UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

PAULO MATHEUS PERUZZO STORRER

PERCEPÇÃO DOS USUÁRIOS DO *TWITTER* EM CURITIBA SOBRE MOBILIDADE NO TRANSPORTE PÚBLICO

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2017

PAULO MATHEUS PERUZZO STORRER

PERCEPÇÃO DOS USUÁRIOS DO *TWITTER* EM CURITIBA SOBRE MOBILIDADE NO TRANSPORTE PÚBLICO

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Nádia Puchalski Kozievitch

CURITIBA
2017

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Câmpus Curitiba
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática
Coordenação do Curso de Bacharelado em Sistemas de Informação

TERMO DE APROVAÇÃO

PERCEPÇÃO DOS USUÁRIOS DO TWITTER EM CURITIBA SOBRE MOBILIDADE NO TRANSPORTE PÚBLICO

por

Paulo Matheus Peruzzo Storrer

Este Trabalho de Conclusão de Curso foi apresentado às **18hs** do dia **01** de **dezembro** de **2017** como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação na Universidade Tecnológica Federal do Paraná - UTFPR - Câmpus Curitiba. O(a)(s) aluno(a)(s) foi(ram) arguido(a)(s) pelos membros da Banca de Avaliação abaixo assinados. Após deliberação a Banca de Avaliação considerou o trabalho _____.

_____ **Nádia Puchalski Kozievitch** (Presidente - UTFPR/Curitiba)	_____ **Alexandre Reis Graeml** (Avaliador 1 – UTFPR/Curitiba)
_____ **Rita Cristina Galarraga Berardi** (Avaliador 2 - UTFPR/Curitiba)	_____ **Leyza Baldo Dorini** (Professor Responsável pelo TCC – UTFPR/Curitiba)
_____ **Leonelo Dell Anhol Almeida** (Coordenador(a) do curso de Bacharelado em Sistemas de Informação – UTFPR/Curitiba)	

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso."

AGRADECIMENTOS

Agradeço primeiramente à Deus pelas oportunidades e também por ter me dado saúde e força para superar as dificuldades.

Aos meus pais e à minha esposa por todo amor, incentivo e apoio.

A Universidade Tecnológica Federal do Paraná pelo ambiente e oportunidade de fazer o curso.

A todos os professores ao longo do curso, em especial à minha orientadora Nádia Puchalski Kozievitch pelo auxílio, correções e oportunidades proporcionadas.

A todos que direta ou indiretamente fizeram parte da minha formação.

“Tudo o que temos de decidir é o que fazer com o tempo que nos é dado.”

— J. R. R. Tolkien

RESUMO

STORRER, P. *Percepção dos Usuários do Twitter em Curitiba Sobre Mobilidade no Transporte Público*. Trabalho de conclusão de curso, 2017.

É nas áreas urbanas que se concentra pouco mais da metade da população mundial. Um desafio das grandes cidades para melhoria da qualidade de vida é a questão sobre mobilidade. Mobilidade envolve tanto o movimento de pessoas e objetos quanto informações. Nesse trabalho será abordada a mobilidade humana. Ao se tratar de mobilidade humana nas grandes cidades também vem a questão do transporte público, que é o meio utilizado pela a maior parte da população para se locomover. Uma cidade com um bom transporte público possibilita não somente melhorias no tráfego, como também no meio ambiente e até na qualidade de vida da população. Uma das formas de medir o quão bom é o transporte público é a partir das opiniões dos usuários. Porém pesquisas de satisfação (p. ex. questionários), mesmo se feitas online, tem uma alta demanda de tempo e esforço, tanto na realização quanto na análise, além de depender de que os usuários respondam a pesquisa. Em contrapartida, os usuários de redes sociais estão a todo momento fornecendo dados, que muitas vezes não são utilizados. A partir dos dados de *Tweets* geolocalizados em Curitiba, foi realizada uma filtragem a fim de obter as informações relevantes na questão de mobilidade, principalmente no transporte público. Com os dados do filtro, uma análise de sentimento foi aplicada, que em conjunto com uma análise de um Sistemas de Informação Geográfica (SIG) possibilitou uma comparação com os dados do trabalho de Kozievitch *et al.* (2015), este que contém informações referentes à cidade de Curitiba, como por exemplo sobre abrangência das linhas de ônibus e horários de pico. Por fim foram levantados os prós e contras sobre a utilização de dados provenientes do *Twitter*, a ferramenta de análise de sentimento escolhida e a disposição gráfica dos dados, como também propostas para trabalhos futuros.

Palavras-chaves: Análise de Sentimento, Dicionário Léxico, Mobilidade, SIG, Transporte Público, *Twitter*.

ABSTRACT

STORRER, P. *Perception of Twitter's Users from Curitiba About Mobility in Public Transportation*. Course Work Conclusion, 2017.

It is in urban areas that more than half of the world population is concentrated. A major challenge for large cities to improve quality of life is related to mobility. Mobility can be considered both for the movement of people, objects and information, in this work it will be approached the human mobility. When dealing with human mobility in big cities also comes the issue of public transportation, which is the medium that most of the population uses to get around. A city with good public transportation is a city with a good quality of life. How to measure how good a public transport is? From users's opinions. However, satisfaction surveys (e. g. questionnaires) have a high demand for time and effort. In contrast, social networking users are constantly providing data, which is often not used. From the geolocated Tweets data in Curitiba, a filtering was done in order to obtain relevant information in the mobility issue, especially in public transportation. With the filter data, a sentiment analysis was applied, that in conjunction with a Geographic Information Systems (GIS) made possible to compare data from the work of Kozievitch *et al.* (2015), which contains information regarding the city of Curitiba, such as on comprehensiveness of bus lines and rush hours. Finally, were raised the pros and cons of using data from the Twitter, the sentiment analysis tool chosen and the graphical layout of the data, as well as proposals for future work.

Key-words: Lexical Dictionary, GIS, Mobility, Public Transportation, Sentiment Analisys, *Twitter*.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

É nas áreas urbanas que se concentra pouco mais da metade da população mundial, com uma projeção de que chegue a 66% até 2050, e também, a maioria das atividades econômicas, governamentais, comerciais e de transporte. O processo de urbanização, na sua maioria, não foi planejado e ameaça o desenvolvimento sustentável com a rápida expansão, poluição e padrões de consumo (NATIONS, 2014), além da alta demanda de serviços públicos como saneamento básico, energia elétrica, saúde, educação e transporte (ZENG *et al.*, 2014).

O transporte público tem um aspecto importante nas grandes cidades, pois oferece um transporte de massa e compartilhado para a população. Além de economicamente reduzir o custo geral dos meios de transporte, garante a toda população a capacidade de se locomover pela cidade, sendo ambientalmente mais eficiente. Contudo, devido ao rápido crescimento da população das cidades, as rotas de transporte público existentes não abrangem a necessidade atual (ZENG *et al.*, 2014), ocorrendo uma concentração maior de rotas em uma parte da cidade do que em outra, mesmo que ambas tenham grande densidade populacional. Em Curitiba, por exemplo, há uma maior concentração de linhas de ônibus no centro, enquanto bairros mais afastados da região metropolitana, mesmo com uma alta densidade populacional, têm uma baixa densidade de linhas de ônibus (KOZIEVITCH *et al.*, 2015).

Para fazer uma análise da mobilidade da população, a fim de trazer informações relevantes para o transporte público, foi proposta a utilização de redes sociais. O uso de redes sociais, além de trazer mais informações para os usuários, quer seja sobre seus amigos, parentes ou notícias do dia a dia, também contêm dados como a localização e o dispositivo utilizado no momento que esse usuário faz alguma tarefa numa rede social (SPENCE; WESTERMAN; HEIDE, 2012). Com os dados provenientes da utilização das redes sociais é possível, por exemplo, obter informações relacionadas à mobilidade humana. Em junho de 2015, 80% dos usuários do Twitter, o acessavam utilizando um dispositivo móvel (CHOI; IMB; HOFSTEDE, 2016), com isso é possível, dependendo do nível de privacidade do usuário, descobrir sua localização aproximada no momento em que utilizou a rede social.Tal informação pode ser usada, por exemplo, para traçar padrões de locomoção (LUO *et al.*, 2016). O *Twitter* também disponibiliza para o público uma opção de capturar os *tweets* (mensagens que são publicadas no *Twitter* pelos usuários) (TwitterSupport, 2016), com um limite de número de *tweets* por tempo de captura e usuários, onde é possível selecionar os *tweets* mais relevantes para o estudo em questão (TwitterAPI, 2016). Tendo em vista essa oportunidade de obter informações de uma rede social, no caso o *Twitter*, muitos trabalhos já relacionaram os *tweets* obtidos pela API disponibilizada com aspectos de mobilidade. Isso porque além do conteúdo (texto) do *tweet*, que já traz muitas informações, o *tweet* também inclui metadados como a localização do usuário.

Para este trabalho são analisados *tweets*, geolocalizados na cidade de Curitiba, coletados pelo projeto EUBra-BIGSEA (EUBra-BIGSEA1, 2016), que é um projeto que visa desenvolver

uma plataforma de acesso via nuvem de gerenciamento e exploração de dados, como por exemplo os *tweets*. Além dos dados já coletados, é possível utilizar a *streaming* disponibilizada do projeto EUBra-BIGSEA (EUBra-BIGSEA1, 2016) para coletar mais *tweets*.

Visando o tema de Cidades Inteligentes, que seria a utilização de tecnologias de informação e comunicação para o planejamento e administração da cidade(WEISS; BERNARDES; CONSONI, 2015), a utilização de dados provenientes e redes sociais pode trazer informações que possam ser utilizadas pelas prefeituras das cidades e também por empresas que gerenciam o transporte público, como por exemplo a Urbanização de Curitiba S/A, que é a empresa responsável pelo transporte público da cidade(URBS, 2017).

1.1 OBJETIVOS

O objetivo desse trabalho é identificar aspectos relevantes sobre a mobilidade no transporte público a partir de dados provenientes dos usuários do *twitter*.

Para atingir o objetivo geral foi adotado os seguintes objetivos específicos:

- Realizar pesquisa sobre termos em mobilidade para filtrar os *tweets*.
- Filtrar os *tweets* em um caso de estudo de Curitiba.
- Aplicar a análise de sentimento nos *tweets*.
- Analisar os *tweets* pela perspectiva de SIG.
- Comparar com os dados de Kozievitch *et al.* (2015).

1.2 JUSTIFICATIVA

Além da relevância da mobilidade para as grandes cidades, a questão da percepção das pessoas sobre os aspectos de satisfação e segurança traz informações que podem ser úteis para os órgãos públicos a fim de um melhor planejamento. A utilização do Twitter para fazer a análise das informações é viável visto que há mais de 313 milhões de usuários ativos mundialmente por mês sendo que 82% destes usuários utilizam um dispositivo móvel, o que é essencial para a análise de mobilidade (AboutTwitter, 2016).

As áreas relacionadas no desenvolvimento do trabalho estão focadas em: 1-Banco de dados; 2-Análise de sentimento; 3-Mobilidade. Essas três áreas principais estão relacionadas conforme a Figura 1.

A partir de fundamentos dessas áreas, utilizar técnicas e abordagens relacionadas a banco de dados para filtrar e estruturar os *tweets* coletados. Após analisar os metadados de cada

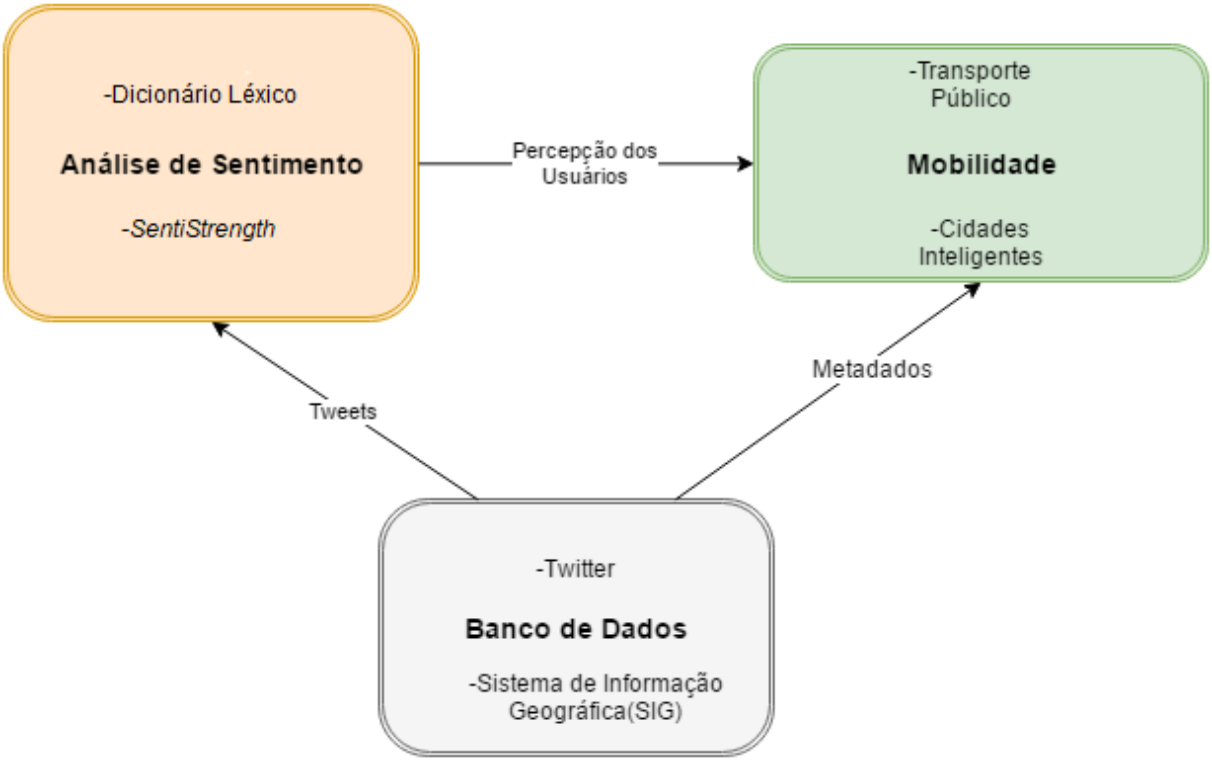

Figura 1 – Áreas Relacionadas

tweet capturado na cidade de Curitiba, e a análise de sentimento do conteúdo dos tweets é esperado verificar a percepção das pessoas sobre o transporte público em Curitiba.

Alguns autores já utilizaram o *Twitter* como fonte de dados para obter informações relativamente precisas de maneira automatizada, O'Connor *et al.* (2010) coletaram *tweets* relacionados à confiança dos consumidores, aprovação presidencial e eleições. Também com essa coleta foram feitas pesquisas diretamente com a população sobre esses assuntos. Após a utilização de técnicas de análise de sentimento, descritas pelos autores como relativamente simples, foi feita uma comparação com os resultados das pesquisas. O resultado foi que, mesmo com a utilização de uma técnica simples, essa análise replicou o resultado das pesquisas feitas, ou seja, ao invés de realizar pesquisas diretamente com as pessoas indagando sobre assuntos específicos, que é uma forma de coleta de informações lenta, a análise de sentimento de textos de redes sociais (p. ex. Twitter), é uma forma rápida e relativamente precisa de se obter os mesmos resultados.

Além disso, o uso do *Twitter* em relação à mobilidade humana também é explorado. Luo *et al.* (2016) utilizaram os dados de localização obtidos dos *tweets* para explorar características demográficas em um caso de estudo na cidade de Chicago. Fizeram uma análise do sobrenome para análise étnica e uma análise do primeiro nome para definição de gênero e também idade, que foi obtida levando em consideração tendências de nomes por geração, foram identificados os residentes de Chicago e então extraídos os *tweets* dos residentes locais. As descobertas mostram que a distribuição dos centros de atividades (de onde provinha a maior parte dos *tweets*) se alinha bem com o desenvolvimento socioeconômico da cidade. Outro trabalho que relaciona os *tweets*

com a mobilidade é de Hawelka *et al.* (2013), que coletaram cerca de um ano de *tweets* de todo mundo e verificaram o país de origem dos usuários, quando os usuários postassem um *tweet* fora do seu local de origem entrariam para uma análise, a fim de descobrir padrões globais da mobilidade humana.

Ainda sobre a mobilidade humana, Frank *et al.* (2015) tiveram uma abordagem diferenciada, se preocuparam com a localização do usuário num espaço de tempo menor, durante o dia e durante os dias da semana. Além da localização, foi feita uma análise no conteúdo do *tweet* para descobrir de onde provinha. Para cada indivíduo foram identificadas as localidades frequentemente visitadas. Para isto eram necessários mais de 50 *tweets* com uma distância de menos de 25 metros. Dividiram em grupos de "dormindo, casa, trabalho e lazer" e distribuídos durante as horas do dia. Foi descoberto que a maioria dos *tweets* parte de casa ou do trabalho. E a análise baysiana foi feita separando mensagens relacionadas a trabalho ou a casa. *Tweets* de casa, por exemplo, utilizam gírias, enquanto os que são feitos do trabalho utilizam uma linguagem mais formal.

Juntamente com os dados que forem obtidos nesse trabalho, é possível uma comparação com os dados de Kozievitch *et al.* (2015), que utilizaram um SIG, relacionando dados das linhas de ônibus de Curitiba, obtendo um mapa de densidade das linhas, pontos e terminais das ruas. Uma vez que já têm informações sobre a cidade de Curitiba provenientes de outras fontes além do *twitter*(p. ex. cartões de usuário), essa comparação proporcionaria resultados relevantes na questão de utilização de redes sociais para obter informações referentes à mobilide, mais especificamente ao transporte público.

2 MÉTODO

O objetivo principal do trabalho é identificar aspectos relevantes para a mobilidade na cidade de Curitiba a partir dos *Tweets* geolocalizados coletados durante o período de março a junho de 2016. Para isso foram seguidos os seguintes passos metodológicos: 1) revisão teórica, 2) pesquisa com usuários, 3) filtro dos *tweets*, 4) análise de sentimento, 5) análise SIG e 6) comparação com Kozievitch *et al.* (2015). Os passos estão representados da Figura 2.

Figura 2 – Método

2.1 REVISÃO TEÓRICA

Nessa etapa foram verificados e escolhidos métodos para a análise de sentimento e também demonstração dos resultados em nível de banco de dados geográficos. Os métodos mais presentes na literatura foram analisados e escolhidos de acordo com a complexidade de aplicação para o cenário em questão, como também da precisão, levando em consideração o conhecimento técnico do aplicador do método.

