COLETA E SELEÇÃO DE PERFIS DE USUÁRIOS DE REDES SOCIAIS PARA CONTRIBUIÇÃO VOLUNTÁRIA

GUILHERME VEZULA MATEVELI

COLETA E SELEÇÃO DE PERFIS DE USUÁRIOS DE REDES SOCIAIS PARA CONTRIBUIÇÃO VOLUNTÁRIA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Clodoveu Augusto Davis Junior
Coorientadora: Mirella Moura Moro

Belo Horizonte
Maio de 2018

Mateveli, Guilherme Vezula

M425c Coleta e seleção de perfis de usuários de redes sociais para contribuição voluntária / Guilherme Vezula Mateveli. — Belo Horizonte, 2018
xxii, 42 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Clodoveu Augusto Davis Junior
Coorientadora: Mirella Moura Moro

1. Computação - Teses. 2. Crowdsourcing. 3. Crowdsourcing criativo. 4. Redes sociais on-line. 5. Colaboração online. I. Orientador. II. Coorientadora. III. Título.

CDU 519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

COLETA E SELEÇÃO DE PERFIS DE USUÁRIOS DE REDES SOCIAIS PARA CONTRIBUIÇÃO VOLUNTÁRIA

GUILHERME VEZULA MATEVELI

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROFA. MIRELLA MOURA MORO
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO CÉSAR MACHADO PEREIRA
Departamento de Ciência da Computação - UFMG

DRA. JUNIA LUCIO DE CASTRO BORGES
LAB-URB - Escola de Arquitetura - UFMG

Belo Horizonte, 11 de maio de 2018.

Dedico este trabalho a todos aqueles que de alguma forma estiveram e estão próximos de mim durante todo esse tempo, fazendo esta caminhada valer cada vez mais a pena.

Agradecimentos

Inúmeros são os nomes que contribuíram para que eu chegasse até aqui. Primeiramente agradeço a Deus por mais essa oportunidade de crescimento e amadurecimento em minha vida. Sem Ele eu não teria forças para concluir esse projeto.

Agradeço aos meus familiares pelo apoio e compreensão, que mesmo de longe torceram por mim.

Agradeço aos professores que me ajudaram a aprimorar ainda mais meus conhecimentos, em especial aos Professores Clodoveu e Mirella pela orientação e paciência nos momentos de indecisão.

Agradeço aos amigos de longa data e àqueles que foram surgindo ao longo dessa trajetória que estavam sempre presentes nas horas boas e ruins da minha vida.

Agradeço à agencia de formento da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa durante o mestrado.

E, para terminar, agradeço a todos os meus amigos *online* (gringos e brasileiros), que mesmo não os conhecendo pessoalmente, de longa distância me apoiaram nas horas de desespero.

"O mundo é muito grande e estamos em muitos lugares diferentes, por isso precisamos de muitos pontos de vista para compreendê-lo."
(Autor desconhecido)

Resumo

Diariamente, milhões de pessoas utilizam recursos e ferramentas que permitem que conteúdo produzido por cidadãos seja publicado na *Web*. Iniciativas de *crowdsourcing* visam coletar esses dados, utilizando cidadãos como sensores humanos. Contudo, essas iniciativas não garantem que cidadãos permaneçam contribuindo, e com o tempo as contribuições se tornam falhas e enviesadas. Nesse contexto, existe interesse em encontrar voluntários a partir das redes sociais, com base nas suas preferências temáticas. Desse modo, o principal objetivo deste trabalho é identificar nas redes sociais interessados em um tema dado, e assim encontrar voluntários que possam atuar em *crowdsourcing* relacionado a temas de sua preferência. Para isso, este trabalho propõe uma metodologia que consiste em realizar um processo interativo para melhorar a coleta de dados sobre potenciais colaboradores, e consequentemente, melhorar o conteúdo coletado. Para verificar a validade da metodologia, foi desenvolvido um estudo de caso que considera um contexto urbano para identificar palavras-chave iniciais e, assim, construir um processo de refinamento. Os resultados mostram que muitos dos dados coletados ainda não estavam relacionados ao tema, mas que mesmo assim o processo de refinamento dos dados levam a melhoramentos em relação à iteração inicial. No entanto, não é possível afirmar que usuários, por postarem comentários relacionados a um contexto temático, teriam nesse tema uma preferência pessoal. Portanto, adaptações no processo de refinamento podem ser realizadas para efetivar as preferências temáticas dos usuários das redes sociais.

Palavras-chave: *Crowdsourcing*, Redes Sociais, Contribuição Voluntária.

Abstract

Nowadays, millions of people use resources and tools that allow user-produced content to be published in the Web. Crowdsourcing initiatives aim at collecting such data, using citizens as human sensors. However, these initiatives do not ensure that citizens keep on contributing, and in time contributions may became flawed and biased.Crowdsourcing projects, in this context, can benefit from finding and incentivizing additional volunteers. Therefore, the main objective of this work is to identify, within social networks, people with interest in a given subject, thereby discovering potential volunteers for crowdsourcing in that theme. To that effect, this work proposed a methodology that consists in interactively improving data collection on potential collaborators. In order to verify the validity of the proposed methodology, a case study is presented, in which an urban context is used as the source of seed keywords on the selected theme, from which an iterative refinement process is started. Results show that much of the collected data were still not related to the theme, but the refinement process leads to improvements relatively to the initial iteration. However, it is not possible to assert that users that post comments related to a given theme are indeed personally interested in it. Therefore, adaptations in the refinement process can be implemented in order to fulfill the thematic preferences of social network users.

Keywords: Crowdsourcing, Social Networks, Volunteered Geographic Information.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Diariamente, milhões de pessoas utilizam recursos e ferramentas que permitem que conteúdo produzido por cidadãos (indivíduos e grupos) seja publicado na *Web*, obtendo imediatamente alcance mundial. São inúmeras as iniciativas de criação de *blogs*, publicação de vídeos domésticos, registro de comentários e opiniões sobre produtos, serviços, livros e filmes, entre muitos outros. O crescimento explosivo de sites que permitem algum tipo de participação dinâmica dos usuários é indicação clara dessa tendência. De modo geral, observa-se que esse fenômeno se relaciona com a necessidade que as pessoas têm de se expressar e de buscar fazer alguma diferença no mundo em que vivem.

Especificamente, a quantidade e a variedade de dados geográficos disponíveis na *Web* para uso pelo cidadão comum têm também aumentado rapidamente. Desde a introdução do *Google Earth*, em 2004, e do *Google Maps*, em 2006, tem crescido o interesse por ferramentas que permitam que as pessoas localizem geograficamente pontos de seu interesse e que tais sistemas ofereçam serviços baseados em posições geográficas.

As facilidades disponíveis a partir dos produtos geográficos têm viabilizado o desenvolvimento de toda uma nova geração de aplicativos geográficos na Web, tais como procurar trajetórias similares para planejar viagens em lugares desconhecidos, ou até reconhecer pontos de interesse localizados em mapas [Ye et al., 2011; Zheng et al., 2010]. O mais interessante é que todas essas facilidades estão disponíveis para qualquer cidadão comum, ou seja, as pessoas estão gradualmente descobrindo o valor intrínseco da existência e disponibilidade gratuita de ferramentas com dados geográficos de qualidade, e aprendendo a usar esses recursos para resolver problemas cotidianos.

A partir de algumas experiências recentes [Chatfield & Brajawidagda, 2014; Barron et al., 2014; Hirata et al., 2015], é possível inferir que muitas pessoas se disporiam

a contribuir, de forma voluntária e sem expectativa de pagamento, para a criação de novos dados e para a correção de erros porventura existentes nos dados geográficos publicados. Afinal, ninguém melhor que um morador de um bairro para apontar erros nos mapas atuais ou para coletar e digitalizar dados sobre algo que está presente em seu cotidiano, como buracos no pavimento, pontos de alagamento, pontos ideais de coleta de lixo ou deficiências na iluminação pública em pontos específicos. Além disso, os dispositivos que essas pessoas carregam constantemente consigo têm o potencial de embutir sensores, capazes de captar ou medir diversos tipos de dados instantâneos sobre o ambiente local. Ou seja, a *coleta de dados voluntária* tem um grande potencial de gerar dados relevantes para pesquisa e desenvolvimento. Gupta & Lee [2010] acreditam que, quando cidadãos podem colaborar com iniciativas de coleta de dados geográficos junto com seus dispositivos, eles podem atuar como *sensores*.

Goodchild [2007] denomina esse fenômeno como *volunteered geographic information* (VGI), indicando que pessoas podem agir como sensores capazes de explorar o ambiente em que vivem e produzir informação útil sobre ele. Genericamente, chamam-se *crowdsourcing* as iniciativas que visam coletar dados com a ajuda de grande quantidade de colaboradores humanos [Cilliers & Flowerday, 2014]. Esse termo tem sido utilizado recentemente em relação a computação humana, computação social, computação nas nuvens e computação móvel [Parshotam, 2013]. *Crowdsourcing* também tem sido utilizado em grandes empresas, como por exemplo, *marketing* [Vukovic, 2009]. VGI pode ser considerado um tipo especial de *crowdsourcing* em que os dados de interesse são geograficamente localizados e a contribuição é voluntária. Nesse contexto, recentes abordagens de *crowdsourcing* consideram solicitações em diversos contextos, propondo tarefas que demandam coleta de dados por voluntários em uma localização específica. Assim, cidadãos que possam assumir tais tarefas e têm acesso ao local obtêm os dados requisitados e completam a tarefa demandada [Mateveli et al., 2015].

As redes sociais também podem ser fonte de um tipo de *crowdsourcing*, considerando que usuários publicam comentários a todo momento relacionado a diversos assuntos. Deste modo, trabalhos relacionados têm sido agregados a estudos das redes sociais para gerar conhecimento [Castro et al., 2017; Klinczak & Kaestner, 2017; Sharifi et al., 2010].

1.1 Motivação

Quando se trata de coleta de dados voluntária, existem diversos problemas no processo de coleta e no uso de dados fornecidos por voluntários que se tornam grandes desafios

para os pesquisadores da área. Os principais problemas são: (i) redução gradativa do interesse dos usuários; (ii) cobertura irregular do tema ou do espaço de interesse; (iii) dúvidas quanto à confiabilidade dos dados e estratégias de validação; e (iv) formas de prover *feedback* das contribuições. Acredita-se que alguns desses problemas possam ser sanados utilizando redes sociais para identificar possíveis voluntários interessados em contribuir com temas de seus interesses.

Como a coleta de dados voluntária não garante a cobertura de espaço e a permanência de contribuições, com o tempo ela se torna falha e enviesada. Nesse contexto, voluntários poderiam ser encontrados a partir das redes sociais para realizarem tarefas com base nas suas preferências, seus interesses, suas necessidades, entre outros. Por consequência, tal estratégia pode significar a melhoria das informações com a realização de tarefas com base em dados sobre os usuários das redes sociais (perfil, conteúdo geolocalizado, localização histórica e comentários) que são motivados a contribuir, ou mesmo que tenham a sua disposição em contribuir priorizada em função de suas preferências individuais.

Portanto, este trabalho propõe um mecanismo de busca por voluntários utilizando os perfis e textos publicados por eles nas redes sociais, identificando possíveis preferências temáticas. Assim, usuários de redes sociais poderão ser convidados a realizar tarefas de seu interesse em assuntos já conhecidos por eles, tornando-se voluntários em projetos de *crowdsourcing*.

1.2 Objetivos

O objetivo geral deste trabalho consiste em identificar e selecionar usuários a partir de suas preferências implícitas nas redes sociais, e assim encontrar voluntários que possam atuar em *crowdsourcing* relacionado a temas de sua preferência. Para isso, os objetivos específicos são:

1. Implementar uma metodologia para buscar perfis de usuários das redes sociais.

2. Criar um processo de coleta de dados que possibilite o melhoramento contínuo da busca de usuários nas redes sociais.

3. Executar a metodologia a partir de um contexto temático urbano.

4. Validar a metodologia.

1.3 Organização do Trabalho

O restante deste trabalho está organizado da seguinte maneira. O Capítulo 2 apresenta os conceitos básicos utilizados na metodologia proposta e discute os trabalhos relacionados relevantes para este trabalho. O Capítulo 3 apresenta e define uma discussão e uma taxonomia para aplicações *crowdsourcing*. O Capítulo 4 detalha a metodologia proposta para melhorar o processo de coleta de dados e categorização de usuários. O Capítulo 5 apresenta um estudo de caso utilizando a metodologia proposta para análise de sua viabilidade, bem como os resultados obtidos. Por fim, o Capítulo 6 apresenta considerações finais, conclusões e propostas para trabalhos futuros.

