1 Revisão Sistemática

Neste capítulo, é desenvolvida uma Revisão Sistemática (RS) com o 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. Sendo assim, esta RS inicia-se com a seção 1.1, sobre o planejamento da Revisão Sistemática; seguida da seção 1.2, referente as questões de pesquisa utilizadas na formulação do problema da RS; a seção 1.3, sobre o processo de coleta dos estudos primários; da seção 1.4 sobre a avaliação dos dados coletados; da seção 1.5 a respeito da análise e interpretação dos estudos selecionados; da seção 1.6 sobre a conclusão e considerações finais; e por fim, da seção 1.7 sobre as limitações da RS e indicações de trabalhos futuros.

1.1 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 (BIOLCHINI et al., 2005). Em seguida, são construídas definições que permitam estabelecer 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 (BIOLCHINI et al., 2005). 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, defini-se a Avaliação de Dados, na qual são selecionados as fontes primárias que deverão ser incluídas na revisão (BIOLCHINI et al., 2005). 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 (BIOLCHINI et al., 2005). E, na quinta etapa são realizados os processos de Conclusão e Apresentação (BIOLCHINI et al., 2005).

1.1.1 Justificativa da Revisão Sistemática

Esta Revisão Sistemática se justifica por não terem sido encontradas revisões 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, 2015a), por outro lado, desenvolveu 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.

1.2 Questões de Pesquisa

Nesta seção, são apresentadas as questões de pesquisa utilizadas para a formulação dos problemas abordados por essa Revisão Sistemática. Por meio delas, busca-se atender os objetivos já mencionados na seção 1.1.1.

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 utilizadas 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 garantia da confiabilidade dos dados processados. 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): *F1 Score* é a principal técnica estatística utilizada para garantir confiabilidade dos dados.

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 as *features* relacionadas ao transporte público?

   O propósito da QP5 é identificar algumas das *features* relacionadas ao Transporte Público. A resposta a essa questão de pesquisa ajudará especialistas no levantamento de *features* relacionadas ao Transporte Público, que podem ser utilizadas em algoritmos de Inteligência Artificial.

   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 de *features* desse contexto.

6. Quais as técnicas de Inteligência Artificial utilizadas no processamento de *tweets*?

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

1.3 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. Constam na Tab. 1 as bases pesquisadas, quantidades de artigos coletados, descartados no processo de filtragem (Fig. 1, descrito na seção 1.4) 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:

***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 1 – 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 Dias

1.4 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), assim como seu respectivo protocolo.