2.2 PESQUISA COM USUÁRIOS

Segundo a página da UFMG (EUBra-BIGSEA2, 2017) do projeto EUBra-BIGSEA (EUBra-BIGSEA1, 2016), responsável pela coleta dos *Tweets* que foram utilizados, algumas relações de palavras e de *hashtags* já foram selecionadas para uma filtragem em relação à mobilidade. A filtragem foi dividida em duas, a do conteúdo do texto do *Tweet* e das *hashtags*.

Através desses dados, que tratam sobre mobilidade na cidade de Belo Horizonte, foram alteradas as palavras-chave para a realidade de Curitiba, além disso, é criado e aplicado um ques-

tionário (Apêndice A) para obter os termos mais utilizados pelos usuários de *twitter* ao se tratar de assuntos envolvendo mobilidade, especificamente o transporte público. Maiores detalhes na Seção 4.2.

2.3 FILTRO DOS *TWEETS*

Nessa etapa foi feita a filtragem e estruturação dos dados. Os dados obtidos do *Twitter* estavam armazenados no formato JSON, abrangem um período de 4 meses e estão geolocalizados na cidade de Curitiba. Após a coleta dos dados do questionário, mencionado na etapa 2.2, e feita a seleção dos dados, descrita na Seção 4.3, os novos *tweets* foram salvos também no formato JSON apenas com os dados de interesse, que seriam o texto, *hashtag*, coordenada e data de criação do *tweet*.

Ainda com os dados no formato JSON, foram criados arquivos de texto separados, que continham apenas o texto de cada *tweet* para utilização na aplicação da ferramenta de análise de sentimento. Cada tipo de classificação foi relacionada com o arquivo JSON original, sendo então, salvos em arquivos JSON separados, de acordo com essa classificação, cada metadado, além do texto em si.

2.4 ANÁLISE DE SENTIMENTO

Para a análise de sentimento foi selecionada a ferramenta *SentiStrength* (SentiStrength, 2016), conforme abordado no referencial teórico (THELWALL *et al.*, 2010), possui um dicionário pré-definido com termos já classificados. Utilizando a versão em JAVA, cujo dicionário é em inglês, há a necessidade de traduzir o dicionário ou traduzir o *tweet*. Para fins de comparação, as duas abordagens são realizadas, onde para a tradução é utilizado o *Google Tradutor* (GoogleTradutor, 2017). O desenvolvimento dessa etapa é descrito na Seção 4.4.

2.5 ANÁLISE SIG

Para visualização dos *tweets* classificados pela análise de sentimento, foi escolhido o *Google Maps* (GoogleMaps, 2017). Essa ferramenta utiliza arquivos no formato *Keyhole Markup Language* (KML), para criação desses arquivos foi primeiramente criado uma planilha a partir dos *tweets* classificados, no formato JSON, e em seguida adicionado o campo ícone para representação gráfica no mapa. Cada ícone apresentado pelas cores verde, vermelho e branco representa as classificações de cada *tweet* positivo, negativo ou neutro, respectivamente.

2.6 COMPARAÇÃO KOZIEVITCH *ET AL.*(2015)

Através da perspectiva SIG, é feita uma comparação com os mapas gerados por este trabalho com os mapas obtidos de Kozievitch *et al.* (2015). A comparação é feita visualmente a partir dos mapas obtidos na representação gráfica dos *tweets* com os mapas das linhas de ônibus e terminais da cidade de Curitiba. O desenvolvimento dessa etapa está na Seção 4.5.

3 FUNDAMENTAÇÃO TEÓRICA

Nessa seção serão abordadas as três principais áreas que foram selecionadas como base fundamental do trabalho: Banco de Dados, Análise de Sentimento e Mobilidade. A Figura 3 ilistra a distrbuição das referências utilizadas no desenvolvimento do trabalho de acordo com o ano de publicação de cada artigo.

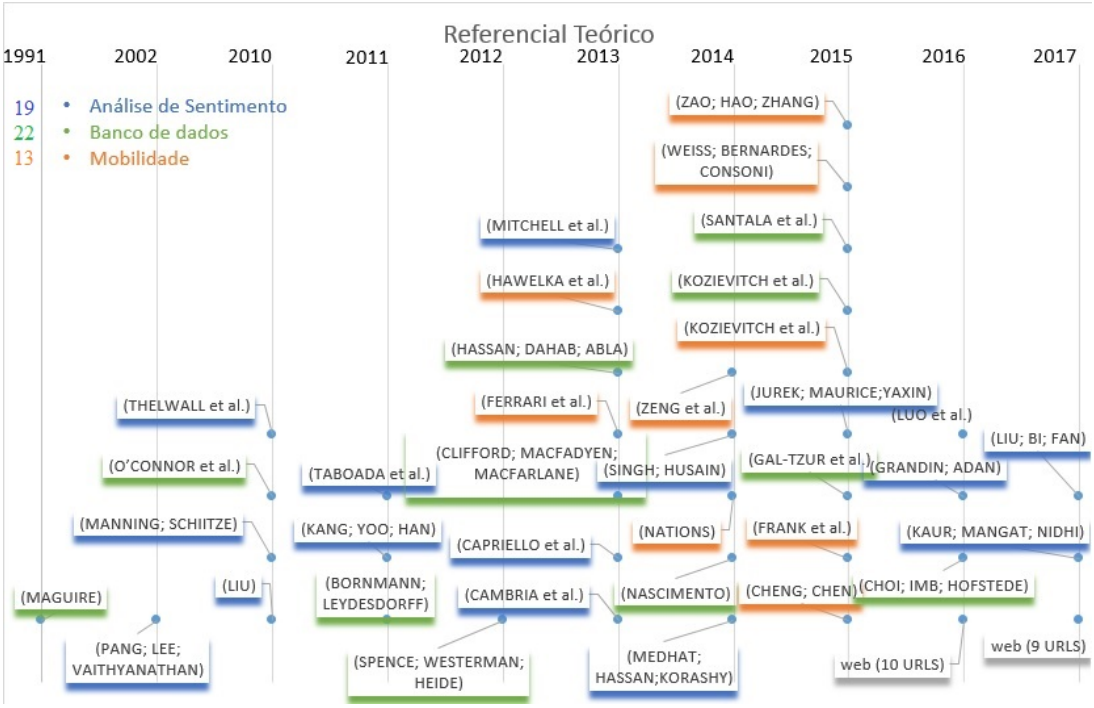

Figura 3 – Relação das Referências.

3.1 ANÁLISE DE SENTIMENTO

Nesta seção serão abordados em profundidade a origem da análise de sentimento, o conceito adotado, algumas técnicas e trabalhos relacionados e a relação com o *Twitter*.

A ciência lingüística estuda o entendimento, a relação entre a diferença do que é escrito para o que a pessoa realmente deseja demonstrar, a relação das estruturas da linguagem, que tipo de coisas as pessoas falam e o que essas coisas dizem sobre o mundo. Com o conhecimento dessa estrutura é possível identificar padrões. O estudo e utilização desses padrões é abordado na área de processamento de linguagem natural (NLP), que, através de métodos estatísticos e análise sintática proveniente da lingüistica, possibilita a criação de um sistema que tome decisões de desambiguação no sentido da palavra, categoria, estrutura sintática e escopo da semântica (MANNING; SCHIITZE, 2010).

Ao tratar do uso de recursos computacionais para entender a linguagem natural é im-

portante ressaltar que, como dizem Cambria *et al.* (2013), "A análise automática de opiniões *online*, na verdade, envolve um profundo entendimento de linguagem natural pelas máquinas, o que ainda estamos muito longe". Alguns algoritmos já retiram informações de textos da internet, como a divisão em partes, contagem de palavras e verificação ortográfica, mas ainda são limitados ao tratar do significado da sentença. Os primeiros trabalhos relacionados à área tentaram classificar uma sentença em positiva ou negativa. Em seguida alguns trabalhos focaram em distinguir os textos que poderiam indicar algum tipo de sentimento dos que não indicavam (CAMBRIA *et al.*, 2013).

Segundo Liu (2010), "A análise de sentimento ou mineração de opinião é o estudo computacional de opiniões, sentimentos e emoções expressadas no texto.", No trabalho de Liu (2010) há um exemplo de um texto de uma qualificação do iPhone:

> "(1) Comprei um iPhone alguns dias atrás. (2) Era um telefone tão bom. (3)A tela era muito boa. (4) A qualidade de voz era limpa também. (5) Apesar da duração da bateria não ser longa, estava bom para mim (6) Entretanto, minha mãe ficou brava comigo pois não contei para ela antes de comprar (7) Ela também achou que o telefone foi muito caro, e quis que eu devolvesse..."(LIU, 2010)

A partir da divisão do texto em partes numeradas foram identificadas que (2), (3), (4), (5) são referentes à qualificação feita pelo autor, enquanto as sentenças (6) e (7) referem a mãe dele. E com isso mostra uma das dificuldades relacionadas à análise de sentimento.

O autor traz a definição de recursos explícitos e implícitos, onde os recursos explícitos estão descritos na sentença como no exemplo citado pelo autor, "A duração da bateria é pequena" está explícito que o recurso é "duração da bateria", já em "O telefone é muito grande", está implícito que o recurso é o tamanho. Define também que uma opinião sobre um recurso pode ser positiva, negativa ou neutra, e com isso define a "sentença opinada", que é uma sentença onde está explícita a qualificação do seu conjunto de opiniões(LIU, 2010).

Taboada *et al.* (2011) apresentam um método para obter o sentimento de um texto que utiliza um dicionário de palavras já com suas orientações de opinião. Há duas abordagens principais para fazer a análise do sentimento: através de uma abordagem léxica ou da classificação do texto (TABOADA *et al.*, 2011). Pela abordagem de classificação são utilizadas técnicas de aprendizado de máquina. A análise pode ser feita pelo classificador de Bayes que dividirá as opiniões em positivas e negativas (PANG; LEE; VAITHYANATHAN, 2002). Já pela abordagem léxica, é utilizado um dicionário de palavras, onde cada palavra já representa, o que seria em uma sentença, uma opinião negativa ou positiva, e então a partir dessa palavra, ou combinações, a sentença é classificada (KANG; YOO; HAN, 2011).

Aprofundando mais essa divisão, Kaur, Mangat e Nidhi (2017) demonstram na Figura 4, o que seria a ramificação dos métodos de abordagem léxica("*Lexicon*") e classificação("*Machine Learning Aproach*").

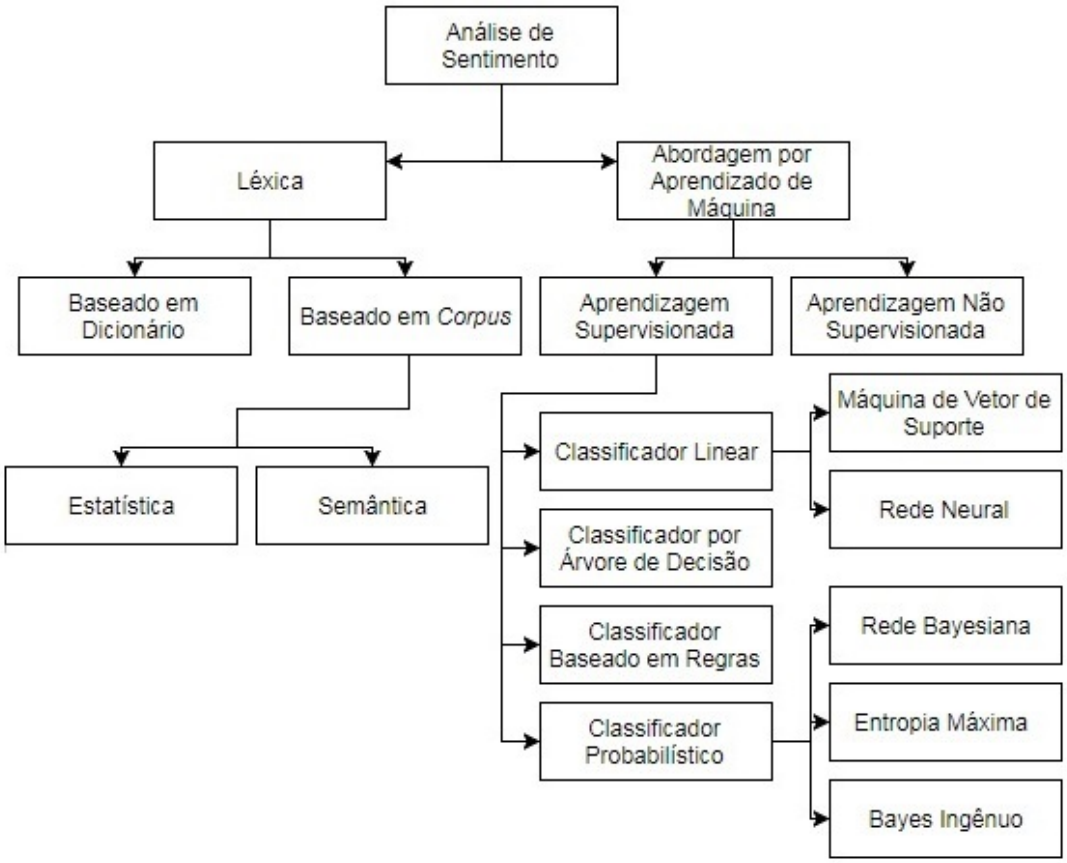

Figura 4 – Técnicas de Classificação de Sentimento(KAUR; MANGAT; NIDHI, 2017).

Na abordagem léxica baseando-se em dicionário, uma lista de palavras é coletada e preparada manualmente com sua classificação onde a maior dificuldade é em relação ao contexto. Um exemplo dessa aplicação seria de Mitchell *et al.* (2013), que utilizaram 10% do total de *tweets* do ano de 2011 dos Estados Unidos, dos quais foram selecionados os que estavam geolocalizados, o que representou cerca de 1% do total de *tweets* daquele ano, aproximadamente 10 milhões de *tweets*. O dicionário foi obtido a partir da Avaliação da Linguagem pelo Turco Mecânico da *Amazon* (Amazon, 2017), que avaliou o sentimento das palavras na escala de 1 (triste) a 9 (feliz). O cálculo da sentença foi feito através do somatório do peso de cada palavra multiplicado pela frequência normalizada dessa palavra. Além disso, palavras que tinham uma pontuação média, entre 4 e 6 foram descartadas do cálculo. No final, obtiveram uma classificação para cada estado com uma média de "felicidade" e esse resultado foi comparado com dados de pesquisas, como por exemplo relacionados ao nível de obesidade. Um resultado dessa relação foi que a "felicidade" diminui enquanto o nível de obesidade aumenta.

Capriello *et al.* (2013) trazem o contexto da opinião de clientes sobre o turismo de fazenda. Fizeram a utilização do método de análise de sentimento baseada em *corpus* semântico. Através da opinião de convidados dos Estados Unidos e do Reino Unido é mostrada a categoria semântica onde foram feitas as comparações entre os termos utilizados por cada cliente.

Além da abordagem semântica baseada em *corpus* , há a abordagem estatística, onde a polaridade de uma palavra é dita positiva se a ocorrência desta palavra é identificada, na maior

parte, em contextos positivos. Além disso, a utilizRação da abordagem baseada em *corpus* não é tão eficiente quanto a baseada em dicionário, pois é difícil abranger todas as palavras existentes. Porém a maior vantagem é em relação a identificação do contexto (MEDHAT; HASSAN; KORASHY, 2014).

Singh e Husain (2014) utilizaram dados de revisões de filmes para trazer uma análise de quatro métodos de análise de sentimento e mineração de opinião, juntamente com suas vantagens e desvantagens das abordagens de aprendizado supervisionado: Classificador Ingênuo de Bayes, Máquina de Vetores de Suporte e Perceptron Multicamadas, da abordagem de aprendizado não supervisionado, o *Clustering*.

A conclusão dos autores Singh e Husain (2014) sobre o método do Classificador ingênuo de Bayes é que é fácil de interpretar e também eficiente computacionalmente, porém as atribuições são feitas para cada palavra, enquanto a sentença por inteiro pode ter um outro sentido. Sobre a máquina de vetores de suporte, que é um modelo de aprendizagem supervisionado, chegaram à conclusão que tem boa performance porém o resultado do modelo é de difícil interpretação, além de ser necessário um pré-processamento de valores faltantes. O perceptron de multicamadas é baseado nos conceitos de redes neurais, onde cada neurônio tem um processamento que, somado aos resultados dos outros neurônios, irá chegar a um resultado final. O resultado foi que o modelo tem uma alta taxa de precisão e pode aprender a partir das entradas e saídas. A maior desvantagem é que requer uma grande quantidade de dados para "treino" da rede. Já o *Clustering* é um método não supervisionado de aprendizagem. Os objetos e instâncias são organizados em classes ou grupos onde os membros de um mesmo *cluster* são parecidos de alguma maneira e os membros de *clusters* diferentes não são similares entre si. A precisão deste método varia muito. A vantagem é que ele oferece classes ou grupos que completam uma medida aproximadamente perfeita, já a desvantagem é que não há uma aprendizagem e o número de grupos é desconhecido.

Liu, Bi e Fan (2017) relatam sobre o classificador por árvore de decisão, que utiliza uma amostragem de treinamento. Ao classificar os textos, ramos são criados com os termos que compõem esse texto. Com a utilização de 3 bases de dados foram aplicados outros métodos de aprendizagem. Como resultado final, em comparação, não apresentou boa precisão ou tempo de execução (LIU; BI; FAN, 2017).

Na questão de ferramentas para fazer a análise de sentimento, Grandin e Adan (2016) relatam o desenvolvimento de um sistema chamado Piegas, que faz a análise automática do sentimento para o idioma português. O sistema requer que seja introduzido o tópico de interesse e então utiliza o Classificador Ingênuo de Bayes para identificar a análise de sentimento dos tweets. O artigo, além do Classificador de Bayes, cita o método de Máquina de Vetores de Suporte como os mais populares. Algumas das dificuldades citadas no artigo referem-se ao uso de gírias que não seguem as normas da língua padrão e, portanto, são parte da margem de erro. Uma necessidade do algoritmo é ter *tweets* já classificados como positivos e negativos sobre vários assuntos.

