Identificação da Reputação de Áreas Urbanas Externas com Dados de Mídias Sociais

Frances A. Santos[1,4], **Thiago H. Silva**[2], **Antonio A. F. Loureiro**[3], **Azzedine Boukerche**[4], **Leandro A. Villas**[1]

[1]Instituto de Computação, Universidade Estadual de Campinas. Campinas, Brasil
[2]Dep. de Informatica, Universidade Tecnológica Federal do Paraná. Curitiba, Brasil
[3]Dep. de Ciência da Computação, Univ. Federal de Minas Gerais. Belo Horizonte, Brasil
[4]SITE, University of Ottawa. Ottawa, Canadá

Abstract. *Learning people's perception that emerges from urban areas has been an interesting multidisciplinary research goal because it has a great potential to ease the hard task of understanding intrinsic characteristics of urban areas, e.g., their reputation. One common way to do that is by exploring traditional data collection approaches, e.g., interviews. However, traditional methods do not scale easily, difficulting the execution of this type of analysis for a large number of places. To overcome this challenge, we propose an alternative method that explores data from location-based social networks (LBSNs), where a large number of users act as social sensors sharing a considerable amount of opinions about urban areas. Our novel method, namely REP-Map, supports the uncovering and mapping of the reputation of urban outdoor areas. REP-Map explores spatial and semantic aspects in messages shared on LBSNs to identify significant reputation of outdoor areas. Studying outdoor areas of Chicago, we show, through a survey with volunteers, that our method has the potential to correctly capture the reputation that users consider regarding those areas.*

Resumo. *Aprender a percepção das pessoas que emerge das áreas urbanas tem sido um objetivo de pesquisa multidisciplinar, pois oferece um grande potencial para facilitar a difícil tarefa de compreender as características intrínsecas das áreas urbanas, por exemplo, sua reputação. Para isso, comumente, são exploradas abordagens tradicionais de coleta de dados, como entrevistas. No entanto, tais métodos não escalam facilmente, dificultando a execução desse tipo de análise para um grande número de lugares. Para superar esse desafio, propomos um método alternativo que explora dados de redes sociais baseadas em localização (LBSNs). O nosso método inovador, chamado de REP-Map, trata da descoberta e mapeamento da reputação das áreas urbanas externas, explorando aspectos semânticos e espaciais em mensagens compartilhadas em LBSNs. Estudando áreas externas de Chicago, mostramos, através de uma pesquisa com voluntários, que nosso método pode capturar a reputação que os usuários consideram em relação a essas áreas urbanas externas.*

1. Introdução

Áreas urbanas externas, tais como parques, praças e bosques, podem oferecer às pessoas a oportunidade de terem experiências diversas, como relaxar depois do trabalho, brincar com crianças ou animais de estimação, exercitar e assim por diante. Com isso, a avaliação da mesma área por diferentes pessoas pode desencadear diferentes percepções entre seus visitantes. Por exemplo, lugares com atrações turísticas podem ser apreciados pelos turistas, mas podem ser considerados muito cheios e barulhentos pelos residentes ao realizar suas rotinas diárias. Nesse sentido, entender a reputação das áreas urbanas, o que poderia melhorar sua descrição e alavancar novos serviços e aplicações, é uma tarefa difícil.

Reputação é definida pelo dicionário de Oxford como: "As crenças ou opiniões que geralmente são mantidas sobre alguém ou alguma coisa". Essa definição está alinhada com a visão de diferentes pesquisadores, variando da economia à sociologia [Marsden and Lin 1982, Weigelt and Camerer 1988]. Todos reconhecem que as reputações são, na verdade, percepções de algo mantido por observadores externos. Por essa razão, aprender sobre a percepção das pessoas que emerge das áreas da cidade é um passo importante para essa tarefa. Com isso, a fim de descobrir uma reputação de área, é necessário agregar corretamente as percepções expressas por diferentes usuários.

Pesquisas de campo e caminhadas sensoriais são abordagens tradicionais usadas para compreender a semântica intrínseca das áreas urbanas, sendo uma ferramenta poderosa para capturar a percepção detalhada das pessoas sobre os lugares [Quercia et al. 2015]. No entanto, essas estratégias podem consumir muito tempo, porque normalmente demandam uma alta quantidade de tempo de observação e entrevistas com os participantes para coletar uma quantidade considerável de amostras de percepção. Isso dificulta a realização desse tipo de análise para um grande número de locais. Para superar esse desafio, propomos um método alternativo que explora dados de redes sociais baseadas em localização (LBSNs), por exemplo, Foursquare, Flickr, Instagram e Twitter.

