajeto. Os IDs de trajetos são indicados no arquivo <i>routes.txt</i> . Se você tem diversos trajetos com os mesmos atributos de tarifa, crie uma linha no arquivo <i>fare_rules.txt</i> para cada trajeto.
<i>origin_id</i>	Opcional	Associa o ID da tarifa a um ID de zona de origens. Os IDs de zona são indicados no arquivo <i>stops.txt</i> . Se há vários IDs de origem com os mesmos atributos, crie uma linha no arquivo <i>fare_rules.txt</i> para cada ID de origem.
<i>destination_id</i>	Opcional	Associa o ID da tarifa a um ID de zona de destino. IDs de zona são indicados no arquivo <i>stops.txt</i> . Se há vários IDs de destino com os mesmos atributos de tarifa, cria-se uma linha no arquivo <i>fare_rules.txt</i> para cada ID de destino.
<i>contains_id</i>	Opcional	Associa o ID da tarifa a um ID de zona ID, indicado no arquivo <i>stops.txt</i> . O ID da tarifa é, então, associado a itinerários que transmitem cada zona de <i>contains_id</i> .

Fonte: Google Transit (adaptada)¹

Tabela 21 – Detalhamento dos campos do arquivo *shapes.txt* da GTFS

<i>shape_id</i>	Obrigatório	Contém um ID que identifica uma forma.
<i>shape_pt_lat</i>	Obrigatório	Associa a latitude de um ponto de forma ao ID de uma forma. O valor do campo deve ser uma latitude WGS 84 válida. Cada linha do arquivo <i>shapes.txt</i> representa um ponto de forma em sua definição de formas.
<i>shape_pt_lon</i>	Obrigatório	Associa a longitude de um ponto de forma ao ID de uma forma. O valor do campo deve ser uma longitude WGS 84 de valor de -180 a 180. Cada linha do arquivo <i>shapes.txt</i> representa um ponto de forma em sua definição de formas.
<i>shape_pt_sequence</i>	Obrigatório	Associa a latitude e a longitude de uma forma de um ponto de formas com sua ordem sequencial juntamente com a forma. Os valores de <i>shape_pt_sequence</i> devem ser números inteiros positivos e devem aumentar com a viagem.
<i>shape_dist_traveled</i>	Opcional	Quando usado no arquivo <i>shapes.txt</i> , o campo <i>shape_dist_traveled</i> posiciona um ponto de forma como uma distância percorrida juntamente com uma forma a partir do primeiro ponto de forma. O campo <i>shape_dist_traveled</i> representa uma distância real percorrida ao longo do trajeto em unidades como, por exemplo, pés ou quilômetros. Esta informação permite que o planejador de viagens determine o quanto da forma deve ser desenhado ao mostrar parte de uma viagem no mapa. Os valores usados para <i>shape_dist_traveled</i> devem aumentar juntamente com <i>shape_pt_sequence</i> . As unidades usadas para <i>shape_dist_traveled</i> no arquivo <i>shapes.txt</i> devem corresponder às unidades usadas para este campo no arquivo <i>stop_times.txt</i> .

Fonte: Google Transit (adaptada)¹

Tabela 22 – Detalhamento dos campos do arquivo *frequencies.txt* da GTFS

Nome do campo	Condisional	Descrição
<i>trip_id</i>	Obrigatório	Contém um ID que identifica uma viagem à qual a frequência especificada de serviço se aplica. Os IDs de viagem são indicados no arquivo <i>trips.txt</i> .
<i>start_time</i>	Obrigatório	Especifica o horário em que o serviço começa com a freqüência especificada. Para horários após a meia-noite, insira-os como um valor maior que 24:00:00 no horário local HH:MM:SS para o dia em que a programação das viagens começa.
<i>end_time</i>	Obrigatório	Especifica o horário em que o serviço muda para uma frequência diferente (ou é interrompido), na primeira parada da viagem. Para horários após a meia-noite, insira-os como um valor maior que 24:00:00 no horário local HH:MM:SS para o dia em que a programação das viagens começa.

Tabela 22 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>headway_secs</i>	Obrigatório	<p>Indica o horário entre as saídas da mesma parada (intervalo entre as viagens) deste tipo de viagem, durante o intervalo de tempo especificado por <i>start_time</i> e <i>end_time</i>. O valor do intervalo de tempo entre duas viagens deve ser inserido em segundos.</p> <p>Períodos em que intervalos entre as viagens são definidos (as linhas no arquivo <i>frequencies.txt</i>) não devem ser sobrepostos para a mesma viagem, uma vez que é difícil determinar o que deve ser inferido de dois intervalos de viagem sobrepostos. No entanto, um período de intervalo entre viagens pode começar exatamente no mesmo horário em que outro termina.</p>

Tabela 22 - continuação da página anterior

Nome do campo	Condisional	Descrição
<i>exact_times</i>	Opcional	<p>Determina se viagens baseadas em frequência devem ser programadas com exatidão com base nas informações especificadas dos intervalos entre as viagens.</p> <p>Os valores válidos deste campo são: 0 ou (vazio) - Viagens baseadas em frequência não são programadas com exatidão. Este é o comportamento padrão; 1 - Viagens baseadas em frequência são programadas com exatidão. Para uma linha no <i>frequencies.txt</i>, as viagens são programadas com início com <i>trip_start_time =start_time + x * headway_secs</i> para todos x em (0, 1, 2, ...), em que <i>trip_start_time < end_time</i>. O valor de <i>exact_times</i> deve ser o mesmo para todas as linhas de <i>frequencies.txt</i> com o mesmo <i>trip_id</i>. Se <i>exact_times</i> for igual a 1, e uma linha de <i>frequencies.txt</i> tiver um <i>start_time</i> igual a <i>end_time</i>, nenhuma viagem deverá ser programada. Quando <i>exact_times</i> é 1, deve-se escolher um valor <i>end_time</i> que seja maior que o último horário de início da viagem programada, mas menor que o último horário de início da viagem desejada + <i>headway_secs</i>.</p>

Fonte: Google Transit (adaptada)¹

Tabela 23 – Detalhamento dos campos do arquivo
transfer.txt da GTFS

Nome do campo	Condisional	Descrição
<i>from_stop_id</i>	Obrigatório	Contém um ID que identifica uma parada ou uma estação onde começa uma conexão entre trajetos. Os IDs de paradas são indicados no arquivo <i>stops.txt</i> . Se a ID de parada se refere a uma estação que contém várias paradas, essa regra de baldeação se aplica a todas as paradas nesta estação.
<i>to_stop_id</i>	Obrigatório	Contém um ID que identifica uma parada ou uma estação onde termina uma conexão entre trajetos. Os IDs de paradas são indicados no arquivo <i>stops.txt</i> . Se a ID de parada se refere a uma estação que contém várias paradas, essa regra de baldeação se aplica a todas as paradas nesta estação.
<i>transfer_type</i>	Obrigatório	Especifica o tipo de conexão para o par (<i>from_stop_id, to_stop_id</i>) especificado. Os valores válidos deste campo são: 0 ou (vazio) <ul style="list-style-type: none"> - Este é um ponto de baldeação recomendado entre dois trajetos; 1 - Este é um ponto de baldeação programado entre dois trajetos; 2 - Essa baldeação exige um tempo mínimo entre a chegada e a partida para garantir uma conexão. O tempo necessário para a baldeação é especificado por <i>min_transfer_time</i>; 3 - Não é possível fazer baldeações entre trajetos neste local.

Continua na próxima página

Tabela 23 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>min_transfer_time</i>	Opcional	Quando uma conexão entre trajetos exige um tempo entre a chegada e a partida (<i>transfer_type=2</i>), o campo <i>min_transfer_time</i> define o período de tempo que deve estar disponível em um itinerário para permitir uma baldeação entre trajetos nestas paradas. O <i>min_transfer_time</i> deve ser suficiente para que um passageiro típico se desloque entre as duas paradas, incluindo um tempo extra para variação na programação em cada trajeto. O valor de <i>min_transfer_time</i> deve ser inserido em segundos e deve ser um número inteiro positivo.

Fonte: Google Transit (adaptada)¹

Tabela 24 – Detalhamento dos campos do arquivo

feed_info.txt da GTFS

Nome do campo	Condisional	Descrição
<i>feed_publisher_name</i>	Obrigatório	Contém o nome completo da organização que publica o <i>feed</i> . Pode ser o mesmo que aquele definido pelos valores de <i>agency_name</i> no arquivo <i>agency.txt</i> . Aplicativos que utilizam GTFS podem exibir este nome ao concederem atribuições relacionadas aos dados de um <i>feed</i> específico.
<i>feed_publisher_url</i>	Obrigatório	Contém o URL do website da organização que está publicando o <i>feed</i> . Pode ser o mesmo que um dos valores de <i>agency_url</i> no arquivo <i>agency.txt</i> .
<i>feed_lang</i>	Obrigatório	Contém um código de idiomas IETF BCP 47 que especifica o idioma padrão usado para o texto neste <i>feed</i> . Esta configuração ajuda os consumidores de GTFS a escolherem regras para o uso de letras maiúsculas e minúsculas e outras configurações específicas do idioma para o <i>feed</i> .

Continua na próxima página

Tabela 24 – continuação da página anterior

Nome do campo	Condisional	Descrição
<i>feed_start_date / feed_end_date</i>	Opcional	<p>O <i>feed</i> fornece informações completas e confiáveis sobre a programação de um serviço, no período entre o início do dia <i>feed_start_date</i> e o final do dia <i>feed_end_date</i>. As datas nos dois dias estão no formato AAAAMMDD, assim como no arquivo <i>calendar.txt</i>, ou são deixadas em branco se não estiverem disponíveis. A data <i>feed_end_date</i> não deve preceder a data <i>feed_start_date</i>, se ambas forem fornecidas.</p> <p>Os provedores de feeds são encorajados a oferecerem dados de programação fora desse período a fim de informarem sobre possíveis serviços no futuro, mas os consumidores de <i>feed</i> devem estar conscientes de seu status não autorizado. Se <i>feed_start_date</i> ou <i>feed_end_date</i> se estendem além das datas do calendário ativo definidas nos arquivos <i>calendar.txt</i> e <i>calendar_dates.txt</i>, o <i>feed</i> se torna uma afirmação explícita de que não há serviços para as datas entre <i>feed_start_date</i> ou <i>feed_end_date</i> que não estão incluídas nas datas do calendário ativo.</p>

Continua na próxima página

Tabela 24 - continuação da página anterior

Nome do campo	Condisional	Descrição
<i>feed_version</i>	Opcional	O editor de <i>feeds</i> pode especificar uma sequência que indique a versão atual do <i>feed</i> GTFS. Os aplicativos que utilizam GTFS podem exibir este valor para ajudar os editores de <i>feed</i> a determinar se foi incorporada a versão mais recente do <i>feed</i> .