Levando em consideração a ampla utilização da análise de sentimento em que grande maioria apenas define o resultado final da sentença, positiva ou negativa, Jurek, Maurice e Yaxin (2015) propõem uma análise de sentimento melhorada, em que, ao invés de apenas decidir se uma sentença é positiva ou negativa, coloca-se uma medida de sentimento, ou seja, o quanto uma sentença é positiva ou negativa. Além disso, a partir dos *tweets* coletados, foram filtrados temas envolvendo eventos proporcionados pela liga de defesa inglesa (*English defence league* - EDL), ocorridos no Reino Unido e analisados aspectos de violência, comportamentos anti-sociais e prisões.

Uma ferramenta chamada *SentiStrength* (SentiStrength, 2016), propõe a classificação do sentimento de textos informais. Thelwall *et al.* (2010) relatam o desenvolvimento da ferramenta com a utilização da rede social *MySpace*. O núcleo do algorítimo conta com uma coleção de termos positivos e negativos, classificados com base nas opiniões das pessoas. Após o desenvolvimento e aplicação do algorítimo, é observada uma precisão de cerca de 60% para emoções positivas e de 73% para emoções negativas, utilizando uma escala que varia de 1 a 5. Uma possibilidade da ferramente é a de treinamento, a partir de que podem vir a ser alterados os valores padrões de sentimento dos termos pré-definidos.

A partir dos trabalhos pesquisados, é possível verificar que há diversas técnicas e abordagens em relação à análise de sentimento. Para desenvolvimento foi escolhida a ferramenta *SentiStrength*, entretanto, essa ferramenta para a versão em JAVA utiliza o idioma inglês. Nesse caso é preciso criar um novo dicionário em português, a partir da tradução do dicionário já disponível em inglês, e inclusive realizar a tradução das sentenças em português para o inglês.

3.2 BANCO DE DADOS

Um banco de dados pode ser utilizado para armazenar *tweets*. O que precisa ser decidido é quais metadados serão necessários antes da criação do banco de dados (GNIP, 2016). Segundo o *Twitter* há duas APIs: *Streaming* e Transferência de Estado Representacional *REST*. Com a API *REST* é possível ler e escrever dados do *Twitter*, criar um novo *Tweet*, ler o perfil do autor e dos seguidores. Já a API de *streaming* é para monitorar os *Tweets* em tempo real. O *Streaming* pode ser usado para coleta de *stream* público, que busca dados públicos dos usuários. Pelo *Twitter*, é possível selecionar apenas um usuário para ser feita a coleta dos dados, ou por *site*, essa restrição é indicada quando relacionada a servidores que, se conectam ao *Twitter* em nome de muitos usuários (TwitterAPI, 2016).

Os dados são coletados no formato JSON (*JavaScript Object Notation* - Notação de Objetos JavaScript) que é uma formatação leve de troca de dados (JSON, 2016). A Tabela 1, adaptada de JSONTwitter (2016) mostra quais são os metadados que podem ser obtidos de um *tweet* através da API.

Uma vez com os dados estruturados, é possível a utilização destes em uma aplicação

Tabela 1 – Tabela Dados *Tweet* JSON

Campo	Tipo	Descrição
coordinates	Coordinates	Representa a localização geográfica do *twitter* no formato geoJSON.
created_at	*String*	O tempo UTC que o tweet foi criado.
entities	*Entities*	As entidades que foram retiradas do texto do *tweet* (p. ex. *hashtags*, urls)
id	*Int64*	O número de identificação do *tweet*.
lang	*String*	Indica a língua do texto do *tweet*.
place	*Places*	Indica que um *tweet* está associado a(mas não necessariamente se originando) um lugar.
text	*String*	É o texto do *tweet*.
user	*Users*	Informações do usuário que publicou o *tweet*.

SIG. Segundo MAGUIRE (1991), há três visões relacionadas a SIG: visão de mapa, visão da base de dados e visão relacionada a análise espacial. A visão de mapa se refere à cartografia, à disponibilização dos dados de uma maneira gráfica em um mapa. A visão da base de dados é a fundamentação do SIG. Deve estar estruturada de uma maneira que facilite as consultas. A visão de análise espacial é a análise e modelagem das informações espaciais. Além disso, MAGUIRE (1991) descreve que o SIG é composto por quatro elementos básicos: *hardware*, *software*, dados e *liveware*. O *hardware* pode ser qualquer tipo de computador. Já para o *software* há cada vez novos desenvolvimentos e opções mais sofisticadas. Os dados são um recurso custoso para a base de dados que pode, em alguns casos, exceder a capacidade do *hardware* e *software*. Já o último elemento, *liveware*, está relacionado com as pessoas responsáveis pelo *design*, implementação e utilização do SIG.

Um exemplo de utilização de SIG está no trabalho de Hassan, Dahab e Abla (2013), que através da aplicação ArcGIS, desenvolveram um sistema que verificava a localização de ambulâncias, corpo de bombeiros e unidades de polícia, a fim de melhor direcionar aqueles que fossem requeridos para as áreas de necessidade. Para isso, levaram em consideração a rota até o local junto com o tempo de chegada, pois poderia existir uma viatura mais próxima, porém, com um caminho congestionado, devido ao tráfego. Com isso uma mais distante poderia chegar antes, por não existir tráfego na rota (HASSAN; DAHAB; ABLA, 2013).

Em um estudo mais específico referente a mesma localização dos dados desse trabalho tem-se o trabalho de Kozievitch *et al.* (2015), que utilizaram do SIG, mais especificamente SIG-T(*Geographic information systems for transportations* - Sistemas de informações geográficas para transporte), e relacionou os dados das linhas de ônibus na cidade de Curitiba, obtendo mapas de densidade das linhas, pontos e terminais nas ruas da cidade. A Figura 5 é um dos mapas produzidos e publicados no artigo de Kozievitch *et al.* (2015).

Já no trabalho de Mitchell *et al.* (2013), é possível verificar a utilização de dados provenientes do *twitter*, onde utilizaram de SIG para mostrar graficamente o nível de felicidade de cada estado dos Estados Unidos, através da análise de 10 milhões de *tweets* geolocalizados du-

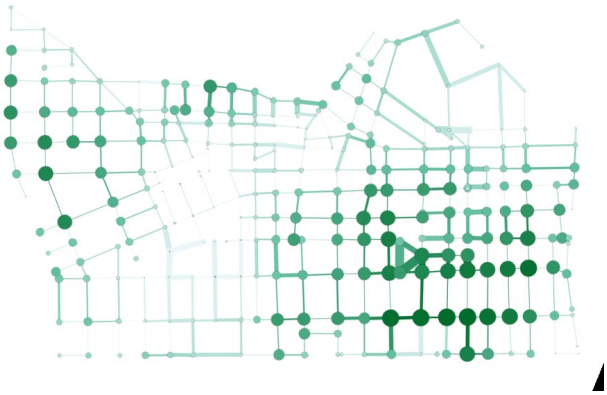

Figura 5 – Centralidade do autovetor das linhas de ônibus no centro. Círculos maiores e mais escuros indicam uma maior densidade.(KOZIEVITCH *et al.*, 2015)

rante o ano de 2011. Como resultado foi obtido um mapa que continha cada *tweet* e seu nível de felicidade, exposto de maneira gráfica utilizando um degradê que variou de azul (mais triste) a vermelho (mais feliz). Também distribuiu graficamente a ocorrência de cada nível de felicidade para cada estado. A Figura 6 foi um dos resultados obtidos (MITCHELL *et al.*, 2013).

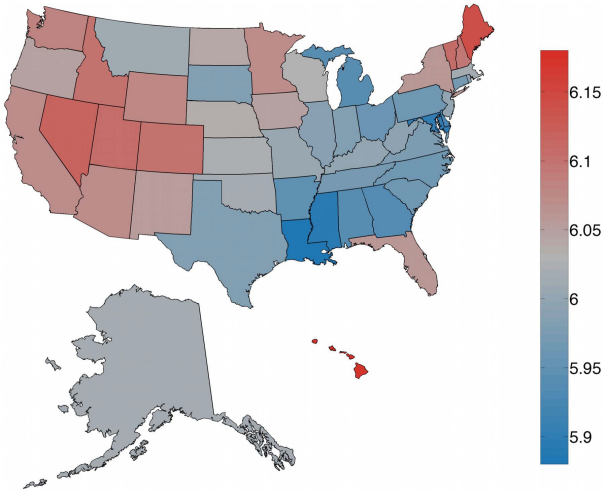

Figura 6 – Nível de felicidade nos Estados Unidos baseados nos *tweets*.(MITCHELL *et al.*, 2013)

Ao utilizar de um SIG, que refere-se a localização geográfica, nos casos citados, também trata de mobilidade do ser humano. Os trabalhos relacionam dados do Twitter, analisam da perspectiva de análise de sentimento ou então SIG, e no caso de (MITCHELL *et al.*, 2013) relacionam com dados de índices de obesidade por exemplo. O trabalho em questão se diferenciaria no idioma dos dados, no tema e cidade de estudo, como também nos dados de comparação com Kozievitch *et al.* (2015). Entretanto, as técnicas e métodos podem ser semelhantes e adaptados para o trabalho.

Santala *et al.* (2015) relatam outro caso de uso de dados provenientes de mídia social, nesse caso voltado para *Place Branding*, que é a criação de uma "marca" para um lugar, especiamente para a cidade de Curitiba. Com o auxílio de SIG obtiveram um mapa que demonstrava

regiões nas quais muitos usuários realizaram “check-ins”, porém eram regiões ainda sem nenhuma marcação específica, sendo assim, novos lugares em potencial. O trabalho em questão demonstra a utilidade dos dados provenientes de mídia social, como também no aproveitamento da abordagem SIG para análise.

Na questão de alternativas de sistemas SIG, encontra-se o trabalho de Bornmann e Leydesdorff (2011), que analisam a distribuição geográfica das maiores citações de artigos científicos. Consideram o total de artigos e então criam um valor de referência a partir daqueles mais citados em cada cidade ou região. A visualização dos resultados é feita através do *Google Maps* (GoogleMaps, 2017) conforme na Figura 7, que representa a contribuição científica de artigos relacionado a área de química em 2008.

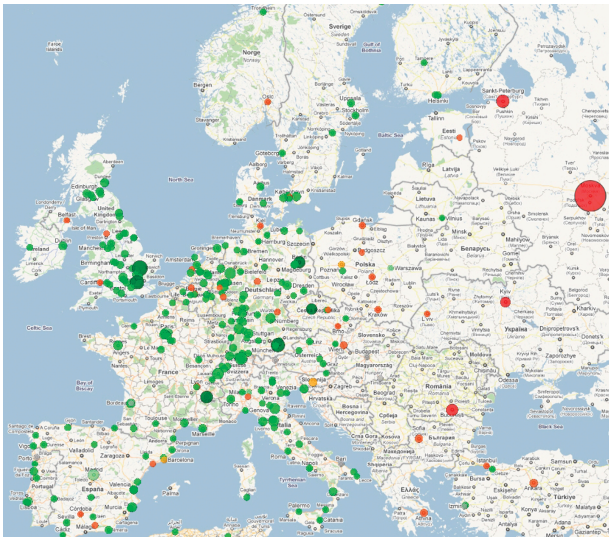

Figura 7 – Artigos de química mais citados nas cidades da Europa em 2008 (BORNMANN; LEYDESDORFF, 2011).

Na questão da ferramenta SIG utilizada, uma possibilidade do *Google Maps* é a utilização da interface via *web* como também do *Google Earth* (GoogleEarth, 2017). A versão *web* é mais básica e não requer qualquer instalação de outros programas, mas é necessário ter uma conta, que pode ser criada gratuitamente. Já o *Google Earth* possui mais recursos, como por exemplo a visualização 3-D de mapas, mas requer a instalação do programa. Os recursos disponibilizados são semelhantes á outros SIGs(CLIFFORD; MACFADYEN; MACFARLANE, 2013).

Referente a análise de sentimento, não há um banco de dados na língua portuguesa que contemple a maioria das palavras e expressões para realizar uma análise de sentimento precisa. Seria possível a criação de um novo dicionário, através da união de bases de dados parciais, no próprio idioma português, também com bases de dados em outras línguas, utilizando a tradução das palavras, e por fim, a aplicação de métodos para unificar os dicionários (NASCIMENTO, 2014). Pode ser feita também uma replicação da base de dados dos *tweets* com a tradução para o inglês.

3.3 MOBILIDADE

A mobilidade pode ser considerada tanto para o movimento de pessoas, objetos e informações pelo mundo quanto para os processos diários como a locomoção de casa para o serviço e momentos de lazer (ZAO; HAO; ZHANG", 2015). Para os fins deste trabalho será considerada apenas a mobilidade humana, cujos dados são obtidos através da publicação dos *Tweets* por dispositivos móveis, e a localização identificada no *Tweet* no campo "*coordinates*".

Hawelka *et al.* (2013) estudaram a mobilidade a nível de migração e turismo, utilizando os dados do *Twitter* e verificando o país de origem de cada usuário pela análise de onde provinha a maioria dos *Tweets*. Cada vez que esse usuário enviasse um *Tweet* de outro país era considerado na estatística. Validaram os resultados com as estatísticas globais de turismo e comprovaram que a utilização do *Twitter* para identificação de padrões globais de mobilidade é viável. Já Luo *et al.* (2016) utilizaram os dados do *Twitter* para verificar características da mobilidade humana na cidade de Chicago. Uma parte a ser ressaltada nesse trabalho foi a filtragem dos *Tweets*. Não foi utilizada toda a base capturada. Foram desconsideradas, por exemplo, pessoas que não tinham pelo menos 1 *Tweet* por semana (LUO *et al.*, 2016). Frank *et al.* (2015) realizaram um trabalho diferenciado ao analisar a mobilidade na escala de tempo de no máximo uma semana. Com os dados dos *Tweets* e diferenciação de algumas palavras chaves e padrões de linguagem (isto é, Formal ou Informal), foi possível verificar quando as pessoas mandavam um *Tweet* de casa ou do trabalho.

O sistema de transporte público é um serviço essencial que tem um grande impacto na mobilidade das cidades. Fazer uma análise entre os dados, oficiais e reais, das rotas, destinos, origens, tempo de rota e de espera, traz informações relevantes sobre a mobilidade no transporte público. A utilização de um mapa com as informações coletadas pode demonstrar graficamente quais as regiões com os maiores problemas (ZENG *et al.*, 2014).

Quando trata-se de mobilidade pode-se levar em consideração o aspecto de acessibilidade. Ferrari *et al.* (2013) mediram a acessibilidade no transporte público de Londres através de dados de mobilidade. Os dados foram coletados do Planejador de Rota, que é uma ferramenta que traça rotas e permite a seleção de rotas que contenham acessibilidade a cadeira de rodas, e também do Cartão Oyster(TfL, 2017) de Origem e Destino, que para os ônibus têm apenas a informação do início da jornada, enquanto para os outros tipos de transporte (p. ex. metrôs) tem a informação do início e fim. Um dos resultados obtidos foi que para pessoas com restrição de rota para acessibilidade com cadeira de rodas, o tempo de jornada é significativamente maior que para rotas sem essa restrição.

Cheng e Chen (2015) relatam que a acessibilidade pode ser medida através da distância ou o tempo que uma pessoa leva para sair de um lugar para outro. Um estudo de caso da cidade de Taipei e Kaohsiung, utilizando métodos conceituais, analisou a percepção da acessibilidade, mobilidade e conectividade do sistema de transporte público. Alguns problemas são citados,

como “condições de tráfego estressantes para os pedestres andando para estação de trânsito”, “as linhas de ônibus têm muitas paradas”, “o ônibus não está no horário” e outros relatos referentes ao custo e dificuldade de entender os sinais e instruções para fazer uma ligação também aparecem. A busca da opinião dos usuários é uma das maneiras de se identificar um problema e propor uma solução que afetaria diretamente esse usuário (CHENG; CHEN, 2015).

O estudo do aspecto de mobilidade leva também ao conceito de Cidades Inteligentes. Com base no trabalho de Weiss, Bernardes e Consoni (2015), que traz algumas das principais definições, Cidades Inteligentes são aquelas que utilizam da técnologia de informação e comunicação para administração e planejamento da cidade. O planejamento, a utilização dos dados provindos de diversas fontes, a fim de propor inovações com o intuito de melhorar a qualidade de vida das pessoas, são parte desse conceito. Com isso, o sistema de transporte público tem um papel fundamental. Kozievitch *et al.* (2015) em um estudo de caso de Curitiba foram analisadas na cidade as rotas de ônibus e, com isso, partes em que essas rotas não alcançam, rotas redundantes e também excesso de tráfego. Tais informações, relacionadas com dados de percepções dos usuários do transporte público, podem trazer propostas melhor direcionadas à solução dos problemas encontrados.

4 DESENVOLVIMENTO

Nesse capítulo será descrito o desenvolvimento do trabalho: Tecnologias e Implementação, Pesquisa com Usuário, Seleção dos Dados, Classificação dos *Tweets* e Representação Gráfica.

4.1 TECNOLOGIAS E IMPLEMENTAÇÃO

Nessa seção são descritas as tecnologias, softwares e ferramentas utilizados para realizar a coleta, análise da estrutura, realização dos filtros, análise de sentimento e a representação gráfica dos dados obtidos. A Figura 8 representa a ordem em que cada assunto será abordado.

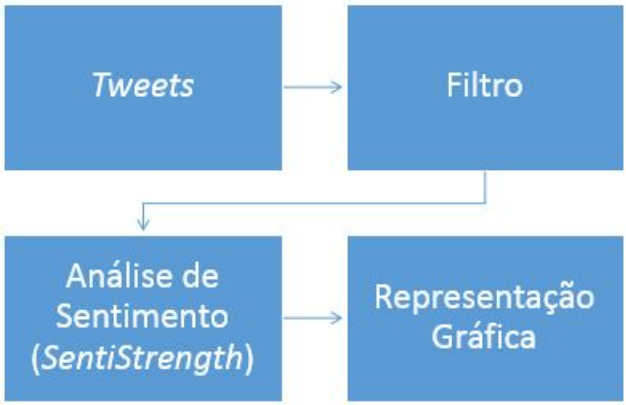

Figura 8 – Tecnologias e Implementação.