Explorar LBSNs para essa tarefa é interessante por diversas razões. Uma das principais é que um grande número de usuários atuam como sensores sociais compartilhando uma quantidade considerável de conteúdo, incluindo suas opiniões, sobre áreas urbanas. Além disso, o conteúdo é rico, podendo ter diferentes significados, variando de publicidade comercial a comentários gerais sobre quase tudo. No entanto, embora as LBSNs possam oferecer uma quantidade enorme de dados, o que potencialmente pode ajudar no problema de escalabilidade dos métodos tradicionais para coletar as percepções urbanas, a sua exploração para isso não é trivial, pois a extração de características urbanas úteis expressadas pelos usuários nesses sistemas é uma tarefa desafiadora. Apesar disso, estudos anteriores (descritos na Seção 2) encontraram evidências de que é possível extrair do conteúdo de LBSNs percepções relevantes dos usuários em relação às áreas urbanas.

Com isso, uma questão importante que orienta o nosso estudo é: podemos explorar o conteúdo compartilhado por usuários em LBSNs para identificar reputações significativas de áreas urbanas externas? Embora existam iniciativas que mostrem que é possível extrair a percepção urbana das mídias sociais em domínios específicos, no melhor do nosso conhecimento, ainda falta na literatura uma abordagem para descobrir reputações de áreas. Para este fim, este estudo propõe um novo arcabouço, descrito na Seção 3, que considera que a reputação das áreas urbanas externas é formada pela combinação de percepções fornecidas por pessoas através de mensagens, onde uma área urbana pode ter reputações heterogêneas simultaneamente. Para isso, usamos dados de uma LBSN popular para descobrir e mapear a reputação de áreas urbanas externas, etapas discutidas na Seção 4.1.

Nossas principais contribuições podem ser resumidas em:

- Dicionário hierárquico, chamado REP-dicionário, que contém as principais palavras utilizadas pelas pessoas para qualificar suas experiências em áreas urbanas ao ar livre das cidades, sendo útil para extrair a percepção urbana do usuário dos dados de LBSN;
- Um algoritmo de agrupamento não-supervisionado para identificar o conteúdo compartilhado pelos usuários que são similares espacialmente e semanticamente;
- Validação dos resultados. Ao estudar a cidade de Chicago, nos Estados Unidos, mostramos que os dados compartilhados por usuários do Twitter (escritos em inglês) trazem valiosas informações sobre as áreas urbanas e podem ser usados para extrair suas reputações, ajudando a entender melhor essas áreas em vários aspectos. Nós validamos a reputação extraída de diferentes áreas urbanas externas realizando uma pesquisa com voluntários,

indicando que a nossa abordagem apresenta resultados praticamente iguais aos apontados pelos usuários (resultados mostrados na Seção 4.2).

Finalmente, na Seção 5, concluímos o estudo e apresentamos direções futuras.

2. Trabalhos Relacionados

Na literatura, é possível encontrar propostas que visam extrair características importantes em relação às áreas urbanas com base na percepção expressada por pessoas de diferentes maneiras, incluindo o modo *offline*, por exemplo, através de passeios sensoriais, ou *online*, por exemplo, através de dados de redes sociais. Apesar das fontes *offline* fornecerem dados bem específicos e ricos de participantes envolvidos e focados em tarefas de sensoriamento, elas não escalam facilmente [Quercia et al. 2015]. Para superar esse problema, surgiram propostas de extração de características urbanas que exploram o conhecimento de multidões compartilhadas em fontes *online* sobre semânticas que as áreas urbanas possuem, com o intuito de mapear a percepção dos usuários para diferentes fins. As fontes *online* mais comumente utilizadas são mídias sociais, principalmente as redes sociais baseadas em localização (LBSNs), que demonstraram ser úteis para a compreensão e resolução de diferentes problemas [Cranshaw et al. 2012, Quercia et al. 2014b, Quercia et al. 2015, Silva et al. 2013].

Uma abordagem comum para o mapeamento da percepção dos usuários é explorar modelos de tópicos, como *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003]. Dada uma grande coleção de documentos, por exemplo, *tweets*[1], contendo palavras que descrevem as percepções dos usuários sobre diferentes tipos de conteúdos, esses modelos têm sido usados para extrair a estrutura temática oculta, os chamados tópicos. Com base nos tópicos descobertos usando documentos de fontes *online*, as áreas urbanas podem ser interpretadas de maneiras diferentes, como de acordo com atividades socioculturais [Steiger et al. 2016], funcionalidades [Yuan et al. 2015], ou pontos de interesse [Jiang et al. 2016].