Capítulo 2

Conceitos e Trabalhos Relacionados

Vistos como "sensores humanos", cidadãos comuns e leigos em tecnologia podem se tornar recursos essenciais para aumentar a quantidade de informação disponível e para melhorar a qualidade do que já existe.

As redes sociais são outro exemplo desse fenômeno atual de geração de conteúdo. Segundo Santos [2013], as redes sociais têm um papel importante na sociedade, pois possibilitam a comunicação e o suporte informativo. Outra característica é a forma como são organizadas, que considera um conjunto de pessoas que se relacionam. Nos últimos anos, cidadãos passaram a produzir mais conteúdo *online*.

Nesse contexto, este Capítulo apresenta conceitos importantes que são necessários para a extração do conhecimento nas redes sociais, considerando como podem ser utilizados em prol de projetos de contribuição voluntária por cidadãos.

2.1 *Volunteered Geographic Information*

Segundo Goodchild [2007], *volunteered geographic information* (VGI) indica que pessoas podem agir como sensores capazes de explorar o ambiente em que vivem e produzir informação útil sobre ele. Existem vários outros nomes para o que foi definido como VGI. De acordo com Craglia [2007] são aspectos ligados à geografia referentes ao fenômeno mais abrangente chamado *user-generated content*, nome genérico atribuído ao que está sendo chamado recentemente de *Web* 2.0 [O'Reilly, 2007; Maguire, 2007], tais como *wikis* e *blogs*.

Com outro enfoque, baseado em conteúdo informacional de acesso público (*Public Commons*) esse fenômeno também é denominado "wikificação" dos sistemas de informação geográficos [Onsrud et al., 2004; Hardy, 2007; Sui, 2008], no sentido de

chamar a atenção para conteúdos fornecidos livremente sem motivação econômica ou qualificação formal para isso.

Um aspecto interessante da contribuição voluntária geográfica é que existem precedentes, inclusive em projetos de pesquisa científica. Um exemplo é o projeto *National Map Corps*, do *United States Geological Survey* (USGS), a agência de mapeamento nacional norte-americana [Bearden, 2007]. O USGS iniciou esse projeto na década de 90, chegando a reunir mais de três mil voluntários em 2001, sendo o mesmo ativo até hoje. A ideia era reunir notas de correção de mapas baseadas na observação dos voluntários, e orientar a atualização a partir disto. Mas o grande número de contribuições causou problemas para os cartógrafos do órgão, que não conseguiram resolver todos os problemas e retornar a solução ao autor da contribuição, como era o planejado. Recentemente, o projeto passou a receber apenas contribuições via *web*[1] e que fossem adequadamente acompanhadas de coordenadas obtidas por GPS. Mesmo assim existe um *backlog* de contribuições que precisam ser tratadas. Na recepção de contribuições via *web*, o projeto foi inspirado em uma iniciativa da NASA, denominada criativamente "Clickworkers"[2].

Devido a popularização das aplicações geográficas móveis, diversas aplicações têm sido desenvolvidas utilizando a localização do usuário como elemento central [Davis Jr et al., 2013; Koswatte et al., 2015]. Por exemplo, Davis Jr et al. [2013] apresentam um *framework* que possibilita a criação de diversas aplicações VGI baseadas na *Web* e baseadas em aplicações *mobile*.

2.2 Contribuição Voluntária

Os mecanismos que levam as pessoas a despender tempo e esforço para produzir de graça algo que pode ou não ser usado por outros ainda não são claros [Budhathoki et al., 2008; Goodchild, 2007]. No entanto, a regra básica dos mecanismos *wiki* (e originaria do desenvolvimento colaborativo de software livre) parece ser válida para uma ampla gama de situações.

Se um conjunto de informações disponível livremente na *web* for de interesse para um grande número de pessoas, uma parcela razoável dessas pessoas tende a retribuir o serviço prestado por meio de uma predisposição a colaborar para que o serviço melhore para seus pares. Na área de Comunicação, esse fenômeno tem sido estudado, conhecido como "ação coletiva" [Bimber et al., 2005]. Tal crescimento de conteúdo vem pelo rápido desenvolvimento dos *smartphones* e de plataformas *crowdsourcing* como vídeos, fotos

[1] http://ims.er.usgs.gov/vfs/faces/index.jspx
[2] http://clickworkers.arc.nasa.gov/top

e áudios [Chen et al., 2014]. Ou seja, essas plataformas possibilitam a coleta de dados com a ajuda de colaboradores humanos [Cilliers & Flowerday, 2014; Chen et al., 2014].

Genericamente, Cilliers & Flowerday [2014] definem *crowdsourcing* como iniciativas que visam coletar dados com a ajuda de grande quantidade de colaboradores humanos. Quando os cidadãos utilizam sensores ambientais acoplados a dispositivos móveis para gerar conteúdo, o termo é adaptado para *crowdsensing*. Ou seja, esses dados passam a ser medições realizadas pelos sensores embutidos em *smartphones* ou *tablets*.

Os dispositivos móveis possuem um conjunto crescente de sensores, como acelerômetro, bússola digital, giroscópio, GPS, microfone e câmera, que permitem o surgimento de grupos e de sensores em escala comunitária [Lane et al., 2010]. Por exemplo, Chowdhury et al. [2014] apresentam técnicas para usar *smartphones* de cidadãos, configurados como uma rede de sensores sem fio para coletar dados ambientais. Nesse contexto, diversas aplicações utilizam o conteúdo geolocalizado (via GPS e outros sensores), denominadas como aplicações VGI.

Uma abordagem de *crowdsourcing* considera um solicitante que propõe tarefas que demandam coleta de dados por voluntários em uma localização específica. Cidadãos que possam assumir tais tarefas e têm acesso ao local obtêm, então, os dados requisitados e completam a tarefa demandada. Nesse contexto, Chen et al. [2014] apresentam o *gMission*, uma plataforma que introduz novas técnicas para coleta de dados por tarefa, incluindo sensoriamento geográfico, detecção de trabalhadores (usuários que assumem uma tarefa) e recomendação de tarefas.

Outra abordagem é a utilização de redes sociais no contexto de contribuição voluntária. Tendo em vista que usuários se juntam às redes sociais, publicam seus perfis ou qualquer outro conteúdo e criam relacionamentos com outros usuários a quem eles são associados [Mislove et al., 2007], o conteúdo gerado por eles pode conter contribuições indiretas. Desse modo, na próxima seção são apresentados os principais conceitos de redes sociais e uma contextualização das redes sociais na contribuição voluntária.

2.3 Redes Sociais

O surgimento da *Web* 2.0 possibilitou que plataformas online evoluíssem para algo com mais conteúdo e disponível para todos. Atualmente, as redes sociais *online* são plataformas que possuem características e públicos variados, tornando-se uma área produtiva para pesquisa [Brandão & Moro, 2016]. Ainda, segundo Mislove et al. [2007], as redes sociais *online* são um dos mais populares sites da Internet.

Na atual situação da sociedade em relação à *Internet* e a interação entre indivíduos, é difícil encontrar alguém que não participe de pelo menos uma rede social. Ou seja, cada vez mais pessoas estão adaptando suas vidas em torno da Internet e das redes sociais. Assim, elas aumentam seus capitais sociais e aumentam o convívio com amigos e parentes que vivem perto ou distantes uns dos outros [Wellman et al., 1996].

Muito antes do surgimento da *Internet*, as redes sociais, bem como suas propriedades, já vinham sendo estudadas em Ciências Sociais [Mislove et al., 2007; Newman, 2003; Scott, 2017]. Nos anos 1930, sociólogos já consideravam a importância do estudo dos padrões entre as pessoas para entender melhor o funcionamento da sociedade [Newman, 2003]. Esses estudos são fundamentais para a descoberta de informação, principalmente quando o assunto aborda uma área social. Partindo desse pressuposto, qualquer sociedade ou interação social pode ser mapeado como uma rede social, sendo usada para revelar todo o tipo de informação nela contida [Brandão & Moro, 2016].

Nesse contexto, define-se como Rede Social (RS) um conjunto ou grupo de pessoas com determinados padrões de contato ou de interações entre elas mesmas [Newman, 2003; Scott, 2017; Wasserman & Faust, 1994]. Quando pessoas ou organizações são conectadas através de computadores, pode ser definido também como uma Rede Social, ou seja, um conjunto de pessoas conectadas em um conjunto de relações sociais (amizade, co-trabalho, troca de informação) [Garton et al., 1997]. Assim, considerando uma classificação de alto nível em relação as essas interações e conexões entre usuários, uma RS pode ser classificada como *Online* e *Offline*.

As Redes Sociais *Online* (RSOn) são *Web sites* ou *Web services* que permitem que pessoas interajam entre si, não necessariamente cara-a-cara [Brandão & Moro, 2016; Scott, 2017], como *Facebook*[3] e *Twitter*[4], por exemplo. Ao contrário, as Redes Sociais *Offline* (RSOff) são caracterizadas pela ausência de comunicação *online* entre usuários, e mesmo assim pode existir uma relação entre eles, que é o caso de Redes Sociais Profissionais [Brandão & Moro, 2016].

Outro aspecto interessante das redes sociais é a utilização de informação espacial. Tais redes são chamadas de Redes Sociais Baseadas em Localização (*Location-Based Social Networks*, LBSN), ou seja, são redes sociais com conteúdo geolocalizado. Portanto, usuários podem expandir suas estruturas sociais com novas experiências do mundo real derivadas de suas localizações [Zheng, 2012]. Atualmente, aplicações geográficas móveis têm se tornado muito populares, permitindo a constituição de LBSN como em sistemas de recomendação, pontos de interesse e desastres ambientais [Berjani & Strufe, 2011; Gao et al., 2015; Hirata et al., 2015; Pat et al., 2015]. Portanto, a localização

[3]http://www.fb.com
[4]http://www.twitter.com

desempenha um papel central em LBSN [Ye et al., 2011].

Diferente da *Web*, que é organizada em torno de conteúdo, uma RS é organizada em torno de usuários [Mislove et al., 2007]. Assim, usuários podem também contribuir com informação relevante indiretamente. Por exemplo, Chatfield & Brajawidagda [2014] apresentam estudos de *tweets* recolhidos através do uso de *hashtags* específicas sobre o tornado que atingiu a cidade de Oklahoma em 20 de maio de 2013. Assim, o Twitter serviu como meio de comunicação de cidadão para cidadão e entre órgãos governamentais e cidadãos.

Para a exploração desse conteúdo provido pelas redes sociais, técnicas vêm sendo exploradas para identificar conceitos, sentimentos, tópicos e similaridades antes ou depois da construção da estrutura de uma rede social [Brandão & Moro, 2016]. Assim, a próxima Seção apresenta uma análise sistemática de trabalhos relacionados ao processamento de textos em redes sociais e outras outras variedades de fontes de dados. O foco principal é apresentar técnicas de ponderação de termos para o processamento dos dados ali obtidos.

2.4 Trabalhos Relacionados

Uma variedade de trabalhos vêm sendo publicados com o objetivo de extrair informação de dados da *Web* [Chatfield & Brajawidagda, 2014; Elmessiry et al., 2017; Hirata et al., 2015; Mislove et al., 2007]. Parte desses trabalhos utilizam técnicas de processamento de linguagem natural para extrair informação [Castro et al., 2017; Klinczak & Kaestner, 2017; Sharifi et al., 2010]. Assim, para relacionar este trabalho com a literatura, trabalhos relacionados a técnicas de ponderação de termos são apresentados a seguir.

