
UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ
CAMPUS LUIZ MENEGHEL - CENTRO DE CIÊNCIAS TECNOLÓGICAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ALLAN VICTOR PIRES

# ESTUDO DE CASO PARA ESTIMAR O BEM ESTAR DA POPULAÇÃO EM CIDADES PARANAENSES POR MEIO DE REDES SOCIAIS

**BANDEIRANTES–PR**

**2017**

ALLAN VICTOR PIRES

# ESTUDO DE CASO PARA ESTIMAR O BEM ESTAR DA POPULAÇÃO EM CIDADES PARANAENSES POR MEIO DE REDES SOCIAIS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Norte do Paraná para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Bruno Squizato Faiçal

**BANDEIRANTES–PR**

**2017**

ALLAN VICTOR PIRES

# ESTUDO DE CASO PARA ESTIMAR O BEM ESTAR DA POPULAÇÃO EM CIDADES PARANAENSES POR MEIO DE REDES SOCIAIS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Norte do Paraná para obtenção do título de Bacharel em Ciência da Computação.

**BANCA EXAMINADORA**

---

Prof. Dr. Bruno Squizato Faiçal
Universidade Estadual do Norte do Paraná
Orientador

---

Prof. Dr. André Luis Andrade Menolli
Universidade Estadual do Norte do Paraná

---

Prof. Me. Luiz Fernando Legore do Nascimento
Universidade Estadual do Norte do Paraná

Bandeirantes–PR, 16 de novembro de 2017

*Este trabalho é dedicado aos meus pais, Atair e Ivanete, que sempre me apoiaram e acreditaram na minha capacidade.*

# AGRADECIMENTOS

Inicialmente, quero começar agradecendo a Deus que todos os dias da minha vida me deu forças para que eu nunca desistisse, por todos os momentos de inspiração e fé.

Ao meu orientador, Prof. Dr. Bruno Squizato Faiçal, por seu apoio e amizade, além da sua dedicação, competência e paciência.

Aos professores da banca Prof. Dr. André Luis Andrade Menolli e Prof. Me. Luiz Fernando Legore do Nascimento, por todos os conselhos, ideias e por me auxiliarem no desenvolvimento do trabalho. Também agradecer ao Prof. Me. Wellington Della Mura e Prof. Dr. Márcio Massashiko Hasegawa, por sempre me auxiliarem, pelos conselhos, pelo companheirismo e grandes ideias.

Aos professores do Centro de Ciências Tecnológicas, que de alguma forma contribuíram para minha formação.

As amizades que foram feitas durante esses anos, principalmente as amizades do curso de Ciência da Computação.

Aos meus amigos, que me apoiaram e que sempre estiveram ao meu lado durante esta longa caminhada, em especial meus amigos Lucas, Vitor, Gustavo e Mateus, também aos meus primos Luan e Maria Letícia, que muitas vezes compartilhei momentos de tristezas, alegrias, angústias e ansiedade, mas que sempre estiveram ao meu lado me apoiando e me ajudando.

E principalmente, aos meus pais, Atair Leopoldo Pires e Ivanete das Neves Pires, que são os maiores incentivadores dos meus estudos, por sempre acreditarem em mim e que sem eles, isso não seria possível. Se possível, também gostaria de pedir desculpas aos meus pais pelos vários momentos em que estive ausente.

*"A vida é uma sequência de encontros inéditos com o mundo,*
*e portanto ela não se deixa traduzir em fórmulas de nenhuma espécie".*
*(Clóvis de Barros Filho)*

PIRES, ALLAN V.. **Estudo de caso para estimar o bem estar da população em cidades paranaenses por meio de redes sociais**. 51 p. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual do Norte do Paraná, Bandeirantes–PR, 2017.

# RESUMO

Informações sobre o desenvolvimento das cidades e a disponibilidade dos serviços urbanos oferecidos por estas são comumente utilizadas por administradores públicos (governantes) para tomar decisões em busca de melhorias. Estas informações são baseadas nas informações coletadas por um processo denominado Censo Demográfico realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Contudo, novos resultados (oriundos de um novo censo) são divulgados com a periodicidade de 10 anos. Esse período de latência é o principal motivo da obsolescência das informações divulgadas. Com objetivo de reduzir o intervalo de disponibilização destas informações e permitir que estas estejam sempre atualizadas este trabalho investigou o processamento de textos curtos disponibilizados em redes sociais por usuários residentes em cidades brasileiras para extrair o bem-estar em residir em suas respectivas cidades. Tal informação é manipulada como um índice comparativo entre as cidades monitoradas, o que permitiria ordená-las conforme o seu respectivo bem-estar. O desenvolvimento deste trabalho possibilitou observar que existem evidências que é possível alcançar o objetivo almeja, mas que ainda existem lacunas em áreas importantes, entre elas: a classificação da emoção de textos curtos no idioma português. Nesse sentido, pode-se citar como contribuições deste trabalho (i) o desenvolvimento de uma plataforma de coleta de textos curtos da rede social Twitter com base em cidades, (ii) análise de estudos com propostas que apresentem contextos semelhantes a este trabalho, (iii) a disponibilização de um conjunto de textos curtos oriundos de usuários de 13 cidades, com início em 30/09/2017 e término em 13/11/2017, resultando em aproximadamente 870 mil instâncias, (iv) estudo inicial sobre a estimação do bem-estar com base na fórmula da média geométrica, e (v) mapeamento dos obstáculos que compõem propostas com natureza semelhante a deste estudo.

**Palavras-chave**: Mineração de Dados. Análise de Sentimento. Twitter. Bem-Estar.

PIRES, ALLAN V.. **Case study to estimate the well-being of the population in cities of Paraná through social networks**. 51 p. Final Project (Bachelor of Science in Computer Science) – State University Northern of Parana , Bandeirantes–PR, 2017.

# ABSTRACT

Information about the development of cities and the availability of urban services offered by them are commonly used by public administrators to make decisions for improvements. This information is based on the information collected by a process called the Demographic Census conducted by the Brazilian Institute of Geography and Statistics (IBGE). However, new results from a new census are published every 10 years. In order to reduce the interval of availability of this information and allow it to be updated this work investigated the processing of short texts available on social networks by users residing in Brazilian cities to extract well-being in living in their respective cities. Such information is manipulated as a comparative index between the monitored cities, which would make it possible to organize them according to their respective well-being. Through the development of this work is possible to observe that there is evidence that it is possible to achieve the desired goal, but that there are still gaps in important areas, among them: the classification of emotion of short texts in Portuguese language. In this respect, we can cite as contributions of this work (i) the development of a platform to collect short texts from social network Twitter based on cities, (ii) analysis of studies with proposals that present contexts similar to this work, ( iii) the availability of short text set from users of 13 cities, started on 09/30/2017 and ended on 11/13/2017, resulting in approximately 870 thousand instances; (iv) an initial study on the estimation of well-being based on the geometric mean formula, and (v) mapping the obstacles that compose proposals with a similar nature to this study.

**Keywords**: Data Mining. Sentiment Analysis. Twitter. Well-Being

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| API | *Application Programming Interface* |
| HTTP | *Hyper Text Transfer Protocol* |
| IBEU | Índice de Bem-Estar Urbano |
| IBGE | Instituto Brasileiro de Geografia e Estatística |
| IDH | Índice de Desenvolvimento Humano |
| IDHM | Índice de Desenvolvimento Humano Municipal |
| INCT | Instituto Nacional de Ciência e Tecnologia |
| JSON | *JavaScript Object Notation* |
| KDD | *Knowledge Discovery from Data* |
| KDT | *Knowledge Discovery from Text* |
| PNAD | Pesquisa Nacional por Amostra de Domicílios |

# SUMÁRIO

# 1 INTRODUÇÃO

O uso das redes sociais têm se tornado uma importante fonte de comunicação entre as pessoas. Milhões de usuários compartilham diariamente seus pensamentos sobre diversos assuntos, possibilitando a análise e mineração de dados. Segundo Araujo et al. (2012), uma possível aplicação é a avaliação da polarização do sentimento (positivo, neutro e negativo) que os usuários apresentam sobre temas variados (tais como, política, esportes e opiniões sobre produtos).

Pesquisadores tem investigado o grande volume de dados que a rede social Twitter produz diariamente (RIOS et al., 2017). O qual permite estudos no sentido de entender o comportamento e o sentimento sobre assuntos específicos (ARAMPATZI; BURGER; NOVIK, 2016; ARAUJO et al., 2012; RIOS et al., 2017; SOUSA, 2012).

O Twitter é uma rede social no qual permite com que seus usuários possam enviar mensagens curtas denominadas *tweets* com no máximo de 140 caracteres. No ano de 2016, o Twitter teve em torno de 313 milhões de usuários ativos mensalmente e aproximadamente de 82% de usuários móveis ativos (TWITTER, 2017a). Este cenário possibilita a obtenção de informações geradas pelos seus respectivos usuários. Adicionalmente, a geolocalização dos dispositivos dos usuários pode ser considerado para análises que consideram informações relacionadas ao espaço-tempo do usuário.

Considerando a velocidade com que as informações são disponibilizadas nas redes sociais e seu respectivo processamento, tal abordagem é vista como uma alternativa a índices atuais. O Índice de Desenvolvimento Humano Municipal (IDHM) e o Índice de Bem-Estar Urbano (IBEU) são índices baseados no Censo Demográfico realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), com intervalos periódicos de 10 anos (ESTATÍSTICA, 2013). Apesar de ser uma importante fonte de informações, tal periodicidade torna as informações obsoletas para análises com fortes requisitos temporais.

Nesse sentido, explorar a mineração de dados em redes sociais permite que informações latentes relacionadas aos índices existentes (IDHM e IBEU) sejam extraídas em menor espaço de tempo. Tal condição permite que as informações sejam menos prejudicadas pelas características temporais. Segundo Sousa (2012), as informações obtidas pelas redes sociais permite descobrir conhecimentos significantes em relação aos meios sócio-cultural e políticos.