Fonte: Google Transit (adaptada)¹

Apêndice D – Linhas de ônibus impactadas por eventos de exceção

Tabela 25 – Linhas de ônibus impactadas por eventos de exceção

Código da linha	Total de eventos de exceção	Letreiro
35276	1663	PÇA. RAMOS DE AZEVEDO / TERM. CAMPO LIMPO
32953	1606	TERM. PINHEIROS / TERM. JD. ÂNGELA
32805	1462	TERM. PRINC. ISABEL / CHÁC. SANTANA
32772	1423	TERM. PRINC. ISABEL / TERM. STO. AMARO
32846	1420	METRÔ BRÁS / TERM. GRAJAÚ
32869	1339	PINHEIROS / GRAJAÚ
34134	1335	METRÔ ANA ROSA / MORRO GRANDE
33966	1296	METRÔ VL. MARIANA / TERM. PARELHEIROS
33943	1295	TERM. STO. AMARO / VARGEM GRANDE
34694	1245	PARAÍSO / TERM. CAMPO LIMPO
32874	1210	METRÔ JABAQUARA / PQ. RES. COCAIA
34660	1184	ACLIMAÇÃO / TERM. CAMPO LIMPO
33272	1150	PÇA. RAMOS DE AZEVEDO / JD. JOÃO XXIII
34210	1129	LGO. SÃO FRANCISCO / TERM. VARGINHA
32838	1105	PÇA. DA SÉ / PQ. RES. COCAIA
32829	1076	TERM. BANDEIRA / TERM. CAPELINHA
34233	1061	TERM. BANDEIRA / TERM. VARGINHA
33640	1056	TERM. CACHOEIRINHA / JD. PRINCESA
33377	1054	PERDIZES / AEROPORTO
33042	1022	PÇA. DA SÉ / JD. IV CENTENÁRIO
34209	953	METRÔ JABAQUARA / TERM. VARGINHA
34144	937	PÇA. DA SÉ / CID. UNIVERSITÁRIA
32827	934	TERM. BANDEIRA / TERM. CAPELINHA
33675	892	ITAIM PAULISTA / VL. CALIFÓRNIA
33258	875	LGO. DA PÓLVORA / JD. MARIA LUIZA
33275	871	METRÔ ANA ROSA / JD. GUARAÚ
33382	840	METRÔ SANTANA / CPTM JARAGUÁ
34861	834	METRÔ STA. CECÍLIA / TERM. STO. AMARO
33075	831	LAPA / IPIRANGA
34149	821	METRÔ PARAÍSO / VL. ANASTÁCIO

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34098	818	TERM. PQ. D. PEDRO II / CID. UNIVERSITÁRIA
33253	789	METRÔ BELÉM / JD. BONFIGLIOLI
34085	771	TERM. BANDEIRA / JD. VAZ DE LIMA
33123	759	TERM. BANDEIRA / TERM. STO. AMARO
35229	738	TURISMO / CIRCULAR
35174	733	TERM. PQ. D. PEDRO II / TERM. STO. AMARO
32871	721	PINHEIROS / VL. SÃO JOSÉ
33908	706	TERM. STO. AMARO / TERM. PARELHEIROS
35160	697	TERM. PQ. D. PEDRO II / TERM. GRAJAÚ
32826	694	TERM. PQ. D. PEDRO II / TERM. JOÃO DIAS
35148	676	METRÔ VL. MADALENA / TERM. SACOMÃ
33224	667	METRÔ VL. MARIANA / TERM. PIRITUBA
32837	664	PÇA. DO CORREIO / SESC/ORION
32849	663	LGO. SÃO FRANCISCO / VL. SÃO JOSÉ
34108	660	METRÔ VL. MARIANA / TERM. LAPA
33117	649	POMPÉIA ATÉ VL. ROMANA / SACOMÃ
33277	646	TERM. PRINC. ISABEL / COHAB RAPOSO TAVARES
32769	645	LGO. SÃO FRANCISCO / TERM. CAPELINHA
33357	641	METRÔ ANA ROSA / VL. BRASILÂNDIA
34109	641	METRÔ ANA ROSA / METRÔ BARRA FUNDA
34907	641	SOCORRO / JD. APURÁ
33443	640	ANA ROSA / METRÔ SANTANA
35196	639	TERM. PQ. D. PEDRO II / METRÔ BUTANTÃ
32816	621	TERM. PQ. D. PEDRO II / TERM. STO. AMARO
32814	613	TERM. BANDEIRA / TERM. STO. AMARO
33476	609	PÇA. DO CORREIO / TERM. CACHOEIRINHA
34840	606	ANHANGABAÚ / SHOP. CONTINENTAL

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34832	604	TERM. PRINC. ISABEL / RIO PEQUENO
33200	599	PÇA. RAMOS DE AZEVEDO / CID. D'ABRIL
32939	590	LGO. SÃO FRANCISCO / JD. ÂNGELA
35280	583	TERM. PQ. D. PEDRO II / TERM. PINHEIROS
33280	582	PÇA. RAMOS DE AZEVEDO / JD. JOÃO XXII-I/EDUC.
34393	579	PÇA. DO CORREIO / TERM. SAPOPEMBA
33544	578	PINHEIROS / PARAISÓPOLIS
34062	578	TERM. BANDEIRA / JD. LUSO
34064	578	PQ. IBIRAPUERA / JD. MIRIAM
32885	577	ACLIMAÇÃO / TERM. PRINC. ISABEL
34761	570	TERM. PINHEIROS / EST. STO. AMARO/GUIDO CALOI
33763	568	PÇA. JOÃO MENDES / JD. VL. FORMOSA
33425	567	CID. UNIVERSITÁRIA / METRÔ SANTANA
32877	562	METRÔ JABAQUARA / GRAJAÚ
34050	549	PQ. D. PEDRO II / CID. ADEMAR
35068	545	METRÔ BARRA FUNDA / TERM. PQ. D. PEDRO II
33121	543	TERM. PRINC. ISABEL / TERM. STO. AMARO
33964	541	TERM. STO. AMARO / JD. HERPLIN
32813	540	PÇA. DA SÉ / CHÁC. SANTANA
34196	540	SOCORRO / LAPA
32892	533	ACLIMAÇÃO / TERM. PRINC. ISABEL
32855	531	TERM. STO. AMARO / JD. ICARAÍ
34218	531	TERM. BANDEIRA / TERM. GUARAPIRANGA
34884	529	BUTANTÃ / TERM. PQ. D. PEDRO II
33643	528	TERM. CACHOEIRINHA / PQ. DE TAIPAS
34883	515	TERM. PINHEIROS / TERM. PQ. D. PEDRO II
33264	513	EST. DA LUZ / JD. BOA VISTA
34049	511	TERM. GUARAPIRANGA / JD. MIRIAM