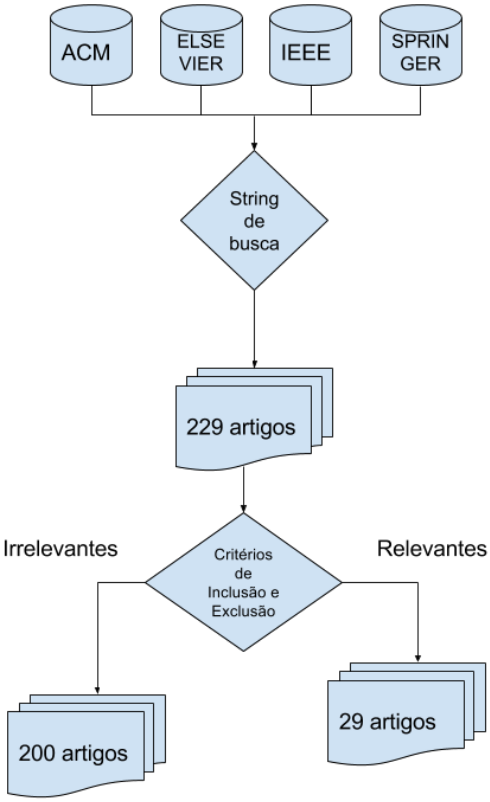

Figura 1 – Processo de Filtragem

Fonte: Felipe Dias

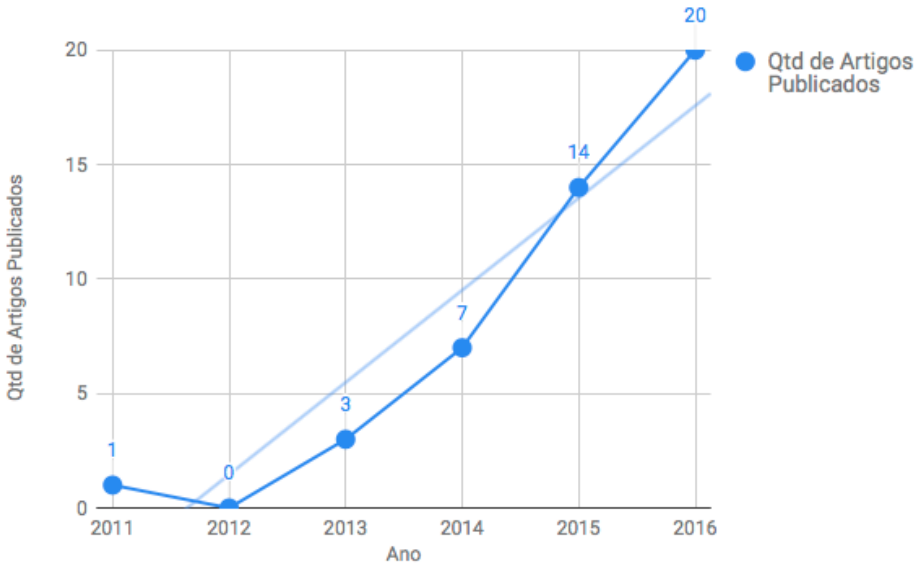

Figura 2 – Quantidade de artigos publicados por ano

Fonte: Felipe Dias

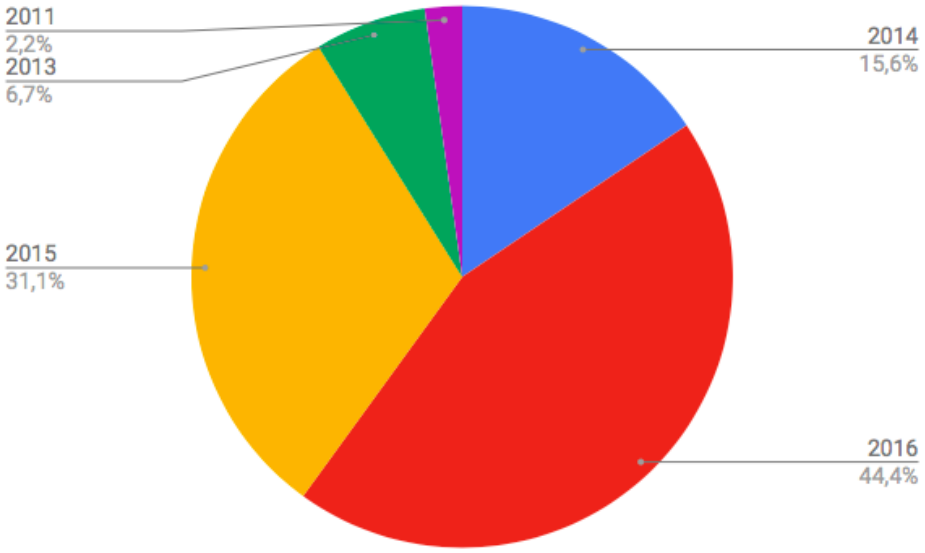

Figura 3 – Porcentagem dos artigos publicados por ano

Fonte: Felipe Dias

Figura 4 – Nuvem de palavras das *keywords* dos artigos selecionados

Fonte: Felipe Dias

Após o processo de condução, alguns dos metadados dos artigos selecionados foram sintetizados.

Sendo assim, a Fig. 4 apresenta uma nuvem de *tags* sintetizando as *keywords* dos estudos primários selecionados; a Fig. 2 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. 3.

1.5 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.

1.5.1 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); (YOUSAF et al., 2014);
5. **Padrões demográficos** (FARSEEV et al., 2015); (GUTEV; NENKO, 2016); (STEIGER; ALBUQUERQUE; ZIPF, 2015b); (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) observa que os *tweets* contendo informações sobre geolocalização são normalmente

publicados em áreas relacionados 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 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).