Com o objetivo de criar um sumário de *posts* em tópicos tendentes no *Twitter*, Sharifi et al. [2010] coletam um número grande de *tweets* (durante vários dias no mesmo horário durante a noite) e criam um sumário que representa todos os *posts* sobre determinado tópico. Para isso, eles implementam um algoritmo TF-IDF híbrido considerando o TF (*Term Frequency* [Luhn, 1957]) como todos os *tweets* em um documento e o IDF (*Inverse Document Frequency* [Sparck Jones, 1972]) considera uma sentença como um documento. Como o objetivo é encontrar um sumário de um *post* que represente um tópico, eles ainda fazem uma normalização por sentença. Para comparação eles utilizaram um algoritmo de *Phrase Reinforcement* chamado *ROUGE-N* [Lin & Hovy, 2003], e os resultados mostram que o TF-IDF híbrido produz melhores sumários.

Outro trabalho tenta encontrar os efeitos em diferentes métodos de ponderação de termos [Zaman & Brown, 2010]. Assim, eles aplicaram diferentes métodos (TF-IDF, *Log-Entropy* e Frequência) em *datasets*. Os resultados mostram que o TF-IDF teve um performance melhor quando o *dataset* se tornou maior. Nesse caso, Zaman & Brown [2010] não utilizaram dados de redes sociais, mas uma base de dados de testes do Los Angeles Times (TREC-8).

Em busca também de estudar e comparar métodos de classificação de textos, os trabalhos de Zhang et al. [2011] e Kido et al. [2014] comparam métodos como TF-IDF e LSI (*Latent semantic Indexing* [Deerwester et al., 1990]) em diferentes bases de dados. Zhang et al. [2011] consideram como base periódicos acadêmicos com mais de 20 categorias e Kido et al. [2014] consideram micro-blogs, mais especificamente o *Twitter*. Em ambos os trabalhos, o LSI se demonstrou melhor que o TF-IDF, que traz uma qualidade semântica melhor e uma melhor performance. Mas, quando se trata de qualidade estatística e contabilização dos termos, o TF-IDF apresenta-se como melhor opção.

Uma pesquisa mais recente [Wang et al., 2017] mostra que o TF-IDF não é melhor que o método desenvolvido por eles (Boosting Feature), que considera a preferência do usuário por um tópico através de uma probabilidade e acrescenta essa *feature* ao TF-IDF para o processamento. Os experimentos mostram que o método pode aumentar a acurácia da classificação. Nesse caso, o objetivo da pesquisa era classificar o conteúdo do Twitter em 14 categorias.

Entretanto, outras duas pesquisas recentes que também utilizam o Twitter como base de dados [Castro et al., 2017; Klinczak & Kaestner, 2017] mostram que o TF-IDF obtém bons resultados. Castro et al. [2017] propuseram um *framework* aplicando técnicas de processamento de textos e *Machine Learning* (ML) não supervisionados para inferir um alinhamento entre a RS e as eleições parlamentares da Venezuela. Para isso, os *tweets* foram agrupados e depois foi criada uma matriz (documento x termo) utilizando um TF-IDF normalizado e outro que, além da normalização, utiliza uma função logarítmica. Assim, foi aplicada uma técnica de *analysis-based* para identificar as *features* e então aplicar o algoritmo *k-means* para agrupamento.

A outra pesquisa utiliza o TF-IDF com sucesso em uma RS envolve uma descoberta de temas-chaves a partir das redes sociais, caracterizadas por termos relevantes a um determinado contexto, e sua evolução ao longo do tempo [Klinczak & Kaestner, 2017]. Os autores utilizam uma matriz (*tweets* x termos) gerada a partir do TF, TF-IDF ou ponderação booleana. Em seguida, agrupam os termos utilizando algoritmos como *k-means* [Han et al., 2011], *k-medoids* [Han et al., 2011] ou NMF (*Non-Negative Matrix Factorization* [Lee & Seung, 2000]). Assim, os melhores resultados foram adquiridos

pelos algoritmos TF-IDF (em relação a matriz) e NMF (em relação ao agrupamento).

Por fim, Elmessiry et al. [2017] avaliam a triagem de pacientes em relação às reclamações dos pacientes com os médicos. Para a extração de *features* são utilizados os métodos TF e TF-IDF. Para a classificação utilizam um método de ML para detectar se uma queixa era relacionada a uma prática médica ou a um medicamento. Com o TF-IDF para extrair as *features*, os classificadores se mostraram mais eficientes. Para essa pesquisa os autores utilizaram uma base de dados real com mais 14.000 pacientes com queixas relacionadas a 768 médicos.

2.5 Considerações

Esta dissertação busca transformar essa contribuição indireta em uma contribuição direta, onde os usuários participariam efetivamente em aplicações *crowdsourcing*. Nesse contexto, as redes sociais são consideradas em um contexto de contribuição voluntária, considerando que os usuários são cidadãos que contribuem indiretamente através de seus perfis. Mas para isso, é necessário um melhoramento contínuo de coleta de dados e, assim, encontrar usuários que estão interessados em uma temática específica, dado que um dos problemas centrais para *crowdsourcing* é encontrar e motivar voluntários para cumprir tarefas de coleta de dados [Hirata et al., 2015].

A partir dos trabalhos relacionados, conforme Seção 2.4, observa-se a ampla aplicação de processamento de textos nas redes sociais [Castro et al., 2017; Klinczak & Kaestner, 2017; Sharifi et al., 2010]. Acredita-se que o uso de técnicas de ponderação de termos em redes sociais possa contribuir para melhorar a coleta de dados, e assim, encontrar possíveis usuários para contribuição voluntária em aplicações *crowdsourcing*, sanando uma parte do problema. Para melhor justificar o desenvolvimento dessa dissertação de mestrado, o Capítulo 3 apresenta estudos iniciais sobre *crowdsourcing* e propõe uma taxonomia para essas aplicações.

Capítulo 3

Definição do Problema

Os diversos projetos envolvendo *crowdsourcing* e *crowdsensing* apresentam terminologia e conceitos variados, em função das diferentes formas de colaboração do usuário e de coleta de informação. Assim, este Capítulo discute sobre os termos e as técnicas empregadas, apresenta uma taxonomia para classificá-las, define os principais desafios em aplicações de *crowdsourcing* e justifica o uso de Redes Sociais para sanar parte do problema.

3.1 Classificação de *Crowdsourcing* e *Crowdsensing*

Mateveli et al. [2015] propõem uma classificação das técnicas de *crowdsourcing* e *crowdsensing* de acordo com duas dimensões: a forma de captura dos dados e o nível de colaboração do usuário, conforme a Figura 3.1. Cada quadrante informa um contexto para a combinação das dimensões.

- ◇ ***Crowdsourcing* Ativo (CSOA):** o usuário fornece informação conscientemente, tipicamente de forma voluntária, sabendo como a mesma será utilizada. Exemplos incluem o *OpenStreetMap* e o projeto *Flooding Points* [Hirata et al., 2015].

- ◇ ***Crowdsourcing* Passivo (CSOP):** a informação é extraída a partir de material postado pelo usuário para outras finalidades, ou seja, dados úteis eventualmente contidos no material postado para outras finalidades são usados sem o conhecimento explícito do usuário. Um exemplo comum é a coleta automática de *tweets* associados a *hashtags* específicos [Chatfield & Brajawidagda, 2014].

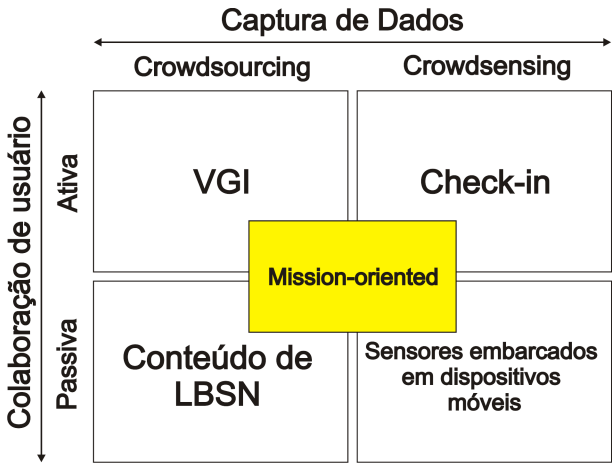


Figura 3.1: Classificação de crowdsourcing e crowdsensing com base na colaboração do usuário e no método de captura de dados

- ◇ ***Crowdsensing* Ativo (CSEA):** o usuário fornece ativamente informação que é capturada por sensores embutidos em seu dispositivo móvel. Um exemplo é o ato de fazer *check-in* em um ponto de interesse com posição determinada por GPS [Ballatore et al., 2010; Mytilinis et al., 2015]. Outro exemplo é um aplicativo para coleta de dados de ruído urbano usando *smartphones* [Davis Jr et al., 2013].

- ◇ ***Crowdsensing* Passivo (CSEP):** a informação é capturada por sensores em dispositivos móveis, mas sem a necessidade de interação com o usuário. Exemplos incluem a função de captura de trajetória em ferramentas como o *Waze* e a captura de outros dados através de sensores embarcados em *smartphones* [Chowdhury et al., 2014].

Na Figura 3.1 foi inserido um elemento central em super posição aos quadrantes para representar atividades *mission-oriented*, ou seja, aplicações VGI que partem de um conjunto de tarefas que precisam ser realizadas e buscam voluntários para executá-las, como é o exemplo do aplicativo *gMission* [Chen et al., 2014]. Ele atua em três das quatro dimensões da figura. Para atuar também em *crowdsourcing* passivo falta a integração do *gMission* com as Redes Sociais.

Observamos inicialmente que Sistemas de Recomendação (*Recommendation Systems*, RecSys) poderiam ser aplicados ao contexto de contribuição voluntária. Mateveli et al. [2015] assumem que sistemas de recomendação geralmente trabalham com *crowdsourcing* passivo, pois têm como fonte de dados um LBSN que considera conteúdos ou perfis de usuários similares para realizar a recomendação. Desse modo, a seguir são apresentados desafios que possibilitam a aplicação de RecSys em *crowdsourcing*.

3.2 Desafios: Possibilidades de Aplicação de Sistemas de Recomendação em Contribuição Voluntária

Contribuições *mission-oriented* podem ser solicitadas dentro do escopo das quatro dimensões, buscando obter o máximo de colaboração dos cidadãos. Como apresentado anteriormente, os problemas de coleta e uso de dados podem ser resolvidos com sistemas de recomendação. Desse modo, a partir desses problemas, Mateveli et al. [2015] destacam que três principais desafios podem, em princípio, ser enfrentados associando sistemas de recomendação e ambientes de contribuição voluntária (ver Figura 3.2):

1. **Relação voluntário-tarefa:** esse desafio consiste em identificar os melhores cidadãos para executar uma determinada tarefa, ou seja, recomendar a tarefa para voluntários que realmente tenham interesse em realizá-la. Isso é possível considerando o perfil do usuário, o conteúdo geolocalizado e dados históricos das redes sociais.

2. **Relação voluntário-região:** esse desafio consiste em identificar as regiões que possuem baixa cobertura e recomendar os melhores voluntários para contribuir com a captura dos dados nessas regiões. Ou seja, baseado no perfil do usuário, na localização histórica do usuário e conteúdo das redes sociais, identificar os voluntários que teriam conhecimentos sobre essas regiões para contribuir com as tarefas.

3. **Relação voluntário-tema:** esse desafio consiste em identificar os melhores voluntários para verificar a confiabilidade das contribuições de acordo com os respectivos temas. Ou seja, as contribuições são direcionadas para os voluntários considerados especialistas em determinados temas. Mesmo esse desafio não estando ligado diretamente ao problema de desinteresse, a partir do momento que usuários são identificados a partir de temas de interesse e/ou especialidade, e os mesmos contribuam, esse problema, indiretamente, é mitigado.