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33952	505	AEROPORTO / CONJ. HAB. PALMARES
33502	504	TERM. PQ. D. PEDRO II / SÃO MATEUS
32825	503	TERM. BANDEIRA / TERM. JOÃO DIAS
33924	502	TERM. STO. AMARO / JD. ORION
35087	496	METRÔ SANTANA / TERM. CACHOEIRINHA
33336	494	PINHEIROS / IMIRIM
34397	494	METRÔ BELÉM / JD. WALKIRIA
34463	488	METRÔ SANTANA / TERM. CACHOEIRINHA
32876	487	METRÔ JABAQUARA / CENTRO SESC
34045	485	TERM. PRINC. ISABEL / JD. MIRIAM
34107	483	TERM. PQ. D. PEDRO II / PQ. DA LAPA
33934	480	SHOP. INTERLAGOS / CANTINHO DO CÉU
33936	480	SHOP. INTERLAGOS / JD. LUCÉLIA
33128	473	TERM. BANDEIRA / SOCORRO
34391	472	METRÔ BELÉM / TERM. SAPOPEMBA
33129	471	TERM. BANDEIRA / VL. CRUZEIRO
35145	470	TERM. PQ. D. PEDRO II / TERM. SÃO Mateus
33538	467	PAULISTA / PARAÍSÓPOLIS
34083	467	PINHEIROS / VALO VELHO
32815	466	TERM. PINHEIROS / TERM. STO. AMARO
35178	458	TERM. PINHEIROS / TERM. STO. AMARO
35069	456	TERM. PINHEIROS / CACHOEIRINHA
33389	451	TERM. PINHEIROS / METRÔ TUCURUVI
34014	445	SHOP. ARICANDUVA / HOSP. IPIRANGA
32882	442	METRÔ JABAQUARA / JD. STA. BARBARA
33131	442	HOSP. DAS CLÍNICAS / TERM. STO. AMARO
33391	442	METRÔ JABAQUARA / METRÔ SANTANA
34395	442	TERM. PRINC. ISABEL / TERM. SAPOPEMBA
34061	440	PQ. IBIRAPUERA / JD. MIRIAM
34789	440	METRÔ ARMÊNIA / SHOP. MORUMBI
35278	440	METRÔ STA. CRUZ / TERM. LAPA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34211	437	TERM. STO. AMARO / TERM. VARGINHA
33191	436	ITAIM BIBI / TERM. PIRITUBA
33126	433	TERM. BANDEIRA / INOCOOP CAMPO LIMPO
33236	433	TERM. BANDEIRA / JD. JAQUELINE
34831	432	TERM. BANDEIRA / JD. PAULO VI
33479	430	TERM. BANDEIRA / TERM. PQ. D. PEDRO II
33564	430	HOSP. DAS CLÍNICAS / JD. DAS PALMAS
34139	427	TERM. BANDEIRA / CEASA
33457	417	METRÔ VL. MADALENA / PQ. EDÚ CHAVES
33506	416	TERM. PQ. D. PEDRO II / SÃO MATEUS
33648	416	TERM. CACHOEIRINHA / JD. DAMASCENO
33363	415	PÇA. JOÃO MENDES / JD. MIRIAM
35083	408	TERM. PINHEIROS / TERM. CACHOEIRINHA
33915	403	TERM. GRAJAÚ / PQ. STA. CECÍLIA
33914	396	SHOP. INTERLAGOS / JD. SÃO BERNARDO
35109	396	TERM. PINHEIROS / TERM. PQ. D. PEDRO II
33652	392	TERM. CASA VERDE / PQ. TIETÊ
33372	388	PINHEIROS / VL. CLARA
33867	388	VL. PRUDENTE / SÃO MATEUS
34101	387	PÇA. RAMOS DE AZEVEDO / MERCADO DA LAPA
32831	384	LGO. SÃO FRANCISCO / TERM. CAPELINHA
35191	383	TERM. PINHEIROS / TERM. JOÃO DIAS
34685	379	TERM. BANDEIRA / TERM. CAMPO LIMPO
34407	377	DIV. DE SÃO CAETANO / SÃO MATEUS
34788	370	ITAIM BIBI / TERM. GUARAPIRANGA
33239	368	PÇA. RAMOS DE AZEVEDO / PQ. CONTINENTAL
34102	368	PÇA. RAMOS DE AZEVEDO / LAPA
34076	366	TERM. PQ. D. PEDRO II / TERM. GUARAPIRANGA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33284	365	ITAIM BIBI / METRÔ SANTANA
33922	365	TERM. STO. AMARO / JD. SÃO BERNARDO
33034	362	PÇA. D. GASTÃO / JD. MIRIAM
34084	362	TERM. PINHEIROS / COHAB ADVENTISTA
33573	360	TERM. CACHOEIRINHA / PERUS
34195	358	PÇA. RAMOS DE AZEVEDO / APIACÁS
35062	352	JD. PRINCESA / TERM. CACHOEIRINHA
34693	350	METRÔ STA. CRUZ / TERM. CAMPO LIMPO
33910	347	TERM. STO. AMARO / UNISA-CAMPUS 1
33919	347	TERM. GRAJAÚ / JD. CASTRO ALVES
34007	346	ITAIM BIBI / TERM. JD. ÂNGELA
33645	344	TERM. CACHOEIRINHA / COHAB BRASILÂNDIA
33486	341	TERM. PQ. D. PEDRO II / TERM. SÃO MATÉUS
32910	340	TERM. PQ. D. PEDRO II / VL. MARA
33229	340	PÇA. DO CORREIO / TERM. CACHOEIRINHA
35088	340	PEDRA BRANCA / TERM. CACHOEIRINHA
34860	336	METRÔ ANA ROSA / E.T. ÁGUA ESPRAIADA
35058	336	TAIPAS / TERM. CACHOEIRINHA
32836	331	METRÔ SÃO JUDAS / TERM. JOÃO DIAS
33646	328	TERM. CACHOEIRINHA / CPTM JARAGUÁ
32900	321	PÇA. DO CORREIO / SÃO MIGUEL
35175	321	TERM. PQ. D. PEDRO II / TERM. STO. AMARO
33234	320	TERM. PRINC. ISABEL / TERM. CACHOEIRINHA
33596	320	METRÔ BARRA FUNDA / VL. TEREZINHA
33122	317	TERM. PQ. D. PEDRO II / TERM. STO. AMARO
33929	317	CPTM JURUBATUBA / JD. GAIOTAS
32884	316	TERM. PQ. D. PEDRO II / TERM. CASA VERDE
35065	312	JD. CAROMBÉ / TERM. CACHOEIRINHA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33534	310	CARDOSO DE ALMEIDA / MACHADO DE ASSIS
32833	308	HOSP. DAS CLÍNICAS / TERM. JOÃO DIAS
34886	302	TERM. STO. AMARO / JD. ÂNGELA
35011	302	METRÔ - TRIANON - MASP / VL. GOMES
35147	299	TERM. PINHEIROS / TERM. SACOMÃ
33328	295	HOSP. DAS CLÍNICAS / LAUZANE PAULISTA
34008	294	MORUMBI SHOP. / JD. GUARUJÁ
35230	294	TERM. PINHEIROS / TERM. STO. AMARO
34138	293	TERM. PQ. D. PEDRO II / TERM. PINHEIROS
34494	291	SHOP. MORUMBI / BUTANTÃ
32966	288	METRÔ STA. CRUZ / TERM. JD. ÂNGELA
35051	287	TERM. PQ. D. PEDRO II / TERM. LAPA
33916	286	TERM. GRAJAÚ / VL. NATAL
34353	286	TERM. GRAJAÚ / VARGEM GRANDE
34355	286	TERM. GRAJAÚ / DIVISA DE EMBU-GUAÇU
35169	286	JD. NORONHA / TERM. GRAJAÚ
33558	284	STO. AMARO / REAL PQ.
33642	280	TERM. CACHOEIRINHA / VL. PENTEADO
33116	279	RIO PEQUENO / IPIRANGA
33535	278	PÇA. DA REPÚBLICA / STA. MARGARIDA MARIA
33536	278	PÇA. DA REPÚBLICA / GENTIL DE MOURA
33051	276	METRÔ BELÉM / JD. IMPERADOR
33550	276	SHOP. SP MARKET / CAMPO LIMPO
34137	276	METRÔ BARRA FUNDA / TERM. CACHOEIRINHA
33339	272	METRÔ SANTANA / COHAB BRASILÂNDIA
33232	270	ITAIM BIBI / COHAB TAIPAS
33170	269	TERM. PQ. D. PEDRO II / ITAIM PAULISTA
35162	266	TERM. PINHEIROS / METRÔ JABAQUARA
33380	264	METRÔ SANTANA / VL. PENTEADO
33639	264	TERM. CACHOEIRINHA / JD. ELISA MARIA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
35054	264	TERM. LAPA / TERM. CACHOEIRINHA
33361	262	PÇA. DA SÉ / BALN. SÃO FRANCISCO
33359	261	TERM. PRINC. ISABEL / VOITH
33481	260	PÇA. DA SÉ / TERM. VL. CARRÃO
34977	260	TERM. MERCADO / TERM. SÃO MATEUS
32834	259	TERM. PINHEIROS / TERM. CAPELINHA
33561	259	E.T. Água Espraiada / JD. PAULO VI
35180	259	TERM. STO. AMARO / TERM. CAPELINHA
33276	258	METRÔ BARRA FUNDA / JD. ARPOADOR
35164	258	TERM. STO. AMARO / TERM. GRAJAÚ
33475	256	TERM. CACHOEIRINHA / JD. STA. CRUZ
33638	256	TERM. CACHOEIRINHA / JD. PERY ALTO
33114	255	TERM. PINHEIROS / SACOMÃ
33072	254	TERM. STO. AMARO / IPIRANGA
34051	254	PQ. IBIRAPUERA / VL. STA. CATARINA
33241	253	PINHEIROS / JD. ADALGIZA
34100	253	TERM. PRINC. ISABEL / CID. UNIVERSITÁRIA
33146	249	TERM. PQ. D. PEDRO II / JD. CAMARGO VELHO
33151	249	TERM. PQ. D. PEDRO II / OLIVEIRINHA
34758	249	METRÔ PÇA. DA ÁRVORE / JD. ÂNGELA
33243	248	ITAIM BIBI / RIO PEQUENO
33233	246	ITAIM BIBI / JD. NARDINI
33553	244	STO. AMARO / JD. JAQUELINE
33032	243	PQ. IBIRAPUERA / JD. SELMA
33090	240	PÇA. DA REPÚBLICA / SHOP. PLAZA SUL
34033	236	PÇA. RAMOS DE AZEVEDO / TERM. PIRITUBA
32879	235	METRÔ VL. MARIANA / TERM. GRAJAÚ
33933	235	TERM. STO. AMARO / JD. PROGRESSO
34053	234	TERM. STO. AMARO / JD. LUSO
35167	234	JD. LUSO / TERM. STO. AMARO
34834	231	TERM. PINHEIROS / JD. COLOMBO
34136	228	LAPA / TERM. CACHOEIRINHA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33142	227	TERM. PQ. D. PEDRO II / VL. NOVA CURUÇÁ
34669	227	METRÔ CONCEIÇÃO / TERM. CAMPO LIMPO
33112	226	TERM. PQ. D. PEDRO II / JD. SÃO SAVÉRIO
33354	226	TERM. PRINC. ISABEL / COHAB TAIPAS
34246	226	METRÔ STA. CRUZ / TERM. STO. AMARO
33366	221	PÇA. JOÃO MENDES / ELDORADO
33477	221	SHOP. CENTER NORTE / JD. DAMASCENO
34962	217	LGO. DA CONCÓRDIA / SHOP. ARICANDUVA
33398	214	CID. UNIVERSITÁRIA / METRÔ SANTANA
35197	214	TERM. PQ. D. PEDRO II / TERM. PINHEIROS
33968	210	TERM. STO. AMARO / JD. PLANALTO
34090	210	METRÔ VL. MARIANA / TERM. CAPELINHA
35207	209	STA. CECÍLIA / TERM. VL. MARIANA
34240	208	TERM. VARGINHA / TERM. GRAJAÚ
33611	204	LAPA / JD. PERY ALTO
34128	200	PÇA. DO CORREIO / BRASILÂNDIA
35125	200	TERM. PQ. D. PEDRO II / TERM. VL. CARÃO
33348	197	PÇA. DO CORREIO / TAIPAS
34283	195	PÇA. JOÃO MENDES / ELDORADO
32872	193	TERM. STO. AMARO / PQ. AMÉRICA
35177	193	TERM. STO. AMARO / TERM. CAPELINHA
35206	190	METRÔ VL. MARIANA / METRÔ BUTANTÃ
32897	189	LUZ / TERM. A. E. CARVALHO
35179	188	TERM. PINHEIROS / TERM. CAMPO LIMPO
33198	187	PÇA. DO CORREIO / CID. D'ABRIL 3 ^a GLEBA
33206	187	PÇA. RAMOS DE AZEVEDO / MORRO DOCE
34856	186	ITAIM BIBI / TERM. LAPA
33144	184	TERM. PQ. D. PEDRO II / JD. NAZARÉ

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34760	184	EST. STO. AMARO/GUIDO CALOI / TERM. JD. JACIRA
32776	182	METRÔ ANA ROSA / TERM. CAPELINHA
34000	182	TERM. STO. AMARO / JD. SÃO FRANCISCO
33211	181	LGO. DO PAISSANDÚ / JD. LÍBANO
33088	179	PÇA. DA REPÚBLICA / VL. MONUMENTO
35203	179	PQ. CONTINENTAL / TERM. PINHEIROS
34938	177	TERM. PQ. D. PEDRO II / TERM. CID. TIRADENTES
35085	177	TERM. PQ. D. PEDRO II / TERM. CASA VERDE
32903	176	TERM. PQ. D. PEDRO II / JD. DANFER
33089	176	TERM. PQ. D. PEDRO II / VL. GUMERCINDO
33989	176	TERM. STO. AMARO / JD. D. JOSÉ
35274	176	PÇA. RAMOS DE AZEVEDO / TERM. LAPA
33000	175	METRÔ VL. MARIANA / PENHA
33230	174	LGO. DO PAISSANDÚ / TERM. CACHOEIRINHA
32820	172	TERM. STO. AMARO / TERM. CAPELINHA
34969	170	METRÔ BELÉM / TERM. VL. CARRÃO
35022	169	METRÔ BARRA FUNDA / CID. UNIVERSITÁRIA
35176	169	TERM. STO. AMARO / TERM. JD. ÂNGELA
33386	168	METRÔ SANTANA / VL. STA. MARIA
33595	168	METRÔ BARRA FUNDA / JD. DOS FRANCOS
33680	168	PQ. D. PEDRO II / UNIÃO DE VL. NOVA
33953	168	TERM. GRAJAÚ / CANTINHO DO CÉU
33979	167	TERM. STO. AMARO / RIVIERA
34920	167	TERM. STO. AMARO / JD. NAKAMURA
34717	166	LAPA / CAMPO LIMPO
34935	166	METRÔ BELÉM / TERM. SÃO MATEUS
33214	165	LGO. DO PAISSANDÚ / MANGALOT
32909	164	TERM. PQ. D. PEDRO II / TERM. A. E. CARVALHO
33037	164	PQ. IBIRAPUERA / JD. APURÁ