4.1.1 *Tweet*

Os dados utilizados para análise foram coletados pela UFMG, através de uma API (Interface de Programação de Aplicação) disponibilizada pelo *Twitter*. A utilização do *Streaming* para coleta pode ser feita para o *stream* público, que busca dados públicos pelo *Twitter* de vários usuários, ou então por *site*, onde é selecionado apenas um usuário para ser feita a coleta dos dados, o que está relacionado mais a empresas, por exemplo. É importante ressaltar que, para que a coleta dos dados seja feita com sucesso, é necessária uma conexão constante (TwitterAPI, 2016).

Os dados coletados estão no formato Notação de Objetos JavaScript (JSON), que é uma formatação leve de troca de dados. JSON é em formato texto e completamente independente de linguagem, pois utiliza técnicas conhecidas de várias linguagens de programação. Estas propriedades fazem com que JSON seja um formato bom de troca de dados. O JSON possui duas

estruturas: Objeto, composto por pares nome / valor, e *array* (isto é, Vetor/Sequência). A Figura 9 representa um objeto JSON. Nota-se que o objeto começa com a abertura da chave "{", o nome e o valor são separados pelo símbolo de dois pontos ":", e cada par é separado entre si por vírgulas ",", até o fim do objeto com o fechamento da chave "}"(JSON, 2016).

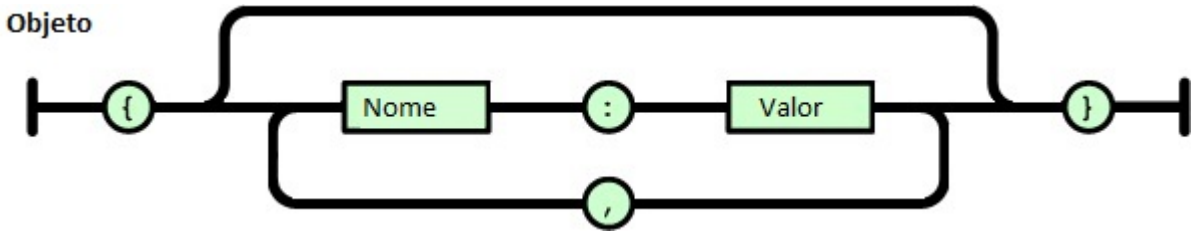

Figura 9 – Objeto JSON (JSON, 2016).

A Figura 10 representa um *array* JSON. O *array* é composto por um único tipo de valor, porém pode conter vários valores, esse valor é, por exemplo, uma instância de um objeto, onde cada instância é separada pela vírgula.

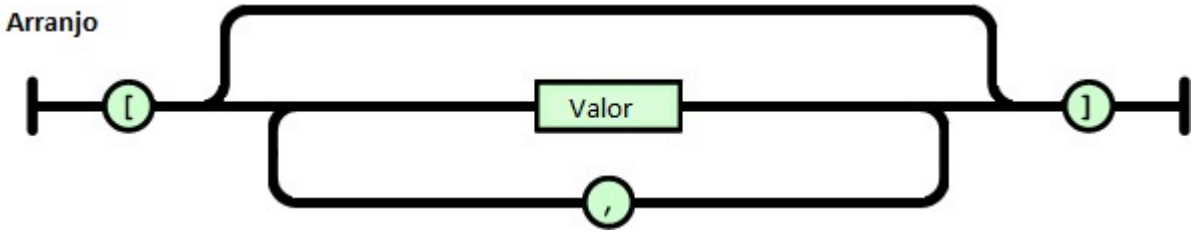

Figura 10 – Array JSON (JSON, 2016).

O texto abaixo apresenta um *Tweet* (Objeto) extraído da base de dados da UFMG (EUBra-BIGSEA2, 2017) no projeto BIGSEA (EUBra-BIGSEA1, 2016).

```
1  { "_id" : 715690118816866304,
2  " contributors " : null ,
3  " control " : { " coletas " : [ { "id" : 349 } ] },
4  " coordinates " : null ,
5  " created_at " : { "\$date" : 1459468795000 },
6  " entities " : { "user_mentions" : [], "symbols" : [], "hashtags" : [], " urls " : [] },
7  " favorite_count " : 0,
8  " favorited " : false ,
9  " filter_level " : "low",
10 "geo" : null ,
11 "id" : 715690118816866304,
12 " id_str " : "715690118816866304",
13 "in_reply_to_screen_name" : null ,
14 " in_reply_to_status_id " : null ,
15 " in_reply_to_status_id_str " : null ,
16 " in_reply_to_user_id " : null ,
17 " in_reply_to_user_id_str " : null ,
18 " is_quote_status " : false ,
19 "lang" : "pt ",
20 "place" : { "country_code" : "BR",
21                         " url " : " https :// api . Twitter .com/1.1/geo/id/6d5542f8d837770d.json",
22                         "country" : " Brasil ",
23                         "place_type" : " city ",
```

```
24                          "bounding_box" : { "type" : "Polygon",
25                                                  "coordinates" : [ [ [ −49.391643, −25.644752 ], [ −49.391643,
                                                        −25.345747 ], [ −49.185278, −25.345747 ], [ −49.185278,
                                                        −25.644752 ] ] ] },
26                          "full_name" : "Curitiba , Brasil ",
27                          " attributes " : {},
28                          "id" : "6d5542f8d837770d",
29                          "name" : "Curitiba " },
30  "retweet_count" : 0,
31  "retweeted" : false ,
32  "source" : "<a href=\" http :// Twitter .com/download/iphone\" rel=\"nofollow\"> Twitter for iPhone</a>",
33  "text" : "Tem gente vindo aq em casa agr compra ingresso kkkkkk merere",
34  "timestamp_ms" : "1459468795862",
35  "truncated" : false ,
36  "user" : { " follow_request_sent " : null ,
37                          "profile_use_background_image" : true ,
38                          "id" : 2361361760, " verified " : false ,
39                          " profile_image_url_https " : "https :// pbs.twimg.com/profile_images/713851221388427265/u−
                                dwD63B_normal.jpg",
40                          " profile_sidebar_fill_color " : "DDEEF6",
41                          " is_translator " : false ,
42                          "geo_enabled" : true ,
43                          " profile_text_color " : "333333",
44                          " followers_count " : 2987,
45                          " protected " : false ,
46                          " location " : "Uberaba, Curitiba ",
47                          " default_profile_image " : false ,
48                          " id_str " : "2361361760",
49                          " utc_offset " : −10800,
50                          " statuses_count " : 74862,
51                          " description " : "Vida eh brisa passageira .. Isa Snap Stteeefs @Coritiba ",
52                          " friends_count " : 684,
53                          " profile_link_color " : "0084B4",
54                          " profile_image_url " : " http :// pbs.twimg.com/profile_images/713851221388427265/u−dwD63B_normal.jpg
                                ",
55                          " notifications " : null ,
56                          " profile_background_image_url_https " : " https :// pbs.twimg.com/profile_background_images
                                /445737650876461056/ylaCa9M_.jpeg",
57                          " profile_background_color " : "C0DEED",
58                          " profile_banner_url " : " https :// pbs.twimg.com/ profile_banners /2361361760/1459217263",
59                          " profile_background_image_url " : " http :// pbs.twimg.com/profile_background_images/445737650876461056/
                                ylaCa9M_.jpeg",
60                          "screen_name" : " stteefs ",
61                          "lang" : "pt ",
62                          " profile_background_tile " : true ,
63                          " favourites_count " : 19891,
64                          "name" : " Stef SML ",
65                          " url " : " http :// ch00sebehappy.tumblr.com",
66                          " created_at " : "Tue Feb 25 16:47:32 +0000 2014",
67                          " contributors_enabled " : false ,
68                          "time_zone" : " Brasilia ",
69                          " profile_sidebar_border_color " : "FFFFFF",
70                          " default_profile " : false ,
71                          " following " : null ,
72                          " listed_count " : 3 } }
```


A partir deste exemplo de objeto JSON é possível verificar (Linha 1) que há um "_id" que é único para cada *Tweet*. O campo "created_at" (Linha 5) indica a data que foi postado. Há dois grandes objetos dentro do objeto *Tweet*: "place" (Linha 20) e "user" (Linha 36). O objeto

"place" (Linha 20) tem as informações relacionadas à geolocalização do *Tweet*, que indica além da cidade (Curitiba), as coordenadas. Após o fim do objeto "place" há o campo "text" (Linha 33) que é o *Tweet* em si. É neste texto que também poderão ser utilizados os *hashtags*, um recurso de palavra chave para facilitar a busca e gerenciamento dos *Tweets* (Hashtag, 2016). Já o objeto "user" (Linha 36) está relacionado ao perfil do usuário que postou o *Tweet*. O usuário também tem um "id" (Linha 38) único, o número de amigos em "friends_count" (Linha 52), o nome escolhido pelo usuário em "name" (Linha 64) e outras informações relacionadas às opções de interface do usuário.

4.1.2 Filtro

Inicialmente, foi desenvolvido o filtro a partir de uma implementação em JAVA. O método principal tem a função de, linha a linha, selecionar os *tweets* que contiverem no seu texto as palavras chaves relacionadas ao transporte público, retirar os metadados que não foram utilizados e então colocar todos os dados filtrados em um arquivo de texto em formato JSON. O processo de obter as palavras-chave para o filtro está descrito na Seção 4.2.

Em uma segunda etapa, foi feita uma visualização preliminar dos dados no formato JSON utilizando o software glogg (Glogg, 2016), conforme exemplo na Figura 11. Com essa visualização foi identificado onde estavam localizados os campos de interesse e também foi possível a verificação, pós aplicação do código, dos resultados obtidos. A partir da descrição do objeto *tweet* no formato JSON disponibilizado pela API do *Twitter* (JSONTwitter, 2016) foram selecionados os seguintes campos de interesse:

- *coordinates*: Coordenadas de onde foi publicado o *tweet*.
- *created_at*: Quando o *tweet* foi publicado.
- *entities*: É um objeto do qual foi selecionado o objeto *hashtag*.
- *text*: É o texto do *tweet*.

O método principal está representado conforme o código do Apêndice B, que tem a função de filtrar os *tweets* pelas palavras-chave e também por textos de *"checkins"*.

O objeto criado, *Tweetobject*, é o objeto principal, representado no Apêndice G, é composto, além do *text* pelos objetos: *coordinates* que está representado no Apêndice C, *created_at* no Apêndice D, *entities* no Apêndice E e *hashtag*, que está contido no *entities*, representado no código do Apêndice F.

Com isso foi possível desconsirar grande parte dos campos que compõem um *tweet*, conforme identificidado na documentação da API(JSONTwitter, 2016), já que não seriam utili-

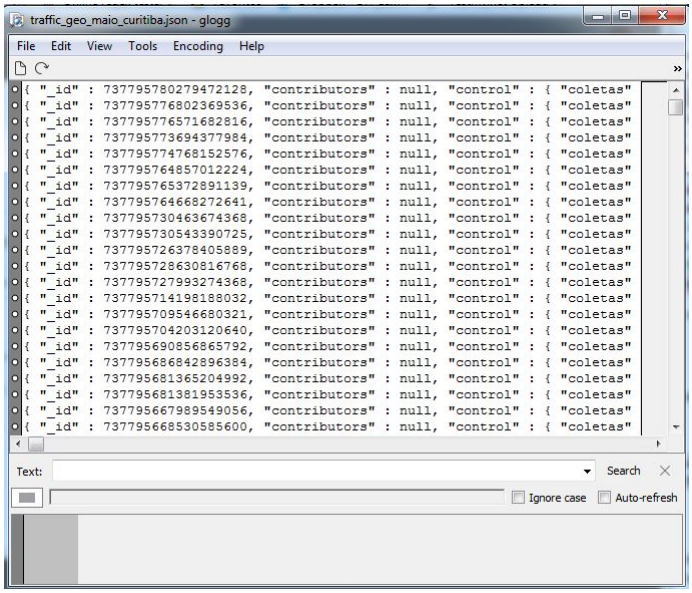

Figura 11 – Glogg.

zados para as análises. Como resultado, os arquivos filtrados são menores e compostos apenas por dados relevantes para o trabalho.

4.1.3 *SentiStrength*

O *SentiStrength* é uma ferramenta de análise de sentimento de maneira automizada que utiliza a abordagem do Classificador de Bayes. Possui uma API em JAVA que foi obtida através de uma licença acadêmica (SentiStrength, 2016).

A API do *SentiStrength* utiliza uma base de dados com arquivos de texto, em inglês, entre elas: (1) *BoosterWordList*, (2) *EmoticonLookupTable*, (3) *EmotionLookupTable*, (4) *EnglishWordList*, (5) *NegatingWordList*, (6) *QuestionWords* e (7) *SlangLookupTable*. Em (1) estão algumas palavras que dão maior intensidade no cálculo do sentimento, podendo modificar o resultado da análise de uma sentença de maneira positiva ou negativa. Em (2) estão alguns *emoticons* que representam valores de -1, 0 e +1 de acordo com a sua representitivade, por exemplo um *emoticon* que representasse o sentimento feliz ":)" seria +1 enquanto um que representasse um sentimento triste ":(" seria -1. Em (3) estão classificadas algumas palavras que por si mesmas já tem um sentimento, por exemplo a palavra *cry* tem um valor negativo de -4 enquanto a palavra *love* tem um valor positivo de +3. Algumas palavras não tem valor assimilado, que são as das listas em (4), (5), (6) e (7) onde (5) pode mudar completamente o sentimento de uma sentença e (6) pode tirar qualquer sentimento relacionado a ela.

Além disso, é possível realizar alguns tipos de classificações: (a) explicativa, (b) por palavras-chave, (c) positiva e negativa separadamente, (d) trinária, (e) binária e (f) em escala. Na classificação (a) é mostrado detalhadamente como foi feito o cálculo da sentença em análise

até chegar ao resultado final da análise de sentimento. Já quando é feito (b), são previamente informadas algumas palavras-chave para análise em questão, desconsiderando as restantes. Em (c) é fornecido um resultado positivo e também negativo, variando de 1 a 5 e de -5 a -1 respectivamente. Derivando do resultado (c) pode obter-se o resultado (d), (e) e (f) onde é feito a soma do resultado positivo com o negativo. A diferença é que (d) possui a classificação "neutra" enquanto (e) é apenas positivo (1) ou negativo (-1). Em (f) é considerado o resultado total da soma, porém com os limites negativo(-4) e positivo(+4).

Uma aplicação em JAVA foi desenvolvida conforme o código do Apêndice H. Para utilização do *SentiStrength* foi selecionado o método de classificação trinário, pois leva em consideração a soma dos resultados individuais positivos e negativos. A aplicação tem como entrada o arquivo com apenas os textos dos *tweets* filtrados, já para a saída tem o resultado da aplicação do método trinário, onde cada conjunto de sentenças (texto do *tweet*) foram salvas em arquivos separados de acordo com o resultado da classificação: bom, ruim ou neutro.

4.1.4 Representação Gráfica

Para representação gráfica dos *Tweets* classificados foi utilizado o *Google Maps* (GoogleMaps, 2017) que permite criar arquivos KML (*Keyhole Markup Language*) e camadas do mapa. O uso do recurso foi para referenciar no mapa os *tweets* que possuem coordenadas em conjunto com as informações de texto e classificação após a análise de sentimento.

4.2 PESQUISA COM USUÁRIO

Essa seção descreve os passos necessários para a obtenção de palavras-chave sobre o transporte público de Curitiba, preferencialmente utilizadas por usuários do *Twitter*. A Figura 12 ilustra, em resumo, os procedimentos realizados e a ordem em que serão abordados.

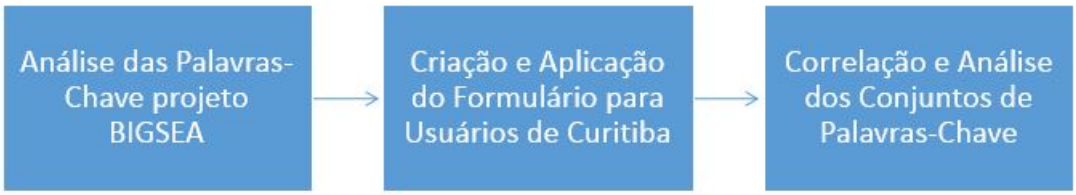

Figura 12 – Procedimentos Palavras-Chave.

O *Twitter* é uma rede social, o corpo do texto de um *tweet* abrange vários temas, então, para a seleção dos textos necessários para a análise sobre mobilidade e transporte público, há a necessidade de aplicar um filtro (GAL-TZUR *et al.*, 2015) que, separe os *tweets* relevantes dos que não convém para o trabalho em questão.

Os filtros utilizados no projeto EUBRA-BigSea da UFMG são divididos em duas partes, a do conteúdo do texto do *Tweet* e a das *hashtags*. O primeiro filtro (texto) é composto pelas seguintes palavras:

> "bhtrans, transito, transito, congestionamento, belo horizonte, bhz, bhte, belzonte, bh, engarrafamento, engarrafado, engarrafada, lentidao, blitzbh, transitobh, pbhonline, nossabh, nordeste_bh, fmcbh, PatrulhaBH, pbh, prefeitura, beaga, beaga, @Transito98FM, @TransitoTrafego, @LeiSeca, @MGLeiSeca, @bhtranstorno, @bhtrans, @BHTrolls, @apoiond, OficialBHTRANS, BRT, VLT, aeromovel, teleférico, metrô, estacao vilarinho, estação vilarinho, terminal itaquera, porto maravilha."

No primeiro filtro note que algumas palavras (p. ex. nordeste_bh, @MGLeiSeca) fazem parte de um contexto da região de Belo Horizonte, cidade que coordena o projeto BIGSEA no Brasil(EUBra-BIGSEA1, 2016). O segundo filtro também contém palavras que só se aplicam a região de Belo Horizonte(p. ex. MarginaldoPinheiros, MigiBertioga, FernaoDias). Enquanto o primeiro filtro utiliza do texto do *tweet*, o segundo é realizado com uma análise das *hashtags* utilizadas, pelos usuários, quando o texto tem algo relacionado com o tráfego composto pelas seguintes palavras:

> "L15, MarginaldoPinheiros, MogiBertioga, PPV, sptrans, CompartilheBonsExemplos, FernaoDias, RaposoTavares, ResumoBOLS, velocidade, Blitz, ciclovia, L11Coral, Laranjeiras, MarginaldoTiete, L4, leisecarj, SeEuVejoDenuncio, Ubatuba, acidente, RECREIO, LEBLON, RadialLeste, LinhaAmarela, RioOlympics2016, ALERTA, L6, RadialOeste, transito, Acidente, L9, litoral, Bertioga, L5, L8, AvenidaPaulista, Rodoanel, Ecopistas, L2, Periscope, RJTV, TAQUARA, ZonaSul, L12, L10, L3, ViaQuatro, Bols, RoubadoRJ, GUARATIBA, Caju, LAGOA, Ramos, AyrtonSenna, L1, SPTrans, LVermelha, VcViu, L7, Ciclovia, ROUBADORJ, Transito, 1info, AvBrasil, DicaManaustrans, ZonaOeste, BOLS, alert, RESUMOBOLS, bols, cptm, metrosp, L11, SP, Estradas, MetroSP, TransitoRJ, ViasExpressas, bicalho, inicio_ponte, fim_bicalho, gaviao_peixoto, tunel, roberto_silveira_tunel, subida, CPTM, sentido_niteroi, sentido_rio, TransitoSP, manilha, subida_ponte, saida_ponte."