Porém, ainda restam desafios importantes em temas relacionados à chamada "Ciência da *Web*", incluindo a capacidade de compreensão e contextualização semântica do conteúdo de textos, com o reconhecimento de termos e expressões em linguagem natural que sejam corriqueiramente utilizados pelas pessoas e, assim, identificar preferências [Hu et al., 2014]. A evolução das técnicas computacionais nessa área pode ajudar no

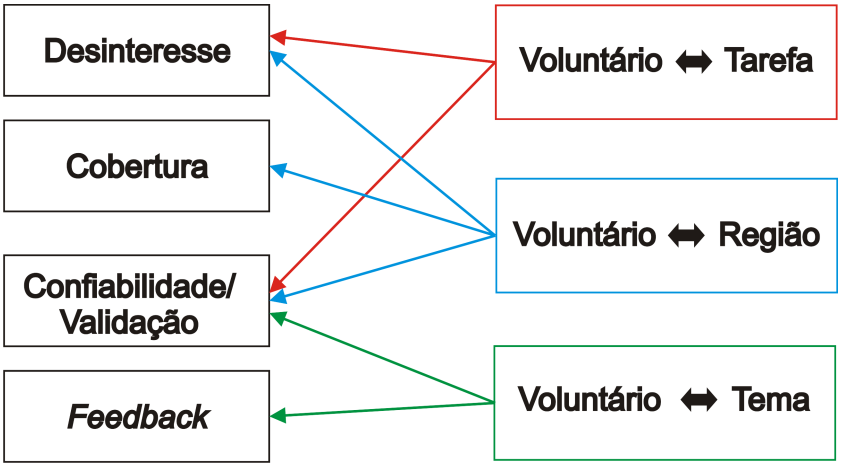


Figura 3.2: Relacionamento Problemas (esquerda) X Desafios (direita)

crowdsourcing passivo, por dar maior precisão semântica ao processo de extração de significado a partir de mensagens e comentários postados nas redes sociais.

3.3 Preferências Temáticas dos Usuários de Redes Sociais

Destaca-se que essa nova forma de obtenção de dados interage com as tendências atuais de participação cidadã e ampla visualização da informação e processos de decisões sobre a gestão e planejamento urbano. Ou seja, a taxonomia apresentada e os desafios abordados já fazem parte da contribuição para a dissertação.

Baseando-se nos desafios abordados na Seção 3.2, onde voluntários realizariam tarefas considerando suas preferências, este trabalho propõe um mecanismo de busca de voluntários a partir de textos publicados por eles nas Redes Sociais, identificando possíveis preferências temáticas a partir de um assunto pré-definido. Assim, pretende-se utilizar técnicas de processamento de texto em comentários das redes sociais, para encontrar voluntários que poderão realizar tarefas de seus interesses em assuntos já conhecidos por eles.

Capítulo 4

Metodologia Proposta

Construir um processo metodológico para categorizar usuários de redes sociais é uma forma de facilitar a busca de pessoas para uma contribuição voluntária ativa, visando o aumento de contribuições por um período maior de tempo. Devido ao grande volume de dados e ao número de usuários, as redes sociais podem conter informação enviesada, o que se torna uma tarefa desafiadora.

Baseados nesses desafios, estudos que visam explorar a descoberta de conhecimento em textos envolvem áreas como mineração de dados, aprendizado de máquina e processamento de linguagem natural [Araújo et al., 2013; Blei et al., 2003; Catae, 2012]. Nas redes sociais o cuidado deve ser redobrado, pois a confiabilidade e a qualidade podem estar comprometidas, não somente pela qualidade e confiabilidade do usuário, mas também pela confiabilidade do conteúdo postado por ele. Outro ponto importante é o processo de validação das preferências dos usuários, pois uma publicação sobre determinado tema não garante que o usuário seja interessado ou especialista no assunto.

Desta forma, a metodologia a seguir propõe uma categorização de usuários que possibilite a identificação de voluntários para a contribuição geográfica ativa. Essa categorização é realizada a partir da coleta, seleção e validação dos dados.

4.1 Visão Geral

Esta seção apresenta uma visão geral da metodologia, que é dividida em três etapas: (1) definição e coleta de dados; (2) seleção e classificação das bases de dados; e (3) validação e processamento dos termos. A Figura 4.1 apresenta uma visão geral da metodologia que compõe o processo de aperfeiçoamento da coleta de dados e da busca por voluntários nas redes sociais. O maior desafio desta metodologia se concentra na

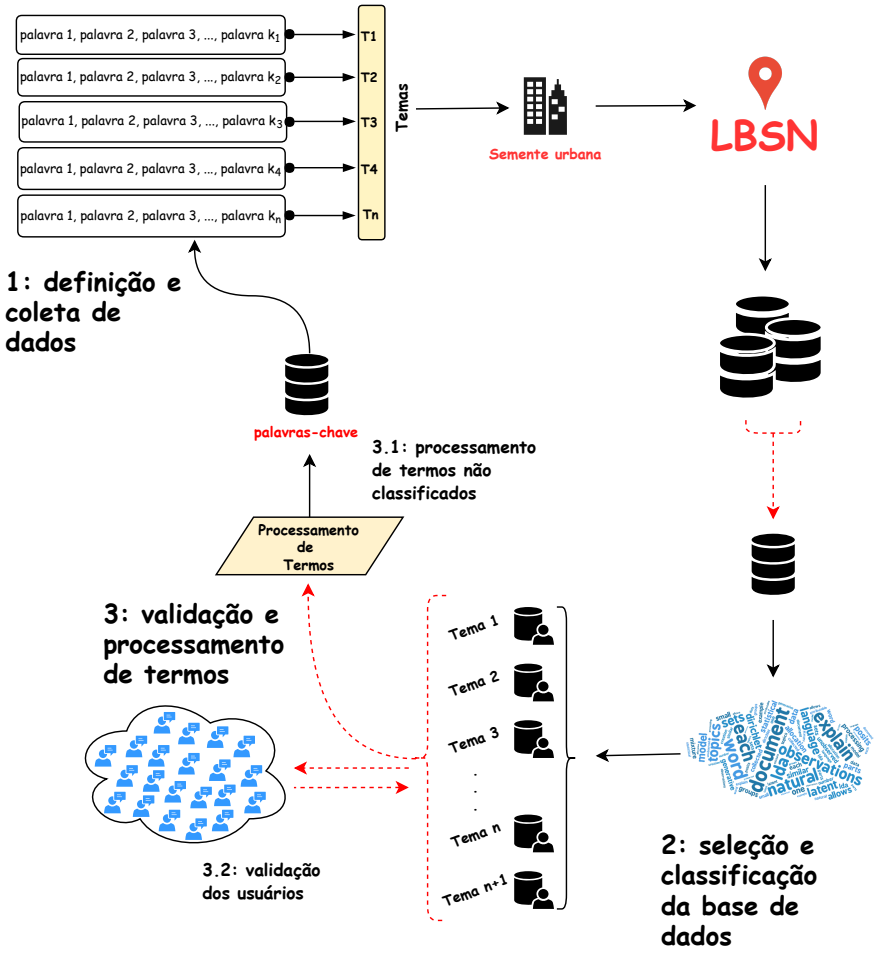


Figura 4.1: Metodologia proposta

etapa de validação e processamento dos termos, pois é subdivida em duas atividades paralelas que exigem mais atenção, principalmente na validação dos resultados com os usuários.

A etapa 1 compreende a definição e coleta de dados, incluindo definir o estudo de caso relacionado a um contexto urbano e definir como é o processo de extração dos dados é estabelecido. Nessa fase, são definidos a rede social utilizada no estudo e os mecanismos de busca, ou seja, quais são os conjuntos de temas e palavras-chave iniciais. Por exemplo, se o estudo de caso selecionado for educação pública, uma possível palavra-chave seria "professores" e uma possível rede social seria o *Facebook*.

A etapa 2 de seleção e classificação das bases de dados consiste em selecionar a base de dados da coleta (total ou parcial) e assim classificá-las de acordo com as palavras-chave contidas nas mensagens. Nessa fase, comentários são classificados de acordo com o contexto temático. Por exemplo, comentários são classificados, de acordo com as palavras-chave, entre relacionados aos temas "Saúde" e "Lazer".

Com a base classificada, a etapa 3 consiste em calcular uma pesagem para palavras-chave que não foram classificadas e validar as preferências dos usuários. Isso ocorre em cada conjunto de usuários classificados para cada tema. Essas duas etapas foram subdividas em duas sub-etapas 3.1 e 3.2 pois podem ser realizadas paralelamente.

O intuito é que todas essas etapas sejam realizadas iterativamente, de modo que o número de palavras-chave de cada tema cresça, aumentando a qualidade da coleta e o número de usuários relacionados com o tema. Neste trabalho, essas iterações serão denominadas *Ciclos*, e cada interação será considerada um ciclo de coleta.

Com o objetivo de melhorar o entendimento da metodologia, cada etapa é detalhada nas seções a seguir. Em cada etapa são descritos os objetivos, as entradas necessárias, as principais atividades e as saídas.

4.2 Definição e Coleta de Dados

Esta primeira etapa envolve a definição do contexto urbano e em seguida a coleta de dados das redes sociais, com o objetivo de apresentar um processo de coleta e as estratégias para a obtenção dos dados. Para isso, o primeiro passo é a escolha da rede social como fonte de dados para identificação de usuários (e.g., Facebook, Foursquare, Twitter, Instagram[1]). Também é necessário verificar como cada rede social disponibiliza os dados para a definição da estratégia de coleta de dados. Especificamente, é importante que a rede social escolhida disponibilize dados como as publicações dos usuários, informação do perfil e também a informação geolocalizada dos comentários.

O segundo passo envolve identificar um contexto urbano para que possam ser definidas palavras-chave que serão o ponto de partida para a execução da metodologia. Deste modo, a entrada consiste em definir temas relacionados ao contexto urbano e assim identificar palavras-chave iniciais relacionadas aos temas. São definidos parâmetros que servirão de entrada para a coleta de dados, que também denominamos como *Sementes Urbanas*, que são o conjunto de palavras-chave que possibilitam a busca de comentários relacionados a um contexto urbano. Neste caso, são representativas para identificar um conteúdo a ser coletado.

As atividades constituem o processo da coleta de dados da rede social. A primeira atividade é definir o método de coleta, configurando um *crawler*[2] (a coleta vai ser

[1] https://www.instagram.com/

[2] *Crawler* é um programa de computador que navega pela *web* de uma forma automatizada com o objetivo de coletar conteúdo.

realizada via API[3], ou simplesmente coletar páginas Web). A segunda atividade é definir os atributos coletados (comentário, perfil, geolocalização, etc). A terceira e última atividade é definir o armazenamento dos dados (em gerenciador relacional ou NoSQL). A partir dessas atividades é gerada a saída, que compreende a base de dados das redes sociais.

4.3 Seleção e Classificação das Bases

Esta etapa considera a seleção e classificação das bases de dados, com o objetivo de classificar o conteúdo coletado de acordo com os temas e suas respectivas palavras-chave. Desse modo, os dados são sumarizados para que apenas os conteúdos relevantes sejam utilizados, para não gerar volumes inapropriados de processamento.

A entrada desta etapa é formada pelos dados coletados a partir da rede social, que é a base de dados da etapa anterior (Seção 4.2). A partir disso, duas atividades são desenvolvidas. Primeiro, o comentário coletado na etapa anterior é classificado de acordo com as palavras-chave, ou seja, se determinada palavra-chave está contida no comentário, o mesmo é relacionado ao tema ao qual aquela palavra-chave pertence. Esse passo é realizado para todos os temas, ou seja, um mesmo comentário pode estar relacionado a mais de um tema. O objetivo é classificar os comentários nos n temas definidos no contexto urbano.

Na segunda atividade, todos os comentários não classificados são resumidos em um tema adicional, em que os comentários não estão relacionados a nenhum contexto urbano, ou pelo menos, as palavras-chave não foram suficientes para classificá-los em algum tema. Por exemplo, o termo "festa" não está relacionado a nenhum tema, então esse termo é classificado como "nenhum". A partir dessa classificação de comentários, os usuários podem ser agrupados de acordo com os temas. Portanto, as saídas desta etapa são as bases de dados separadas por tema.

4.4 Processamento dos Termos e Validação

A próxima etapa envolve a validação e processamento dos termos com o objetivo de calcular uma ponderação de termos para encontrar novas palavras-chave e validar a classificação realizada na etapa anterior (Seção 4.3). Para isso, essa etapa foi dividida em duas sub-etapas diferentes: pesagem e validação.