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33130	164	METRÔ ANA ROSA / TERM. STO. AMARO
33516	160	METRÔ BRESSER / CID. TIRADENTES
33575	160	CEM. VL. NOVA CACHOEIRINHA / PIRI-TUBA
34668	156	TERM. STO. AMARO / TERM. CAMPO LIMPO
35192	156	JD. PLANALTO / TERM. JOÃO DIAS
34498	153	TERM. STO. AMARO / TERM. JD. ÂNGELA
33956	152	TERM. STO. AMARO / JD. ICARAÍ
34043	152	METRÔ STA. CRUZ / CPTM AUTÓDROMO
34110	150	JAGUARÉ / CITY JARAGUÁ
35072	150	METRÔ BARRA FUNDA / CONEXÃO PETRÔ-NIO PORTELA
33343	149	PÇA. DO CORREIO / JD. GUARANI
34086	149	METRÔ SÃO JUDAS / PQ. STO. ANTONIO
32975	148	TERM. PQ. D. PEDRO II / TERM. A. E. CARVALHO
33656	148	METRÔ BARRA FUNDA / JD. TEREZA
34386	148	TERM. PQ. D. PEDRO II / TERM. SÃO MIGUEL
35104	147	TERM. PQ. D. PEDRO II / TERM. A. E. CARVALHO
33314	144	SHOP. CENTER NORTE / VL.NOVA CACHOEIRINHA
33070	142	METRÔ SAÚDE / VL. LIVIERO
33237	142	METRÔ BARRA FUNDA / RIO PEQUENO
35103	142	TERM. PQ. D. PEDRO II / TERM. A. E. CARVALHO
34396	138	TERM. PQ. D. PEDRO II / TERM. SAPO-PEMBA
35084	138	METRÔ VL. MADALENA / METRÔ SANTANA
33255	137	PAULISTA / COHAB EDUCANDÁRIO
34659	136	TERM. PINHEIROS / TERM. CAMPO LIMPO
33427	135	PÇA. DO CORREIO / VL. SABRINA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33462	134	PÇA. DO CORREIO / PQ. EDÚ CHAVES
34928	133	TERM. PQ. D. PEDRO II / E.T. ITAQUERA
35209	133	COHAB RAPOSO TAVARES / TERM. PINHEIROS
33667	132	CACHOEIRINHA / COHAB ANTÁRTICA
35050	132	TERM. PQ. D. PEDRO II / TERM. LAPA
33468	130	PÇA. DO CORREIO / JD. BRASIL
33095	129	TERM. PQ. D. PEDRO II / ZOOLÓGICO
34852	129	TERM. STO. AMARO / JD. CAIÇARA
33472	128	LUZ / CANGAÍBA
34133	128	TERM. CASA VERDE / VL. PENTEADO
33240	126	TERM. LAPA / RIO PEQUENO
34439	126	JD. ITÁPOLIS / TERM. SACOMÃ
35252	126	TERM. STO. AMARO / E.T. ÁGUA ESPRAIADA
33870	124	OBJETIVO UNIP / VL. DAS MERCÊS
33543	122	PINHEIROS / PQ. ARARIBA
35165	121	TERM. STO. AMARO / TERM. GRAJAÚ
33441	120	MUSEU DO IPIRANGA / VL. SABRINA
33474	120	PENHA / METRÔ SANTANA
35171	117	UNISA / TERM. GRAJAÚ
35172	117	VARGEM GRANDE / TERM. GRAJAÚ
33039	116	TERM. STO. AMARO / VL. IMPÉRIO
34650	116	TERM. PQ. D. PEDRO II / TERM. PENHA
32858	114	TERM. STO. AMARO / JD. GRAUNA
33576	114	TERM. PIRITUBA / RECANTO DOS HUMILDES
33986	114	STO. AMARO / JD. JANGADEIRO
35161	110	TERM. STO. AMARO / TERM. GRAJAÚ
33581	109	METRÔ BARRA FUNDA / JD. VISTA ALEGRE
34132	109	METRÔ BARRA FUNDA / PENTEADO
32934	108	TERM. PQ. D. PEDRO II / JD. SÃO PAULO
34941	108	TERM. PQ. D. PEDRO II / TERM. CID. TIRADENTES
32956	106	TERM. STO. AMARO / TERM. JD. JACIRA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
32964	106	TERM. STO. AMARO / JD. ARACATI
34273	106	TERM. STO. AMARO / TERM. GRAJAU
33017	104	CERET / JD. HELENA
33571	104	JD. PRIMAVERA / CPTM VL. AURORA
34140	104	TERM. PRINC. ISABEL / TERM. PINHEIROS
33985	103	STO. AMARO / VALO VELHO
35061	103	PERUS / TERM. PIRITUBA
33105	101	METRÔ JABAQUARA / SHOP. PLAZA SUL
35139	101	HOSP. STA. MARCELINA / METRÔ ITAQUERA
34838	100	BUTANTÃ / VL. SÔNIA
35143	100	TERM. PQ. D. PEDRO II / TERM. SÃO Mateus
33566	99	LAPA / PERUS
33540	96	HOSP. DAS CLÍNICAS / JD. ROSA MARIA
33794	96	METRÔ CARRÃO / SHOP. ARICANDUVA
33981	96	STO. AMARO / VL. GILDA
34193	96	TERM. LAPA / MORRO GRANDE
32798	95	TERM. JOÃO DIAS / JD. INGÁ
34745	93	ITAIM BIBI / JD. MIRIAM
32773	92	TERM. JOÃO DIAS / JD. IBIRAPUERA
32799	92	TERM. JOÃO DIAS / TERM. CAPELINHA
33176	92	LAPA / JARAGUÁ
33274	92	HOSP. DAS CLÍNICAS / JD. JOÃO XXIII
33848	92	METRÔ BELÉM / VL. INDUSTRIAL
35173	92	ELDORADO / TERM. STO. AMARO
33577	90	LAPA / CAPELA DA LAGOA
33601	90	TERM. PIRITUBA / COHAB BRASILÂNDIA
34024	90	TERM. PIRITUBA / JD. DONÁRIA
34027	90	TERM. PIRITUBA / CID. D'ABRIL 3 ^a GLEBA
34029	90	TERM. PIRITUBA / CPTM VL. AURORA
34035	90	METRÔ BARRA FUNDA / TERM. PIRITUBA
34114	90	TERM. PIRITUBA / JD. RINCÃO
34205	90	TERM. PIRITUBA / PQ. DE TAIPAS
34937	90	METRÔ PENHA / TERM. CID. TIRADENTES

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
35059	90	JD. DONÁRIA / TERM. PIRITUBA
35265	90	CONEXÃO VL. IÓRIO / PERUS
35081	89	TERM. PQ. D. PEDRO II / METRÔ TUCU-RUVI
35202	89	JD. JOÃO XXIII / TERM. PINHEIROS
32954	88	TERM. STO. AMARO / JD. NAKAMURA
33040	88	TERM. STO. AMARO / VL. GUACURI
33251	88	METRÔ BARRA FUNDA / PINHEIROS/VILA IDA
33333	88	CEASA / METRÔ SANTANA
35163	87	TERM. PQ. D. PEDRO II / METRÔ JABAQUARA
32923	86	CERET / TERM. A. E. CARVALHO
32944	86	TERM. STO. AMARO / TERM. CAPELINHA
33356	86	PÇA. DO CORREIO / PEDRA BRANCA
34939	86	TERM. PQ. D. PEDRO II / TERM. SÃO Mateus
35271	86	METRÔ JABAQUARA / TERM. GUARAPI-RANGA
33610	85	CORREIO / PQ. VL. MARIA
33863	85	METRÔ TATUAPÉ / VL. CALIFÓRNIA
33876	85	METRÔ BELÉM / PQ. BANCÁRIO
34835	85	TERM. PINHEIROS / RIO PEQUENO
34903	85	TERM. PINHEIROS / CONEXÃO VL. IÓRIO
32824	84	STO. AMARO / CAPÃO REDONDO
32863	84	TERM. GRAJAÚ / PQ. RES. COCAIA
33859	84	METRÔ BRESSER / JD. ITÁPOLIS
33913	84	TERM. GRAJAÚ / JD. ELLUS
33935	84	TERM. GRAJAÚ / ILHA DO BORORÉ
33954	84	TERM. GRAJAÚ / JD. PRAINHA
33957	84	TERM. GRAJAÚ / JD. LUCÉLIA
33961	84	TERM. GRAJAÚ / JD. GAIOTAS
33993	84	TERM. GRAJAÚ / PQ. COCAIA
35110	84	TERM. PQ. D. PEDRO II / METRÔ ITAQUERA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33245	83	METRÔ - TRIANON - MASP / PQ. CONTINENTAL
34804	83	E.T. ÁGUA ESPRAIADA / TERM. GRAJAÚ
33079	82	PÇA. ALMEIDA JR. / VL. EMA
33879	81	IBIRAPUERA / JD. ELBA
34200	81	LGO. DO PAISSANDÚ / TERM. PIRITUBA
33325	80	SHOP. CENTER NORTE / COHAB ANTÁRTICA
33412	80	METRÔ SANTANA / CACHOEIRA
33988	80	TERM. STO. AMARO / JD. CAPELINHA
34058	80	TERM. STO. AMARO / METRÔ JABAQUARA
34427	79	PÇA. DO CORREIO / TERM. SACOMÃ
34837	79	TERM. PINHEIROS / JD. D'ABRIL
32913	78	METRÔ TATUAPÉ / VL. CISPER
33807	78	METRÔ ITAQUERA / RECANTO VERDE SOL
33810	78	METRÔ ITAQUERA / JD. LARANJEIRA
33831	78	METRÔ ITAQUERA / COHAB BARRO BRANCO
34380	78	METRÔ ITAQUERA / CID. TIRADENTES
33657	77	METRÔ BARRA FUNDA / JD. GUARANI
34812	76	METRÔ BUTANTÃ / TERM. CAMPO LIMPO
35156	74	JD. ITÁPOLIS / TERM. SACOMÃ
33093	73	TERM. PQ. D. PEDRO II / JD. PLANALTO
33226	73	PÇA. DO CORREIO / TERM. CASA VERDE
33165	72	METRÔ TATUAPÉ / JD. ROMANO
33342	72	PÇA. DO CORREIO / JD. PAULISTANO
33702	72	METRÔ ITAQUERA / JD. NAZARÉ
33735	72	METRÔ ITAQUERA / JD. ALTO PAULISTANO
33795	72	METRÔ ITAQUERA / JD. LIMOEIRO
33835	72	METRÔ ITAQUERA / CID. TIRADENTES
33548	71	SHOP. MORUMBI / JD. INGÁ
34836	71	TERM. PINHEIROS / COHAB EDUCANDÁRIO
34414	70	MOEMA / TERM. SACOMÃ
35080	70	TERM. PINHEIROS / METRÔ SANTANA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
35246	70	TERM. PINHEIROS / METRÔ SANTANA
33173	69	METRÔ ITAQUERA / JD. CAMARGO VELHO
33482	69	PÇA. DA SÉ / PÇA. SILVIO ROMERO
33987	69	STO. AMARO / JD. TRÊS ESTRELAS
32991	68	METRÔ ARTUR ALVIM / JD. HELENA
33549	68	STO. AMARO / JD. INGÁ
33990	68	STO. AMARO / VALO VELHO
34191	68	METRÔ BARRA FUNDA / VL. ZATT
35144	68	TERM. PQ. D. PEDRO II / TERM. SACOMÃ
35146	68	TERM. PQ. D. PEDRO II / TERM. SACOMÃ
34990	67	METRÔ BRESSER / CONJ. MANOEL DA NÓ-BREGA
32781	66	TERM. JOÃO DIAS / CAPÃO REDONDO
32912	66	METRÔ TATUAPÉ / ERMELINO MATA-RAZZO
33758	66	SÃO MIGUEL PAULISTA / TERM. CID. TIRADENTES
33873	66	TERM. NORTE METRÔ CARRÃO / VL. INDUSTRIAL
33972	65	STO. AMARO / PQ. INDEPENDÊNCIA
33973	65	STO. AMARO / VL. CALÚ
33974	65	STO. AMARO / JD. NAKAMURA
33975	65	STO. AMARO / PQ. CEREJEIRA
33111	64	TERM. AMARAL GURGEL / JD. DA SAÚDE
33984	64	STO. AMARO / JD. DAS ROSAS
34501	64	TERM. GRAJAÚ / JD. ELIANA
34747	64	HOSP. PEDREIRA / CID. DUTRA
35168	64	JD. GAIOTAS / TERM. GRAJAÚ
35200	64	CEASA / TERM. PINHEIROS
33460	63	LIBERDADE / VL. MEDEIROS
33846	63	SHOP. ARICANDUVA / JD. SÃO FRANCISCO
35149	63	METRÔ SANTANA / TERM. SACOMÃ
33365	62	PÇA. JOÃO MENDES / DIV. DIADEMA
34010	61	STO. AMARO / JD. CAPELA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34942	61	TERM. PQ. D. PEDRO II / INÁCIO MONTEIRO
32787	60	TERM. CAPELINHA / JD. DAS ROSAS
32791	60	TERM. CAPELINHA / PQ. FERNANDA
33406	60	SHOP. CENTER NORTE / VL. ALBERTINA
33514	60	TERM. PQ. D. PEDRO II / VL. DALILA
33777	60	METRÔ ITAQUERA / JD. SÃO JOÃO
34031	60	TERM. LAPA / TERM. PIRITUBA
34409	60	PÇA. ALMEIDA JR. / TERM. SAPOPEMBA
34560	60	METRÔ SANTANA / PEDRA BRANCA
34940	60	TERM. PQ. D. PEDRO II / JD. MARÍLIA
35053	60	TERM. LAPA / TERM. PIRITUBA
35111	60	OLIVEIRINHA / TERM. A. E. CARVALHO
35275	60	TERM. LAPA / TERM. PIRITUBA
32800	59	TERM. JOÃO DIAS / JD. NOVO ORIENTE
32943	59	SHOP. INTERLAGOS / JD. HERCULANO
32990	58	METRÔ ARTUR ALVIM / PQ. D. JOÃO NERY
33139	58	TERM. ARICANDUVA / CID. KEMEL
33311	57	SHOP. D / JD. PERY ALTO
33819	57	METRÔ CARRÃO / 3A. DIVISÃO
33904	57	SHOP. MORUMBI / METRÔ CONCEIÇÃO
34826	57	SHOP. MORUMBI / TERM. CAMPO LIMPO
33312	56	METRÔ SANTANA / LAUZANE PAULISTA
33455	56	TERM. VL. CARRÃO / JAÇANÃ
33635	56	PINHEIROS / METRÔ BARRA FUNDA
33784	56	METRÔ TATUAPÉ / VL. STA. ISABEL
34857	56	TERM. PINHEIROS / LAPA
35032	56	METRÔ SANTANA / LAUZANE PAULISTA
33287	55	TERM. AMARAL GURGEL / JD. PERY ALTO
33488	55	CIRCULAR / TERM. VL. CARRÃO
34059	55	METRÔ ANA ROSA / JD. MIRIAM
33043	54	METRÔ CONCEIÇÃO / SHOP. SP MARKET
33711	54	METRÔ PENHA / JD. DO CASTELO
34387	54	TERM. SÃO MATEUS / TERM. SÃO MIGUEL
32932	53	TERM. SÃO MATEUS / JD. HELENA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34048	53	LGO. SÃO FRANCISCO / JD. SELMA
34077	53	STO. AMARO / VALO VELHO
34684	53	SHOP. MORUMBI / TERM. CAMPO LIMPO
35204	53	CDHU BUTANTÃ / TERM. JOÃO DIAS
33044	52	VL. PRUDENTE / PQ. BANCÁRIO
33837	52	METRÔ ARTUR ALVIM / SHOP. ARICANDUVA
33320	51	METRÔ SANTANA / JD. ANTÁRTICA
33321	51	METRÔ SANTANA / JD. PERY
33556	51	STO. AMARO / PARAISÓPOLIS
34943	51	TERM. PQ. D. PEDRO II / TERM. VL. CARRÃO
33009	50	METRÔ TATUAPÉ / CID. PEDRO JOSÉ NUNES
33160	50	METRÔ ARTUR ALVIM / JD. DAS OLIVEIRAS
33439	50	TERM. AMARAL GURGEL / VL. SABRINA
33781	50	NOVA AMERICA / METRÔ ARTUR ALVIM
34882	50	METRÔ ARTUR ALVIM / CONJ. ENCOSTA NORTE
34056	49	METRÔ CONCEIÇÃO / CID. JÚLIA
34127	49	PÇA. DO CORREIO / FREGUESIA DO Ó
32926	48	TERM. SÃO MATEUS / TERM. A. E. CARVALHO
33136	48	TERM. PENHA / JD. DAS OLIVEIRAS
33158	48	METRÔ VL. MATILDE / CID. KEMEL II
33299	48	LAPA / COHAB ANTÁRTICA
33346	48	TERM. LAPA / JD. DOS CUNHAS
33585	48	METRÔ SANTANA / JD. ALMANARA
33609	48	LAPA / LAUZANE PAULISTA
33653	48	LAPA / VL. TEREZINHA
33822	48	METRÔ TATUAPÉ / TERM. VL. CARRÃO
33991	48	STO. AMARO / JD. SÃO BENTO NOVO
34036	48	TERM. LAPA / ITABERABA
35033	48	LAPA / MANDAQUI
35093	48	JD. ANTÁRTICA / METRÔ SANTANA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33326	47	LAPA / METRÔ SANTANA
33982	47	STO. AMARO / JD. MACEDÔNIA
33614	46	TIETÊ / JOVA RURAL
34419	46	TERM. MERCADO / TERM. SACOMÃ
34757	46	TERM. JOÃO DIAS / JD. MARACÁ
34839	46	METRÔ BUTANTÃ / PQ. CONTINENTAL
34867	46	STO. AMARO / PARAISÓPOLIS
35193	46	JD. VAZ DE LIMA / TERM. JOÃO DIAS
33440	45	METRÔ SANTANA / VL. CONSTANÇA
33539	45	BROOKLIN NOVO / REAL PQ.
33021	44	TERM. SÃO MATEUS / GUAIANAZES
33307	44	METRÔ SANTANA / VL. DIONÍSIA
33461	44	LIBERDADE / PQ. EDÚ CHAVES
33731	44	SHOP. ARICANDUVA / VL. MINERVA
35102	44	JD. PERY ALTO / METRÔ SANTANA
35126	44	TERM. VL. CARRÃO / METRÔ ITAQUERA
33728	43	PQ. SÃO RAFAEL / SHOP. ARICANDUVA
33827	43	METRÔ CARRÃO / RES. STA. BÁRBARA
33868	43	SHOP. ARICANDUVA / FAZENDA DA JUTA
33878	43	METRÔ CARRÃO / JD. VERA CRUZ
34936	43	METRÔ CARRÃO / TERM. SÃO MATEUS
34964	43	METRÔ CARRÃO / JD. NOVA VITÓRIA
35150	43	TERM. PQ. D. PEDRO II / TERM. SACOMÃ
35266	43	METRÔ BELÉM / PQ. EDÚ CHAVES
33077	42	BOM RETIRO / PQ. SÃO LUCAS
33906	42	METRÔ CONCEIÇÃO / PQ. PRIMAVERA
34394	42	TERM. PQ. D. PEDRO II / TERM. SAPO-PEMBA
34453	42	TERM. VL. CARRÃO / METRÔ CONCEIÇÃO
34791	42	CID. UNIVERSITÁRIA / METRÔ BUTANTÃ
34853	42	CID. UNIVERSITÁRIA / METRÔ BUTANTÃ
34960	42	TERM. PENHA / CPTM JOSÉ BONIFÁCIO
34965	42	METRÔ ITAQUERA / JD. SÃO FRANCISCO
33110	41	METRÔ SÃO JUDAS / JD. CLÍMAX
33555	41	CAMPO BELO / PARAISÓPOLIS