Ao visitar áreas urbanas, sentimentos podem florescer nas pessoas, potencialmente incentivando-as a compartilhar sua percepção da área ao seu redor. Nesse sentido, várias abordagens exploram a polaridade do sentimento [Gonçalves et al. 2013] expressa nas mensagens dos usuários em mídias sociais para, por exemplo, a criação de mapas de sentimento de áreas urbanas [Flaes et al. 2016] e na proposição de serviços sensíveis ao sentimento, como o rastreamento da felicidade comunitária [Quercia et al. 2012], recomendação de locais [Yang et al. 2013] e recomendação de rotas [Kim et al. 2014].

Considerando que as experiências vividas por pessoas nas áreas urbanas são frutos de suas percepções sensoriais (ou seja, visão, audição, tato, paladar e olfato), existem alguns estudos focados em compreender os efeitos que as características visuais e olfativas presentes no ambiente urbano refletem sobre a percepção das pessoas. Quanto à percepção visual, Quercia et al. [Quercia et al. 2014a] usou uma abordagem de *crowdsourcing* para colecionar as percepções das pessoas sobre fotos tiradas nas ruas de Londres. Considerando como possíveis percepções as opções “belas”, “felizes”e “silenciosas”, os autores encontraram correlação positiva entre cores, textura e manchas visuais de fotos com as três percepções. Também com base em percepções visuais, em [Naik et al. 2014] é proposta uma abordagem computacional para predizer a segurança percebida das cidades explorando imagens de diferentes locais disponíveis para estudo.

Quanto aos efeitos de odores urbanos na percepção das pessoas, Henshaw [Henshaw 2013] conduziu sessões de caminhadas sensoriais (*smellwalks*), para colecionar a percepção dos cidadãos em relação ao cheiro em áreas urbanas distintas da cidade de Doncaster, Inglaterra. Como resultado, os odores urbanos foram classificados em 11 categorias, que foi complementada por

[1]Mensagens compartilhadas no Twitter por usuários.

[Quercia et al. 2015] depois de realizar *smellwalks* em outras cinco cidades dos EUA e da Europa. Isso permitiu a criação de um dicionário com palavras relacionadas ao cheiro urbano, que é usado para descobrir mensagens relacionadas a percepções de odor em dados de mídias sociais, para fornecer mapas de odor à escala da cidade sem depender de alto envolvimento de pessoas [Quercia et al. 2015].

O estado da arte mostra que LBSNs são uma fonte rica de dados, facilitando o processo de compreensão de vários aspectos das cidades. Neste trabalho, nós também aproveitamos essa valiosa fonte para estudar um novo aspecto das áreas urbanas: a reputação. Além disso, nós propomos uma abordagem escalável para realizar essa tarefa, que tem potencial para ser aplicada a várias áreas geográficas, uma vez que a criação do dicionário, a coleta da percepção das pessoas, e a extração da reputação das áreas urbanas não exige levantamentos de campo demorados ou uma grande quantidade de usuários envolvidos.

3. Metodologia

Nesta seção, descrevemos a metodologia do arcabouço proposto para descobrir e mapear a reputação de áreas externas a partir de percepções de pessoas expressas em textos livres compartilhados em LBSNs. Nosso arcabouço, ilustrado na Figura 1, consiste em cinco fases principais para atingir o seu objetivo. Em seguida, discutimos as principais etapas de nossa metodologia.

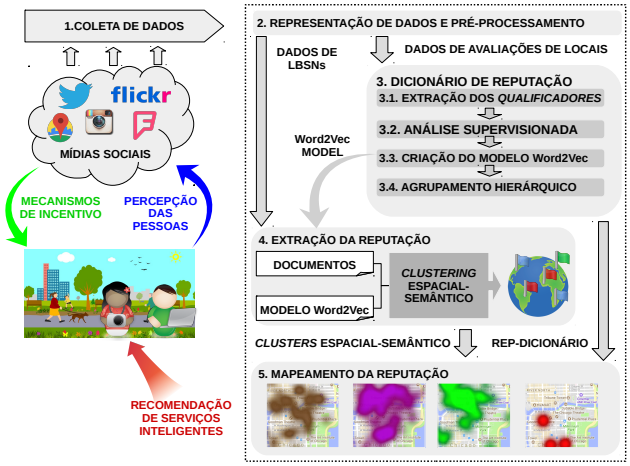

Figura 1. Visão geral do arcabouço para extração de reputações.