A parte dos estudos relacionados a Segurança Pública, há exemplos dos *tweets* sendo utilizados para inferir padrões demográficos. Por exemplo, em (FARSEEV et al., 2015); (GKIOTSALITIS; 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) estudou 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; ALBUQUERQUE; ZIPF, 2015b), no qual processou *tweets* 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 a 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 esse 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, USA. 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, relaciona-se ao processamento dos eventos que acontecem na cidade (idealmente em tempo real, como sugere (SOOMRO; KHAN; HASHAM, 2016)). Um dos estudos encontrados sobre esse assunto, foi o realizado por (ANANTHARAM et al., 2015), no qual desenvolveu 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 Baia de São Francisco, USA.

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) inferiu 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) utilizou 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 (NYC), Londres (Reino Unido) e Madrid (Espanha).

1.5.2 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, USA (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) utilizou *tweets* para analisar o impacto dos ataques terroristas em Paris (2015) nos padrões de mobilidade referentes ao uso de transporte público. Semelhantemente, (ITOH et al., 2016) desenvolveu 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) propôs 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), analisou-se 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, incentivamento assim o uso desse modal de transporte. Também relacionado aos pontos de interesse, (MAGHREBI et al., 2015) os utilizou para identificar padrões das atividades humanadas (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.

1.5.3 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 performance** (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. ***F1 Score*** (ANANTHARAM et al., 2015); (CHEN et al., 2016);
4. ***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. ***Linear Regression*** (GUTEV; NENKO, 2016); (BENDLER et al., 2014); (NI; HE; GAO, 2016); (GUO et al., 2016);
8. ***Local Indicators of Spatial Association*** (LISA) (STEIGER; ALBUQUERQUE; ZIPF, 2015b);
9. ***Local Moran's*** (STEIGER; ALBUQUERQUE; ZIPF, 2015b);

10. ***Maximum likelihood estimation*** (MUKHERJEE et al., 2015);
11. ***Seasonal Autoregressive Integrated Moving Average*** (SARIMA) (NI; HE; GAO, 2016);
12. ***Optimization and Prediction with hybrid loss function*** (NI; HE; GAO, 2016).

Em (GUTEV; NENKO, 2016), os autores utilizaram Regressão Linear para analisar a demanda por bicicletas de acordo com as localizações extraídas dos *tweets*. Enquanto que (BENDLER et al., 2014), 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. (NI; HE; GAO, 2016), por outro lado, uniram Regressão Linear com a técnica *Seasonal Autoregressive Integrated Moving Average*, propondo uma abordagem baseada em otimização paramétrica e convexa, chamada *Optimization adn Prediction with hybrid loss function* e adequada para modelagem utilizando séries temporais.

Devido 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, (YOUSAF et al., 2014) agruparam usuários de acordo com a *Cosine similarity*, unindo pessoas com interesses em comuns 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; ALBUQUERQUE; ZIPF, 2015b) usaram a técnica LISA na identificação de *clusters* espaciais e valores esporádicos espaciais, obtendo assim os locais

com atividades sociais. Além disso, também utilizaram *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 *Iverse Coefficient of variation* para verificar a dispersão negativa da distribuição de um conjunto de variáveis.

1.5.4 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; PELECHRINIS, 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; ALBUQUERQUE; ZIPF, 2015b); (GALTZUR 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).

1.5.5 *Features* relacionadas ao transporte público (QP5)

Nesta seção, encontram-se a seguir as *features* extraídas dos estudos primários, relacionadas ao transporte:

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. **Metereológicas**;

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

1.5.6 Técnicas de Inteligência Artificial utilizadas no processamento de *tweets* (QP6)

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

1. ***Bayes classificatio****n* (MATA; CLARAMUNT, 2015);
2. ***C5.0 algorithm*** (ZAGAL; MATA; CLARAMUNT, 2016);
3. ***Conditional Random Field (CRF) with Logistic Regression*** (ANANTHARAM et al., 2015);
4. ***Event extraction based on tweet hashtahs*** (NI; HE; GAO, 2016);
5. ***Latent Dirichlet Allocation*** **(LDA)** (FARSEEV et al., 2015); (ABBASI et al., 2015); (HASAN; UKKUSURI, 2014); (Di Lorenzo et al., 2013);
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 Inteligência Artificial, 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*.