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
32780	40	TERM. CAPELINHA / VALO VELHO
33370	40	LGO. CAMBUCI / AMERICANÓPOLIS
33551	40	STO. AMARO / JD. TABOÃO
33796	40	METRÔ VL. MATILDE / SHOP. ARICANDUVA
33803	40	JD. SÃO JOÃO / METRÔ ARTUR ALVIM
33805	40	METRÔ GUILHERMINA/ESPERANÇA / SHOP. ARICANDUVA
34399	40	SHOP. ARICANDUVA / MASCARENHAS DE MORAIS
34689	40	METRÔ PENHA / JD. KERALUX
34847	40	BUTANTÃ / PQ. IPÊ
34849	40	BUTANTÃ / JD. GUARAÚ
34851	40	BUTANTÃ / JD. INGÁ
33015	39	METRÔ TATUAPÉ / VL. SANTANA
32987	38	CONJ. JOSÉ BONIFÁCIO / PENHA
33067	38	METRÔ VL. MARIANA / JD. MARIA ESTELA II
33100	38	METRÔ VL. MARIANA / JD. CLÍMAX
33135	38	TERM. A. E. CARVALHO / CID. KEMEL II
33140	38	TERM. A. E. CARVALHO / CONJ. ENCOSTA NORTE
33190	38	TERM. PINHEIROS / VL. PIAUÍ
33473	38	PQ. D. PEDRO II / PQ. NOVO MUNDO
34619	38	TERM. MERCADO / TERM. VL. PRUDENTE
33058	37	TERM. PQ. D. PEDRO II / PQ. STA. MADALENA
32927	36	METRÔ ITAQUERA / VL. MARA
33001	36	METRÔ PENHA / GUAIANAZES
33783	36	METRÔ BRESSER / UNIV. SÃO JUDAS TADEU
33787	36	METRÔ ARTUR ALVIM / SHOP. ARICANDUVA
34576	36	SÃO MIGUEL / JD. MABEL
34652	36	TERM. CAMPO LIMPO / JD. DAS ROSAS

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34692	36	TERM. CAMPO LIMPO / PQ. DO ENGENHO
35052	36	TERM. LAPA / TERM. PIRITUBA
33628	35	MOOCA / CEM. PQ. DOS PINHEIROS
33726	35	SHOP. ARICANDUVA / COHAB JOSÉ BONIFÁCIO
33761	35	METRÔ TAMANDUATEÍ / SHOP. ARICANDUVA
34983	35	METRÔ CARRÃO / JD. STO. ANDRÉ
32922	34	PENHA / VL. PARANAGUÁ
33375	34	METRÔ VERGUEIRO / ELDORADO
34968	33	METRÔ TATUAPÉ / TERM. CID. TIRADENTES
35208	33	STA. CECÍLIA / TERM. VL. MARIANA
35270	33	TERM. GUARAPIRANGA / JD. GUARUJÁ
33027	32	CPTM GUAIANAZES / TERM. SÃO MIGUEL
33028	32	CPTM GUAIANAZES / SÃO MIGUEL
33143	32	TERM. ARICANDUVA / VL. CURUÇÁ
33167	32	METRÔ ITAQUERA / JD. CAMARGO NOVO
33670	32	SÃO MIGUEL / JD. DAS OLIVEIRAS
34405	32	VL. ALPINA / METRÔ BRESSER
34418	32	TERM. SÃO MIGUEL / ITAIM PAULISTA
35028	32	METRÔ TUCURUVI / CACHOEIRA
35097	32	CACHOEIRA / METRÔ SANTANA
35124	32	VL. CISPER / TERM. PENHA
32797	31	TERM. JOÃO DIAS / JD. CAPELINHA
33078	30	PÇA. ALMEIDA JR. / PQ. STA. MADALENA
33743	30	HOSP. SAPOPEMBA / JD. PALANQUE
33992	30	STO. AMARO / JD. LÍDIA
34926	30	E.T. ITAQUERA / COHAB FAZENDA DO CARMO
35079	30	TERM. PQ. D. PEDRO II / METRÔ TUCURUVI
35082	30	TERM. PQ. D. PEDRO II / METRÔ TUCURUVI
35166	30	MORUMBI SHOP. / METRÔ JABAQUARA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34527	29	E.T. ÁGUA ESPRAIADA / METRÔ CONCEIÇÃO
33030	28	LAR ESC. SÃO FRANCISCO / METRÔ VL. MARIANA
33157	28	METRÔ PENHA / JD. ROMANO
33470	28	METRÔ SANTANA / TERM. PENHA
33983	28	STO. AMARO / JD. MITSUTANI
34237	28	METRÔ PENHA / JD. DANFER
34818	28	METRÔ BUTANTÃ / JD. JOÃO XXIII
33599	27	SHOP. CENTER NORTE / CEM. DO HORTO
33615	27	SHOP. CENTER NORTE / JD. FONTÁLIS
32907	26	TERM. ARICANDUVA / BURGO PAULISTA
32978	26	TERM. ARICANDUVA / JD. COIMBRA
32981	26	TERM. ARICANDUVA / VL. SÃO FRANCISCO
33451	26	METRÔ BELÉM / CENTER NORTE
33970	26	TERM. GUARAPIRANGA / CHÁC. STA. MARIA
34239	26	PENHA / JD. VL. NOVA
35105	26	TERM. ARICANDUVA / TERM. A. E. CARVALHO
35106	26	TERM. ARICANDUVA / TERM. SÃO MIGUEL
33450	25	TERM. PRINC. ISABEL / PQ. VL. MARIA
34423	25	PQ. BELÉM / TERM. SACOMÃ
32782	24	TERM. CAPELINHA / VALO VELHO
32802	24	TERM. CAPELINHA / JD. GUARUJÁ
33174	24	TERM. LAPA / SOL NASCENTE
33222	24	COHAB TAIPAS / PERUS
33387	24	SHOP. CENTER NORTE / JD. VISTA ALEGRE
33418	24	METRÔ SANTANA / VL. MARIETA
33600	24	METRÔ SANTANA / VL. ROSA
33632	24	METRÔ TUCURUVI / JD. MARINA
33636	24	METRÔ JD. SÃO PAULO / VL. AMÉLIA
33682	24	METRÔ ARTUR ALVIM / VL. AMERICANA
33793	24	PQ. SAVOY CITY / METRÔ ARTUR ALVIM