3.1. Coleta de Dados

Sabendo que utilizar conteúdo coletado a partir de LBSNs para extrair percepções urbana é difícil, pois o conteúdo pode ser muito ruidoso, primeiro coletamos dados sobre **avaliações** de lugares. Alguns *sites* de mídias sociais, como Google Places, TripAdvisor e Foursquare, permitem aos usuários fazerem avaliações contendo suas opiniões pessoais sobre os lugares. Nesses canais, os usuários podem compartilhar críticas sobre qualquer local já disponível no sistema, por exemplo, um restaurante específico, a qualquer momento, não necessariamente enquanto ele está fisicamente presente no local formulando sua opinião. Normalmente, esses dados contêm a percepção detalhada dos usuários sobre os lugares e são semanticamente bem definidos. Por estas razões, esse tipo de dado é útil para aprender o vocabulário usado pelos usuários para expressarem suas percepções.

Para este fim, recolhemos avaliações públicas do **Foursquare** (Tips[2]) e **Google Places**. Os dois sites definem conjuntos discretos de categorias que especificam o tipo de lugares[3][4], o que

[2]As revisões de lugares neste sistema recebem o nome de *Tips*.
[3]Categorias do Foursquare: https://goo.gl/cSFas4.
[4]Categorias do Google Places: https://goo.gl/ACd1AT.

permite a seleção de apenas comentários sobre lugares ao ar livre. Apesar de ser uma amostra relativamente pequena em comparação a outros dados compartilhados em LBSNs, as avaliações trazem detalhes ricos sobre a percepção dos usuários, sendo importantes para alcançar nosso objetivo.

Com o intuito de testar nossa abordagem para extração de reputação, coletamos dados do Twitter, que é um serviço de *microblogging online*, onde os usuários podem, entre outras coisas, compartilhar mensagens curtas com tamanho máximo de 280 caracteres. Usando a API do Twitter é possível coletar *tweets* de áreas de interesse. Parte desse conjunto de *tweets* é geolocalizado.

Portanto, temos dois conjuntos de dados distintos: **(i) conjunto de dados de avaliações de locais**, que contém dados e metadados do Google Places e Foursquare (Tips), usado para aprender as principais palavras relacionadas à reputação de áreas urbanas externas e **(ii) conjunto de dados de LBSN**, que contém dados e metadados do Twitter, usado para mapear a reputação de lugares ao ar livre.

3.2. Representação de Dados e Pré-processamento

Para combinar dados e metadados das fontes *online* heterogêneas, cada documento, isto é, uma mensagem compartilhada por um usuário, deve ter uma sequência de caracteres contendo o texto livre escrito pelo usuário, um *timestamp* representando o momento em que o documento foi criado e uma *geotag* referente à localização geográfica do local ou localização atual do usuário. Mais formalmente, nossos dados podem ser representados da seguinte forma:

Definição 3.2.1. Um documento doc é determinado por uma tupla $doc = (id, s, \tau, g)$, onde id é um identificador único, s é uma lista de frases que compreende todo o conteúdo informado pelo usuário, $\tau \in \mathbb{R}$ é um *timestamp* e $g \in \mathbb{R} \times \mathbb{R}$ é uma *geotag* que denota as coordenadas geográficas, expressas por latitude e longitude.

Desta forma, nossos conjuntos de dados são representados por duas coleções de documentos conforme definido acima. Em seguida, precisamos executar um passo de pré-processamento nas frases de cada documento para transformar as frases brutas (ou seja, texto original fornecido pelo usuário) em um vetor de palavras. Para isso, removemos das frases tudo que não forma uma palavra, ou seja, números, *Uniform Resource Locators* (URLs), caracteres especiais e pontuação. Em seguida, dividimos cada frase em um vetor de palavras, ou seja, *tokens*, utilizando expressão regular. Além disso, também removemos os *tokens* que são considerados ruídos, como palavras irrelevantes (*stop words*), palavras com apenas um ou dois caracteres, e palavras relacionadas com anúncios de trabalho.

3.3. Dicionário de Reputação

Com base na coleção de documentos $\mathcal{D}_R$ com as avaliações de pessoas sobre lugares, ou seja, o conjunto de dados de avaliações de locais, estamos interessados em descobrir quais palavras são frequentemente usadas por pessoas para qualificar suas experiências em áreas urbanas externas e, assim, construir um dicionário abrangente de reputação, que é chamado de REP-dicionário.