Algumas expressões utilizadas, como por exemplo, bhz, @bhtrans, são específicas para a região de Belo Horizonte. Já no caso da cidade de Curitiba o objetivo era realizar a substituição para por exemplo cwb, @urbs, que são os equivalentes para a cidade de estudo. No *hashtag* um exemplo de equivalência para LinhaAmarela, que está relacionado à avenida do Rio de Janeiro é a LinhaVerde, que é um trecho da BR 116 em Curitiba.

Um formulário foi criado no *Google Forms* (GoogleForms, 2017), conforme o apêndice A, para obter informações dos usuários do *Twitter* sobre as palavras chaves para falar de algum assunto relacionado com transporte público da cidade de Curitiba. Esse questionário foi

respondido pelos estudantes da turma CSB30 de banco de dados da UTFPR campus Curitiba no dia 21/09/2017.

A primeira pergunta do formulário é se a pessoa utiliza o *Twitter*. Caso positivo é perguntado se a pessoa mora na cidade de Curitiba. A partir dessas perguntas também é possível obter uma relação, entre as pessoas que preencheram o formulário, quantos são usuários do *Twitter*, e mais especificamente moram em Curitiba. Após a resposta afirmativa a essas duas perguntas é solicitado que seja feita uma escolha de palavras-chave que poderiam ser utilizadas para tratar do assunto transporte público. Além de algumas opções pré-selecionadas com base no projeto da UFMG era permitido que o usuário adicionasse novas respostas.

O resultado do formulário contou com um total de 54 respostas. Dessas 37 são de usuários do *Twitter* e destes 35 moram em Curitiba, então, por fim 35 respostas selecionaram e/ou indicaram novas palavras-chave em relação ao transporte público conforme a Figura 13.

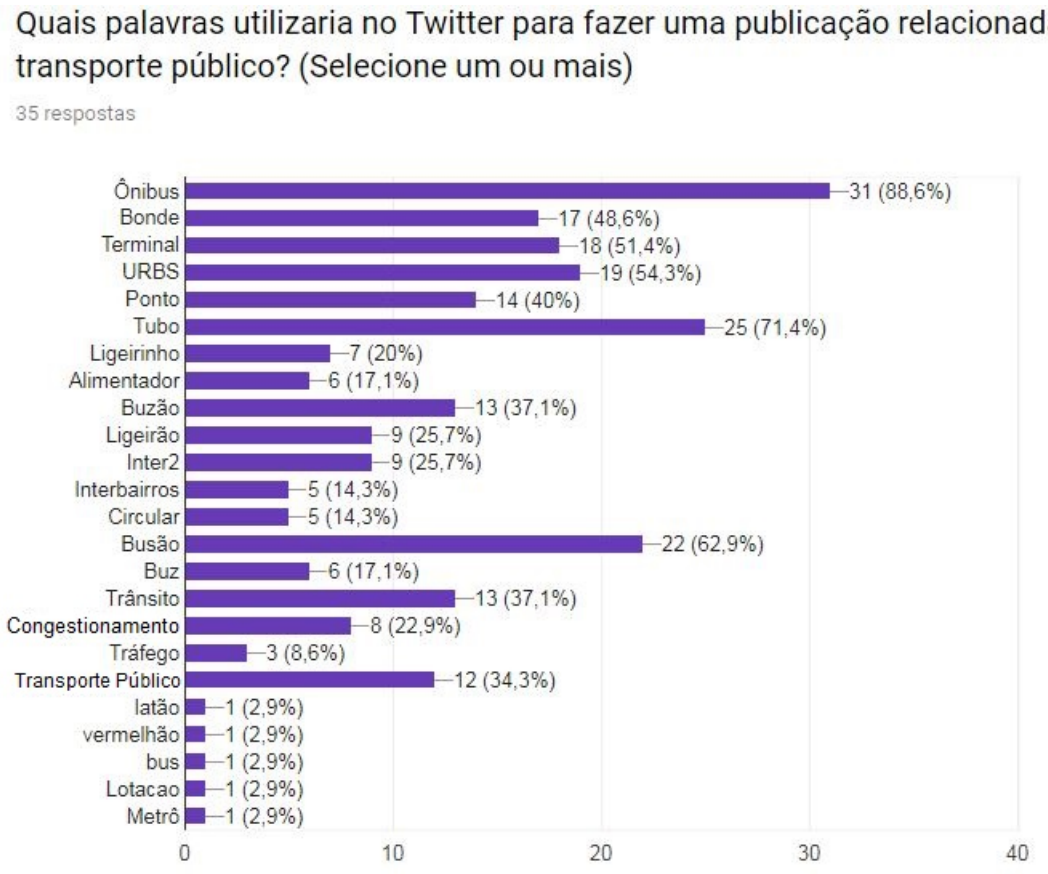

Figura 13 – Resultado formulário aplicado aos usuários.

Assim como no projeto BIGSEA (EUBra-BIGSEA1, 2016), que continha palavras-chave relacionadas ao contexto da cidade de Belo Horizonte, o resultado da pesquisa evidenciou palavras regionais da cidade de Curitiba, como “Tubo”, “Ligeirão” e “Vermelhão”. A partir das palavras obtidas no resultado do formulário, foram geradas novas palavras com variações na forma escrita, com acentos, sem acentos e com erros de português.

Com base no resultado obtido e as variações das palavras-chave, foi realizado o filtro

tanto para o conteúdo de texto de um *tweet*, quanto para as *hashtags* utilizando as seguintes palavas: “bonde, bus, busao, busão, buz, buzao, buzão, circular, inter 2, interbairros, latao, latão, ligeirao, ligeirão, ligeirinho, ligerao, ligerão, lotacao, lotaçao, lotação, metrô, onibos, ônibos, onibus, ônibus, terminal, transito, transporte público, tubo, urbs, vermelhao, vermelhão”.

Algumas das palavras obtidas na pesquisa não foram utilizadas por poder ter contextos diferentes, como por exemplo a palavra “Ponto” que, segundo o dicionário Michaelis (Michaelis, 2017), no contexto pretendido tem um significado de “Local onde os veículos coletivos param para o embarque e desembarque de passageiros”, já em contextos diferentes poderia ter, por exemplo, o significado de “Sinal ou marca, arredondado e pequeno”, ou então, “União cirúrgica de duas superfícies com agulha e fio”, que trariam *tweets* fora do contexto pretentido.

4.3 SELEÇÃO DOS DADOS

Nessa seção está descrita a origem dos *tweets* utilizados, a seleção dos dados necessários e a aplicação do filtro para seleção dos textos de *tweets* relacionados ao tema mobilidade e transporte público. A Figura 14 representa a ordem de ação e demonstração dos procedimentos realizados.

Figura 14 – Seleção dos Dados.

Os dados utilizados para análise foram coletados pela UFMG, através de uma API (Interface de Programação de Aplicação) disponibilizada pelo *Twitter*. A coleta foi durante o período de março a junho de 2016 e foram coletados *Tweets* geolocalizados na cidade de Curitiba.

Dos metadados do *tweet* foram selecionadas as coordenadas, a data de criação, as *hashtags* e o texto. A partir deste texto, um filtro foi desenvolvido e aplicado, foram selecionados textos, ou *hashtags*, que continham alguma das seguintes palavras: “bonde, bus, busao, busão, buz, buzao, buzão, circular, inter 2, interbairros, latao, latão, ligeirao, ligeirão, ligeirinho, ligerao, ligerão, lotacao, lotaçao, lotação, metrô, onibos, ônibos, onibus, ônibus, terminal, transito, transporte público, tubo, urbs, vermelhao, vermelhão”.

Foi percebido que, mesmo com os filtros das palavras-chave, há muitos *Tweets* que não estão relacionados ao tema mobilidade ou transporte público. De um total de 2.402.081 *tweets*, obteve-se 6.303 cujo o texto ou *hashtag* continha alguma das palavras-chave utilizadas. A Figura 15 representa as palavras que obtiveram um íncide de incidência maior ou igual a 1% dentre as utilizadas.

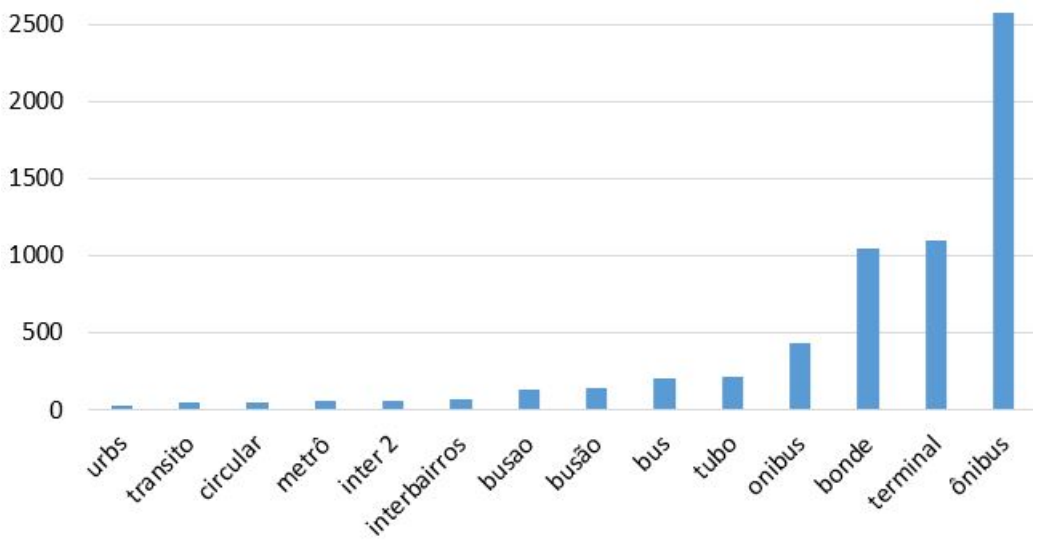

Figura 15 – Filtro por Palavras-Chave.

Uma amostra do arquivo original pode ser identificada no Anexo A enquanto uma amostra do resultado obtido após aplicação do filtro encontra-se no Anexo B.

4.4 CLASSIFICAÇÃO DOS *TWEETS*

Essa seção descreve a aplicação do SentiStrength para a tradução dos *tweets* e para a tradução do dicionário da ferramenta. Demonstra também o resultado dessa aplicação individualmente para as palavras-chave com as três maiores incidências dentre as utilizadas na realização do filtro.

A Figura 16 ilustra um gráfico percentual do total de *tweets* analisados em bons, 24%, ruins, 25%, e neutros, 51%. Nesta etapa foi feita a tradução do corpo de texto do *tweet* em si permanecendo o dicionário do *SentiStrength* original. Já a Figura 17 ilustra um gráfico percentual do total de *tweets* analisados em bons, ruins, ambos aproximadamente 19% e neutros, 62%. Nesta etapa foi feita a tradução do dicionário do *SentiStrength* permanecendo o corpo do texto do *tweet* original.

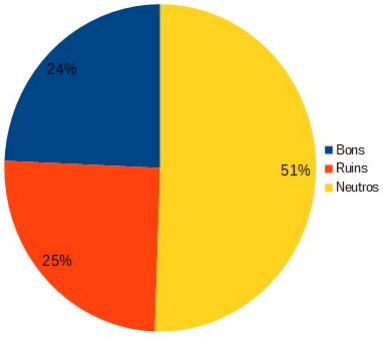

Figura 16 – Percentual Classificativo - Tradução do *Tweet*.

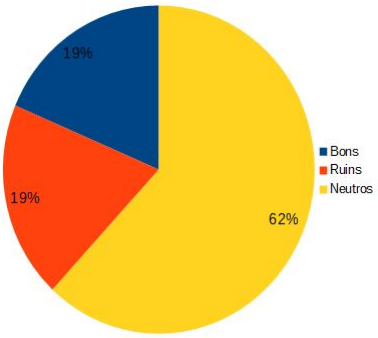

Figura 17 – Percentual Classificativo - Tradução do Dicionário *SentiStrength*.

Na verificação manual do resultado é perceptível que, além das frases sem relação ao transporte público, a análise de sentimento pode inverter os sentidos, ou seja, uma frase positiva pode ser classificada como negativa. Além disso, as traduções automáticas são normalmente literais e também não levam em consideração gírias ou erros de ortografia, o que pode ser uma das causas de que, ao fazer a verificação manual, algumas frases são classificadas de maneira incorreta.

Uma modificação no código fonte do apêndice H foi feita para obter o resultado da análise de sentimento com referência a palavras-chave. Essa abordagem foi realizada para os três maiores índices de incidência das palavras-chave utilizadas no filtro dos *tweets*: bonde, ônibus e terminal.

Para a palavra-chave "bonde" foram gerados dois gráficos para os resultados da análise de sentimento. O primeiro gráfico foi gerado baseado na tradução da sentença. Para isso, o termo "*tram*" foi o filtro utilizado, visto que seria a tradução automática de "bonde". A Figura 18 mostra o resultado do *SentiStrength* como bons, 18%, ruins, 12%, e neutros, 69%. O segundo gráfico, Figura 19, foi gerado com a tradução do dicionário, mantendo a frase original, o termo de filtro foi exatamente a palavra-chave "bonde" e o resultado foi de 10% bons, 13%ruins e 78% neutros.

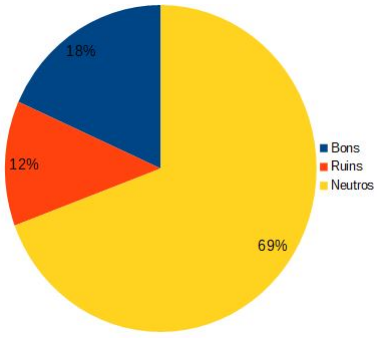

Figura 18 – Percentual Classificativo - "tram".

A palavra-chave "ônibus" foi a que teve maior índice de incidência dentre as palavras-chave utilizadas. Da mesma forma que apresentado para palavra-chave "bonde", há dois gráficos, gerados após a aplicação do *SentiStrength*. O primeiro gráfico de resultado, Figura 20, com a tradução da frase, onde então do termo de filtro foi "*bus*", teve o índice de 17% bons, 11% ruins

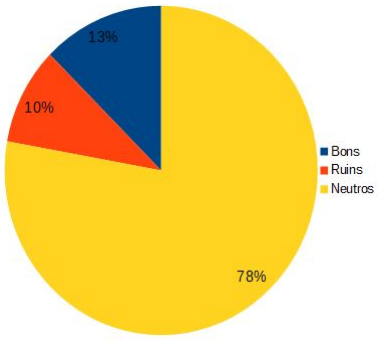

Figura 19 – Percentual Classificativo - “bonde”.

e 73% neutros. O segundo gráfico de resultado, tradução do dicionário. representado na Figura 21, com o termo de filtro “ônibus” teve como índice 15% bons, 7% ruins e 78% neutros.

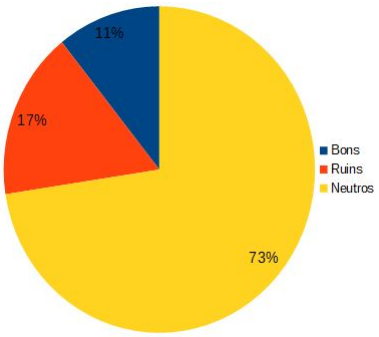

Figura 20 – Percentual Classificativo - “bus”.

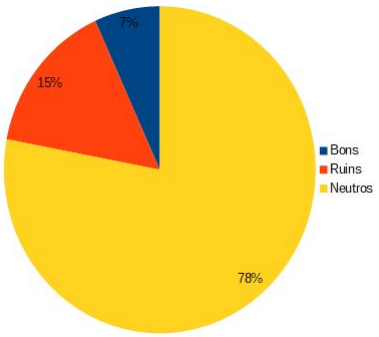

Figura 21 – Percentual Classificativo - “ônibus”.

A palavra-chave “terminal” tem uma particularidade: é um dos casos em que o contexto deve ser levado em consideração para ser obtida uma tradução mais adequada. Como o contexto não é levado em consideração, a palavra-chave “terminal” foi utilizada como filtro para ambos os gráficos gerados. A Figura 22 representa o resultado da aplicação do *SentiStrength* ao realizar a tradução das frases onde 13% são bons, 10% ruins e 77% neutros. A Figura 23 representa os resultados obtidos para aplicação do *SentiStrength* com a tradução do dicionário, onde 12% são bons, 10% ruins e 78% neutros.

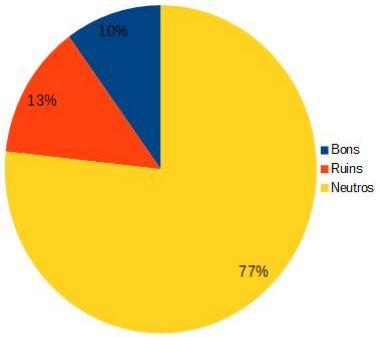

Figura 22 – Percentual Classificativo - Tradução *Tweet* “terminal”.

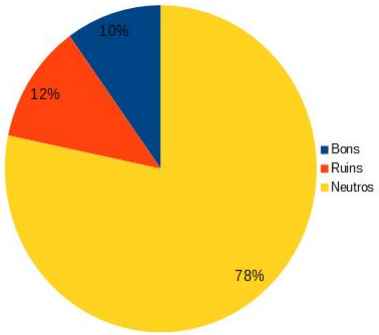

Figura 23 – Percentual Classificativo - Tradução Dicionário “terminal”.

Note que há uma diferença entre os resultados obtidos a partir da tradução das frases e da tradução dos dicionários. A tradução dos dicionários trouxe em geral um aumento no índice neutro.