[3]API é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de *software* ou plataforma baseado na *web*.

As sub-etapas podem ser realizadas simultaneamente, sem uma depender da outra. Deste modo, as duas sub-etapas têm como entrada as bases de dados classificadas por temas, mas possuem atividades diferentes.

4.4.1 Processamento de Termos não Classificados

Esta sub-etapa consiste em calcular um peso para um termo que não está relacionado a nenhum tema. O objetivo é encontrar novas palavras-chave para compor o processo de coleta de dados.

A principal atividade desta sub-etapa é identificar palavras que estão contidas nos comentários classificados mas que ainda não fazem parte do banco de dados de palavras-chave. Para isso, uma ou mais técnicas de ponderamento de textos são utilizadas para calcular um peso para esses termos. Deste modo, esses termos podem se tornar palavras-chave para um novo ciclo de coleta.

Durante esse processo, outra atividade envolve identificar palavras-chave que fazem parte do contexto temático, porém são termos genéricos que podem estar presentes em outro contexto. Por exemplo, a palavra "Paz" está totalmente relacionada ao tema anti-guerra, porém um usuário pode publicar algo relacionado ao termo sossego utilizando essa palavra. Portanto, seria uma forte candidata a sair do banco de palavras. Ao final desta sub-etapa, um novo conjunto de palavras-chave compõem a saída da mesma.

4.4.2 Validação dos Usuários

Esta sub-etapa consiste em enviar os comentários classificados para um grupo de pessoas. Três atividades são realizadas: (i) escolher o perfil desse grupo de pessoas; (ii) determinar qual o meio de comunicação/divulgação; e (iii) coletar e analisar os resultados.

A escolha do grupo de pessoas é importante, pois influencia no resultado final. Assim, duas coisas devem ser levadas em consideração: qualidade ou quantidade. Acredita-se que, o número de pessoas a contribuir será menor se o perfil a ser considerado for muito específico. Mas, se a especialidade não for considerada, pode aumentar o número de pessoas aptas a contribuir, porém com acurácia reduzida.

Determinado o grupo de pessoas que vão participar da validação, a próxima atividade é determinar o meio de comunicação com elas (através das redes sociais, boca-a-boca, etc.). Essa atividade também é importante pois determina o alcance da coleta nas redes sociais.

Por fim, a última atividade desta sub-etapa é coletar e analisar os resultados. Ou seja, os usuários que postaram comentários que tiverem relação com o tema são fortes candidatos a terem interesse no contexto relacionado. Assim, a saída dessa sub-etapa é um banco de dados de possíveis candidatos a participarem de contribuição voluntária ativa.

4.5 Ciclo de Coleta de Dados

O plano é que o processo metodológico seja iterativo, ou seja, que as etapas de 1 a 3 sejam realizadas diversas vezes. Como o objetivo é aumentar o número de palavras-chave de cada tema e assim melhorar a qualidade da coleta de dados, o principal critério de parada é a redução de palavras-chave relevantes para a realização de um próximo ciclo.

Melhorar o processo de coleta de dados significa melhorar a qualidade dos comentários. Assim, aumenta a possibilidade de um usuário realmente ser interessado pelo contexto temático em questão.

Como mencionado anteriormente, cada iteração é considerada um ciclo de coleta, que a cada nova interação produz um acréscimo de palavras-chave para o ciclo subsequente. Ao realizar vários ciclos, a ideia é que a qualidade do conteúdo seja melhor que o ciclo inicial. Portanto, a metodologia proposta neste trabalho poderia ser aplicada para encontrar potenciais candidatos a voluntários em *crowdsourcing* ativo.

Capítulo 5

Estudo de Caso

Para verificar a validade da metodologia proposta, este capítulo apresenta um estudo de caso que considera um contexto urbano, com o objetivo de identificar palavras-chave iniciais e, assim, construir um processo de refinamento. Esse processo consiste em melhorar o processo de coleta e, consequentemente, melhorar o conteúdo coletado.

O objetivo é que usuários possam ser categorizados por temas identificando potenciais voluntários em aplicações *crowdsourcing*. Para isso, o estudo de caso é detalhado a seguir, demonstrando passo a passo como realizar esse processo.

5.1 Contexto Urbano, Palavras-chave e Coleta de Dados

Para este estudo de caso, o contexto urbano considera serviços que os governos federal, estadual e municipal prestam para a sociedade no dia-a-dia. Assume-se que usuários tenham interesse em temas do seu cotidiano. Desse modo, a Tabela 5.1 apresenta uma lista de temas e suas respectivas palavras-chave, que são definidas dedutivamente como termos que estão relacionados aos temas.

Observa-se que algumas palavras-chave estão incompletas. A ideia é usar todas as variações daquela palavra-chave para coleta. Isso pode ser feito utilizando um mecanismo de *stemming*[1]. Assim, aumenta a possibilidade de se obter comentários relacionados ao tema segurança.

A próxima etapa é a coleta de dados, que é realizada através da rede social Twitter. Ela foi escolhida por diversos motivos: pela robustez da API, que possibilita o uso de diversos recursos como, por exemplo, o perfil do usuário que a utiliza; por ser

[1]Processo de reduzir palavras flexionadas ou derivadas.

Tabela 5.1: Lista de temas e respectivas palavras-chave relacionadas

Temas	Palavras-chave
Saúde	saúde, upa, socorro, hospital, ambulância
Transporte	transporte, ônibus, busão, roleta, busu, catraca
Segurança	guarda, polícia, militar, roub, morte, assalt, pm, medo
Limpeza	lixo, polui, gari, esgoto, bueiro, sujo, sujeira

Tabela 5.2: Coleta de dados inicial

Início da coleta	11/08/2017
Fim da coleta	12/08/2017
Tweets coletados	262.048

uma rede social popular entre diversos usuários; por ser uma LBSN, ou seja, uma rede social baseada em localização; e por possibilitar a coleta por palavras-chave.

Para a coleta de dados são considerados os *tweets* publicados em língua portuguesa. Como não existe um polígono geográfico para determinar a localização, determinar a língua na coleta de dados é importante para que não ocorra ruído no momento de processamento e análise dos dados.

Como definido anteriormente no Capítulo 4, uma coleta inicial é realizada utilizando essas palavras-chave, denominada como *Ciclo 0*. A Tabela 5.2 apresenta as informações do Ciclo 0, como a data inicial da coleta, a data final da coleta e a quantidade de *tweets* coletados. O período de coleta para todos os ciclos é de aproximadamente um dia, pois o volume de comentários gerados por dia é muito grande, portanto é suficiente para o estudo de caso.

Para armazenar a coleta de dados é necessário um banco de dados robusto, pois o número de *tweets* coletados a cada instante é muito grande, como observado na Tabela 5.6 onde o número *tweets* coletados passou de duzentos mil em 24 horas. Deste modo, o sistema de gerenciamento de banco de dados (SGBD) escolhido é o *MongoDB*[2] que é um SGBD *NoSQL* que possibilita o armazenamento de coleções de arquivos. Para o estudo de caso, o *MongoBD* é utilizado para armazenar objetos JSON retornados pela API do *Twitter*, que contêm todos os atributos associados ao *tweet*.

5.2 Estruturação e Classificação dos Dados

Para essa etapa, os dados são estruturados no banco de dados relacional PostgreSQL. Neste caso, apenas os dados necessários para a classificação são filtrados para essa nova

[2]https://www.mongodb.com

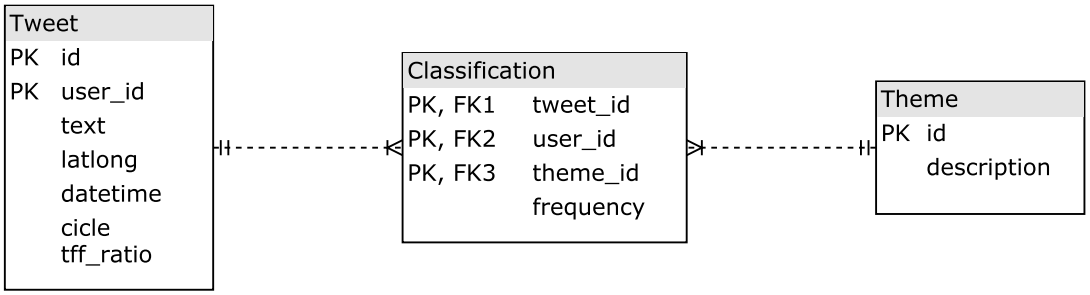


Figura 5.1: Esquema de banco de dados

base. A Figura 5.1 apresenta o esquema de banco de dados criado no PostgreSQL. O ponto principal desse esquema é a tabela *Classification*, que possibilita armazenar a classificação dos *tweets* de acordo com um ou mais temas, onde a coluna *frequency* armazena o número de palavras daquele *tweet* que estão relacionadas ao tema em que ele foi classificado.

Com o objetivo de selecionar *tweets* de usuários que tenham um grau de influência relevante, uma filtragem é realizada para que o *tweet* seja utilizado na classificação. Para determinar a influência de um usuário, nesse estudo de caso é utilizado o *tff* (*Twitter Follower-Followee Ratio*), que reflete a relação *followers*/*following* de um usuário para determinar sua popularidade [Bigonha, 2012].

De acordo com Krishnamurthy et al. [2008] e Leavitt et al. [2009], se a relação se aproxima do infinito (↑ *followers*, ↓ *following*) os usuários tendem a ser celebridades, fontes de notícias ou muito populares. De outra forma, se a relação se aproxima a 1 (*followers* $\simeq$ *following*) os usuários possuem reciprocidade, ou seja, são usuários comuns que utilizam a rede social no dia-a-dia. Outro caso, se a relação se aproxima de zero (↓ *followers*, ↑ *following*), os usuários podem ser classificados como fontes de *spam* ou robôs, ou seja, o número de seguidos é muito maior que de seguidores. Bigonha [2012] utiliza essa métrica combinada com outras para identificar usuários influentes determinando como os mais relevantes aqueles com *tff* mais elevado. Neste estudo de caso, usuários são considerados relevantes quando $0,2 \leq tff \leq 2,0$. São considerados usuários ideais, aqueles que possuem até 20% menos seguidores do que seguindo ou que possuem o dobro de seguidores. Esses valores foram definidos de maneira empírica a partir de observações em perfis de alguns usuários.

Por fim, com o intuito de facilitar o processo de classificação dos dados coletados, uma limpeza é realizada nos *tweets* (remoção de acentos, caracteres especiais, *stopwords*, *links* e emojis). Assim, um *tweet* é classificado a partir das palavras que estão contidas nele e que são palavras-chave relacionada ao tema. Além disso, para incluir o *tweet* no processamento de termos, apenas são aceitos comentários de usuários

considerados relevantes de acordo com a faixa de 20%.

5.3 Algoritmo de Processamento de Termos

Para o processamento de termos que não fazem parte do conjunto de palavras-chave, este estudo de caso utiliza um algoritmo TF-IDF. Neste caso, para calcular o TF utiliza-se o conceito apresentado por Sharifi et al. [2010], onde todos os *tweets* compreendem um único documento. Já para o IDF, considera-se que cada *tweet* é um único documento, pois o IDF apresentado por Sharifi et al. [2010] não se aplica ao nosso contexto, pois são baseados em sumários, onde consideram um documento como uma sentença. Deste modo, para determinar o peso de um termo, a formula usada é:

$$TF\_IDF = tf * \log_2 \frac{N}{1 + df} \tag{5.1}$$

onde tf é a frequência de um termo em um documento (todos os *tweets* como um documento), N é o número total de documentos (número total de *tweets*), e df é o número de documentos que contém o termo.

Após a classificação das bases, e utilizando o TF-IDF nos termos que não pertencem à base de palavras-chave, obtem-se um *ranking* de termos. Esse *ranking* tem como objetivo auxiliar na identificação de novas palavras-chave para a base, o número de termos não classificados é muito grande, como mostra a Tabela 5.3, o que era esperado. Então, para definir novas palavras-chave é necessário que seja feita manualmente a seleção de termos que não foram classificados, mas que estão relacionados ao tema. O TF-IDF auxilia colocando peso maior em termos que possuem uma frequência grande mas que não aparecem em todos os documentos.