Portanto, o objetivo geral do trabalho é utilizar a área de mineração de dados para investigar um índice representativo do bem-estar social, de forma a propor um índice alternativo no qual reduzirá o período de divulgação de informações em relação ao bem-estar. Tal índice permitiria um acesso amigável à informação atualizada para a po-

pulação e aos governantes políticos. Este estudo é classificado como "mineração de dados sociotécnicos", pois envolve uma avaliação humana e não apenas técnicas estatísticas e de aprendizado de máquina para a realização do estudo (DODDS; DANFORTH, 2010).

Espera-se que os resultados obtidos nesse trabalho apresente evidências que é possível empregar tal abordagem para analisar o bem-estar da população em relação a cidade que reside. Para isso são considerados a polarização do sentimento dos textos publicados por usuários da rede social que residem nas respectivas cidades. Considerando-se assim que os usuários ativos na rede social residentes em cada cidade monitorada são representativos em relação a população geral do município.

É importante ressaltar que este estudo busca contribuir para reduzir a escassez de estudos neste contexto para cidades brasileiras. Assim, as cidades consideradas no estudo de caso para este trabalho pertencem ao estado do Paraná, mesmo estado da universidade que fornece a infraestrutura para a realização deste estudo. Por fim, é considerado o idioma português para os textos analisados por ser o idioma nativo e o mais empregado pelo grupo de usuário considerados.

Este trabalho está organizado conforme segue. O Capítulo 2 contém a fundamentação teórica sobre os conceitos técnicos e teóricos necessários para realização do trabalho. O Capítulo 3 apresenta os trabalhos relacionados. O Capítulo 4 descreve a proposta de estudo apresentada. O Capítulo 5 apresenta sobre o desenvolvimento e as verificações realizadas. O Capítulo 6 descreve os resultados obtidos pela pesquisa. Finalmente, o Capítulo 7 apresenta as considerações finais e a descrição de possíveis trabalhos futuros.

# 2 FUNDAMENTAÇÃO TEÓRICA

Tendo o aumento significativo do uso das redes sociais, a questão de analisar os sentimentos e/ou mineração de opiniões se tornou um assunto de estudo para muitas pesquisas com intuitos diferentes (GONÇALVES; DORES; BENEVENUTO, 2013), como compartilhamento de informações em tempo real em (CHA et al., 2012), eventos relacionados com tragédias e revoluções em (SAKAKI; OKAZAKI; MATSUO, 2010; GOMIDE et al., 2011) e previsão de ataques terroristas (TUMASJAN et al., 2010; DIAKOPOULOS; SHAMMA, 2010). Para o decorrer deste trabalho, é preciso introduzir alguns conceitos importantes para obter conhecimento do assunto e que os mesmos estejam bem definidos.

As próximas seções irão abordar sobre a Mineração de Dados, Mineração de Textos, Redes Sociais e Mídias Sociais e Índices Urbanos.

## 2.1 Mineração de Dados

A tecnologia da informação tem nos permitido coletar enormes quantidades de dados em vários campos, visto que a internet tem produzido muito conteúdo diariamente e analisar os conteúdos produzidos tornou-se uma necessidade. Sendo assim é preciso da mineração de dados, pois como um campo multidisciplinar, a mineração de dados trabalha em conjunto com várias áreas como estatística, aprendizado de máquina, inteligência artificial, reconhecimento de padrões, visualização de dados (HAN; KAMBER; PEI, 2012).

A mineração de dados, também conhecida como descoberta de conhecimento a partir dos dados (do inglês, *Knowledge Discovery from Data* - KDD) permite descobrir padrões úteis a partir de quantias maciças de dados. A conversão de dados maciços em informações e conhecimentos úteis envolve duas etapas: (1) padrões de mineração presentes nos dados e (2) interpretar esses padrões de dados em seus domínios problemáticos para transformá-los em informações e conhecimentos úteis, como por exemplo, reduzir riscos, e custos, melhorar o relacionamento com o cliente (YE, 2014).

O processo de descoberta de conhecimento contém uma série de passos, como: seleção, pré-processamento, transformação, mineração de dados e interpretação. A Figura 1, demonstra uma visão de como funciona e o relacionamento de ambas as fases:

Abaixo iremos exemplificar mais sobre cada etapa do ciclo do KDD para que ao fim tenha-se o conhecimento descoberto.

- Seleção: Nesta fase é o primeiro processo para que haja o descobrimento dos dados, assim é definido todas as possíveis características que irão fazer parte da análise. O

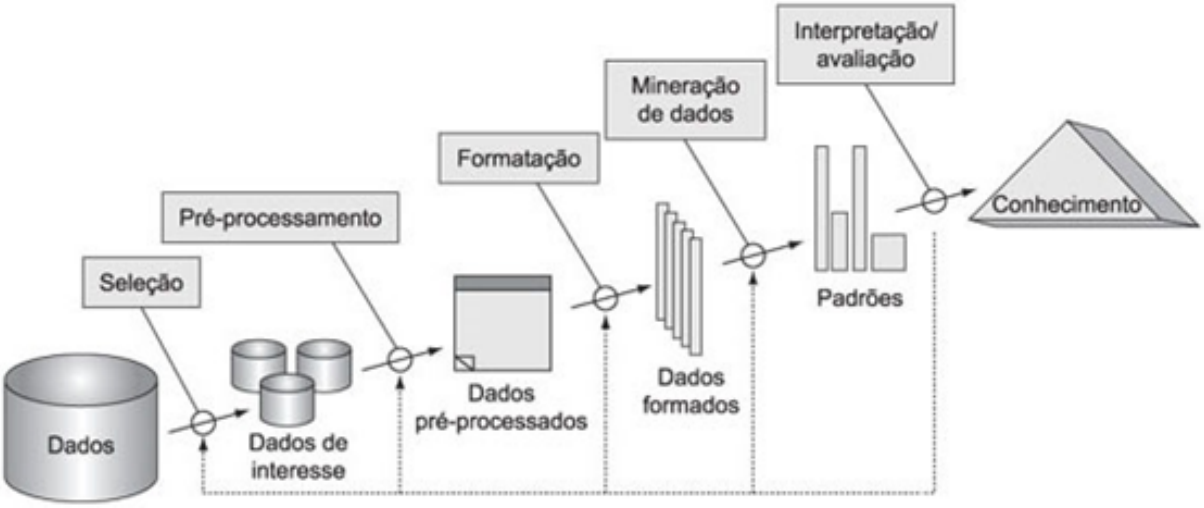


Figura 1 – Ciclo do processo de KDD

Fonte: Adaptado de (FAYYAD; PIATETKSY-SHAPIRO; SMYTH, 1996)

processo de seleção é complexo, visto que os dados podem ser obtidos de diversas fontes e seres dados estruturados ou não-estruturados, o que impactará no resultado.

- Pré-Processamento: Nesta fase inclui operações básicas, como remover ruídos ou valores se for possível, de forma a deixar os dados com qualidade para melhorar o desempenho.

- Transformação: Nesta fase inclui realizar processos para armazenar e transformar os dados para que os algoritmos sejam utilizados.

- Mineração de Dados: Nesta fase inclui procurar por padrões de interesse em uma representação. Segundo Fayyad, Piatetksy-Shapiro e Smyth (1996), "a mineração de dados envolve montagem de modelos ou determinação de padrões de dados observados".

- Interpretação: Nesta fase é realizado a interpretação e avaliação do conhecimento adquirido através da mineração de dados, assim como a visualização dos padrões extraídos.

Nesse contexto, Rezende (2005) define mineração de textos, ou também conhecida como por descoberta de conhecimento por textos (do inglês, *Knowledge Discovery from Text*), como "um conjunto de técnicas e processos que descobrem conhecimento inovador nos textos", assim como para Aranha e Passos (2006) "consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos".

A Figura 2 demonstra as etapas a serem realizadas para realizar a mineração de textos, desde a etapa dos documentos disponíveis até a validação dos resultados obtidos através das etapas realizadas.

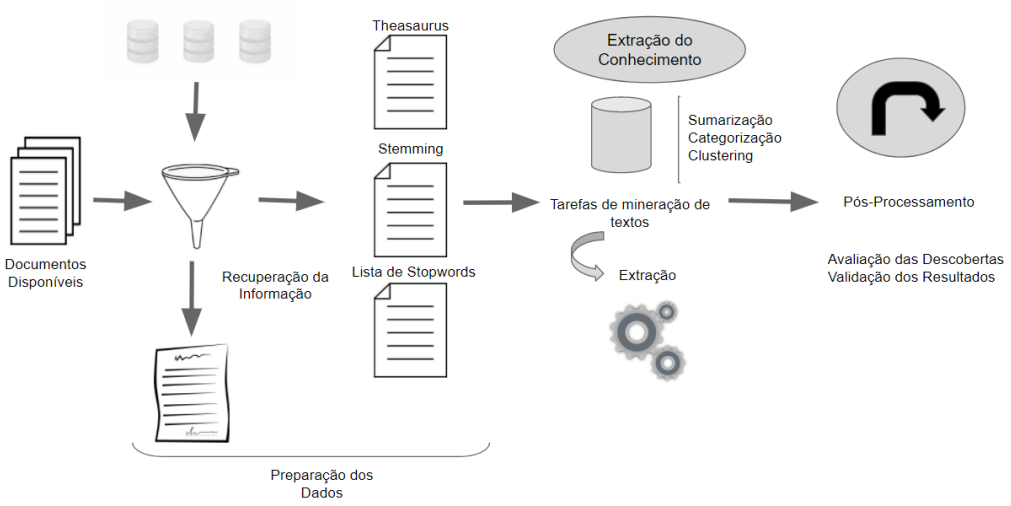


Figura 2 – Etapas da Mineração de Textos

Fonte: Adaptado de (REZENDE, 2005)

De acordo com Rezende (2005) há duas formas para analisar dados textuais, sendo que ambas podem ser utilizadas em conjunto ou separadamente. A primeira é a análise semântica que busca a funcionalidade dos termos nos textos, e a segunda a análise estatística que é baseada em frequência dos termos encontrados no texto.