Para calcular a precisão da aplicação da ferramenta para a tradução do dicionário e do texto, foram selecionados, para as respectivas abordagens, os *tweets* com o termo “ônibus”, e consequentemente o termo “bus”. Também levando em consideração a utilização das coordenadas para visualização em um mapa, foram selecionados apenas *tweets* com coordenadas disponíveis.

Um total de 141 *tweets* foram selecionados para a palavra-chave “ônibus”. Em seguida, esses *tweets* foram analisados manualmente em relação ao sentimento, para então realizar uma comparação com a abordagem de aplicação do *SentiStrength* para tradução do dicionário. A Figura 24 representa a comparação entre os gráficos percentuais do resultado da ferramenta e da análise manual. Para a palavra-chave representante de “ônibus”, na abordagem de tradução do texto, “bus”, foram selecionados 107 *tweets* com coordenadas para comparação, conforme na Figura 25.

Note-se que devido à tradução, os *tweets* selecionados a partir das palavras-chave “ônibus” e “bus”, são em números diferentes, o que indica que a tradução automática pode ter resultados diferentes do correto. Além disso, a quantidade de análises classificadas pelo *SentiStrength* como neutras para a tradução do texto é significativamente maior.

Em relação à precisão das abordagens, a tradução do dicionário possui uma precisão

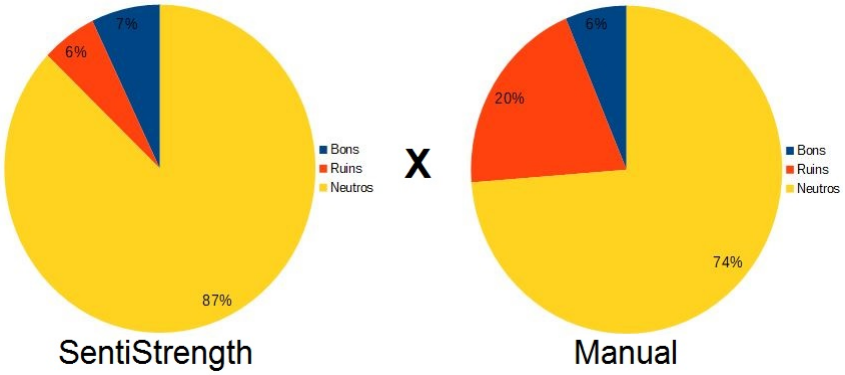

Figura 24 – Percentual Classificativo - "ônibus" - *SentiStrength* x Manual.

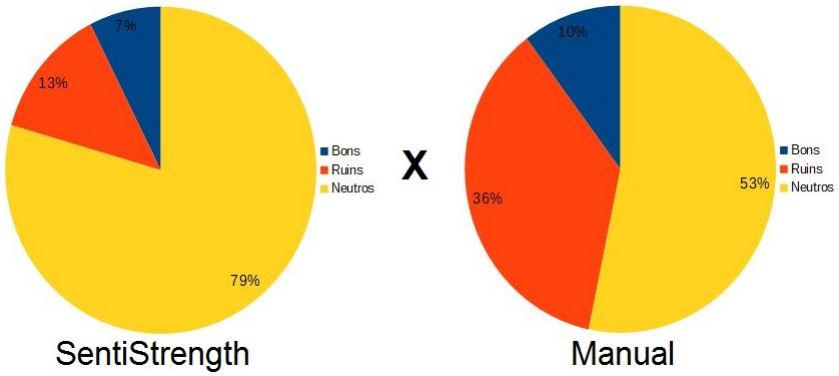

Figura 25 – Percentual Classificativo - "bus" - *SentiStrength* x Manual.

de aproximadamente 75%, enquanto a tradução do *tweet* possui uma precisão de 62%. Esse resultado pode indicar que a tradução individual de cada palavra foi mais precisa que a tradução de uma sentença completa, muitas vezes composta por gírias e abreviações, as quais foram desconsideradas pelo traduor automático.

4.5 MAPA CLASSIFICATIVO E COMPARAÇÃO

Nessa seção são representadas algumas classificações dos *Tweets*, divididas em palavras-chave, em um mapa da cidade de Curitiba e a comparação com o trabalho de Kozievitch *et al.* (2015).

Foram selecionados somente os *tweets* que continham a palavra-chave "ônibus" e que possuiam coordenadas. Posteriormente eles foram classificados pelo *SentiStrength* em bons, ruins ou neutros. Cada classificação é um marcador no mapa, onde os marcadores verdes representam os bons, os vermelhos os ruins e os brancos, os neutros, conforme a Figura 26. A Figura 27 identifica exemplos para cada tipo de classificação.

É perceptível que a maioria dos *tweets* selecionados para representação estão localizados na região central da cidade. Também foi observado que, apesar do filtro do "*check-in*", ainda há alguns relatos semelhantes, os quais o *SentiStrength* classificou em maior parte como

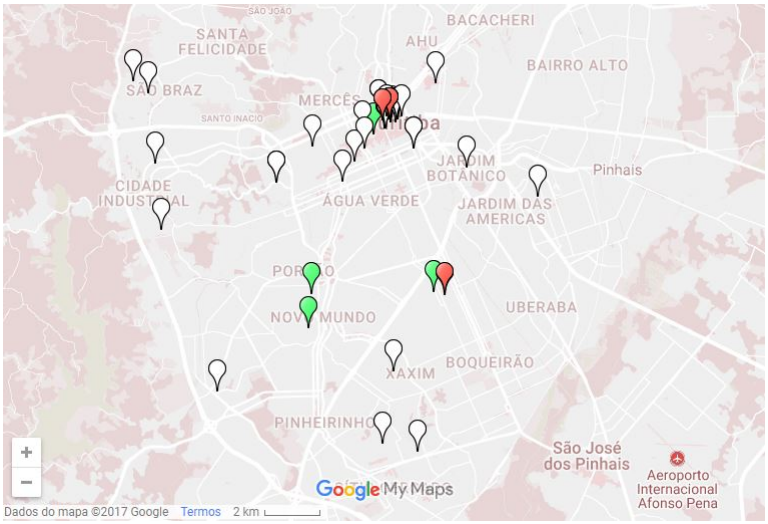

Figura 26 – Representação Classificativa - Palavra-Chave "Ônibus".

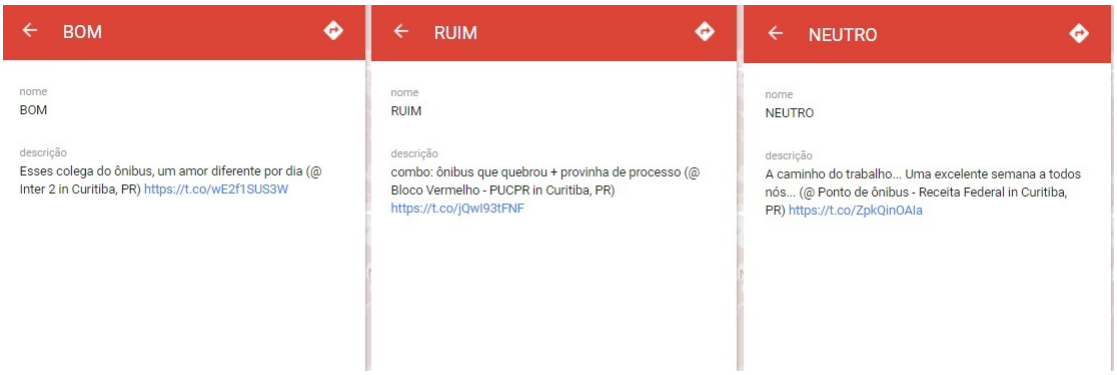

Figura 27 – Detalhes dos Marcadores - Palavra-Chave "Ônibus".

neutros.

Kozievitch *et al.* (2015) trazem o tema de cidades inteligentes ao tratar mais especificamente sobre o transporte público da cidade de Curitiba. A partir da perspectiva SIG, foram apresentados dados provenientes de diversas fontes, onde então obteve-se mapas representativos de cada linha de ônibus, ruas e também de densidade das linhas de ônibus com um foco na região central da cidade. A Figura 28 gerada por Kozievitch *et al.* (2015) apresenta as linhas de ônibus (em azul), terminais (em vermelho) e 11 categorias de ônibus, onde é possível perceber uma grande concentração de linhas de ônibus da região central, assim como foi verificada uma concentração de *tweets* na Figura 26.

Dois mapas foram gerados a partir da seleção de todos os *tweets*, filtrados pelas palavras-chave sobre transporte público que possuíam coordenadas, a Figura 29 representa, acima, os *tweets*, enquanto abaixo estão representados, além dos *tweets*, os terminais da cidade.

É perceptível que há uma semelhança entre os pontos onde estão localizados os terminais com alguns onde estão localizados os *tweets* e também a maior concentração na região central. Em comparação com a Figura 28, apenas visualmente por esses gráficos, não é possível verificar uma relação com os *tweets* e os caminhos das linhas de ônibus representadas.

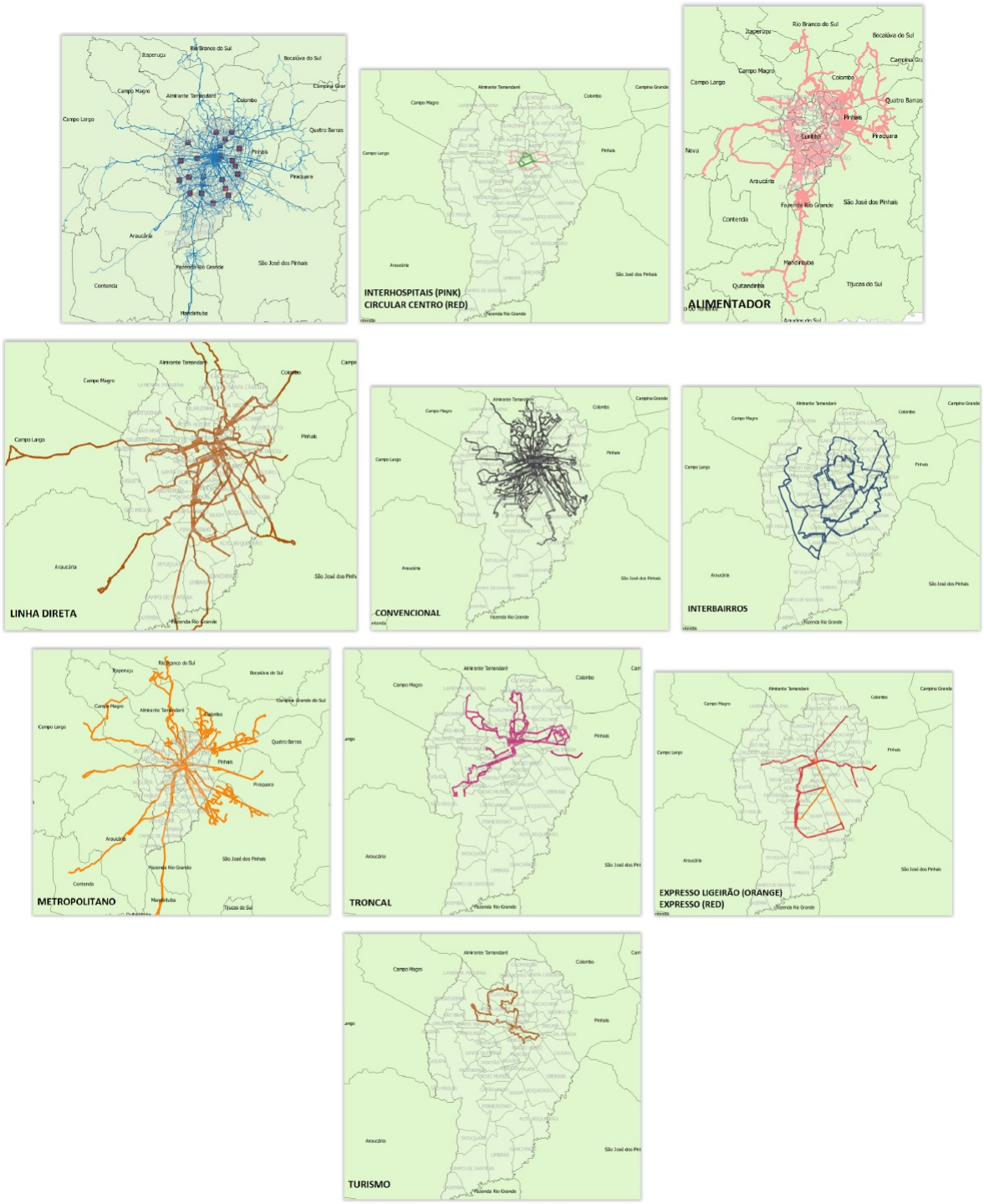

Figura 28 – Representação das Linhas de ônibus, terminais e categorias de ônibus (KOZIEVITCH *et al.*, 2015).

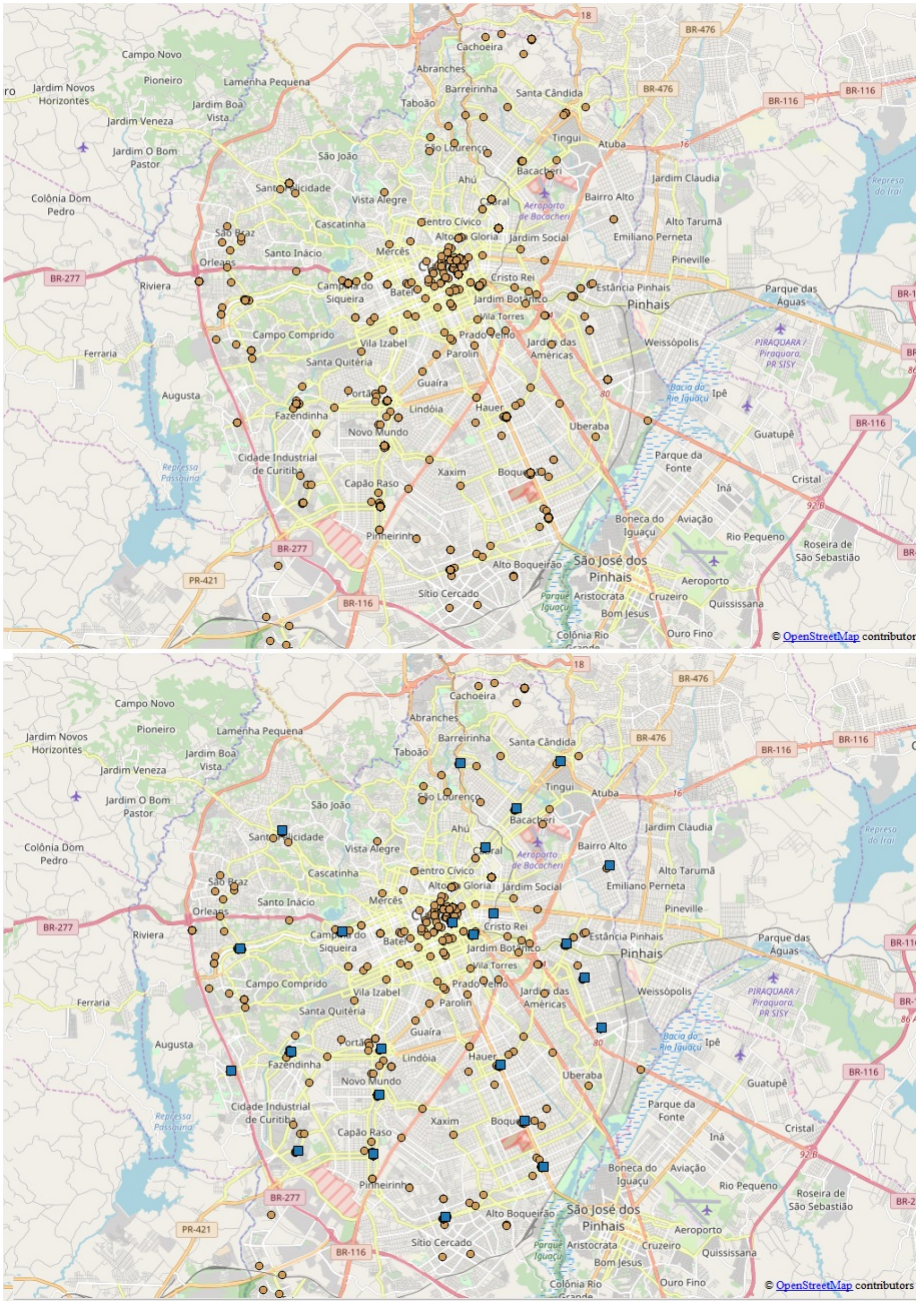

Figura 29 – Representação dos *tweets* com coordenadas filtrados.

5 CONCLUSÃO

Este trabalho propôs identificar aspectos relevantes sobre a mobilidade no transporte público a partir de dados provenientes dos usuários do *twitter*, em um estudo de caso da cidade de Curitiba. Para isso foram propostos os seguintes objetivos específicos: (I) realizar pesquisa sobre termos em mobilidade para filtrar os *tweets*, (II) filtrar os *tweets* em um caso de estudo de Curitiba, (III) aplicar a análise de sentimento nos *tweets*, (IV) analisar os *tweets* pela perspectiva SIG, (V) comparar com os dados de Kozievitch *et al.* (2015).

Para atingir os objetivos, primeiramente foi levantado um referencial teórico abordando as três principais áreas à serem tratadas no trabalho: Mobilidade, Análise de Sentimento e Banco de Dados. A mobilidade humana foi escolhida como o foco do trabalho, principalmente, referente ao transporte público. Em relação a análise de sentimento foram encontrados métodos e técnicas, como também algumas aplicações prontas, que, a partir de um texto, analisam se a pessoa é favorável ou contra algo ou como ela estava se sentindo no momento. Cada estratégia apresenta vantagens e desvantagens. Optou-se pela utilização da ferramenta *SentiStrength*. Já em banco de dados, os temas abordados podem ser separados em três grandes partes: banco de dados de *Tweets*, os quais foram armazenados e tratados em arquivo JSON; dicionário de palavras (isto é, dicionário com os pesos de sentimentos de cada palavra), proveniente da ferramenta *SentiStrength* e sistema de informação geográfica. Sobre Sistemas de Informação Geográfica (SIG), uma vez que os *Tweets* estavam filtrados e analisados na parte de sentimentos, a partir dos metadados de posição do *Tweet* foi possível uma análise em termos de SIG.