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
33882	24	STO. AMARO / JABAQUARA
34429	24	TERM. SACOMÃ / PQ. BRISTOL
34454	24	VL. MATIAS / IPIRANGA
34657	24	TERM. CAMPO LIMPO / VALO VELHO
34680	24	TERM. CAMPO LIMPO / PQ. DO LAGO
34909	24	CONEXÃO VL. IÓRIO / PERUS
35030	24	METRÔ PARADA INGLESA / HORTO FLORESTAL
35091	24	LGO. DO PERY / METRÔ TUCURUVI
35155	24	JD. CELESTE / TERM. SACOMÃ
35190	24	VALO VELHO / TERM. CAPELINHA
35263	24	TERM. PIRITUBA / SOL NASCENTE
33266	23	LAPA / JD. D'ABRIL
33448	23	METRÔ BARRA FUNDA / JD. FONTÁLIS
33020	22	COHAB II / JD. HELENA
33103	22	METRÔ SAÚDE / JD. MARIA ESTELA
33671	22	SÃO MIGUEL / JD. ROBRU
33672	22	SÃO MIGUEL / JD. CAMPOS
33708	22	METRÔ PENHA / LIMOEIRO
33736	22	SÃO MATEUS / GUAIANAZES
33897	22	E.T. ÁGUA ESPRAIADA / JD. SELMA
34952	22	TERM. SÃO MATEUS / JD. LIMOEIRO
34959	22	TERM. SÃO MATEUS / METALÚRGICOS
34979	22	MUSEU DO IPIRANGA / SÃO MATEUS
35130	22	TERM. SÃO MATEUS / TERM. CID. TIRADENTES
33269	21	METRÔ BARRA FUNDA / JD. JOÃO XXIII
33432	21	METRÔ BELÉM / SHOP. CENTER NORTE
33633	21	SANTANA / CENTER NORTE
32795	20	TERM. CAPELINHA / JD. SÃO BENTO
33186	20	TERM. LAPA / VL. PIAUÍ
33668	20	METRÔ BARRA FUNDA / JD. PERY ALTO
33699	20	METRÔ PENHA / VL. SÍLVIA
33877	20	METRÔ SAÚDE / VL. MORAES
33994	20	STO. AMARO / JD. UNIVERSAL

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34016	20	LAPA / METRÔ BARRA FUNDA
34483	20	METRÔ TATUAPÉ / PQ. SÃO LUCAS
34656	20	TERM. CAMPO LIMPO / JD. GUARUJÁ
35183	20	VALO VELHO / TERM. CAPELINHA
35157	19	VL. PRUDENTE / METRÔ VL. MARIANA
32902	18	METRÔ ARTUR ALVIM / TERM. A. E. CARVALHO
32994	18	METRÔ ARTUR ALVIM / JD. ROBRU
34443	18	TERM. PQ. D. PEDRO II / JD. CELESTE
34841	18	BUTANTÃ / JD. MARIA LUIZA
34930	18	BUTANTÃ / VL. DALVA
35112	18	VL. CISPER (CPTM USP) / TERM. A. E. CARVALHO
35201	18	TERM. PINHEIROS / TERM. LAPA
33811	17	METRÔ ITAQUERA / SÃO MATEUS
32789	16	TERM. CAPELINHA / JD. JANGADEIRO
32969	16	TERM. CAPELINHA / TERM. JD. JACIRA
33337	16	METRÔ SANTANA / HOSP. CACHOEIRINHA
33624	16	TATUAPÉ / JD. BRASIL
33802	16	METRÔ ARTUR ALVIM / CID. LIDER
34763	16	JD. ÂNGELA / JD. HORIZONTE AZUL
34766	16	JD. ÂNGELA / JD. VERA CRUZ
35010	16	METRÔ SÃO JUDAS / AEROPORTO
35115	16	ERMELINO MATARAZZO / TERM. PENHA
33310	15	METRÔ SANTANA / JD. ANTÁRTICA
33315	15	METRÔ SANTANA / PEDRA BRANCA
33316	15	METRÔ SANTANA - CIRCULAR / CONJ. DOS BANCÁRIOS
33625	14	METRÔ TATUAPÉ / JD. TREMEMBÉ
33629	14	METRÔ TATUAPÉ / PQ. NOVO MUNDO
33698	14	CANGAÍBA / METRÔ GUILHERMINA/ESPERANÇA
33705	14	METRÔ ITAQUERA / UNIÃO DE VL. NOVA
33893	14	HOSP. SÃO PAULO / JD. MIRIAM
34105	14	HOSP. DAS CLÍNICAS / LAPA

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34258	14	JD. ALFREDO / TERM. GUARAPIRANGA
35108	14	METRÔ ITAQUERA / TERM. SÃO MIGUEL
35151	14	TERM. SACOMÃ / TERM. SAPOPEMBA
35187	14	PQ. DO LAGO / TERM. GUARAPIRANGA
35188	14	JD. UNIVERSAL / TERM. CAPELINHA
33056	13	MOOCA / PQ. STA. MADALENA
33788	13	METRÔ PENHA / COHAB JOSÉ BONIFÁCIO
34811	13	TERM. GUARAPIRANGA / PQ. DO LAGO
33101	12	METRÔ VL. MARIANA / JD. SÃO SAVÉRIO
33188	12	LAPA / PQ. SÃO DOMINGOS
33804	12	METRÔ ARTUR ALVIM / JD. NSA. SRA. DO CARMO
34260	12	JABAQUARA / SHOP. INTERLAGOS
34437	12	TERM. SACOMÃ / JD. MARIA ESTELA
35067	12	MORRO GRANDE / METRÔ BARRA FUNDA
35092	12	PEDRA BRANCA / METRÔ SANTANA
35218	12	LAPA / CONEXÃO VL. IÓRIO
33874	11	METRÔ STA. CRUZ / SACOMÃ
34433	11	TERM. SACOMÃ / VL. BRASILINA
34974	11	METRÔ VL. PRUDENTE / PQ. SAVOY CITY
33026	10	TERM. VL. CARRÃO / GUAIANAZES
33637	10	TERM. PARADA INGLESA / JD. HEBRON
34021	10	TERM. LAPA / REMÉDIOS
34944	10	TERM. VL. CARRÃO / ITAQUERA
34949	10	TERM. VL. CARRÃO / COHAB JUSCELINO
35055	10	VL. PIAUÍ / TERM. LAPA
35129	10	METRÔ ITAQUERA / TERM. VL. CARRÃO
35141	10	SAVOY/DALILA / TERM. VL. CARRÃO
35153	10	JD. PLANALTO / TERM. SACOMÃ
34945	9	TERM. VL. CARRÃO / GUAIANAZES
35015	9	LGO. DA CONCÓRDIA / JD. FILHOS DA TERRA
32796	8	TERM. CAPELINHA / JD. VALE DAS VIRTUDES

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
32893	8	TERM. PQ. D. PEDRO II / TERM. VL. PRUDENTE
32988	8	CHÁC. BELA VISTA / METRÔ PENHA
32992	8	METRÔ ARTUR ALVIM / CPTM JOSÉ BONIFÁCIO
33018	8	TERM. A. E. CARVALHO / VL. PROGRESSO
33182	8	LAPA / PERUS
33421	8	METRÔ SANTANA / JD. FONTÁLIS
33515	8	TERM. PENHA / TERM. SÃO MATEUS
33569	8	PERUS / MORRO DOCE
33578	8	BOM RETIRO / JD. ELISA MARIA
33686	8	METRÔ VL. MATILDE / CEM. DA SAUDADE
33847	8	JD. DA CONQUISTA / HOSP. SÃO MATEUS
33852	8	TERM. PQ. D. PEDRO II / JD. COLORADO
33865	8	JD. STO. ANDRÉ / HOSP. SÃO MATEUS
33866	8	DIV. DE MAUÁ / HOSP. SÃO MATEUS
34654	8	TERM. CAMPO LIMPO / JD. HELGA
34880	8	CONJ. CHAPARRAL / METRÔ PENHA
35013	8	LAPA / JD. BOA VISTA
35099	8	JD. CAMPO LIMPO / METRÔ SANTANA
35128	8	TERM. PENHA / TERM. SÃO MATEUS
35210	8	PQ. ARARIBA / TERM. CAPELINA
33057	6	VL. PRUDENTE / VL. CALIFÓRNIA
33094	6	VL. PRUDENTE / VL. INDUSTRIAL
33613	6	CARANDIRU / JOVA RURAL
33695	6	VL. RUI BARBOSA / METRÔ VL. MATILDE
34430	6	TERM. SACOMÃ / VL. ARAPUÁ
34901	6	METRÔ BARRA FUNDA / LIMÃO
35035	6	SANTANA / VL. NOVA GALVÃO
35154	6	VL. ARAPUÁ / TERM. SACOMÃ
33741	5	METRÔ BELÉM / JD. ITÁPOLIS
33045	4	VL. PRUDENTE / VL. INDUSTRIAL
33180	4	LAPA / PQ. MORRO DOCE
33244	4	SESC POMPÉIA / PQ. CONTINENTAL

Continua na próxima página

Tabela 25 – continuação da página anterior

Código da linha	Total de eventos de exceção	Letreiro
34402	4	METRÔ ALTO DO IPIRANGA / CONJ. HAB. HELIÓPOLIS
34436	4	TERM. SACOMÃ / VL. ARAPUÁ
34440	4	TERM. SACOMÃ / JD. PATENTE
35060	4	MORRO DOCE / TERM. LAPA
33371	3	STO. AMARO / METRÔ JABAQUARA
33790	3	VL. DALILA / METRÔ VL. MATILDE
35023	3	TERM. PQ. D. PEDRO II / METRÔ SANTANA
33429	2	TERM. PRINC. ISABEL / PQ. EDÚ CHAVES
33896	2	METRÔ CONCEIÇÃO / JD. APURÁ
34872	2	SHOP. D / PQ. EDU CHAVES
35049	2	METRÔ TIETÊ / VL. MEDEIROS
33049	1	SHOP. METRÔ TATUAPÉ / JD. GUAIRACÁ
33426	1	SHOP. D / JD. PRIMAVERA
33884	1	JABAQUARA / VL. GUACURI
33885	1	STO. AMARO / JD. LUSO
33905	1	METRÔ JABAQUARA / REFÚGIO STA. TE-REZINHA
34291	1	METRÔ SÃO JUDAS / JD. UBIRAJARA
34842	1	BUTANTÃ / CDHU MUNCK

Apêndice E – Matrizes de confusão

Figura 19 – Matriz de confusão relacionada a classificação dos *tweets* em eventos de exceção por meio do algoritmo Regressão Logística