A próxima etapa do estudo de caso é iniciar o processo iterativo da metodologia, que implica no refinamento das palavras-chave e no melhoramento da coleta de dados. Como o intuito deste trabalho é melhorar a coleta para assim adquirir melhores resultados em relação à caracterização de usuários, o processo iterativo de refinamento das palavras-chave e validação fazem parte dos resultados.

5.4 Resultados

Para continuar com o estudo de caso, é necessário iniciar o processo iterativo, ou seja, iniciar novos ciclos de coleta. A partir da aplicação do processamento de termos, novas palavras-chaves devem ser definidas. Porém, como observado anteriormente na

Tabela 5.3: Vocabulário das bases por temas do Ciclo 0

	Total de palavras				
Base de dados/Temas	Saúde	Transporte	Segurança	Limpeza	Nenhum
Saúde	232	5	51	19	19.172
Transporte	8	37	44	12	10.936
Segurança	18	6	436	30	28.415
Limpeza urbana	5	5	38	176	14.092

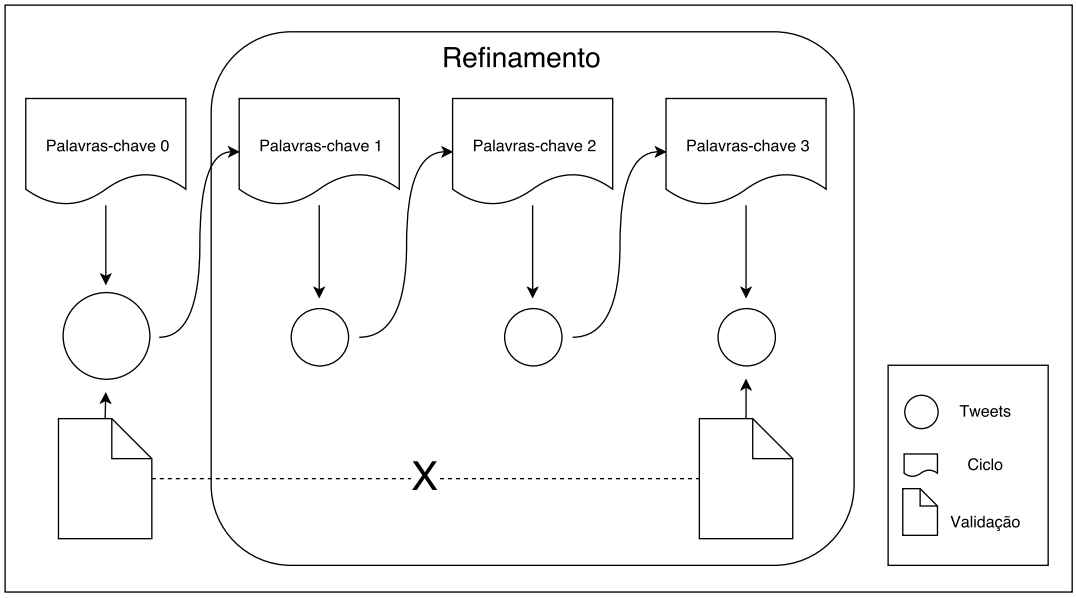


Figura 5.2: Processo de refinamento das palavras-chave

Tabela 5.3, o volume de vocabulário gerado por todos os temas é muito grande e há necessidade de aplicar filtros pós-coleta (desambiguar, excluir contextos não relevantes, etc.).

Com o intuito de diminuir esse volume e agilizar o processo de definição de novas palavras-chave, coleta, classificação e validação, para os próximos ciclos de coleta, apenas um tema relacionado ao contexto urbano for utilizado. Neste caso, o tema escolhido é *Segurança*, pois é o tema que obteve mais resultados na coleta.

5.4.1 Refinamento das Palavras-Chave

Como mencionado anteriormente, o refinamento das palavras-chave consiste em um processo iterativo de coleta de dados com o intuito de melhorar o conteúdo coletado. Para esse processo são utilizados três novos ciclos, onde cada ciclo possui um conjunto de palavras-chave, que com a coleta de dados, gera um novo conjunto de *tweets*, como mostra a Figura 5.2.

Termos	TF-IDF
vida	0,02266
venezuela	0,01842
trump	0,01826
onde	0,01612
temer	0,01113
sofrer	0,00904
eua	0,00704
fbi	0,00702
army	0,00638
armys	0,00617
avancado	0,00596
entra	0,00521
ditadura	0,00510
considera	0,00510
intervencao	0,00495
ruim	0,00479
rua	0,00464
poder	0,00461
sonhos	0,00461
coragem	0,00445
triste	0,00417
cair	0,00404
viver	0,00400
dor	0,00391
globo	0,00387

(a) Termos de 1 a 25

Termos	TF-IDF
ladrao	0,00384
lula	0,00378
cruz	0,00332
cogita	0,00325
soltar	0,00324
ditador	0,00304
civil	0,00300
doria	0,00298
usa	0,00291
recado	0,00289
forca	0,00282
coreia	0,00276
confundido	0,00262
sufocar	0,00259
libertados	0,00244
enorme	0,00241
traficante	0,00240
ansiedade	0,00235
fuzil	0,00231
mudar	0,00225
criminoso	0,00223
solucao	0,00221
nervosa	0,00216
jimin	0,00215
musica	0,00215

(b) Termos de 26 a 50

Tabela 5.4: *Top* 50 dos termos ponderados do Ciclo 0

O refinamento é iniciado com a definição de novas palavras-chave a partir da seleção manual de termos com o auxilio da ponderação utilizando o TF-IDF. A Tabela 5.4 apresenta o *Top* 50 dos termos não classificados do Ciclo 0. Como o vocabulário é muito grande, o número de *stopwords* também é grande. Assim, durante a aplicação da ponderação dos termos, diversas *stopwords* são removidas com o objetivo de limpar o *ranking* e facilitar a escolha das palavras-chave para o ciclo posterior. A partir do *ranking*, pode-se observar que alguns termos poderiam fazer parte da lista de novas palavras-chave, por exemplo, a palavra "criminoso". Outra observação que foge do contexto urbano, mas que chama a atenção, são os termos que se relacionam a uma notícia internacional do dia 11/08/2017 (trump, venezuela, coreia, army) que estão ligadas diretamente à notícia de ameaças do presidente norte-americano Donald Trump à Coreia do Norte, e também sobre uma possível ação militar na Venezuela.

Tabela 5.5: Palavras-chaves adicionadas por ciclo do tema Segurança

Ciclo	Palavras-chave
1	ladra, civil, traficante, fuzil, criminos
2	mort, denunci, violen, justica, crime, cadeia, delegad
3	morr, preso, federal, bandido, suspeito, arma, disparo, drogas

Tabela 5.6: Dados da coleta no processo de refinamento

Ciclo	Data de início	Data de fim	Tweets coletados
1	14/08/2017	15/08/2017	169.663
2	17/08/2017	18/08/2017	184.783
3	19/08/2017	20/08/2017	194.473

A partir dos *Top* 50 termos da Tabela 5.4, pode-se realizar uma seleção de novos termos relevantes para compor o banco de dados de palavras-chave e, assim, iniciar o processo iterativo. A Tabela 5.5 apresenta termos adicionados no banco de palavras-chave durante os ciclos. Os ciclos posteriores seguem a mesma ideia do Ciclo 0 de um processo de *stemming* de algumas dessas palavras para aumentar o campo de coleta. Os resultados mostram que, com a ajuda do processamento de termos utilizando o algoritmo da Seção 5.3, termos foram adicionados ao banco de palavras-chave. Para este estudo de caso decidiu-se manter todas as palavras-chave adicionadas anteriormente.

De acordo com as palavras-chave de cada ciclo, uma nova coleta de dados é realizada. No processo interativo, um ciclo depende do outro. Portanto, a coleta realizada no Ciclo 0 gera as palavras-chave do Ciclo 1 e, assim, sucessivamente. A Tabela 5.6 apresenta os dados da coleta de cada ciclo do processo de refinamento. Como mencionado na Seção 5.1, cada ciclo corresponde a, aproximadamente, um dia de coleta. Para o processo iterativo, apenas palavras-chave relacionadas ao tema Segurança foram utilizadas.

5.4.2 Avaliação e Validação dos Resultados

A validação é realizada em uma só etapa, utilizando a coleta do Ciclo 0 e a do Ciclo 3. Na Seção 4.4 foi mencionado que as etapas de pesagem e validação não dependem uma da outra. Como os ciclos possuem um período pequeno de coleta, seria mais eficiente se a validação fosse realizada ao final, pois além de avaliar os dados coletados, a validação avalia também a eficácia do processo de refinamento.

Um grupo de usuários são convidados a realizar o processo de validação. Esses usuários não possuem um perfil específico, são usuários das redes sociais Twitter, Face-

(a) Tela principal (b) Sobre a avaliação (c) Avaliação

Figura 5.3: *Screenshots* do *website* de avaliação

book e WhatsApp[3], usadas como meio de divulgação para atingir o máximo de pessoas possíveis.

A avaliação é realizada através de um *website*[4] divulgado via *link* através das redes sociais. A Figura 5.3 apresenta as três principais telas de interação com o usuário. Na tela principal (Figura 5.3a e 5.3b) encontra-se o título do *website* e logo em seguida uma breve descrição de como a avaliação funciona. A avaliação é composta por um conjunto de *tweets*, aos quais o usuário dá uma nota de 0 até 5, indicando se concorda que o *tweet* está relacionado ao tema Segurança. Assim que o voluntário entra na tela de avaliação (ver Figuras 5.3c) é feita uma pergunta: "Quão relacionado está esse *tweet* com o tema Segurança?" (0 - totalmente fora do tema e 5 - totalmente dentro do tema). No instante em que o botão "Próximo" é clicado, um novo *tweet* aparece para ser avaliado, e assim sucessivamente. Um usuário pode avaliar quantos *tweets* puder ou achar necessário. Para finalizar a avaliação o usuário necessita apenas fechar o *website*. A avaliação ficou disponível por 14 dias (06/02/2018 até 20/02/2018) e foi acessado por de 43 usuários.

Para que seja possível a avaliação e posteriormente validação dos resultados, é utilizada uma amostragem de aproximadamente 5% de cada base no caso do Ciclo 0. Como as bases possuem tamanhos diferentes, a base de amostragem ficou dividida em

[3]https://www.whatsapp.com
[4]https://guilhermevezula.000webhostapp.com/

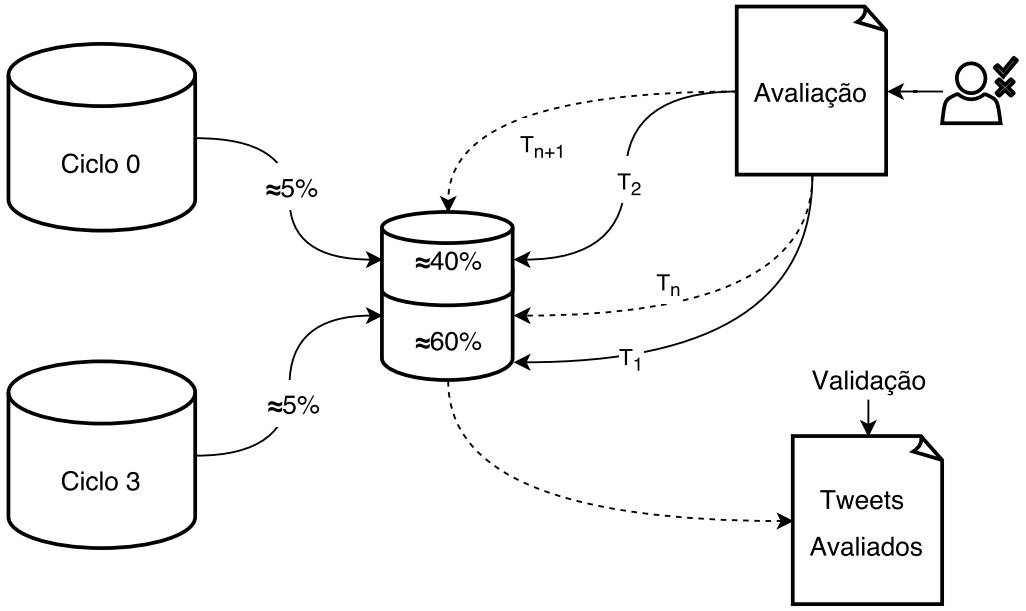


Figura 5.4: Esquema de validação dos tweets

Tabela 5.7: Dados da etapa de validação

Ciclo	Amostragem	Tweets avaliados	Avaliações	(%)
0	4458	469	504	≈11,3
3	6562	500	520	≈7,9

aproximadamente 40% e 60% para o Ciclo 0 e Ciclo 3 respectivamente. Para que os *tweets* fossem igualmente avaliados, considerando que são selecionados aleatoriamente, um esquema foi criado para troca entre *tweets* de cada ciclo. Esse esquema é apresentado na Figura 5.4. Como a base do Ciclo 3 é maior, quando um usuário inicia uma avaliação, o primeiro *tweet* avaliado é do Ciclo 3, e a partir disso, os *tweets* vão sendo alternados entre os ciclos.