**Análise Semântica**

A análise semântica aplica-se técnicas de linguística, visto que é fundamentada na área de processamento de linguagem natural, assim é utilizado as análises para buscar a importância das palavras nos diversos tipos de textos. Segundo Rezende (2005) é preciso utilizar algum modo de conhecimento para que a análise seja realizada, como:

1. **Conhecimento Morfológico:** Conhecimento da estrutura, da forma e das inflexões das palavras.
2. **Conhecimento Sintático:** Conhecimento estrutural das listas de palavras e como as palavras podem ser combinadas para produzir sentenças.
3. **Conhecimento Semântico:** O que as palavras significam independentes do contexto, e como significados mais complexos são formados pela combinação de palavras.
4. **Conhecimento Pragmático:** O conhecimento do uso da língua em diferentes contextos, e como o significado e a interpretação é afetada pelo contexto.

5. **Conhecimento do Discurso:** Como as sentenças imediatamente precedentes afetam a interpretação da próxima sentença.

6. **Conhecimento do Mundo:** Conhecimento geral do domínio ou o mundo que a comunicação da linguagem natural se relaciona.

#### Análise Estatística

Nesse tipo de análise estatística é verificado a quantidade das palavras repetidas no texto, tendo assim a definição da importância daquela palavra no contexto (REZENDE, 2005). Portanto, o conhecimento por meio desta análise é adquirido através dos passos de:

1. **Codificação dos dados:** A codificação dos dados se faz através de seleção de características informativas para que tenha algum critério em relação aos dados. Caso haja perca de informações relevantes nessa etapa, não há como recuperar depois.

2. **Estimativa dos dados:** Nessa etapa é utilizado um algoritmo de aprendizado ou um método para obter um modelo para os dados já codificados.

3. **Modelos de representação dos dados:** Nessa etapa é utilizado uma abordagem chamada Saco de Palavras (do inglês, *Bag of Words*), no qual é verificado a quantidade de vezes que a palavra é repetida dentro do texto sem se preocupar com a ordem das palavras e como ela está estruturada no texto.

## 2.2 Mídias Sociais e Redes Sociais

Cada vez mais as pessoas estão conectadas à Internet, por meio das mídias e redes sociais, para que possam ter acesso ao que está acontecendo no mundo. Mas em redes sociais ou mídias sociais? Há uma confusão referente as definições dos termos, de forma a acreditar que ambas são a mesma coisa, mas segundo Ciribeli e Paiva (2011) "a mídia social que dá suporte às redes sociais na internet".

Segundo Social (2014) "as mídias sociais têm várias características que as diferem fundamentalmente das mídias tradicionais, como jornais, televisão, livros ou rádio. A começar pela falta de espaço. Elas não são finitas: não há um número determinado de páginas ou horas específicas destinadas à produção de conteúdo". Os conteúdos gerados podem ser tal como imagens, textos e gráficos, pois trata-se de um universo repleto de novos conteúdos instantaneamente. A Figura 3 demonstra o que é uma mídia social e seu comportamento.

Figura 3 – Exemplo de interação em uma mídia social

Fonte: Autor

Existem diversos tipos de mídias sociais, para diferentes públicos e assuntos, seja com foco em profissionais, amizades, relacionamentos, pesquisas, entre os mais diversos assuntos com postagem de mensagens instantâneas e textos, compartilhamento de vídeos, áudios e imagens (CIRIBELI; PAIVA, 2011).

As redes sociais são um espaço de interação entre pessoas, grupos ou empresas na qual se relacionam diretamente com um ou mais indivíduos que buscam um relacionamento por interesses em comum e/ou exposição de opiniões. Elas são acessíveis a qual lugar, desde que tenha uma conexão válida com a Internet. A Figura 4 demonstra o que é uma rede social, tal como seus relacionamentos.

Figura 4 – Exemplo de interação em uma rede social

Fonte: Autor

Tendo uma base do que é uma mídia/rede social em 2006 foi lançado por Jack

Dorsey, o Twitter, com o intuito de ser um serviço de mensagens curtas para internet. O Twitter é uma rede social que permite a disseminação de suas ideias, opiniões e entre os diversos posicionamentos sobre os mais variados tipos de assuntos em tempo real, desde que cada mensagem - *tweet* - tenha no máximo 140 caracteres, o que significa que essas postagens são realmente fáceis de escrever e ler, com a finalidade de responder a seguinte pergunta "O que está acontecendo?"(REAGAN, 2010). No ano de 2016, segundo Twitter (2017a), o Twitter teve em torno de 313 milhões de usuários ativos mensalmente e em torno de 82% de usuários móveis ativos, o que nos possibilita um ambiente de pesquisa amplo para obter uma massa de dados em instantes, além da possibilidade de utilizar a geolocalização dos dispositivos móveis desde que o usuário tenha habilitado.

E porque as pessoas usam o Twitter? Segundo Reagan (2010) é porque a rede social permite conhecer novas pessoas, se manter informado com os assuntos mais populares, acompanhar e interagir com as celebridades, pode promover seu negócio e entre as diversas vantagens que a rede social pode proporcionar. A Figura 5 demonstra um exemplo de como é um *tweet*.


Figura 5 – Exemplo de *Tweet*

Fonte: (O'REILLY; MILSTEIN, 2011)

O Twitter funciona em grande parte porque ele se encaixa na própria rotina (RUSSEL, 2011), pois todas as atualizações são realizadas no perfil do usuário em tempo real e podem serem postadas através do *smartphone* e/ou do computador. É o lugar perfeito para transmitir mensagens rápidas, curtas e que podem ter um grande alcance.

## 2.3 Índices Urbano

### 2.3.1 Índice de Desenvolvimento Humano Municipal

O índice de desenvolvimento humano (IDH) foi difundido no início da década de 1990, pela Organização das Nações Unidas (ONU), com o propósito de ser uma medida que verifica o desenvolvimento de um país com base no relacionamento de três pontos de desenvolvimento humano - renda, educação e saúde - e também com o propósito de ser um índice que iria oferecer um contraponto ao Produto Interno Bruto (PIB), que apenas considera o aspecto econômico (BRASIL, 2017). A ONU tornou-se capaz de demonstrar através do IDH para os governantes de países em desenvolvimento de que o crescimento não está exclusivamente vinculado ao aumento do PIB. O IDH é referência para verificar o desenvolvimento humano de um determinado país ou região, sendo publicado anualmente possibilitando análises para verificar se as pessoas estão vivendo bem e saudável e seu índice varia de 0 (valor mínimo) até 1 (valor máximo).

O governo federal do Brasil é um dos pioneiros a realizar o índice de desenvolvimento municipal (IDHM), que é um ajuste metodológico ao IDH. Ambos não podem ser comparados pois embora utilizem os mesmo indicadores para realizar os seus índices, o contexto é diferente visto que o IDH visa medir o desenvolvimento dos países e o IDHM o desenvolvimento dos municípios (BRASIL, 2013). Realizar análises municipais é importante para o desenvolvimento do município, visualizando aspectos como longevidade, educação e renda, sendo assim possível contribuir para o progresso humano e de condições de vida, segundo (BRASIL, 2013) o IDHM é muito importante pois acaba popularizando o desenvolvimento centrado em pessoas, possibilita a comparação entre os municípios brasileiros a curto e longo prazo e estimula os governantes de nível municipal a implantar novas práticas de melhoria de vida.

O prazo para a divulgação de novos resultados para o IDHM depende de novos dados coletados do censo através do IBGE (Instituto Brasileiro de Geografia e Estatística) que acontece aproximadamente a cada 10 anos (ESTATÍSTICA, 2013). Os censos servem para que possa conhecer melhor o país, estados e municípios, e por meio dele é possível medir a densidade populacional e o perfil da população. O cálculo do índice é realizado através dos dados coletados dos últimos Censos Demográficos (1991, 2000 e 2010), promovendo a estagnação da informação durante o período em que não é realizado nenhum censo e não permitindo acompanhar o desenvolvimento periodicamente.

A fórmula realizada pelo IBGE para calcular o IDHM é realizada conforme a Figura 6 demonstra. O cálculo é feito por meio de uma média geométrica com 3 variáveis (longevidade, educação e renda) na qual são agrupados e dão resultado final, o IDHM.

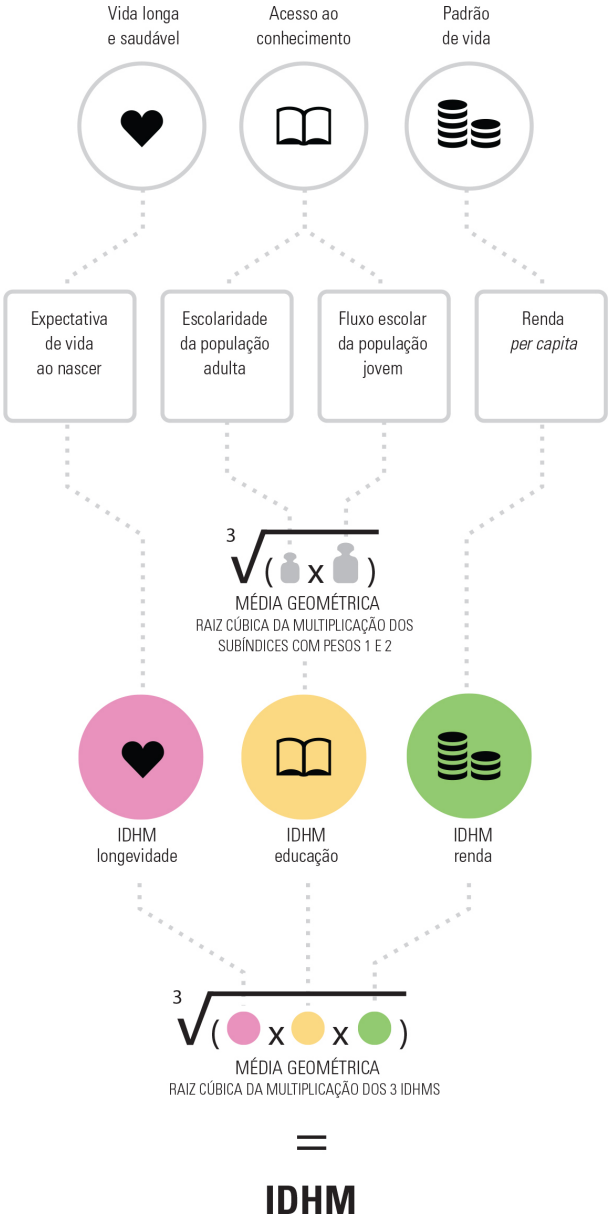


Figura 6 – Fórmula do cálculo do IDHM

Fonte: (BRASIL, 2013)

### 2.3.2 Índice de Bem-Estar Urbano

O Índice de Bem-Estar Urbano (IBEU-Municipal) é um índice desenvolvido segundo Metrópoles (2016), "com o propósito de oferecer a atores governamentais, universidades, movimentos sociais e sociedade civil em geral o mais novo instrumento para avaliação e formulação de políticas urbanas para o país brasileiro". Foi proposto pelo Instituto Nacional de Ciência e Tecnologia (INCT) em Observatório das Metrópoles com apoio da CNPq. Ainda segundo Metrópoles (2016), a metodologia para o cálculo do índice é feito por meio de informações do Censo Demográfico de 2010, no qual totaliza 5.565 municípios. E mesmo assim ainda possível refletir as condições conforme demonstra as análises existentes da Pesquisa Nacional por Amostra de Domicílios (PNAD) e também

pelas análises do Instituto Brasileiro de Geografia e Estatística (IBGE).