Em (MATA; CLARAMUNT, 2015), por sua vez, aplicou-se 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 a melhor performance 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 relacionas 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*.

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; ALBUQUERQUE; ZIPF, 2015b). (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.

1.6 Conclusão

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, concentram-se em sua maioria na identificação de pontos de interesse, utilizando-os em diferentes contextos, tais como o de Turismo, Mobilidade. Além disso, abordam também as problemáticas relacionadas a 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*, *F1 Score* tem sido a principal técnica estatística utilizada para garantir a confiabilidade do dados, confirmando a terceira hipótese (HP3). Apesar disso, a HP3 não considerou outras técnicas importantes, como a *Linear Regression*, amplamente utilizada nos estudos analisados. Referente as técnicas de Inteligência Artificial, 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.

1.7 Limitações e indicações de trabalhos futuros

As limitações dessa Revisão Sistemática, situam-se na redução do escopo de pesquisa, considerando apenas os estudos que utilizam *tweets* na caracterização de problemas urbanos e relacionados ao transporte público. Sendo assim, é importante ressaltar que existem outras fontes de dados que podem ser utilizadas visando a mitigação de problemas urbanos.

A parte das limitações, indica-se como trabalho futuro a exploração de *features* relacionadas ao Transporte Público, pois essas podem contribuir no melhor entendimento dos eventos capazes de impactar a mobilidade urbana, assim como a construção de sistemas que sejam capazes não apenas de coletar, mas também de processar *tweets* em tempo real, contribuindo assim para melhores tomadas de decisão.

Referências[1]

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 5 vezes nas páginas 8, 9, 15, 16 e 18.

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 7 vezes nas páginas 8, 10, 13, 14, 15, 16 e 17.

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 5 vezes nas páginas 8, 10, 13, 14 e 15.

BIOLCHINI, J. et al. Techincal report rt-es 679/05: Systematic review in software engineering. *COPPE/UFRJ, 2005Rio de Janeiro*, 2005. Citado 2 vezes nas páginas 1 e 2.

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 8 e 15.

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 2.

CHEN, L. et al. Dynamic Cluster-Based Over-Demand Prediction in Bike Sharing Systems. *UBICOMP*, p. 841–852, 2016. Citado 9 vezes nas páginas 8, 10, 11, 12, 13, 14, 15, 16 e 17.

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 8 e 9.

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 8, 10, 15, 16 e 18.

DIAS, F. *Repositório contendo os artefatos da Revisão Sistemática*. 2017. Disponível em: ⟨https://github.com/fcas/dissertacao⟩. Citado na página 5.

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 6 vezes nas páginas 8, 9, 15, 16, 17 e 18.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

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 6 vezes nas páginas 8, 11, 13, 14, 15 e 17.

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 6 vezes nas páginas 11, 12, 14, 15, 16 e 17.

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 8, 9 e 13.

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 8, 9, 12, 13 e 15.

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 6 vezes nas páginas 8, 9, 11, 13, 14 e 17.

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 7 vezes nas páginas 8, 9, 11, 12, 13, 14 e 15.

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 8, 9, 16 e 18.

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 4 vezes nas páginas 11, 12, 15 e 16.

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 2.

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 4 vezes nas páginas 8, 10, 15 e 16.

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 8, 9, 11 e 12.

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 5 vezes nas páginas 8, 9, 15, 16 e 17.

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 7 vezes nas páginas 8, 10, 12, 13, 14, 15 e 17.

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 6 vezes nas páginas 11, 12, 13, 14, 16 e 18.

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 8 e 10.

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 8, 10 e 15.

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 2.

STEIGER, E.; ALBUQUERQUE, J. P. de; ZIPF, A. *An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data*. 2015. 809–834 p. Citado 6 vezes nas páginas 8, 9, 13, 14, 15 e 18.

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 8 e 9.

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 na página 10.

WEN, X.; LIN, Y.-R.; PELECHRINIS, K. PairFac: Event Analytics through Discriminant Tensor Factorization. *Cikm*, p. 519–528, 2016. Citado 8 vezes nas páginas 8, 9, 11, 12, 13, 15, 16 e 17.

YOUSAF, 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 8, 11, 13, 14 e 15.

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 8, 9, 16 e 17.