A Tabela 5.7 apresenta a quantidade de *tweets* utilizados para a etapa de validação, bem como a quantidade de tweets avaliados pelos voluntários. Antes de selecionar a amostragem, uma filtragem foi realizada de acordo com os usuários relevantes (Seção 5.2) e *tweets* que realmente estão relacionados ao tema Segurança, totalizando 89.169 *tweets* do Ciclo 0 e 131.124 *tweets* do Ciclo 3. O número de avaliações no processo de validação atingiu 969 *tweets*, avaliados em 14 dias. A partir deste ponto, os dados são analisados para verificar a relevância da metodologia apresentada neste trabalho. Assim, a próxima seção apresenta os resultados dessas análises.

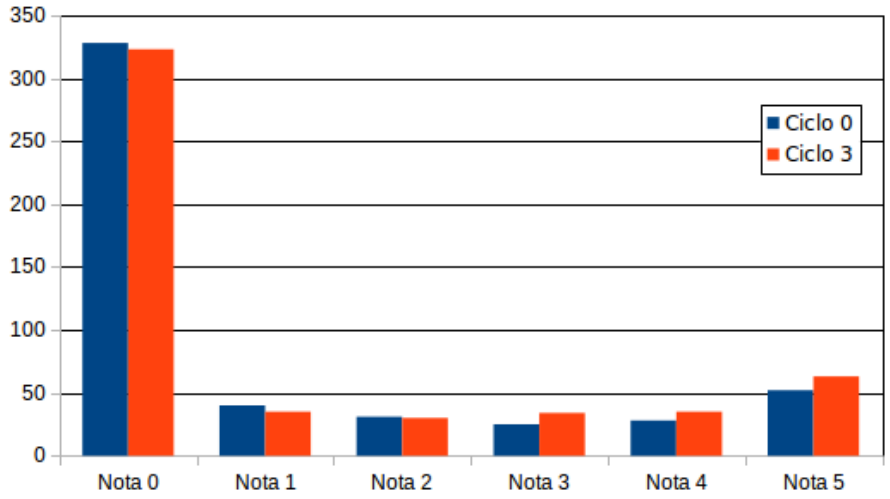


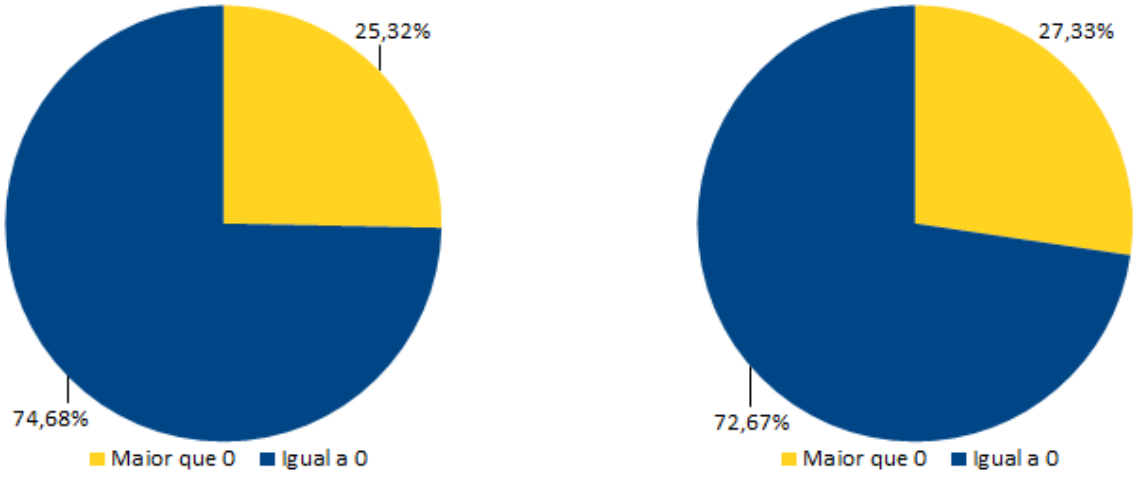
Figura 5.5: Resultado das avaliações separadas por nota

5.4.3 Análise

A primeira análise realizada é do resultado geral das avaliações para cada ciclo. Neste caso, são consideradas as 1.024 avaliações realizadas nos 969 *tweets*, sendo 504 e 520 para os Ciclos 0 e 3 respectivamente (Tabela 5.7). A Figura 5.5 apresenta o resultado das avaliações separando-as por notas. Observa-se que, em ambos os ciclos, muitos *tweets* foram avaliados como 0 (não estão relacionados ao tema), sendo que a diferença numérica entre *tweets* dos Ciclos 0 e 3 é pequena. O mesmo ocorre para avaliações com notas 1 e 2. Porém, *tweets* do Ciclo 3 com notas 3 e acima têm uma melhora em relação ao Ciclo 0.

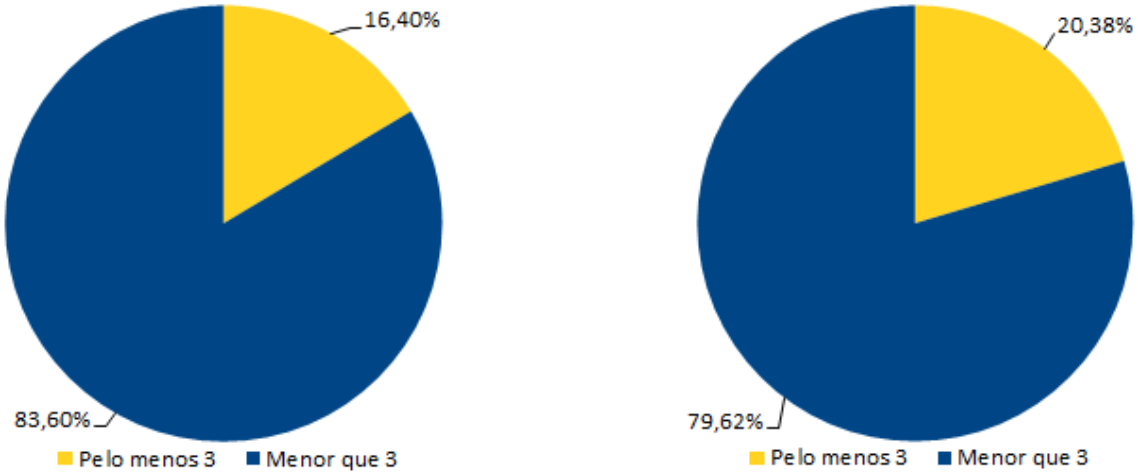
Após uma análise nos comentários avaliados com nota 0, observa-se que existem palavras em comum. Na maioria dos casos em que *tweets* obtiveram nota 0, as palavras mais comuns foram "guarda", "medo" e "morte". Desse modo, são palavras-chave definidas inicialmente para a primeira coleta (Tabela 5.1), como foi definido como Ciclo 0. Como dito anteriormente, durante o processo iterativo nenhuma palavra-chave era removida, apenas acrescentava-se novas. Portanto, a não remoção de termos genéricos no processo iterativo trouxe tweets que não estão relacionados com o tema.

As próximas análises são realizadas a partir da quantidade de *tweets* avaliados. Para complementar os resultados apresentados na Figura 5.5, a Figura 5.6 apresenta a percentagem de *tweets* com notas maiores que 0 para o Ciclo 0 e Ciclo 3 (Figura 5.6a e Figura 5.6b, respectivamente). Neste caso, é considerada a percentagem, pois as duas bases possuem tamanhos diferentes. Assim, é possível uma comparação mais precisa dos resultados. Deste modo, observa-se que a diferença entre os dois ciclos de coleta é de quase 2%. Uma diferença relativamente pequena, mas considerando a quantidade



(a) Ciclo 0 (b) Ciclo 3

Figura 5.6: *Tweets* com notas acima de 0 por ciclo



(a) Ciclo 0 (b) Ciclo 3

Figura 5.7: *Tweets* avaliados com nota pelo menos 3

onde o número de avaliações do Ciclo 3 foram menores que a do Ciclo 0, as diferenças foram relevantes.

Para a última análise são considerados os *tweets* avaliados com notas a partir de 3. Assim, a Figura 5.7 mostra um aumento de quase 4% em relação aos dois ciclos, sendo que a percentagem maior é do Ciclo 3. Esses resultados são complementares aos apresentados pela Figura 5.5. Ou seja, nas três análises realizadas, observa-se uma melhora do Ciclo 3 em relação ao Ciclo 0. Portanto, mostra que a metodologia proposta neste trabalho melhora o processo de coleta e a qualidade dos comentários.

Após a aplicação das avaliações com o intuito de validar a qualidade da coleta e

dos dados coletados, obtem-se um conjunto de *tweets* que estão relacionados ao tema Segurança. Assim, os usuários que os publicaram podem ser considerados potenciais contribuintes para aplicações de *crowdsourcing*/*crowdsensing* ativa.

Capítulo 6

Considerações Finais

Este trabalho foi desenvolvido com o objetivo de identificar usuários das redes sociais a partir de suas preferências implícitas. Para isso, foi utilizado um processo iterativo de coleta de dados para melhorar a qualidade dos dados utilizando um algoritmo de pesagem de termos para facilitar a escolha de novas palavras-chave, além de validar esses dados através de um processo avaliativo com pessoas reais.

Na metodologia foi proposto que a validação fosse realizada em cada nova iteração (ciclo), porém não foi possível realiza-lá pois no estudo de caso proposto seria necessário ter pessoas que pudessem validar um conjunto de *tweets* todos os dias, em ciclos diferentes, até o final do processo iterativo. Assim, foi considerado que a melhor maneira seria realizar a validação ao final de todas as iterações.

Também não foi possível realizar neste trabalho a validação das preferências dos usuários. Os resultados mostram que uma parte dos *tweets* coletados estão relacionados ao tema Segurança, porém não é possível afirmar que usuários, por postarem comentários relacionados a um contexto temático, teriam nesse tema uma preferência pessoal. Ou seja, obteve-se um indicativo de que os usuários poderiam ser potenciais contribuintes em *crowdsourcing*.

6.1 Conclusão

Mesmo com a dificuldade de execução de alguns processos, como a validação por ciclo e a não identificação das preferências dos usuários, os resultados se mostraram positivos. De modo geral, os resultados mostram que os *tweets* coletados no Ciclo 3 foram melhor avaliados em relação ao Ciclo 0. Ao considerar o total de avaliações realizadas (504 para o Ciclo 0 e 520 para o Ciclo 3). Os resultados mostram que com o processo iterativo a coleta teve uma melhora, e os *tweets* estavam mais relacionados ao tema.

Porém, se durante o processo iterativo fossem analisados termos considerados genéricos, a qualidade entre os ciclos 0 e 3 poderia ser maior.