O índice apresenta as condições urbanas de cada município por meio de análises de dimensões como mobilidade, condições ambientais urbanas, condições habitacionais, atendimentos de serviços coletivos e infra-estrutura. A Figura 7 demonstra o mapa do Brasil apresentando os IBEU de cada município brasileiro.

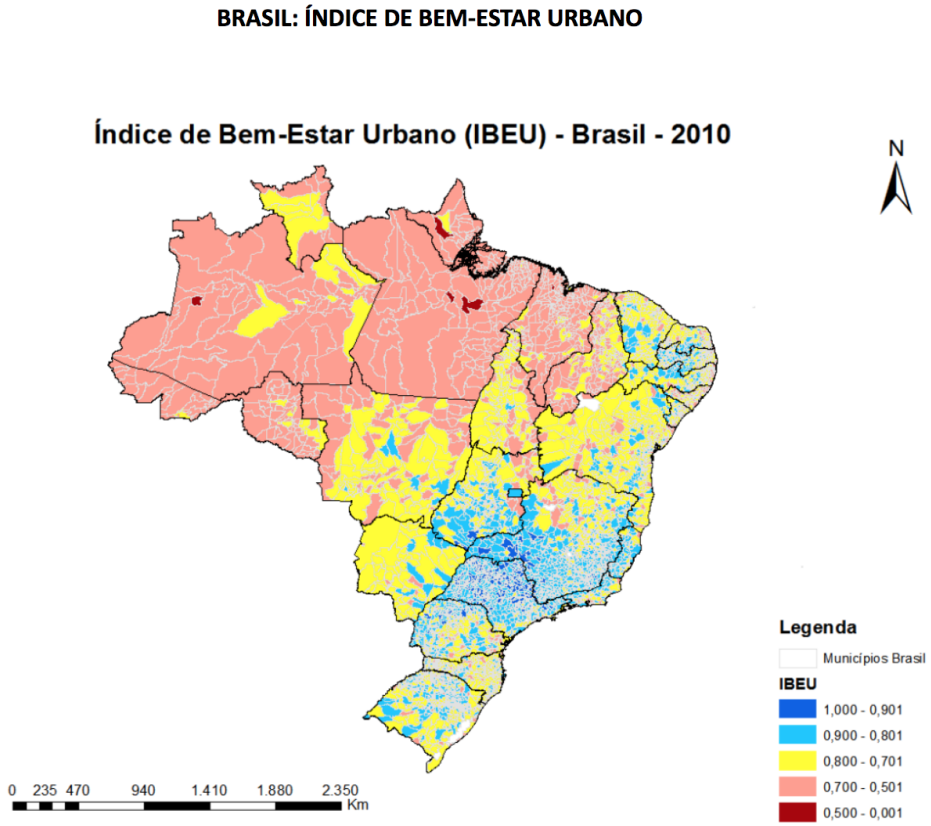


Figura 7 – Mapa do Brasil demonstrando o IBEU-Municipal

Fonte: (METRÓPOLES, 2016)

Para atingir o objetivo proposto, o IBEU foi concebido em dois tipos:

- **IBEU Global:** é calculado para o conjunto de metrópoles do país, o que permite comparar as condições de vida urbana em três escalas: entre as metrópoles, os municípios metropolitanos e entre bairros que integram o conjunto das metrópoles.
- **IBEU Local:** é calculado especificamente para cada metrópole, permitindo avaliar as condições de vida urbana interna a cada uma delas.

Sendo assim, é possível visualizar os desafios enfrentados, no qual os maiores são infraestrutura (pavimentação, calçamento, iluminação pública, etc) e os serviços coletivos

(atendimento adequado de água, esgoto, energia e lixo). Para isso existem algumas análises de dimensões específicas, como:

- **Dimensão de Infraestrutura Urbana:** pode ser compreendido por sete indicadores de análise: iluminação pública, pavimentação, calçada, meio-fio/guia, bueiro ou boca de lobo, rampa para cadeirantes e logradouros. Esses indicadores expressam as condições de infraestrutura na cidade que podem possibilitar melhor qualidade de vida para pessoas, estando relacionados com a acessibilidade e saúde.
- **Dimensão de Serviços Coletivos Urbanos**: Para analisar os serviços públicos essenciais para garantia de bem-estar urbano, o IBEU concebeu quatro indicadores: atendimento adequado de água, atendimento adequado de esgoto, atendimento adequado de energia e coleta adequada de lixo.
- **Dimensão de Condições Habitacionais Urbanas:** As condições habitacionais também constituem uma importante dimensão que influencia o bem-estar das pessoas na cidade. Tal dimensão pode ser apreendida pela situação de adensamento (entendida pela razão número de pessoas no domicílio e número de dormitórios), pelas condições materiais da estrutura habitacional, assim como aglomeração dos domicílios.
- **Dimensão de Condições Ambientais Urbanas:** Para analisar essa dimensão o IBEU concebeu três indicadores: arborização do entorno dos domicílios, esgoto a céu aberto no entorno dos domicílios e lixo acumulado no entorno dos domicílios.
- **Dimensão de Mobilidade Urbana:** avalia o deslocamento casa-trabalho.

# 3 TRABALHOS RELACIONADOS

Com a grande popularização das redes sociais, as pessoas interagem continuamente contando como foi seu dia ou opinando sobre diversos temas, tais como: produtos, política, futebol. A grande quantidade de dados é exponencialmente interessante para obter novos estudos, e um deles é tentar classificar a emoção do texto descrito, de forma a diagnosticar um sentimento - positivo, negativo ou neutro. Nesse contexto, surgiu a área de estudo chamada análise de sentimento, a área se refere à um problema de extração de opinião de uma pessoa referente à algum assunto em contexto (NARAYANAN; LIU; CHOUDHARY, 2009; BAUMGARTEN et al., 2013). Os estudos mais básicos referente a área tentam classificar uma polaridade para cada texto, de forma a rotular como sentimento positivo, neutro e negativo, outros sentimentos também podem ser identificados através da análise de sentimentos, como por exemplo, felicidade, tristeza, raiva e alegria.

A literatura sobre a área de análise de sentimentos apresenta diversos trabalhos que buscam os métodos para a classificação do sentimento como, Li e Li (2011), Zhang et al. (2011), Araujo et al. (2012), Hu et al. (2013) na qual utilizaram métodos de análise de sentimento com base em *emoticons*. Esse método não é preciso utilizar algum idioma específico, pois expressa a mensagem de um sentimento na qual possui o mesmo significado para qualquer idioma. Porém ao ser utilizado nos estudos relacionados à análise de sentimento, o mesmo não possui uma grande abrangência devido ao fato de que é necessário que a sentença obtenha *emoticon*, caso contrário será desconsiderado.

Outras pesquisas já foram realizadas para extrair opiniões e sentimentos através da rede social Twitter. No trabalho de Pak e Paroubek (2010) desenvolveram um trabalho para categorizar *tweets* em positivo e negativo de forma a utilizar o classificador *Naïve-Bayes* na classificação gramatical dos textos, no caso os *tweets* com o auxílio de N-gramas. Assim como em Nascimento et al. (2012) desenvolveram um trabalho de análise de sentimentos com busca em notícias através das *hashtags* e depois agrupadas por áreas - entretenimento, policial e política -, utilizando classificadores de unigrama, octograma e *Naïve-Bayes* para realizar a acurácia e demonstrar a comparação entre os classificadores com base nos sentimentos expressados nos *tweets*.

Em França e Oliveira (2014) realizaram um estudo com análise de sentimentos relacionado aos protestos no Brasil no ano de 2013, utilizando as *hashtags* relacionadas ao movimento de protesto para montar a base de dados e também o modelo estatístico de aprendizagem *Naïve-Bayes* para que os dados fossem treinados e apresentassem uma probabilidade alta de classificação de textos. Com uma abordagem também política, Rios et al. (2017) realizaram um estudo motivados pela questões políticas enfrentadas pelo Brasil, assim utilizando a rede social Twitter para monitorar os usuários que utilizaram a

*hashtag* impeachment (#impeachment) e o perfil da presidenta do Brasil, Dilma Rousseff (@dilmabr), eles desenvolveram um escutador (do inglês, *listener*) que guarda em uma base de dados todos os dados do usuário que utilizam a *hashtag* por meio de um robô chamado TSViz[1] (do inglês, *Time Series Visualization*). O projeto é composto por vários fatores de séries temporais como: análise de frequência, distância e soma da compressão normalizada, análise de sentimentos e detecção de *drift* de conceito usando a análise de quantificação de recorrência cruzando, e assim com o desenvolvimento de um monitor, monitora em tempo real a reação da população à medida que novas notícias de corrupção são divulgadas (RIOS et al., 2017).