Já com os *tweets* geolocalizados na cidade de Curitiba, um formulário foi criado para obter os termos relacionados ao transporte público, atingindo o objetivo específico (I). Em seguida, a partir dos termos obtidos, foi aplicado um filtro na base de dados, alcançando (II). O objetivo específico (III) foi atingido com a seleção e aplicação da ferramenta *SentiStrength*, onde foram escolhidas duas abordagens para a análise de sentimento, traduzindo o *tweet*, que mostrou uma menor precisão, e traduzindo o dicionário da ferramenta, que obteve a melhor precisão entre as abordagens selecionadas. Ao reproduzir graficamente os *tweets* classificados, o objetivo específico (IV) foi alcançado, onde foi possível a seleção de cada *tweet* classificado para obter o relato que levou àquela classificação. Em geral os *tweets* foram classificados como neutros, além de que, o contexto dos *tweets*, apesar de conter alguma das palavras-chave selecionadas, muitas vezes não estavam relacionados com o transporte público, mas sim sobre temas pessoais. Com os *tweets* representados graficamente, foi possível a comparação com os dados de Kozievitch *et al.* (2015), mostrando que os *tweets* filtrados com as palavras-chave selecionadas sobre o transporte público, tem uma concentração maior na região central, além de uma semelhança com a localização dos terminais da cidade. Uma relação com as linhas de ôbibus e os *tweets* não foi possível apenas com os gráficos gerados.

Durante o desenvolvimento do trabalho foram encontradas algumas dificuldades relaci-

onadas principalmente à precisão da ferramenta escolhida para a análise. A tradução automática, tanto do dicionário quanto do texto, não é precisa, impedindo que se obtivesse resultados confiáveis em relação ás classificações. Outra questão é a quantidade de *tweets* não relacionados com o transporte público, onde a aplicação do filtro, feita individualmente para cada palavra-chave selecionada, não foi suficiente para selecionar apenas *tweets* no contexto pretendido. Tendo em vista essas questões, algumas propostas para melhorias e trabalhos futuros seriam: a utilização da possibilidade de treinamento da ferramenta *SentiStrength*; aprimoramento do dicionário da ferramenta no idioma pretendido; utilização de um filtro mais específico, como por exemplo apenas nas *hashtags*; consideração de todos os *tweets* com coordenadas para uma análise SIG e exploração de temas sociais, visto que grande parte dos *tweets* se tratavam de "conversas públicas".

REFERÊNCIAS

AboutTwitter. 2016. Disponível em: <https://about.twitter.com/company>.

Amazon. **Amazon's Mechanical Turk service**. 2017. Disponível em: <https://www.mturk.com/>.

BORNMANN, L.; LEYDESDORFF, L. Which cities produce more excellent papers than can be expected? a new mapping approach, using google maps, based on statistical significance testing. **Journal of the American Society for Information Science and Technology**, Wiley Subscription Services, Inc., A Wiley Company, Hoboken, v. 62, n. 10, p. 1954–1962, October 2011. ISSN 1532-2882.

CAMBRIA, E.; SCHULLER, B.; LIU, B.; WANG, H.; HAVASI, C. Knowledge-based approaches to concept-level sentiment analysis. **IEEE Intelligent Systems**, v. 28, n. 2, p. 12–14, March 2013. ISSN 1541-1672.

CAPRIELLO, A.; MASON, P. R.; DAVIS, B.; CROTTS, J. C. Farm tourism experiences in travel reviews: A cross-comparison of three alternative methods for data analysis. **Journal of Business Research**, v. 66, n. 6, p. 778 – 785, 2013. ISSN 0148-2963. International Tourism Behavior in Turbulent Times. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0148296311003274>.

CHENG, Y.-H.; CHEN, S.-Y. Perceived accessibility, mobility, and connectivity of public transportation systems. **Elsevier**, 2015.

CHOI, K. S.; IMB, I.; HOFSTEDE, G. J. A cross-cultural comparative analysis of small group collaboration using mobile twitter. **Computers in Human Behavior**, n. 65, p. 308 – 318, 2016.

CLIFFORD, J.; MACFADYEN, J.; MACFARLANE, D. Intro to google maps and google earth. **The Programming Historian (2013)**, Editorial Board of the Programming Historian, December 2013. ISSN 2397-2068. Disponível em: <https://doaj.org/article/8ab386823b634cf4bdcc696e9bf9c029>.

EUBra-BIGSEA1. **EUROPE - BRAZIL COLLABORATION OF BIG DATA SCIENTIFIC RESEARCH THROUGH CLOUD-CENTRIC APPLICATIONS**. 2016. Disponível em: <http://www.eubra-bigsea.eu/>.

EUBra-BIGSEA2. **Databases**. 2017. Disponível em: <http://data.ctweb.inweb.org.br>.

FERRARI, L.; BERLINGERIO, M.; CALABRESE, F.; CURTIS-DAVIDSON, B. Measuring public- transport accessibility using pervasive mobility data. **IEEE**, 2013.

FRANK, M. R.; WILLIAMS, J. R.; MITCHELL1, L.; BAGROW, J. P.; DODDS, P. S.; DANFORTH, C. M. Constructing a taxonomy of fine-grained human movement and activity motifs through social media. 2015.

GAL-TZUR, A.; GRANT-MULLER, S. M.; KUFLIK, T.; MINKOV, E.; SHOOR, I.; NOCERA, S. Enhancing transport data collection through social media sources: Methods, challenges and opportunities for textual data. v. 9, p. 407–417, 05 2015.

Glogg. 2016. Disponível em: <http://glogg.bonnefon.org/>.

GNIP. 2016. Disponível em: <http://support.gnip.com/articles/relational-databases-part-1.html>.

GoogleEarth. 2017. Disponível em: <https://www.google.com.br/intl/pt-PT/earth/>.

GoogleForms. 2017. Disponível em: <https://forms.google.com/>.

GoogleMaps. 2017. Disponível em: <https://www.google.com.br/maps>.

GoogleTradutor. 2017. Disponível em: <https://translate.google.com.br/>.

GRANDIN, P.; ADAN, J. M. Piegas: A systems for sentiment analysis of tweets in portuguese. **IEEE**, 2016. ISSN 1548-0992.

Hashtag. 2016. Disponível em: <https://support.twitter.com/articles/49309>.

HASSAN, P. H. A.; DAHAB, D. M. Y.; ABLA, E. H. E. E. Development of an intelligent gis application for spatial data analysis. **International Journal of Computer Science and Information Security**, v. 11, n. 4, 2013.

HAWELKA, B.; SITKO, I.; BEINAT, E.; SOBOLEVSKY, S.; KAZAKOPOULOS, P.; RATTI, C. Geo-located twitter as the proxy for global mobility patterns. 2013.

JSON. 2016. Disponível em: <http://www.json.org/>.

JSONTwitter. 2016. Disponível em: <https://dev.twitter.com/overview/api/tweets>.

JUREK, A.; MAURICE, D.; YAXIN, B. Improved lexicon-based sentiment analysis for social media analytics. **Springer**, 2015.

KANG, H.; YOO, S. J.; HAN, D. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. **Expert Systems With Applications**, Elsevier Ltd, 2011. ISSN 0957-4174.

KAUR, H.; MANGAT, V.; NIDHI. A survey of sentiment analysis techniques. p. 921–925, Feb 2017.

KOZIEVITCH, N. P.; GADDA, T. M. C.; FONSECA, K. V. O.; ROSA, M. O.; GOMES-JR, L. C.; AKBAR, M. Exploratory analysis of public transportation data in curitiba. 2015.

LIU, B. Sentiment analysis and subjectivity. **Handbook of Natural Language Processing**, 2010.

LIU, Y.; BI, J.-W.; FAN, Z.-P. Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. **Expert Systems with Applications**, v. 80, p. 323 – 339, 2017. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417417301951>.

LUO, F.; CAO, G.; MULLIGAN, K.; LI, X. Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of chicago. **Applied Geography**, n. 70, p. 11 – 25, 2016.

MAGUIRE, D. J. An overview and definition of gis. **Geographical information systems: Principles and applications**, 1991.

MANNING, C. D.; SCHIITZE, H. Foundations of statistical natural language processing. **The MIT Press**, 2010.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, v. 5, n. 4, p. 1093 – 1113, 2014. ISSN 2090-4479. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2090447914000550>.

Michaelis. **Ponto**. 2017. Disponível em: <http://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/ponto/>.

MITCHELL, L.; FRANK, M. R.; HARRIS, K. D.; DODDS, P. S.; DANFORTH, C. M. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. **PLOS ONE**, v. 8, 2013.

NASCIMENTO, P. C. Dicionário de polaridades para apoio a análise de sentimento. **COPRE UFRJ**, 2014.

NATIONS, U. World urbanization prospects. In: . Department of Economic and Social Affairs, 2014. ISBN 978-92-1-151517-6. Disponível em: <https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Highlights.pdf>.

O'CONNOR, F.; BALASUBRAMANYAN, R.; ROUTLEDGE, B. R.; SMITH, N. A. From tweets to polls: Linking text sentiment to public opinion time series. **Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media**, 2010.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? sentiment classification using machine learning techniques. **Proceedings of the Conference on Empirical Methods in Natural Language Processing**, p. 79 – 86, 2002.

SANTALA, V.; MICZEVSKI, S.; BRITO, S. A. de; LAO, A.; GADDA, T.; KOZIEVITCH, N.; SILVA, T. H. Making sense of the city: Exploring the use of social media data for urban planning and place branding. 2015.

SentiStrength. 2016. Disponível em: <http://sentistrength.wlv.ac.uk/>.

SINGH, P. K.; HUSAIN, M. S. Methodological study of opinion mining and sentiment analysis techniques. **International Journal on Soft Computing**, v. 5, n. 1, 2014.

SPENCE, P. R.; WESTERMAN, D.; HEIDE, B. V. D. A social network as information: The effect of system generated reports of connectedness on credibility on twitter. **Computers in Human Behavior**, n. 28, p. 199 – 206, 2012. Disponível em: <http://works.bepress.com/patric_spence/3/>.

TABOADA, M.; BROOKE, J.; TOFILOSKI, M.; VOLL, K.; STEDE, M. Lexicon-based methods for sentiment analysis. **Association for Computational Linguistics**, 2011.

TfL. **Transportation for London**. 2017. Disponível em: <https://oyster.tfl.gov.uk/oyster/>.

THELWALL, M.; BUCKLEY, K.; PALTOGLOU, G.; CAI, D.; KAPPAS, A. Sentiment strength detection in short informal text. **Journal of the American Society for Information Science and Technology**, Wiley Subscription Services, Inc., A Wiley Company, v. 61, n. 12, p. 2544–2558, 2010. ISSN 1532-2890. Disponível em: <http://dx.doi.org/10.1002/asi.21416>.

TwitterAPI. 2016. Disponível em: <https://dev.twitter.com/streaming/overview>.

TwitterSupport. 2016. Disponível em: <http://twitter.com>.

URBS. 2017. Disponível em: <https://www.urbs.curitiba.pr.gov.br/>.

WEISS, M.; BERNARDES, R.; CONSONI, F. Cidades inteligentes como nova prática para o gerenciamento dos serviços e infraestruturas urbanos: A experiência da cidade de porto alegre. **Urbe**, Editora CHAMPAGNAT, v. 7, n. 3, p. 310–324, 2015. ISSN 21753369.

ZAO, F.; HAO, H.; ZHANG", M. Sustainable mobility in china and its implications for emerging economies. v. 2, n. 1, p. 6 – 10, 2015.

ZENG, W.; FU, C.-W.; ARISONA, S. M.; ERATH, A.; QU, H. Visualizing mobility of public transportation system. **IEE**, v. 20, n. 12, 2014.

Apêndices

APÊNDICE A – FORMULÁRIO DE PESQUISA DE PALAVRAS-CHAVE

Essa seção apresenta o formulário de pesquisa sobre palavras-chave em relação ao transporte público. A Figura 30 é a primeira pergunta, a Figura 31 a segunda e na Figura 32 é solicitado a seleção e/ou também a criação de novas palavras-chave.