Portanto, as contribuições centrais desta dissertação são a taxonomia apresentada no Capítulo 3, que demonstra os principais desafios para aplicações *crowdsourcing*, e a metodologia proposta no Capítulo 4, que demonstra, com o estudo de caso proposto, que é possível considerar um processo iterativo para melhorar a coleta de dados, e consequentemente, melhorar a qualidade dos dados. Outra contribuição deste trabalho foi a utilização de *crowdsourcing* passivo (CSOP), que demonstra o processo de coleta e classificação das bases a partir de comentários do Twitter.

6.2 Trabalhos Futuros

Uma proposta de trabalho futuro é realmente efetivar as preferências temáticas dos usuários. A ideia é coletar mais informações de um usuário, por exemplo, *hashtags* publicadas, curtidas de um *tweet*, *retweets* e a *timeline* que compõem de todos os comentários realizados por ele. Especificamente, a *timeline* pode ser coletada e assim classificada a partir de palavras-chave, aumentando a possibilidade do usuário estar interessado em um contexto temático. A aplicação dessa ideia no presente trabalho ficou dificultada, pois a coleta da *timeline* utilizando a parte gratuita da API do Twitter é limitada. O número de aquisições com múltiplos usuários é limitada, além de não permitir o acesso completo à *timeline* dos usuários.

Outra proposta de trabalho futuro é automatizar o processo de escolha de palavras-chave, que neste trabalho ainda foi manual, mesmo com o auxilio do algoritmo TF-IDF para pesagem dos termos. A ideia é utilizar técnicas de processamento de linguagem natural e ontologias para que esse processo seja mais automatizado que manual. Assim, mais palavras-chave podem ser identificadas durante o processo iterativo, e mais coletas específicas podem ser realizadas.

Por fim, muitos ainda são os desafios para aplicações relacionadas à contribuição voluntária por cidadãos. Espera-se que, cada vez mais, cidadãos se interessem a contribuir com assuntos que sejam relacionados ao seu dia-a-dia e, assim, contribuir junto aos representantes no governo para solucionar problemas do cotidiano urbano.

Referências Bibliográficas

Araújo, M.; Gonçalves, P. & Benevenuto, F. (2013). Measuring sentiments in online social networks. Em *Proceedings of the 19th Brazilian symposium on Multimedia and the web*, pp. 97--104. ACM.

Ballatore, A.; McArdle, G.; Kelly, C. & Bertolotto, M. (2010). Recomap: An interactive and adaptive map-based recommender. Em *25th ACM Symposium on Applied Computing (SAC)*, pp. 887--891.

Barron, J. P. G.; Manso, M. A.; Alcarria, R. & Gomez, R. P. (2014). A mobile crowdsourcing platform for urban infrastructure maintenance. Em *8th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, pp. 358--363.

Bearden, M. J. (2007). The national map corps. Em *Workshop on Volunteered Geographic Information, University of California*, pp. 13--14, Santa Barbara, California, USA.

Berjani, B. & Strufe, T. (2011). A recommendation system for spots in location-based online social networks. Em *4th Workshop on Social Network Systems*, pp. 4--9.

Bigonha, C. A. S. (2012). Detecção de influência no Twitter baseada em sentimentos. Dissertação de mestrado, Universidade Federal de Minas Gerais, Belo Horizonte, MG, BRA.

Bimber, B.; Flanagin, A. J. & Stohl, C. (2005). Reconceptualizing collective action in the contemporary media environment. *Communication Theory*, 15(4):365.

Blei, D. M.; Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993--1022.

Brandão, M. A. & Moro, M. M. (2016). Social professional networks: A survey and taxonomy. *Computer Communications*, 100:20–31.

Budhathoki, N. R.; Nedovic-Budic, Z. et al. (2008). Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal*, 72(3-4):149--160.

Castro, R.; Kuffó, L. & Vaca, C. (2017). Back to #6d: Predicting Venezuelan states political election results through Twitter. Em *4th International Conference on eDemocracy & eGovernment (ICEDEG)*, pp. 148--153.

Catae, F. S. (2012). Classificação automática de texto por meio de similaridade de palavras: Um algoritmo mais eficiente. Dissertação de mestrado, Universidade de São Paulo, São Paulo, SP, BRA.

Chatfield, A. & Brajawidagda, U. (2014). Crowdsourcing hazardous weather reports from citizens via twittersphere under the short warning lead times of ef5 intensity tornado conditions. Em *47th Hawaii International Conference on System Science (HICSS)*, pp. 2231--2241.

Chen, Z.; Fu, R.; Zhao, Z.; Liu, Z.; Xia, L.; Chen, L.; Cheng, P.; Cao, C. C.; Tong, Y. & Zhang, C. J. (2014). gMission: A general spatial crowdsourcing platform. *40th International Conference on Very Large Databases*, 7(13):1629--1632.

Chowdhury, M.; Patwary, K. H.; Imteaj, A. & Chowdhury, S. (2014). Designing a wireless sensor network using smartphone as data source. Em *9th International Forum on Strategic Technology (IFOST)*, pp. 166–169.

Cilliers, L. & Flowerday, S. (2014). Information security in a public safety, participatory crowdsourcing smart city project. Em *2nd IEEE World Congress on Internet Security (WorldCIS)*, pp. 36--41.

Craglia, M. (2007). Volunteered geographic information and spatial data infrastructures: when do parallel lines converge. Em *Position paper for the VGI Specialist Meeting*, pp. 13--14, Santa Barbara, California, USA.

Davis Jr, C. A.; Vellozo, H. S. & Pinheiro, M. B. (2013). A framework for web and mobile volunteered geographic information applications. Em *14th Brazilian Symposium on Geoinformatics (GeoInfo)*, pp. 147--157.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W. & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391--407.

Elmessiry, A.; Cooper, W. O.; Catron, T. F.; Karrass, J.; Zhang, Z. & Singh, M. P. (2017). Triaging patient complaints: Monte Carlo cross-validation of six machine learning classifiers. *JMIR Medical Informatics*, 5(3):e19.

Gao, H.; Tang, J.; Hu, X. & Liu, H. (2015). Content-aware point of interest recommendation on location-based social networks. Em *29th AAAI Conference on Artificial Intelligence*, pp. 1721--1727.

Garton, L.; Haythornthwaite, C. & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1):0.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Gupta, G. & Lee, W. (2010). Collaborative spatial object recommendation in location based services. Em *39th International Conference on Parallel Processing (ICPP) Workshops*, pp. 24--33.

Han, J.; Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann.

Hardy, D. (2007). Digital commons and the state of our environment. Em *Workshop on Volunteered Geographic Information*. Disponível em: <`http://www.ncgia.ucsb.edu/projects/vgi/docs/position/hardy_paper.pdf`>. Acessado em: 08 mar. 2016.

Hirata, E.; Giannotti, M.; Larocca, A. & Quintanilha, J. (2015). Flooding and inundation collaborative mapping – use of the CrowdMap/Ushahidi platform in the city of São Paulo, Brazil. *Journal of Flood Risk Management*, 11(1):98–109.

Hu, L.; Tong, Q.; Du, Z.; Liu, Y. & Tang, Y. (2014). Location-based service using ontology and collaborative recommendation. Em *2014 International Conference on Information Science, Electronics and Electrical Engineering (ISEEE)*, pp. 653--656.

Kido, G. S.; Junior, S. B. & Moriguchi, S. N. (2014). Comparação entre TF-IDF e LSI para pesagem de termos em micro-blog. Em *10th Simpósio Brasileiro de Sistemas de Informação*. Disponível em: <`https://www.researchgate.net/publication/262804841_Comparacao_entre_TF-IDF_e_LSI_para_pesagem_de_termos_em_micro-blog`>. Acessado em: 17 fev. 2017.

Klinczak, M. N. M. & Kaestner, C. A. (2017). Identificação de temas em redes sociais por meio de técnicas de agrupamento. *Anais do Computer on the Beach*, pp. 090--099.

Koswatte, S.; McDougall, K. & Liu, X. (2015). SDI and crowdsourced spatial information management automation for disaster management. *Survey Review*, 47(344):307–315.

Krishnamurthy, B.; Gill, P. & Arlitt, M. (2008). A few chirps about Twitter. Em *1st Workshop on Online Social Networks*, pp. 19--24.

Lane, N. D.; Miluzzo, E.; Lu, H.; Peebles, D.; Choudhury, T. & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine*, 48(9):140–150.

Leavitt, A.; Burchard, E.; Fisher, D. & Gilbert, S. (2009). The influentials: New approaches for analyzing influence on Twitter. *Web Ecology Project*, 4(2):1--18.

Lee, D. D. & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. Em *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pp. 535--541, Cambridge, MA, USA. MIT Press.

Lin, C.-Y. & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. Em *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 71--78.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Maguire, D. J. (2007). Geoweb 2.0 and volunteered GI. Em *Workshop on Volunteered Geographic Information*. Disponível em: <`http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Maguire_paper.pdf`>. Acessado em: 08 mar. 2016.

Mateveli, G. V.; Machado, N. G.; Moro, M. M. & Jr, C. A. D. (2015). Taxonomia e desafios de recomendação para coleta de dados geográficos por cidadãos. Em *30th Simpósio Brasileiro de Bancos de Dados (SBBD)*, pp. 105--110.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P. & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. Em *7th ACM SIGCOMM conference on Internet measurement*, pp. 29--42.

Mytilinis, I.; Giannakopoulos, I.; Konstantinou, I.; Doka, K.; Tsitsigkos, D.; Terrovitis, M.; Giampouras, L. & Koziris, N. (2015). Modissense: A distributed spatio-temporal and textual processing platform for social networking services. Em *2015 ACM SIGMOD International Conference on Management of Data (SIGMOD/PODS)*, pp. 895--900.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167--256.

Onsrud, H.; Camara, G.; Campbell, J. & Chakravarthy, N. S. (2004). Public commons of geographic data: Research and development challenges. Em *International Conference on Geographic Information Science*, pp. 223--238. Springer.

O'Reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. Em *MPRA Paper*. Disponível em: <`https://ideas.repec.org/p/pra/mprapa/4578.html`>. Acessado em: 8 mar. 2016.

Parshotam, K. (2013). Crowd computing: A literature review and definition. Em *South African Institute for Computer Scientists and Information Technologists Conference*, pp. 121--130.

Pat, B.; Kanza, Y. & Naaman, M. (2015). Geosocial search: Finding places based on geotagged social-media posts. Em *24th International Conference on World Wide Web*, pp. 231--234.

Santos, A. D. P. (2013). Descobrindo eventos locais utilizando análise de séries temporais nos dados do Twitter. Dissertação de mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, BRA.

Scott, J. (2017). *Social network analysis*. Sage Publications, 4th edition. ISBN 978-1-4739-5211-9.

Sharifi, B.; Hutton, M.-A. & Kalita, J. K. (2010). Experiments in microblog summarization. Em *2010 IEEE Second International Conference on Social Computing (SocialCom)*, pp. 49--56.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11--21.

Sui, D. Z. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, 32(1):1--5.

Vukovic, M. (2009). Crowdsourcing for enterprises. Em *2009 World Conference on Services - I*, pp. 686–692.

Wang, Q.; Bhandal, J.; Huang, S. & Luo, B. (2017). Classification of private tweets using tweet content. Em *11th IEEE International Conference on Semantic Computing (ICSC)*, pp. 65--68.

Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Wellman, B.; Salaff, J.; Dimitrova, D.; Garton, L.; Gulia, M. & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual review of sociology*, 22(1):213--238.

Ye, M.; Janowicz, K.; Mülligann, C. & Lee, W. (2011). What you are is when you are: The temporal dimension of feature types in location-based social networks. Em *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 102--111.

Zaman, A. & Brown, C. G. (2010). Latent semantic indexing and large dataset: Study of term-weighting schemes. Em *5th International Conference on Digital Information Management (ICDIM)*, pp. 1--4.

Zhang, W.; Yoshida, T. & Tang, X. (2011). A comparative study of TF-IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758--2765.

Zheng, Y. (2012). Tutorial on location-based social networks. Em *Proceedings of the 21st international conference on World wide web (WWW)*, volume 12, p. 0. Citeseer.

Zheng, Y.; Xie, X. & Ma, W. (2010). Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32--39.