Há pesquisas mais próximas do trabalho que estamos desenvolvendo que tentam classificar o bem-estar de um país, estado e município por meio de coordenadas geográficas e tendo como referencial a utilização das redes sociais em especial o Twitter.

O trabalho de Dodds et al. (2011), objetiva trabalhar com a utilização de Big Data em redes social para a medir e compreender a felicidade das populações em tempo real. Por meio de comparações entre textos e ao nível das palavras descritas, foi possível realizar uma média das palavras, utilizando o projeto *Hedonometer.org*[2]. A base de dados foi coletada através da rede social Twitter, na qual tem um grande potencial para descrever padrões humanos, por exemplo, emocionais e sociais, sendo possível aplicar estudos diversos utilizando-a (DODDS et al., 2011).

Também fazendo uso do projeto *Hedonometer.org*, o estudo de Mitchell et al. (2013) foi desenvolvido para combinar dados de pesquisas tradicionais realizadas pelos EUA e então correlacionar com dados retirados de rede social, de forma a utilizar o uso das palavras com as características do ambiente em que ele estava. Com isso foi possível realizar um estudo para dizer quais as cidades e estados mais felizes e as menos felizes do país americano, e também as palavras mais positivas e negativas ditas na rede social de cada estado e cidade para entender como as palavras influenciam na felicidade em correlação com os dados do censo socioeconômico.

Grande parte dos trabalhos na área de análise de sentimentos está relacionado a algum evento e a ocorrência que o mesmo apresenta nas redes sociais ou realizar estudos para identificar os melhores algoritmos na literatura. Também é possível verificar que existem trabalhos que possuem a classificação do bem-estar da um país, estado o município na qual está muito próximo da proposta de trabalho que estamos desenvolvendo. É possível vislumbrar que o presente trabalho, utiliza textos curtos escritos na língua portuguesa.

[1] http://www.tsviz.com.br/
[2] http://www.hedonometer.org/index.html

# 4 PROPOSTA

A exploração dos sentimentos expressados nos textos curtos publicados em redes sociais permitem extrair informações relevantes sobre o diversos temas, tal como o bem-estar da população em residir em determinadas cidades. Essa abordagem é investigada por outros trabalhos disponíveis na literatura (DODDS; DANFORTH, 2010; DODDS et al., 2011; MITCHELL et al., 2013), mas é importante ressaltar que estes trabalhos se direcionam a cidades não brasileiras.

Portanto, este trabalho investiga um índice para mensurar o bem-estar da população em cidades brasileiras através do sentimento expressado na rede social. Tal proposta visa contribuir para reduzir o período de obsolência da informação sobre o bem-estar das cidades. É importante que os governantes municipais tenham em mãos informações atuais, visto que o resultado do IDHM é divulgado a cada 10 anos por conta do Censo Demográfico e o IBEU-Municipal que também utiliza dados do Censo para realizar seus cálculos e gerar o índice de bem-estar.

Espera-se que o índice apresente um comportamento representativo semelhante ao IBEU-Municipal das respectivas cidades em análises, o que permite ser utilizado como um valor de ordenação. Para simplificar o entendimento e as informações expressadas neste trabalho, o índice investigado é denominado PIRES.

Para que este índice possa ser calculado, é empregado um sistema composto por três componentes. Assim, o sistema deve coletar, pré-processar e aplicar o índice PIRES nos textos publicados pela população em suas respectivas contas na rede social. Os componentes que compõem tal sistema são: (i) coleta de dados, é a realização da coleta de dados de acordo com cada cidade monitorada; (ii) pré-processamento, é a preparação dos dados para serem analisados, realizando a retirada de dados não relevantes para esse trabalho, diminuindo ruídos e melhorando a qualidade dos dados; (iii) aplicação do índice PIRES, é nesse componente que realiza cálculos matemáticos para obter valores proporcionais para cada cidade. A Figura 8 demonstra a arquitetura e a interação dos componentes do sistema utilizado.

É almejado que tal índice possibilite realizar uma análise comparativa entre duas ou mais cidades a fim de definir uma ordem entre as respectivas cidades com base no bem-estar populacional identificado. A distância entre os valores obtidos por cada cidade pode ser empregada como informação referente ao quanto uma cidade é melhor classificada do que outra. Porém, tal valor deve ser utilizado com critério proporcionalista e não como valor absoluto. Em outras palavras, o índice permite uma interpretação como: a cidade A possui uma qualidade de vida duas vezes melhores em relação a cidade B, enquanto a

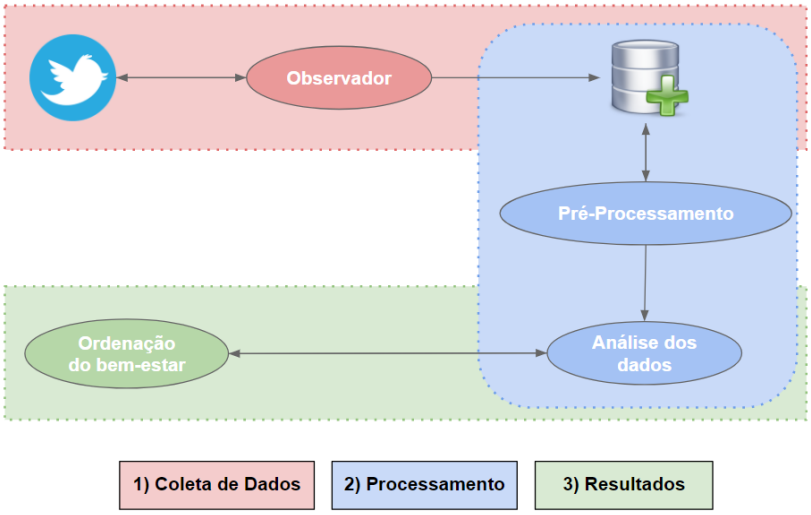


Figura 8 – Arquitetura do sistema empregado para o cálculo do índice PIRES.

Fonte: Autor

cidade C é superior a B e inferior a A. Essa interpretação poderia ser obtida através de um resultado exemplificado na Tabela 1.

Tabela 1 – Exemplo de uma possível ordenação sobre o bem-estar de três cidades realizada com base no índice PIRES e IBEU.

| **Cidade** | **PIRES** | **IBEU** |
|---|---|---|
| (A) Maringá | 0,46684 | 0,90642 |
| (B) Londrina | 0,34234 | 0,89017 |
| (C) Bandeirantes | 0,23342 | 0,51233 |

A Tabela 1 demonstra a ordenação em relação ao bem-estar das cidades exemplificadas, levando em consideração o índice IBEU-Municipal em correlação ao índice PIRES. Verifica-se que a cidade com maior índice bem-estar pode também apresentar maior índice de bem-estar por meio da rede social, considerando a quantidade de iterações positivas dentro de um intervalo temporal de coleta demonstrado pelo cálculo do índice proposto.

# 5 DESENVOLVIMENTO

Para o desenvolvimento do trabalho é necessário dividir em duas partes: a primeira como foi feito a aquisição dos dados pela geolocalização, os pré-processamentos realizados e o cálculo do sentimento dos textos da base de dados adquirida pelo serviços cognitivos da Microsoft; a segunda parte está relacionada a seleção dos dados relevantes e os cálculos realizados para realizar a validação do estudo. Esta seção irá discutir o desenvolvimento do trabalho, descrevendo cada etapa que foi realizado.

## 5.1 Aquisição

A primeira etapa é realizar a aquisição dos dados, assim foi utilizado as opções de desenvolvedor do Twitter para obter autorização de acesso a API[1], a qual possibilita acesso para leitura e escrita de dados. O Twitter disponibiliza dois modos de abordagens de API, a *REST API* e a *Streaming API* (TWITTER, 2017b).

- *REST API*[2]: Fornece acesso programático a dados, com limitação de taxa de janela, ou seja, só é permitido 15 solicitações por janela por *token* de acesso, ou seja, para cada requisição ele cria uma nova requisição HTTP e realiza a coleta de dados de forma retrógada.

- *Streaming API*[3]: Fornece acesso a dados com baixa latência e de forma com que não precise se preocupar com limitação de taxa de janela, pois ela utiliza de uma conexão HTTP persistente aberta e é mais indicada para monitoramento em tempo real.

A aquisição foi realizada por meio da API de *streaming*, pois a mesma permite vantagens em relação a *REST API* e em conjunto com a linguagem *Python*, utilizando filtro de geolocalização por cidade, ou seja, foi filtrado todos os *tweets* que estavam dentro da latitude e da longitude da cidade escolhida. A Figura 9 demonstra mais detalhadamente como é feito o processo de aquisição de *tweets* com base na primeira etapa da Figura 8.

A API realiza conexão com o Twitter e retorna os dados através da delimitação do filtro de geolocalização da cidade, e após esse processo é salvo em uma base de dados apenas guardando informações como o nome do usuário, o texto (*tweet*), a cidade escolhida e o horário realizado sendo particular de cada *tweet*.

[1] www.dev.twitter.com
[2] https://dev.twitter.com/rest/public
[3] https://dev.twitter.com/streaming/overview

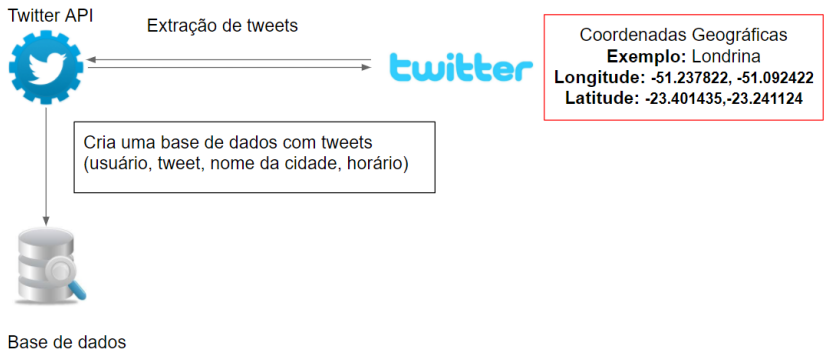


Figura 9 – Processo de aquisição de *tweets*

## 5.2 Pré-Processamento

Após obter a base de dados através da coleta por geolocalização, o segundo passo é realizar o pré-processamento conforme a segunda etapa da Figura 8 e portanto para essa fase foi utilizado a linguagem R[4], uma linguagem muito utilizada para realizar cálculos estatísticos e análise de dados. O processo de pré-processamento é a segunda etapa do ciclo de descoberta de conhecimento a partir dos dados, e consiste em realizar processos para diminuir os ruídos de dados ou valores se for possível, de forma a melhorar a qualidade da base de dados e preparar para que as etapas posteriores possam usufruir de dados mais relevantes (FAYYAD; PIATETKSY-SHAPIRO; SMYTH, 1996).