Figura 30 – Formulário p1

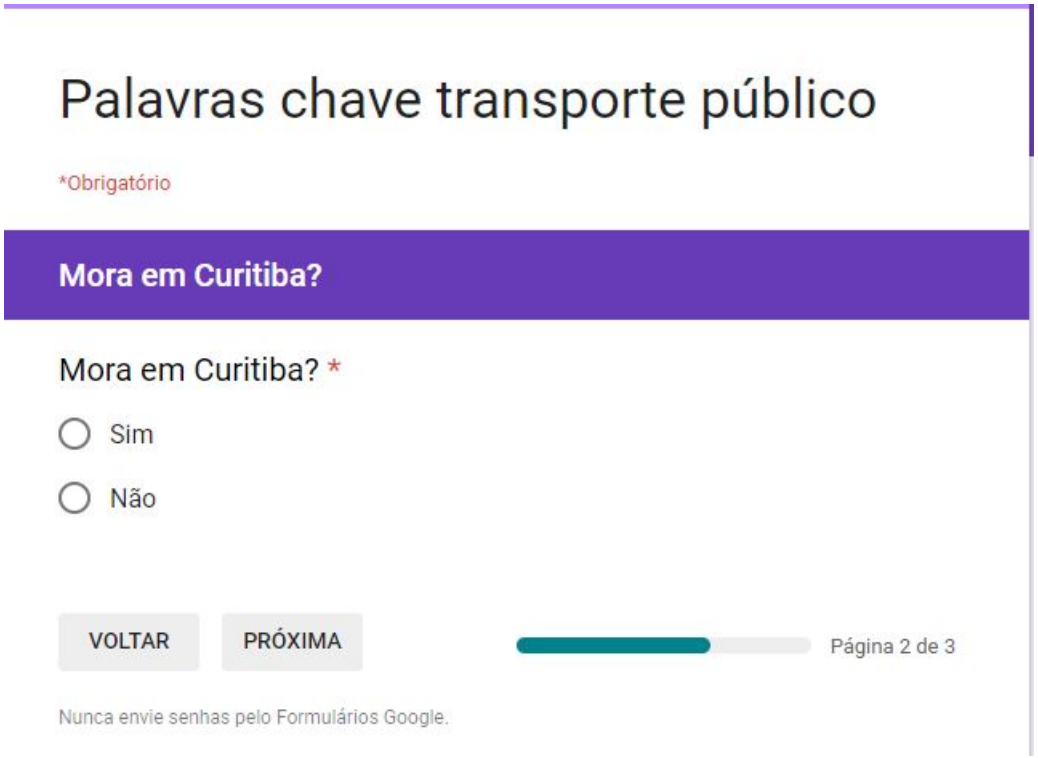

Figura 31 – Formulário p2

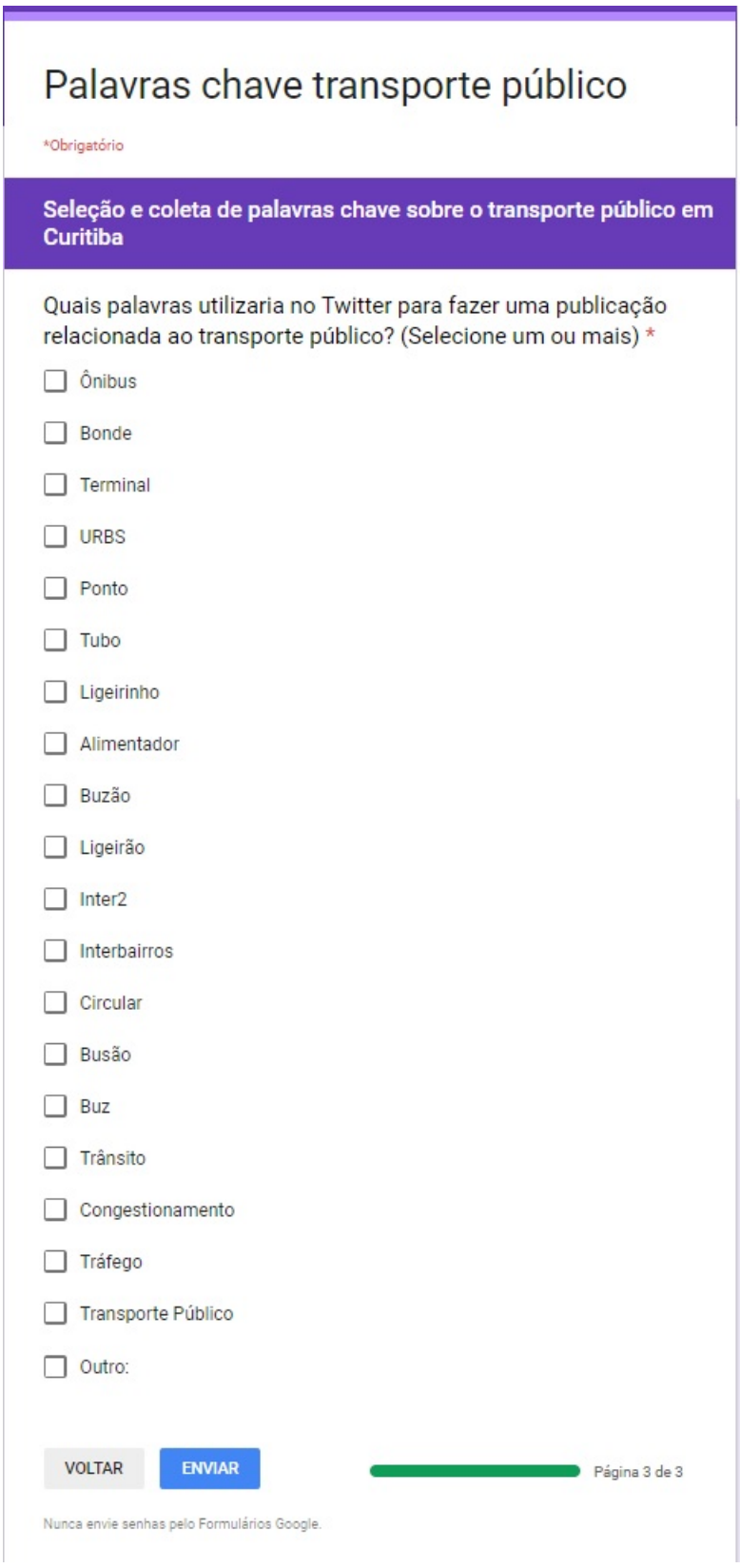

Figura 32 – Formulário p3

APÊNDICE B – *FILTRO*

Essa seção apresenta o código fonte do método principal para o filtro.

```
package filtrotweet;
import java.io.BufferedWriter;
import java.io.File;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.FileWriter;
import java.io.IOException;
import java.io.OutputStreamWriter;
import java.io.Writer;
import java.util.Scanner;
import com.google.gson.Gson;
import com.google.gson.JsonIOException;

public class main {
        private static int i = 0;
        private static boolean condition = true;
        public static void main(String[] args) throws JsonIOException,
            IOException {
                FileInputStream inputStream = null;
                Scanner sc = null;
                String[] filtro = {"bonde","bus","busao","busÃ£o","buz","buzao","
                    buzÃ£o","circular","inter 2","interbairros","latao","latÃ£o","
                    ligeirao","ligeirÃ£o","ligeirinho","ligerao","ligerÃ£o","
                    lotacao","lotaÃ§ao","lotaÃ§Ã£o","metrÃ´","onibos","Ã´nibos","
                    onibus","Ã´nibus","terminal","transito","transporte pÃºblico
                    ","tubo","urbs","vermelhao","vermelhÃ£o"};
                String url_entrada = "Z:$\backslash$TCC1$\
                    backslash$dump_UTFPR_geo$\backslash$traffic_geo_curitiba.json
                    ";
                String url_saida = "Z:$\backslash$TCC1$\backslash$dump_UTFPR_geo$\
                    backslash$abrilsaida.json";
                String url_saidatexto = "Z:$\backslash$TCC1$\
                    backslash$dump_UTFPR_geo$\backslash$abrilsaidaTEXTO.json";
                try (Writer writer = new BufferedWriter(new OutputStreamWriter(new
                     FileOutputStream(url_saida), "utf-8"))) {
                        try (Writer writerTexto = new BufferedWriter(
```

```
        new OutputStreamWriter(new FileOutputStream(
            url_saidatexto), "utf-8"))) {
try {
        inputStream = new FileInputStream(
            url_entrada);
        sc = new Scanner(inputStream, "UTF-8");
        String json;
        Thread t = new Thread(new Runnable() {
                public void run() {
                        while (condition)
                                try {
                                        Thread.sleep
                                            (1000);// 5s
                                        System.out.
                                            println("
                                            Linhas
                                            processadas:
                                             " + i);
                                } catch (Exception e)
                                    {
                                        e.
                                            printStackTrace
                                            ();
                                }
                }
        });
        t.start();
        while (sc.hasNextLine()) {

                json = sc.nextLine();
                Gson gson = new Gson();
                try{
                tweetObject tweet = gson.fromJson(
                    json, tweetObject.class);
                String h;
                try {
                        h = tweet.getEntities().
                            getHashtags()[1].
                            getHashtag().toLowerCase()
                            ;
                } catch (Exception e) {
                        h = "";
```

```
                }
                if (!tweet.getText().toLowerCase().
                    contains("i'm at")) {
                        for (String s : filtro) {
                                if (tweet.getText().
                                    toLowerCase().
                                    contains(" "+s+" ")
                                     || h.contains(s))
                                    {
                                        writer.write(
                                            gson.toJson(
                                            tweet) + "\n
                                            ");
                                        writerTexto.
                                            write(tweet.
                                            getText() +
                                            "\n");
                                        break;
                                }
                        }
                }
                i++;
        } catch (Exception e){
                i++;
        }
}
        System.out.println("Total de linhas
            processadas: " + i);

        // note that Scanner suppresses exceptions
        if (sc.ioException() != null) {
                throw sc.ioException();
        }
        condition = false;
        t.stop();
} finally {
        if (inputStream != null) {
                inputStream.close();
        }
        if (sc != null) {
                sc.close();
        }
```

```
                        }
                    } catch (IOException e) {
                        // TODO Auto-generated catch block
                        e.printStackTrace();
                    }
                }
            }
        }
```

APÊNDICE C – *COORDINATES*

Essa seção apresenta o código fonte do objeto *coordinates*.

```
package filtrotweet;

import java.util.ArrayList;

import com.google.gson.annotations.SerializedName;

public class Coordinites {
	@SerializedName("coordinates")
	Float[] coordinates;
	public Float[] getCoordinates() {
		return coordinates;
	}

	public void setCoordinates(Float[] coordinates) {
		this.coordinates = coordinates;
	}

}
```

APÊNDICE D – *CREATED_AT*

Essa seção apresenta o código fonte do objeto *created_at*.

```
package filtrotweet;

import com.google.gson.annotations.SerializedName;

public class Created_at {
	@SerializedName("\$date")
	String date;
	public String getDate() {
		return date;
	}

	public void setDate(String date) {
		this.date = date;
	}
}
```

APÊNDICE E – *ENTITY*

Essa seção apresenta o código fonte do objeto *entity*.

```
package filtrotweet;

import com.google.gson.annotations.SerializedName;

public class Entity {
	@SerializedName("hashtags")
	Hashtags[] hashtags;
	public Hashtags[] getHashtags() {
		return hashtags;
	}

	public void setHashtags(Hashtags[] hashtags) {
		this.hashtags = hashtags;
	}
}
```

APÊNDICE F – *HASHTAGS*

Essa seção apresenta o código fonte do objeto *hashtags*.

```
package filtrotweet;
import com.google.gson.annotations.SerializedName;
public class Hashtags {
	@SerializedName("text")
	String hashtag;

	public String getHashtag() {
		return hashtag;
	}

	public void setHashtag(String hashtag) {
		this.hashtag = hashtag;
	}
}
```

APÊNDICE G – *TWEETOBJECT*

Essa seção apresenta o código fonte do objeto *tweetobject* que é composto por outros objetos de relevância.

```
package filtrotweet;
import com.google.gson.annotations.SerializedName;
import twitter4j.EntitySupport;
public class tweetObject {
	@SerializedName("coordinates")
	Coordinites coordinates;
	@SerializedName("created_at")
	Created_at created_at;
	@SerializedName("entities")
	Entity entities;
	@SerializedName("text")
	String TweetText;
	public Coordinites getCoordinates() {
		return coordinates;
	}
	public Created_at getCreated_at() {
		return created_at;
	}
	public Entity getEntities() {
		return entities;
	}
	public String getText() {
		return TweetText;
	}
	public void setCoordinates(Coordinites coordinates) {
		this.coordinates = coordinates;
	}
	public void setCreated_at(Created_at created_at) {
		this.created_at = created_at;
	}
	public void setEntities(Entity entities) {
		this.entities = entities;
	}
	public void setText(String text) {
		this.TweetText = text;
	}
}
```

APÊNDICE H – CLASSIFICAÇÃO *SENTISTRENGTH*

Essa seção apresenta o código fonte do método principal para aplicação do *SentiStrength*.

```
package senticlassification;

import java.io.BufferedWriter;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.IOException;
import java.io.OutputStreamWriter;
import java.io.UnsupportedEncodingException;
import java.io.Writer;
import java.util.Scanner;
import uk.ac.wlv.sentistrength.SentiStrength;

public class main {
static boolean condition = true;
private static int i = 0;
	public static void main(String[] args) throws IOException {
		FileInputStream inputStream = null;
		Scanner sc = null;
		String url_entrada = "C:\\filtros\\marcosaidaTEXTO.pt.en.txt";
		String url_bom = "C:\\filtros\\saidaBOM frase.txt";
		String url_ruin = "C:\\filtros\\saidaRUIN frase.txt";
		String url_neutro = "C:\\filtros\\saidaNEUTRO frase.txt";
		SentiStrength sentiStrength = new SentiStrength();
		String ssthInitialisation[] = {"sentidata", "C:/SentiStrength_Data
		    /", "trinary"};
		sentiStrength.initialise(ssthInitialisation);
		inputStream = new FileInputStream(url_entrada);
		sc = new Scanner(inputStream, "UTF-8");
		Writer writer_bom = new BufferedWriter(new OutputStreamWriter(new
		    FileOutputStream(url_bom), "utf-8"));
		Writer writer_ruin = new BufferedWriter(new OutputStreamWriter(new
		     FileOutputStream(url_ruin), "utf-8"));
		Writer writer_neutro = new BufferedWriter(new OutputStreamWriter(
		    new FileOutputStream(url_neutro), "utf-8"));
		String line;
		Thread t = new Thread(new Runnable() {
```

```
                public void run() {
                        while (condition)
                                try {
                                        Thread.sleep(5000);// 5s
                                System.out.println("Linhas processadas:" + i
                                    );
                                } catch (Exception e) {
                                        e.printStackTrace();
                                }
                        }
                });
                t.start();
                int bons=0, ruins=0, neutros=0;
                while (sc.hasNextLine()) {
                        i++;
                        line = sc.nextLine();
                        String[] result = sentiStrength.computeSentimentScores(
                            line).split(" ");
                        if(Integer.valueOf(result[2])==-1){
                                ruins++;
                                writer_ruin.write(line+"\n");
                        }
                        else if((Integer.valueOf(result[2])==0)){
                                neutros++;
                                writer_neutro.write(line+"\n");
                        }
                        else {
                                writer_bom.write(line+"\n");
                                bons++;
                        }
                }
                System.out.println("Total linhas: "+i);
                condition=false;
                t.stop();
                sc.close();
                writer_bom.close();
                writer_neutro.close();
                writer_ruin.close();
                System.out.println("Bons: "+bons+" Ruins: "+ ruins + " Neutros: "+
                    neutros);
        }
}
```

Anexos

ANEXO A – PARTE DE UM ARQUIVO JSON DO TWEET ORIGINAL

Essa seção apresenta dois *tweets* em JSON originais.

{ "_id" : 715690118816866304, "contributors" : null, "control" : { "coletas" : [{ "id" : 349 }] }, "coordinates" : null, "created_at" : { " \$date" : 1459468795000 }, "entities" : { "user_mentions" : [], "symbols" : [], " hashtags" : [], "urls" : [] }, "favorite_count" : 0, "favorited" : false, " filter_level" : "low", "geo" : null, "id" : 715690118816866304, "id_str" : "715690118816866304", "in_reply_to_screen_name" : null, " in_reply_to_status_id" : null, "in_reply_to_status_id_str" : null, " in_reply_to_user_id" : null, "in_reply_to_user_id_str" : null, " is_quote_status" : false, "lang" : "pt", "place" : { "country_code" : "BR", "url" : "https://api.twitter.com/1.1/geo/id/6d5542f8d837770d.json", "country " : "Brasil", "place_type" : "city", "bounding_box" : { "type" : "Polygon", "coordinates" : [[[-49.391643, -25.644752], [-49.391643, -25.345747], [-49.185278, -25.345747], [-49.185278, -25.644752]]] }, "full_name" : "Curitiba, Brasil", "attributes" : {}, "id" : "6d5542f8d837770d", "name" : " Curitiba" }, "retweet_count" : 0, "retweeted" : false, "source" : "Twitter for iPhone ", "text" : "Tem gente vindo aq em casa agr compra ingresso kkkkkk merere", "timestamp_ms" : "1459468795862", "truncated" : false, "user" : { " follow_request_sent" : null, "profile_use_background_image" : true, "id" : 2361361760, "verified" : false, "profile_image_url_https" : "https://pbs. twimg.com/profile_images/713851221388427265/u-dwD63B_normal.jpg", " profile_sidebar_fill_color" : "DDEEF6", "is_translator" : false, " geo_enabled" : true, "profile_text_color" : "333333", "followers_count" : 2987, "protected" : false, "location" : "Uberaba, Curitiba", " default_profile_image" : false, "id_str" : "2361361760", "utc_offset" : -10800, "statuses_count" : 74862, "description" : "Vida ï¿½ brisa passageira .. Isa ? Snap ï¿½ Stteeefs ï¿½ @Coritiba ?", "friends_count" : 684, " profile_link_color" : "0084B4", "profile_image_url" : "http://pbs.twimg.com/ profile_images/713851221388427265/u-dwD63B_normal.jpg", "notifications" : null, "profile_background_image_url_https" : "https://pbs.twimg.com/ profile_background_images/445737650876461056/ylaCa9M_.jpeg", " profile_background_color" : "C0DEED", "profile_banner_url" : "https://pbs. twimg.com/profile_banners/2361361760/1459217263", " profile_background_image_url" : "http://pbs.twimg.com/ profile_background_images/445737650876461056/ylaCa9M_.jpeg", "screen_name" : "stteefs", "lang" : "pt", "profile_background_tile" : true, " favourites_count" : 19891, "name" : "Stef SML ?", "url" : "http:// ch00sebehappy.tumblr.com", "created_at" : "Tue Feb 25 16:47:32 +0000 2014",

"contributors_enabled" : false, "time_zone" : "Brasilia", "
profile_sidebar_border_color" : "FFFFFF", "default_profile" : false, "
following" : null, "listed_count" : 3 } }
{ "_id" : 715690105101545472, "contributors" : null, "control" : { "coletas" : [
{ "id" : 349 }] }, "coordinates" : null, "created_at" : { " \$date" :
1459468792000 }, "entities" : { "user_mentions" : [], "symbols" : [], "
hashtags" : [], "urls" : [] }, "favorite_count" : 0, "favorited" : false, "
filter_level" : "low", "geo" : null, "id" : 715690105101545472, "id_str" :
"715690105101545472", "in_reply_to_screen_name" : "lucasggusmao", "
in_reply_to_status_id" : 715690058389594112, "in_reply_to_status_id_str" :
"715690058389594112", "in_reply_to_user_id" : 1352416244, "
in_reply_to_user_id_str" : "1352416244", "is_quote_status" : false, "lang" :
"pt", "place" : { "country_code" : "BR", "url" : "https://api.twitter.com
/1.1/geo/id/6d5542f8d837770d.json", "country" : "Brasil", "place_type" : "
city", "bounding_box" : { "type" : "Polygon", "coordinates" : [[[
-49.391643, -25.644752], [-49.391643, -25.345747], [-49.185278,
-25.345747], [-49.185278, -25.644752]]] }, "full_name" : "Curitiba,
Brasil", "attributes" : {}, "id" : "6d5542f8d837770d", "name" : "Curitiba"
}, "retweet_count" : 0, "retweeted" : false, "source" : "<a href=\"http://
twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone", "
text" : "eu sou o cancer do grupo da sala", "timestamp_ms" :
"1459468792592", "truncated" : false, "user" : { "follow_request_sent" :
null, "profile_use_background_image" : true, "id" : 1352416244, "verified" :
false, "profile_image_url_https" : "https://pbs.twimg.com/profile_images
/714633255098691590/gPO73lUF_normal.jpg", "profile_sidebar_fill_color" : "
C0DFEC", "is_translator" : false, "geo_enabled" : true, "profile_text_color"
: "333333", "followers_count" : 421, "protected" : false, "location" : "
curitiba", "default_profile_image" : false, "id_str" : "1352416244", "
utc_offset" : -10800, "statuses_count" : 38350, "description" : null, "
friends_count" : 49, "profile_link_color" : "981CEB", "profile_image_url" :
"http://pbs.twimg.com/profile_images/714633255098691590/gPO73lUF_normal.jpg
", "notifications" : null, "profile_background_image_url_https" : "https://
pbs.twimg.com/profile_background_images/671454737628332032/psjO7MVH.jpg", "
profile_background_color" : "022330", "profile_banner_url" : "https://pbs.
twimg.com/profile_banners/1352416244/1458942394", "
profile_background_image_url" : "http://pbs.twimg.com/
profile_background_images/671454737628332032/psjO7MVH.jpg", "screen_name" :
"lucasggusmao", "lang" : "pt", "profile_background_tile" : true, "
favourites_count" : 7436, "name" : "LUCAS", "url" : "http://Instagram.com/
lucasgusmaoo", "created_at" : "Sun Apr 14 17:44:44 +0000 2013", "
contributors_enabled" : false, "time_zone" : "Brasilia", "
profile_sidebar_border_color" : "A8C7F7", "default_profile" : false, "

following" : null, "listed_count" : 9 } }

ANEXO B – PARTE DE UM ARQUIVO JSON DO TWEET FILTRADO

Essa secao apresenta alguns *tweets* em JSON apos o filtro.


```
{"created\_at":{"\$date":"1459468543000"},"entities":{"hashtags":[]},"text":"Sai
    de vermelho hoje. Voltando pra casa, um passageiro do onibus disse que eu
    nao iria entrar, pois estava com uma blusinha vermelha. Pode?"}
{"created\_at":{"\$date":"1459468400000"},"entities":{"hashtags":[]},"text":"
    Minha mae podia me levar la naquela parada amanha n to afim de andar de
    nibus nao"}
{"created\_at":{"\$date":"1459468157000"},"entities":{"hashtags":[]},"text":"[20
    h43 31/03/2016]Fer:vamos no shopping amanha, de onibus [20h44 31/03/2016]
    Paola:E o carro? [20h44 31/03/2016]Fer:N sei dirigir ate la"}
{"created\_at":{"\$date":"1459468094000"},"entities":{"hashtags":[]},"text":"ela
    so nao ta falando ainda e ainda ta comendo e bebendo por tubos"}
{"coordinates":{"coordinates":[-49.25,-25.4167]},"created\_at":{"\$date
    ":"1459467509000"},"entities":{"hashtags":[]},"text":"O desespero esta
    tomando conta agora nao e mais R\$ 30,00 e show, palco, onibus custeio de
    https://t.co/UDJmTOT0jp"}
{"created\_at":{"\$date":"1459467462000"},"entities":{"hashtags":[]},"text":"nao
    e domingo a noite, mas to com essa na cabeca desde o onibus https://t.co/
    XLmLCNfn5w"}
{"created\_at":{"\$date":"1459465350000"},"entities":{"hashtags":[]},"text":"Meu
    boy me buscou no terminal e passou a tarde INTEIRA COMIGO ??????????"}
{"created\_at":{"\$date":"1459464715000"},"entities":{"hashtags":[]},"text":"
    Voce percebe que tem algo errado quando voce dorme em pe no onibus..."}
{"created\_at":{"\$date":"1459464385000"},"entities":{"hashtags":[]},"text":"
    Falando em encarada tinha uma mina do Pedro no msm bonde q eu hj q nem
    disfarca p olhar. Ia na alma"}
{"created\_at":{"\$date":"1459464323000"},"entities":{"hashtags":[]},"text":"eu
    nao acredito que vou ter que pegar onibus sozinha amanha, me recusooooo"}
```