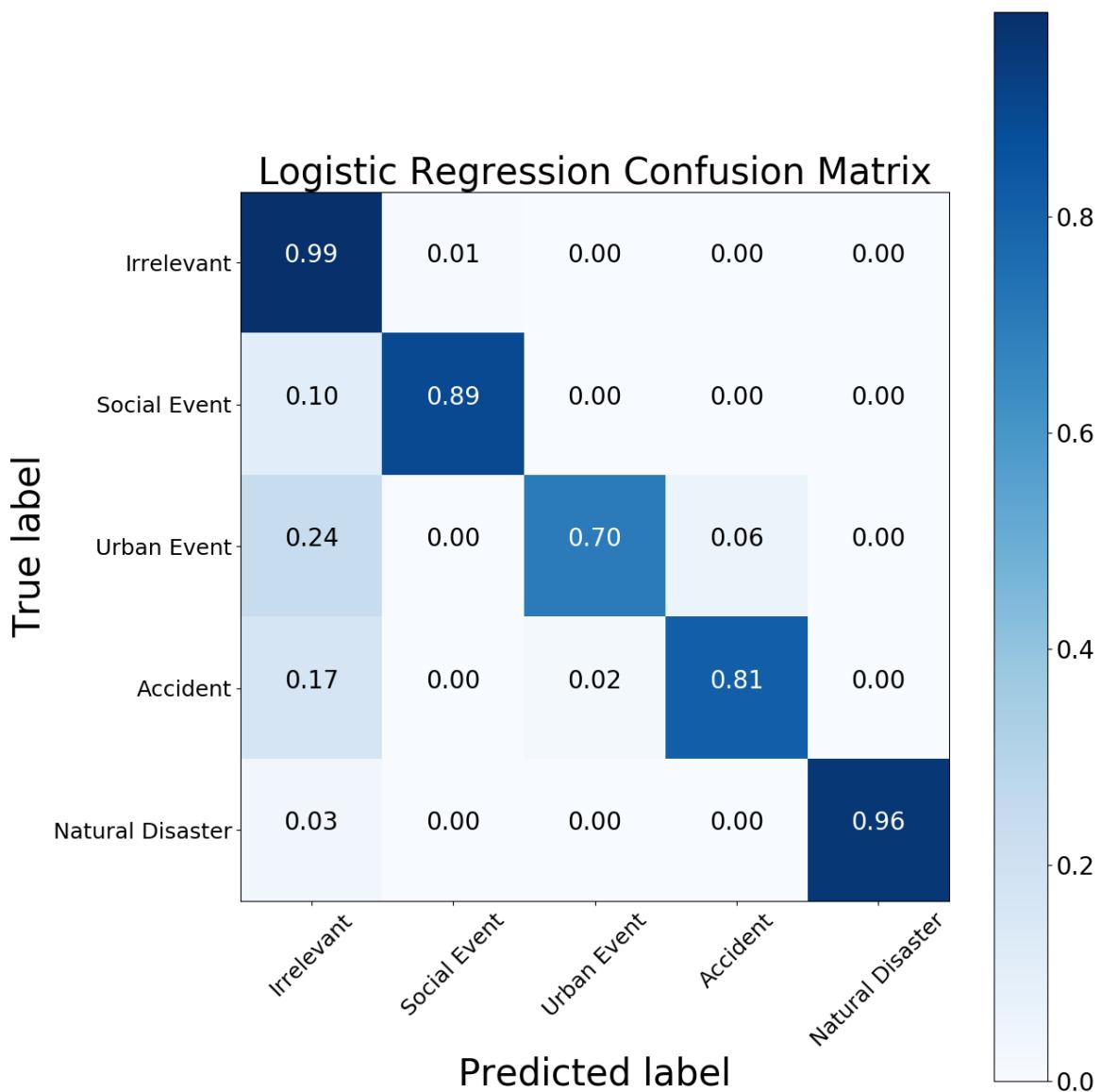


Figura 20 – Matriz de confusão relacionada a classificação dos tweets em eventos de exceção por meio do algoritmo Árvore de Decisão

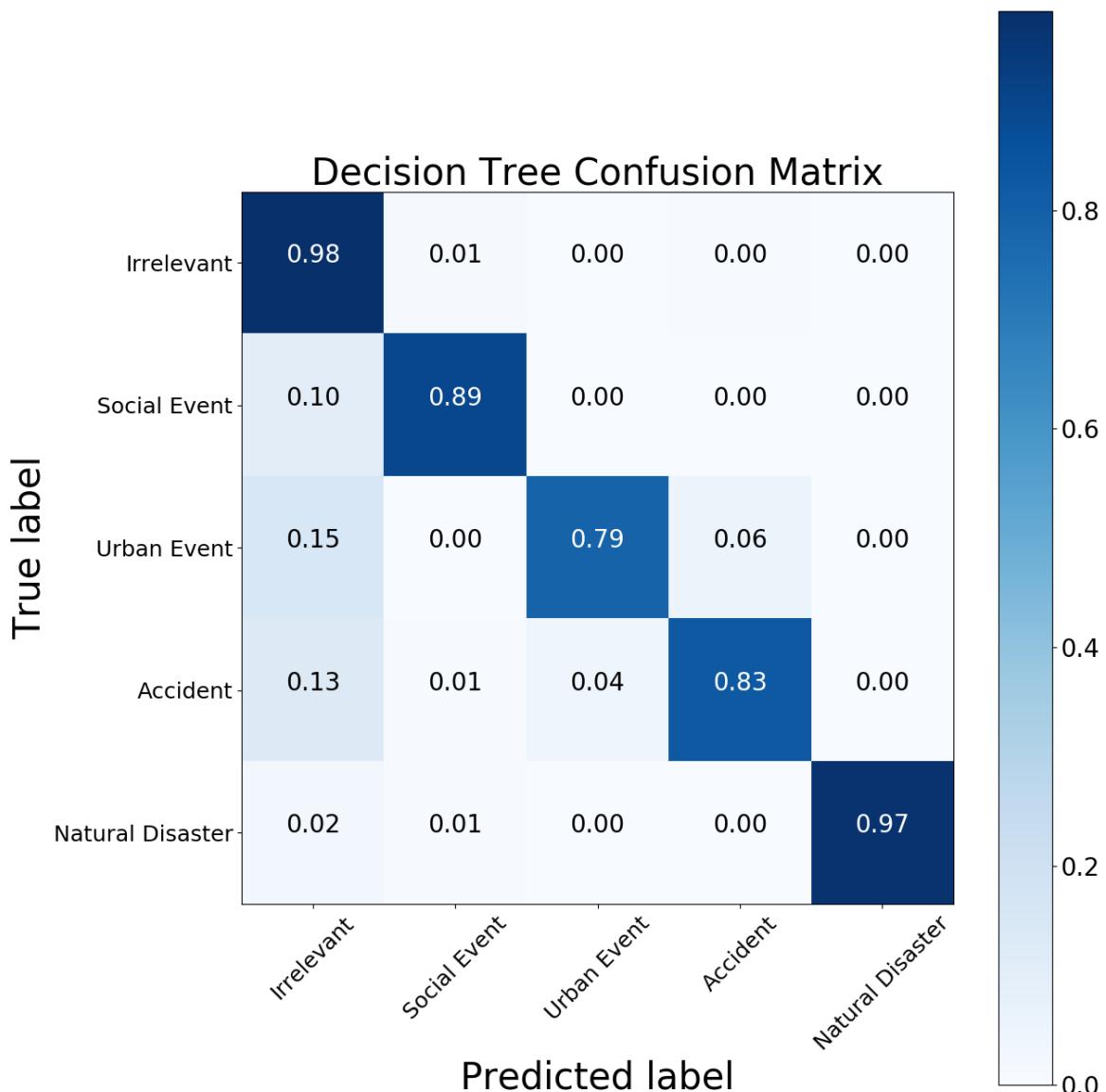


Figura 21 – Matriz de confusão relacionada a classificação dos tweets em eventos de exceção por meio do algoritmo *Multinomial Naive Bayes*

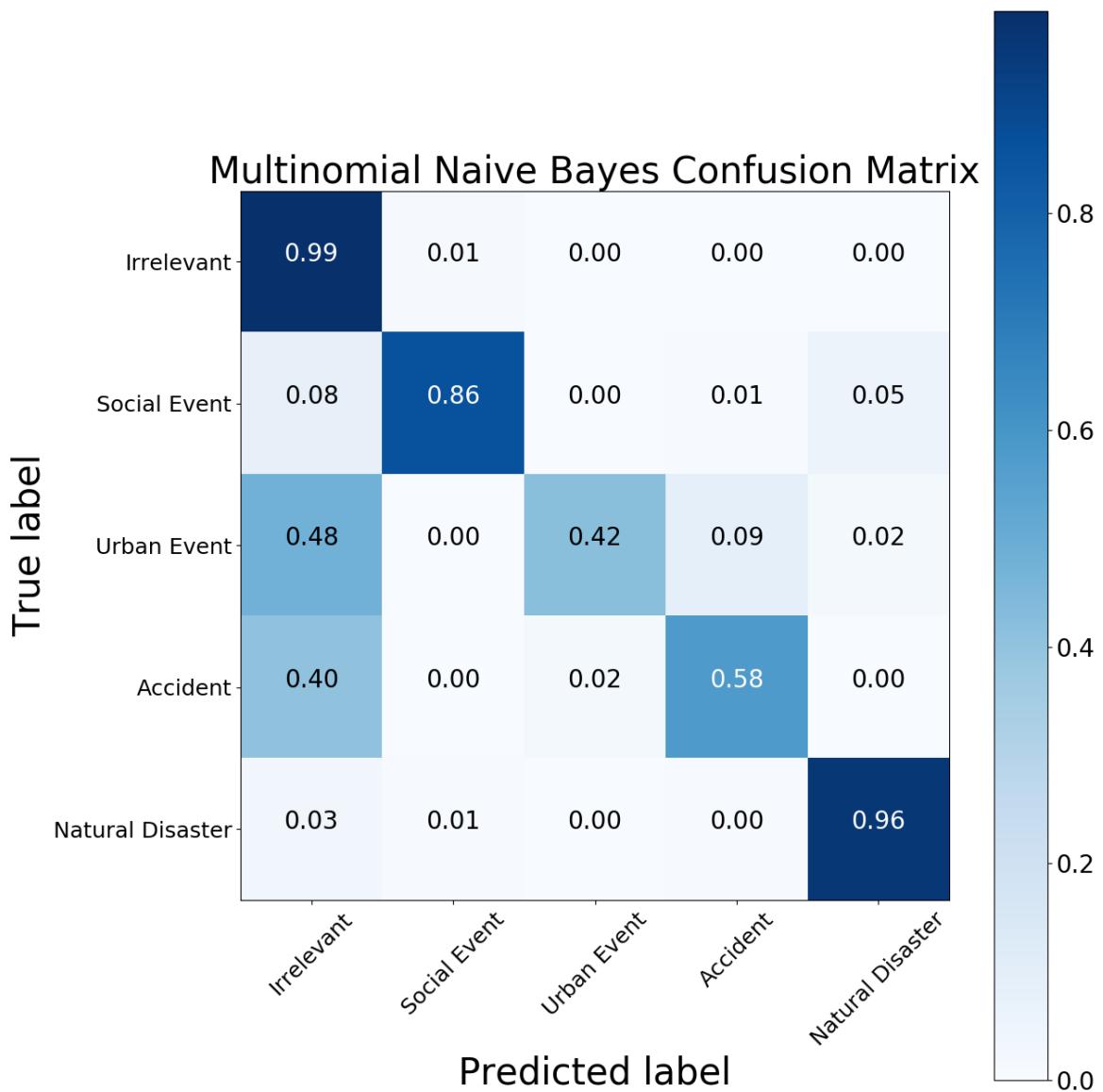


Figura 22 – Matriz de confusão relacionada a classificação dos tweets em eventos de exceção por meio do algoritmo *Gaussian Naive Bayes*