Nesse estudo os processos de pré-processamento consiste em remover *tweets* que sejam relacionados à *check-in* - é uma expressão muito utilizada para o ato de dar entrada, confirmar presença em algum local, - postagens de fotos e vídeos e dados meteorológicos, e realizar a remoção de URLs - endereços de páginas *web* - vinculados juntamente com a mensagem, por não apresentar nenhum sentimento para o contexto desse estudo.

## 5.3 Cálculo do sentimento

Para esse trabalho, foi utilizado uma API de Análise de Textos, denominada *Microsoft Cognitive Services* a qual possui uma limitação de 1000 chamadas por transação[5] e uma outra limitação de 5000 transações por mês. Ela consegue detectar sentimentos, palavras-chave, tópicos e até o idioma do texto. A língua portuguesa até o momento só tem a versão disponível na API para detectar o sentimento, as demais - palavras-chave e tópicos - ainda não possuem versões disponíveis para serem utilizadas no idioma.

A utilização dessa API consiste em utilizar um arquivo de formatação no formato JSON (do inglês, *JavaScript Object Notation*) através do método POST via protocolo

[4] https://www.r-project.org/
[5] https://azure.microsoft.com/pt-br/pricing/details/cognitive-services/text-analytics/

HTTP. Para que haja o cálculo do sentimento, o arquivo JSON precisa ser formatado conforme a documentação da API, então é preciso anexar informações com um identificador único para cada documento, a sigla da linguagem do texto e por fim o texto que será detectado o sentimento. A Figura 10 demonstra um exemplo da formatação do arquivo JSON.

```
"{"documents":
  {["language":"pt",
    "id":"1",
    "text":"Pra quem tem fé a vida nunca tem um fim."]},
  {["language":"pt",
    "id":"2",
    "text":"Fui jantar no Outback, estava maravilhoso."]}
    }"
```

Figura 10 – Modelo de formatação do arquivo JSON para a API de Análise de Textos

Assim que o arquivo estiver formatado conforme a Figura 10, é enviado para a API pelo método de requisição POST, no qual foi projetado para enviar dados anexados na mensagem para os servidores *web* (FIELDING; RESCHKE, 2014). É um dos métodos mais utilizados pelo protocolo HTTP, ele funciona encapsulando as informações junto ao corpo da mensagem HTTP, não possui limitação de tamanho de comprimento da mensagem e também não possui restrições podendo enviar textos quanto dados binários.

A detecção do sentimento das frases é gerado usando técnicas de processamento de linguagem natural avançadas, não precisa de treinamento de dados somente o texto. A pontuação do sentimento varia entre 0 e 1, sendo que quanto mais próximo de 0 é classificado como sentimento negativo e quanto mais próximo de 1 é classificado como sentimento positivo (MICROSOFT, 2017).

| V1 | V2 |
|---|---|
| 1.00000000 | 1 |
| 0.56333240 | 2 |
| 0.50000000 | 3 |
| 0.58590715 | 4 |
| 0.66839750 | 5 |

Figura 11 – Exemplo de saída da API após o cálculo do sentimento

A Figura 11 demonstra o exemplo de saída após realizar a chamada a API para calcular o sentimento, na coluna da esquerda - V1 - são valores calculados pela API enquanto na coluna da direita - V2 - são valores de identificação de cada sentença já definido no arquivo de formatação JSON.

## 5.4 Média do sentimento dos *tweets* dos usuários

Essa etapa é necessária para a validação do estudo visto que pode ocorrer de vários usuários com centenas de *tweets* enquanto outros apenas com algumas dezenas. Para evitar que um usuário muito ativo influencie no resultado da pesquisa, visto que foi possível identificar. A Figura 12 demonstra um exemplo do agrupamento de usuários que foi realizado na cidade de Londrina - PR, podendo visualizar que alguns são mais ativos que outros.

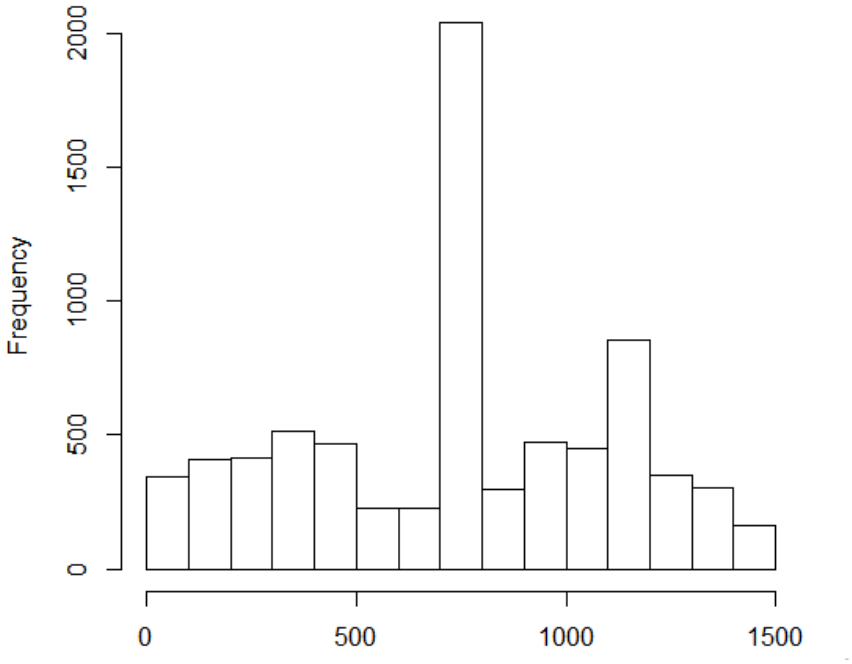


Figura 12 – Histograma de agrupamento por usuário para verificar a quantidade de *tweets* publicados na cidade de Londrina - PR

A Figura 12 apresenta um histograma em relação à base de dados já realizada os processos acima descritos, a base consta com aproximadamente 1500 usuários com milhares de *tweets* realizados. Pode-se perceber que há usuários muito ativos e assim o que faria com que os resultados não fossem tão satisfatórios, sendo assim foi necessário realizar um agrupamento de *tweets* por usuários para que cada usuário tenha apenas um valor médio de seus textos publicados.

## 5.5 Cálculo do índice

Para criar o índice PIRES, foi-se necessário realizar uma seleção dos dados na qual esse processo está relacionado com o processo feito na subseção 4.4, na qual obteve-se um agrupamento dos valores de sentimentos por usuário e realizou uma média do mesmo. Assim para que possamos preparar os dados para aplicar na última etapa foi necessário

realizar uma seleção nos dados de sentimentos. Os sentimentos estão classificados entre 0 e até 1, de forma a que quanto mais próximo de 0 é um sentimento negativo e quanto mais próximo de 1 sentimento positivo. Portanto, para esse estudo realizou uma seleção dos dados de forma a utilizar somente dados de sentimentos positivos para cada cidade, a faixa de valores foram iguais e acima de 0,6.

Para calcular o índice representativo, utilizou-se da média geométrica, na qual é utilizada pelo Índice de Desenvolvimento Humano Municipal (IDHM) para realizar seus cálculos, conforme apresentado no Capítulo 2 por meio da Figura 6. A fórmula do IDHM consiste em realizar a raiz cúbica das três variáveis - educação, longevidade e renda - que são consideradas para o índice (BRASIL, 2013).

Para o cálculo do índice representativo PIRES, a fórmula proposta para a realização da média geométrica é da seguinte maneira:

$$\exp[\frac{1}{n}\sum_{i=1}^{n}\log(x_i)] \tag{5.1}$$

Onde:

- n = número de termos do vetor
- $x_i$ = ao elemento do vetor de sentimentos positivos na posição i

Após realizado o cálculo do índice PIRES, é necessário fazer a ordenação das cidades que apresente o maior valor correspondendo a terceira etapa da Figura 8 que é a etapa dos Resultados.

# 6 RESULTADOS

Iniciamos essa parte discutindo sobre a primeira fase de avaliação, onde é demonstrado a comparação entre os índices de IDHM e do IBEU das cidades de Maringá-PR e Londrina-PR, que a principio foram as primeiras cidades a terem seus *tweets* coletados para a aplicação da viabilidade da proposta. A Tabela 2 demonstra os dados obtidos do IDHM e do IBEU através do Censo Demográfico realizado pelo IBGE no ano de 2010. Pode-se visualizar comparativamente que a cidade de Maringá apresenta um IDHM e um IBEU maior do que a cidade de Londrina, por meio dos dois índices apresentados.

Tabela 2 – Dados relativos das cidades de Maringá-PR e Londrina-PR

| **Cidade** | **IDHM** | **IBEU** |
|---|---|---|
| Maringá | 0,808 | 0,924 |
| Londrina | 0,778 | 0,903 |

A Tabela 3 demonstra as informações referentes a base de dados das cidades de Maringá e Londrina, que foram coletadas para a primeira fase de avaliação do estudo proposto. As bases foram coletadas através de *tweets* geo-localizados, ou seja, as informações só iam para a base se o *tweet* tivesse a sua coordenada geográfica disponível, o que levou a um baixo número de coletadas em relação ao período de coleta.

Tabela 3 – Dados coletados das cidades de Maringá-PR e Londrina-PR na primeira fase de avaliação

| **Cidade** | **Início** | **Término** | **Total** |
|---|---|---|---|
| Maringá | 24/04/2017 | 18/07/2017 | 16,820 |
| Londrina | 24/04/2017 | 18/07/2017 | 12,023 |

Nesse cenário, aplicando a proposta do trabalho com o intuito de investigar que a cidade de Maringá apresente melhor comportamento que a de Londrina, a Tabela 4 demonstra os dados feitos através do índice PIRES, utilizando os dados coletados para as cidades de Maringá-PR e Londrina-PR.

Tabela 4 – Dados relativos ao índice representativo PIRES nas cidades de Maringá-PR e Londrina-PR

| **Cidade** | **PIRES** |
|---|---|
| Maringá | 0,774 |
| Londrina | 0,768 |

A análise possibilitou mostrar com que os dados do índice PIRES, demonstrasse que a cidade de Maringá apresenta um melhor bem-estar por meio das redes sociais durante o mesmo período de coleta que a cidade de Londrina. A Figura 13 demonstra a

aplicação do índice PIRES realizada nos dados processados para as cidades de Londrina e Maringá, pode-se dizer que Maringá apresenta uma maior bem-estar nas redes sociais do que Londrina, analogicamente a mesma comparação acontece no IDHM e no IBEU.

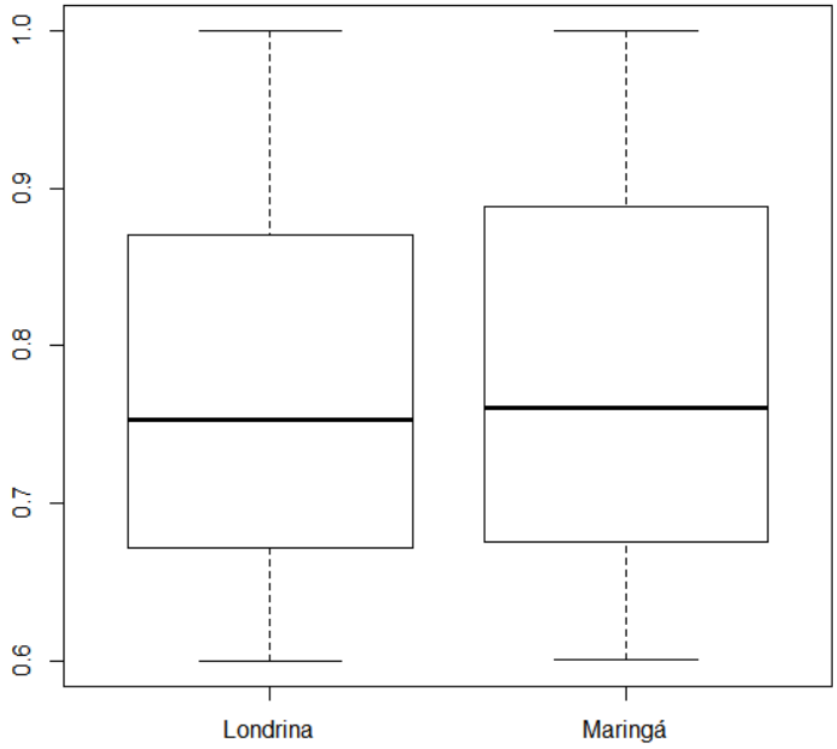


Figura 13 – Representação estatística do índice representativo PIRES nas cidades de Londrina-PR e Maringá-PR

Os resultados obtidos demonstrou a priori, resultados satisfatórios em relação ao cenário descrito, pois conseguiu demonstrar que por meio da rede social o índice PIRES apresentasse o mesmo comportamento ao IDHM e do IBEU. Sem considerar as diversas variáveis existentes que poderiam apresentar influência no resultado, é cabível de ressalva que o índice representativo apresentou um possível avanço para obter informações atualizadas e não estagnadas através dos resultados produzidos pelo Censo.

Obtendo resultados satisfatórios, realizou-se então uma segunda fase de avaliação dos resultados, sendo aplicado a comparação ao índice mais adequado, o IBEU, por se tratar de um índice mais relacionado ao bem-estar das cidades do que o IDHM que leva em consideração a educação, a qualidade de vida e a renda *per capita*. Também foi aumentado o número de cidades a serem monitoradas, além das duas cidades já citadas anteriormente (Londrina e Maringá). Sendo assim, a Tabela 5 demonstra as informações relativas das bases de dados coletadas das cidades paranaenses que foram consideradas para a segunda fase de avaliação do estudo.

Para a segunda fase, é preciso salientar que teve uma mudança de paradigma

Tabela 5 – Dados coletados das cidades paranaense na segunda fase de avaliação

| **Cidade** | **Início** | **Término** | **Total** |
|---|---|---|---|
| Apucarana | 30/09/2017 | 13/11/2017 | 7,766 |
| Arapongas | 30/09/2017 | 13/11/2017 | 8,034 |
| Cascavel | 30/09/2017 | 13/11/2017 | 34,169 |
| Colombo | 30/09/2017 | 13/11/2017 | 15,031 |
| Curitiba | 30/09/2017 | 13/11/2017 | 484,102 |
| Foz do Iguaçu | 30/09/2017 | 13/11/2017 | 45,274 |
| Guarapuava | 30/09/2017 | 13/11/2017 | 22,908 |
| Londrina | 30/09/2017 | 13/11/2017 | 46,854 |
| Maringá | 30/09/2017 | 13/11/2017 | 68,378 |
| Ponta Grossa | 30/09/2017 | 13/11/2017 | 47,143 |
| São José dos Pinhais | 30/09/2017 | 13/11/2017 | 62,598 |
| Toledo | 30/09/2017 | 13/11/2017 | 14,396 |
| Umuarama | 30/09/2017 | 13/11/2017 | 13,676 |
| **TOTAL** | | | **870,339** |

na coleta dos dados. Antes, era feito por *tweets* geo-localizados dos usuários o que nos apresentou um baixo número de dados coletados, pois para esse tipo de paradigma é preciso que os usuários permitam que o Twitter use a localização do *smartphone*. Portanto, objetivou nesse momento que a coleta de dados seja feita por geolocalização da cidade e não mais por *tweets* geo-localizados, o que nos levou a obter uma base de dados para cada cidade bem maior em um curto espaço de tempo do que caso fosse por *tweets* geo-localizados.

Assim com os dados coletados, foi aplicado os cálculos do índice PIRES nas bases de dados das cidades paranaenses. Conforme demonstrado na Tabela 6, é possível visualizar com que a priori que os resultados não foram satisfatórios, por exemplo, a cidade com maior IBEU é a cidade de Maringá, que com a aplicação do índice PIRES, apresentou resultados abaixo do esperado.

Deste modo é apresentado na Figura 14, a demonstração por meio do gráfico de linhas o comportamento pela correlação do índice **PIRES** ao IBEU, para que haja uma melhor avaliação dos dados. É possível de afirmar que os resultados não comportaram conforme era esperado através a possibilidade levantada na proposta.

É necessário discutir que o prazo de coleta de dados da primeira fase foi bem maior do que para segunda fase de avaliação, sendo assim eventos externos acontecidos durante a coleta apresenta uma maior influência em relação a um curto período de espaço-tempo. Por exemplo, um time de futebol que tenha ganhado um campeonato pode apresentar um pico de influência muito alto na amostragem com um período pequeno de coleta do que para uma amostragem com um período grande de coleta. Sendo assim, eventos que fazem com que a população use mais a rede social tem uma grande influência em relação a média geral feita da amostragem sendo necessário esticar o prazo de coleta para que

Tabela 6 – Dados relativos das cidades paranaenses monitoradas no estudo de caso ordenadas pelo índice IBEU

| **Cidade** | **IBEU** | **PIRES** |
|---|---|---|
| Maringá | 0,924 | 0,735 |
| Umuarama | 0,919 | 0,730 |
| Londrina | 0,903 | 0,759 |
| Toledo | 0,901 | 0,731 |
| Arapongas | 0,892 | 0,759 |
| Curitiba | 0,874 | 0,738 |
| Cascavel | 0,871 | 0,757 |
| Apucarana | 0,860 | 0,729 |
| Foz do Iguaçu | 0,852 | 0,731 |
| São José dos Pinhais | 0,811 | 0,760 |
| Ponta Grossa | 0,810 | 0,744 |
| Guarapuava | 0,787 | 0,735 |
| Colombo | 0,740 | 0,758 |

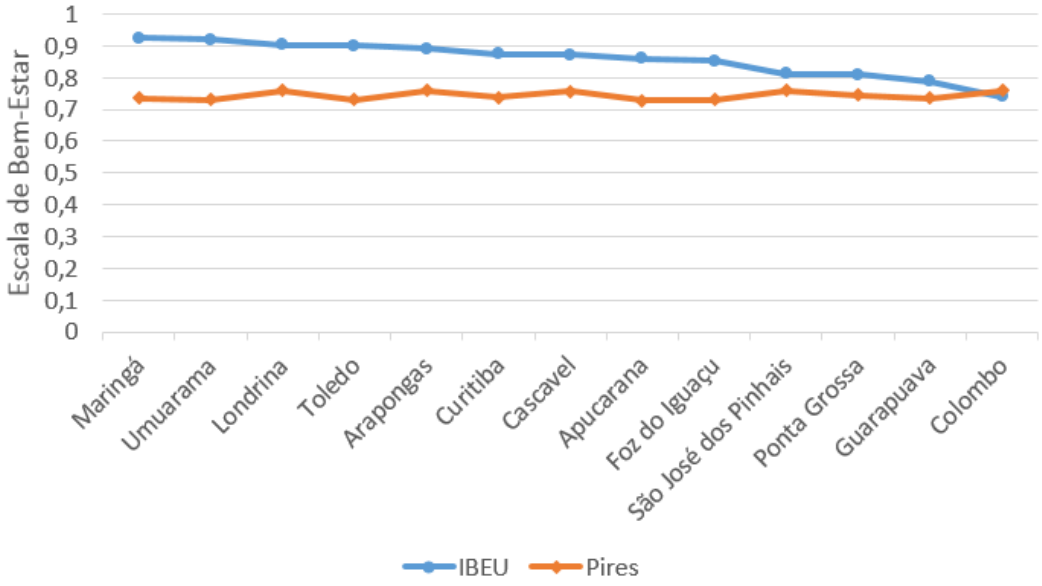


Figura 14 – Representação gráfica da correlação entre o índice representativo PIRES e o IBEU

esse evento acabe apresentando pouca influência.