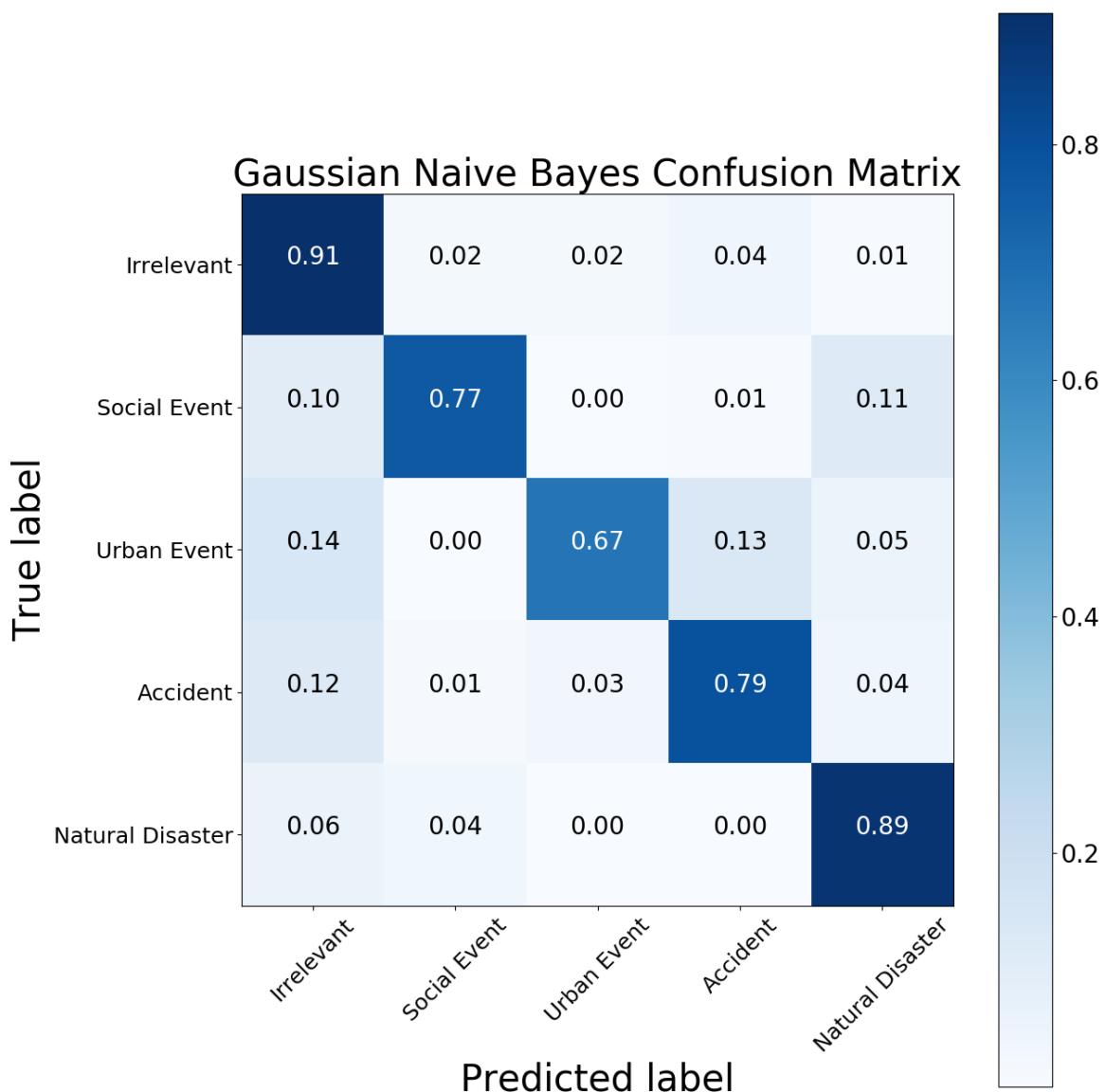


Figura 23 – Matriz de confusão relacionada a classificação dos tweets em eventos de exceção por meio do algoritmo Florestas Aleatórias

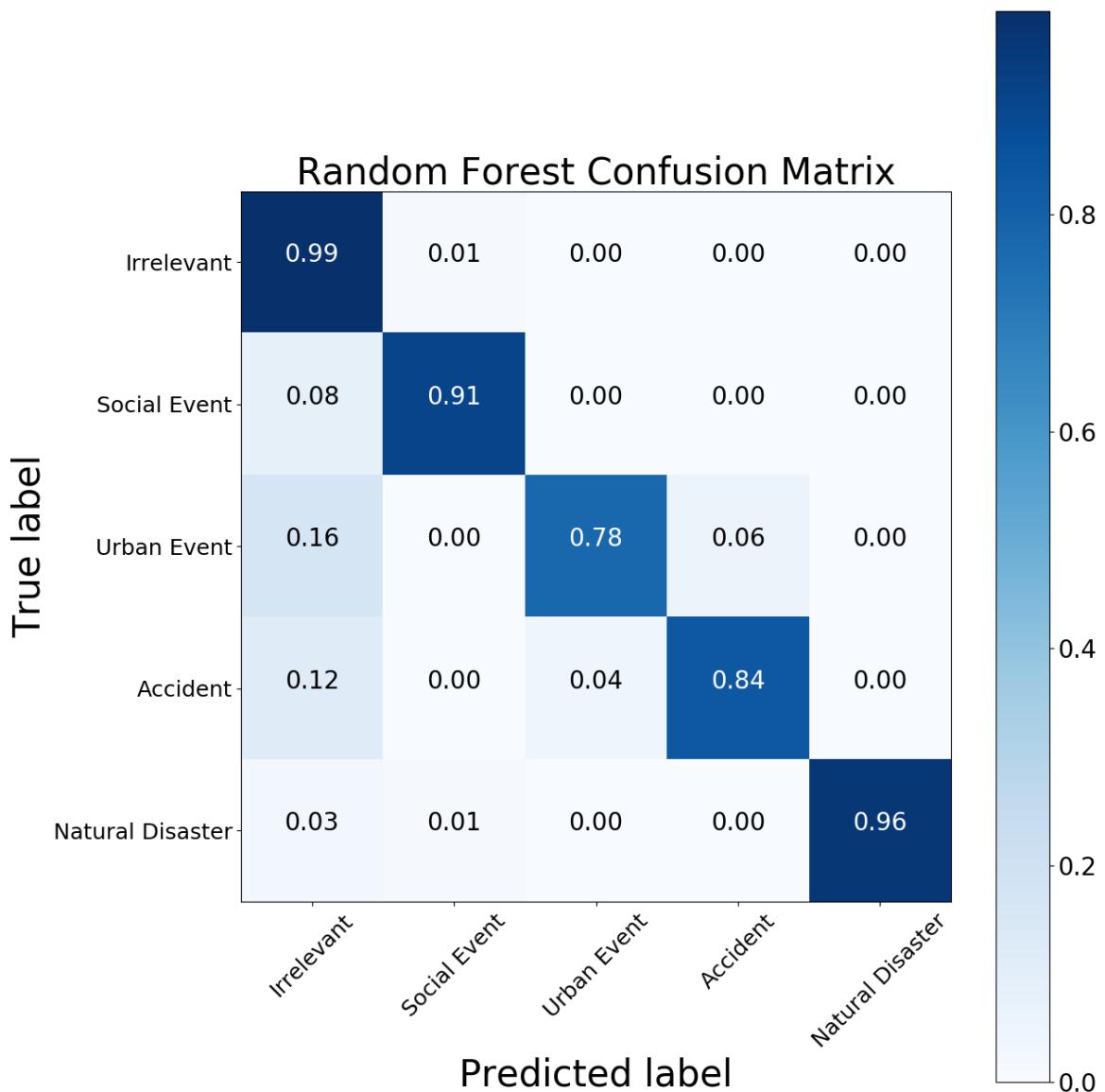
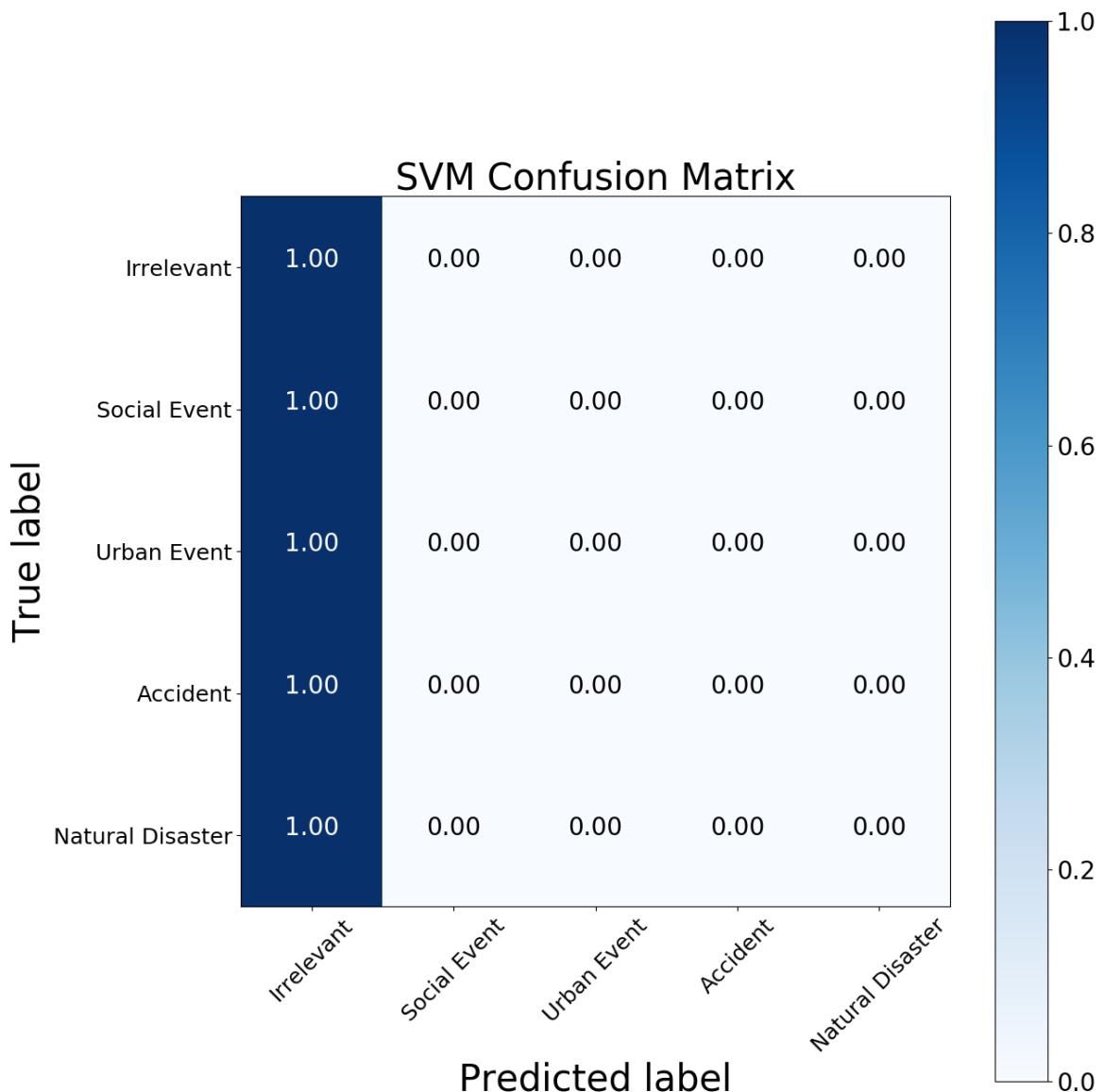


Figura 24 – Matriz de confusão relacionada a classificação dos tweets em eventos de exceção por meio do algoritmo Máquina de Vetores de Suporte



Apêndice F – Parametrizações dos algoritmos