# 7 CONCLUSÃO E TRABALHOS FUTUROS

O presente trabalho objetivou-se realizar um estudo investigativo para a criação de um índice representativo atráves de textos curtos publicados na rede social Twitter, para buscar verificar o bem-estar dos usuários. Desta maneira, correlacionando os dados do índice proposto com o Índice de Bem-Estar Urbano (IBEU) era esperado com que ambos apresentassem um comportamento igualitário demonstrando com que os índices poderiam estar relacionados comportamentais.

Vislumbra-se que este presente trabalho apresenta a caracterização de ser um estudo inicial para realizar um estudo de bem-estar em redes sociais considerando a comparação entre o índice de bem-estar das cidades paranaenses. Além disso, poderá realizar a diminuição do tempo de espera da população e dos governantes municipais sobre o bem-estar das cidades permitindo a tomada de decisão mais rápida e com informação atual da presente situação.

Ao obter resultados satisfatórios na primeira fase de avaliação do trabalho, foi percebido na análise da segunda fase de avaliação apresentada na Tabela **??** que o comportamento dos dados do índice PIRES não apresentavam um comportamento semelhante. É preciso destacar algumas suposições que podem ter influenciado no resultado final, a primeira delas foi o prazo de coletada de dados, que na qual demonstra um pouco mais de um mês, evidenciando que eventos externos tenham uma maior influência em relação ao tempo. Em segundo, a classificação de sentimentos para textos curtos na língua portuguesa é um desafio, pois não há trabalhos evidenciados que tenham realizados, a literatura apresenta que os classificadores trabalham mais com a língua inglesa por se tratar de um língua muito utilizada por diversos países. O método utilizado neste trabalho apresentava várias limitações, tais como: chamadas por transação mensais, quantidade de textos a serem classificados e então sendo preciso utilizar serviços mais robustos que demandaria um custo financeiro para que houvesse a classificação dos textos em língua portuguesa.

Como proposta para trabalhos futuros, será preciso obter um tempo maior de coleta de dados referentes às diversas cidades paranaenses e brasileira, abrindo a possibilidade de uma avaliação mais completa e sucinta sobre a proposta de trabalho. Os trabalhos de [Dodds et al. (2011), Mitchell et al. (2013), Dodds e Danforth (2010), Bollen et al. (2011), Frank et al. (2013), Loff, Reis e Martins (2015)] demonstraram que é possível realizar tal estudo, na qual foi aplicado para as cidades, estados e o país americano. Também é necessário aplicar novos métodos de análise de sentimentos, mais precisamente que permitam a realização da classificação de textos curtos em português para a tentativa de aumentar a acurácia dos resultados e comparar com os resultados já obtidos com o classificador que utilizamos neste trabalho.

# REFERÊNCIAS


ARAMPATZI, E.; BURGER, M. J.; NOVIK, N. Social network sites, individual social capital and happiness. *Journal of Happiness Studies*, Springer, p. 1–24, 2016.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.21529/RESI*, v. 5, n. 2, 2006.

ARAUJO, G. D. de et al. Análise de sentimentos sobre temas de saúde em mídia social. *Journal of Health Informatics*, v. 4, n. 3, 2012.

BAUMGARTEN, M. et al. Keyword-based sentiment mining using twitter. *International Journal of Ambient Computing and Intelligence*, IGI Publishing, v. 5, n. 2, p. 56–69, 2013.

BOLLEN, J. et al. Happiness is assortative in online social networks. *Artificial life*, MIT Press, v. 17, n. 3, p. 237–251, 2011.

BRASIL, A. do Desenvolvimento Humano do. *Índice de Desenvolvimento Humano Municipal*. 2013. Acesso em: 27 de Julho de 2017. Disponível em: <http://www.atlasbrasil.org.br/2013/pt/o_atlas/idhm/>.

BRASIL, P. das Nações Unidas para o Desenvolvimento no. *Desenvolvimento Humano e IDH*. 2017. Http://www.br.undp.org/content/brazil/pt/home/idh0.html. Acesso em 27 de Julho de 2017.

CHA, M. et al. The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, IEEE, v. 42, n. 4, p. 991–998, 2012.

CIRIBELI, J. P.; PAIVA, V. H. P. Redes e mídias sociais na internet: realidades e perspectivas de um mundo conectado. *Revista Mediação*, v. 13, n. 12, 2011.

DIAKOPOULOS, N. A.; SHAMMA, D. A. Characterizing debate performance via aggregated twitter sentiment. In: ACM. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. [S.l.], 2010. p. 1195–1198.

DODDS, P. S.; DANFORTH, C. M. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of happiness studies*, Springer, v. 11, n. 4, p. 441–456, 2010.

DODDS, P. S. et al. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, Public Library of Science, v. 6, n. 12, p. e26752, 2011.

ESTATÍSTICA, I. B. de Geografia e. *História do Censo*. 2013. Disponível em: <http://memoria.ibge.gov.br/sinteses-historicas/historicos-dos-censos/panorama-introdutorio.html>.

FAYYAD, U.; PIATETKSY-SHAPIRO, G.; SMYTH, P. From data mining to knowdledge discovery in databases. 1996.

FIELDING, R.; RESCHKE, J. Hypertext transfer protocol (http/1.1): Message syntax and routing. 2014.

FRANÇA, T. C. de; OLIVEIRA, J. Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013. *Brazilian Workshop on Social Network Analysis and Mining (BRASNAN)*, p. 128–139, 2014.

FRANK, M. R. et al. Happiness and the patterns of life: A study of geolocated tweets. *arXiv preprint arXiv:1304.1296*, 2013.

GOMIDE, J. et al. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In: ACM. *Proceedings of the 3rd international web science conference*. [S.l.], 2011. p. 3.

GONÇALVES, P.; DORES, W.; BENEVENUTO, F. Panas-t: Uma escala psicométrica para medição de sentimentos no twitter. 2013.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. third. [S.l.]: Elsevier Inc., 2012.

HU, X. et al. Unsupervised sentiment analysis with emotional signals. In: ACM. *Proceedings of the 22nd international conference on World Wide Web*. [S.l.], 2013. p. 607–618.

LI, Y.-M.; LI, T.-Y. Deriving marketing intelligence over microblogs. In: IEEE. *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. [S.l.], 2011. p. 1–10.

LOFF, J.; REIS, M.; MARTINS, B. Predicting well-being with geo-referenced data collected from social media platforms. In: ACM. *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. [S.l.], 2015. p. 1167–1173.

METRÓPOLES, I. N. de Ciência e Tecnologia Observatório das. *O Bem-Estar Urbano dos municípios brasileiros — IBEU Municipal*. 2016. Acesso em 06 de Outubro de 2017. Disponível em: <http://observatoriodasmetropoles.net/index.php?option=com_content&view=article&id=1777&Itemid=176&lang=pt#>.

MICROSOFT. *Azure Cognitive Services Documentation*. 2017. Acesso em: 22 de Abril de 2017. Disponível em: <https://opbuildstorageprod.blob.core.windows.net/output-pdf-files/en-us/Azure.azure-documents/live/cognitive-services.pdf>.

MITCHELL, L. et al. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, Public Library of Science, v. 8, n. 5, p. e64417, 2013.

NARAYANAN, R.; LIU, B.; CHOUDHARY, A. Sentiment analysis of conditional sentences. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. [S.l.], 2009. p. 180–189.

NASCIMENTO, P. et al. Análise de sentimento de tweets com foco em notícias. In: *Brazilian Workshop on Social Network Analysis and Mining*. [S.l.: s.n.], 2012.

O'REILLY, T.; MILSTEIN, S. *The Twitter Book*. second. [S.l.]: O'Reilly Media, Inc., 2011.

PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: *LREc*. [S.l.: s.n.], 2010. v. 10, n. 2010.

REAGAN, D. *Twitter Application Development For Dummies*. [S.l.]: John Wiley & Sons, 2010.

REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Manole, 2005.

RIOS, R. A. et al. Analyzing the public opinion on the brazilian political and corruption issues. *Brazilian Conference on Intelligent Systems*, p. 13–18, 2017.

RUSSEL, M. A. *21 Recipes for Mining Twitter*. [S.l.]: O'Reilly Media Incorporated, 2011.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: ACM. *Proceedings of the 19th international conference on World wide web*. [S.l.], 2010. p. 851–860.

SOCIAL, S. de C. *Manual de Orientação para atuação em mídias sociais: Identidade padrão de comunicação digital do Poder Executivo Federal*. 2014. Acesso em: 29 de Agosto de 2017. Disponível em: <http://www.secom.gov.br/pdfs-da-area-de-orientacoes-gerais/internet-e-redes-sociais/secommanualredessociaisout2012_pdf.pdf>.

SOUSA, G. L. S. de. *Tweetmining: Análise de Opinião Contida em Textos Extraídos do Twitter*. 2012. Http://repositorio.ufla.br/handle/1/5417. Acesso em 25 de Outubro de 2017.

TUMASJAN, A. et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, v. 10, n. 1, p. 178–185, 2010.

TWITTER. *Company/About*. 2017. Acesso em: 11 de Jun. de 2017. Disponível em: <http://about.twitter.com/company/>.

TWITTER. *Twitter Developers*. 2017. Acesso em: 10 de Abril de 2017. Disponível em: <http://dev.twitter.com>.

YE, N. *Data Mining - Theories, Algorithms, and Examples*. [S.l.]: Taylor & Francis Group, 2014.

ZHANG, K. et al. Ses: Sentiment elicitation system for social media data. In: IEEE. *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. [S.l.], 2011. p. 129–136.