Para este fim, realizamos a classificação[5] da categoria gramatical em cada frase $s \in doc, \forall doc \in \mathcal{D}_R$, atribuindo a provável classe gramatical entre substantivo, adjetivo, verbo, pronome, etc, para cada palavra de s. Depois de rotular todas as palavras, extraímos o conjunto de palavras consideradas qualificadoras (ou seja, rotuladas como adjetivo) e que ocorreram pelo menos 20 vezes em $\mathcal{D}_R$ (como avaliado empiricamente, esse é um limite relativamente baixo para excluir qualificadores incomuns). Depois disso, pedimos a três supervisores para analisar, de

[5]Usando o *Tagger Perceptron*: http://www.nltk.org/api/nltk.tag.html.

forma independente, esse conjunto de palavras para gerar um subconjunto que efetivamente qualifica uma área urbana externa, de acordo com a opinião deles. Em seguida, combinamos os três conjuntos de palavras, onde os qualificadores restantes devem estar em pelo menos dois deles. Como resultado, obtemos um conjunto contendo 88 palavras em inglês. Essas palavras expressam adjetivos presentes em nosso conjunto de dados e que são comumente usados pelos usuários para qualificar áreas urbanas externas. Ao empregar nossa metodologia em outros conjuntos de dados contendo avaliações de usuários, temos a possibilidade de enriquecer o nosso dicionário com novas palavras, uma vez que o idioma está em constante evolução.

Depois disso, organizamos as palavras do conjunto obtido em categorias, de acordo com a similaridade sintática e semântica delas, usando para isso o modelo Word2Vec. O Word2Vec é um modelo baseado em rede neural usado para aprender as representações vetoriais de palavras que contêm muitas regularidades linguísticas e padrões [Mikolov et al. 2013]. O modelo recebe como entrada um *corpus* (ou seja, todas as frases em documentos de $\mathcal{D}_R$), um tamanho de janela ws, uma contagem mínima de ocorrência da palavra *minCount* e um hiper-parâmetro m que representa o número de características. Em seguida, o modelo cria um vocabulário de n palavras únicas, denotado por W, a partir de *corpus*, onde cada palavra $w \in W$ deve ocorrer pelo menos *minCount* vezes em *corpus*. Ao usar um método de aprendizado profundo com uma única camada oculta, o modelo *skip-gram* ou o modelo *Continuous Bag-Of-Words* (CBOW), e alguma função de ativação somente nos neurônios de saída (não em neurônios da camada oculta), *softmax* hierárquico ou amostragem negativa, o modelo Word2Vec calcula para cada par de palavras $w_1, w_2 \in W : w_1 \neq w_2$, a probabilidade de encontrá-las "próximas" em frases de *corpus*. Duas palavras são consideradas próximas se tiverem no máximo $ws - 1$ posições entre elas.

Como existem n palavras únicas e m características na camada oculta, após o treinamento, o modelo Word2Vec produz $n \times m$ pesos na camada oculta, que são chamados de vetores de palavras. Tais vetores, um para cada palavra, permitem predizer se duas palavras diferentes têm contextos semelhantes. O modelo CBOW e a amostragem negativa tiveram melhor desempenho em nossos experimentos e, portanto, foram empregados em nosso arcabouço. Em seguida, usamos o agrupamento hierárquico para agrupar palavras de acordo com a semelhança de seus vetores de palavras. Ao experimentar diferentes critérios de lincagem (*linkage*) e métricas de similaridade/distância, descobrimos que o critério de lincagem "*complete*" e a similaridade do cosseno produziram um resultado melhor, por isso, mantivemos essa configuração.

Após esses passos obtivemos o que chamamos de REP-dicionário, representando o resultado do agrupamento hierárquico, onde as palavras estão em inglês, uma vez que o dicionário é utilizado para extrair reputações em documentos escritos em língua inglesa. Uma ilustração desse dicionário é mostrada na Figura 2. A forma como a imagem é organizada é baseada no dendrograma resultante do processo de agrupamento. Depois de realizar um corte no dendrograma, cada cor representa um agrupamento resultante que contém qualificadores com maior similaridade entre eles. Na visualização, escolhemos uma palavra que representa um determinado ramo (subgrupo) de uma categoria, por exemplo, o ramo *Fabulous*, *Brilliant* e *Excellent*, da categoria verde é representado pela palavra *Excellent*. Observe que a categoria verde também possui outros três ramos, representados por *Vibrant*, *Memorable* e *Iconic*. Para simplificar a rotulação da categoria, escolhemos apenas duas palavras que descrevem todos os ramos do grupo, caso haja mais. Com isso, o nome que representa a categoria verde é *Vibrant* & *Memorable*, em vez das quatro palavras.

Como podemos ver, existem sete categorias principais, que podem ser usadas para atribuir reputação a algum lugar ao ar livre. Este dicionário é um passo importante para descobrir a reputação de lugares externos com base em dados de LBSN ruidosos. Todas as palavras que compõem o REP-dicionário são apresentadas na Tabela 1. Apresentamos cada categoria a seguir

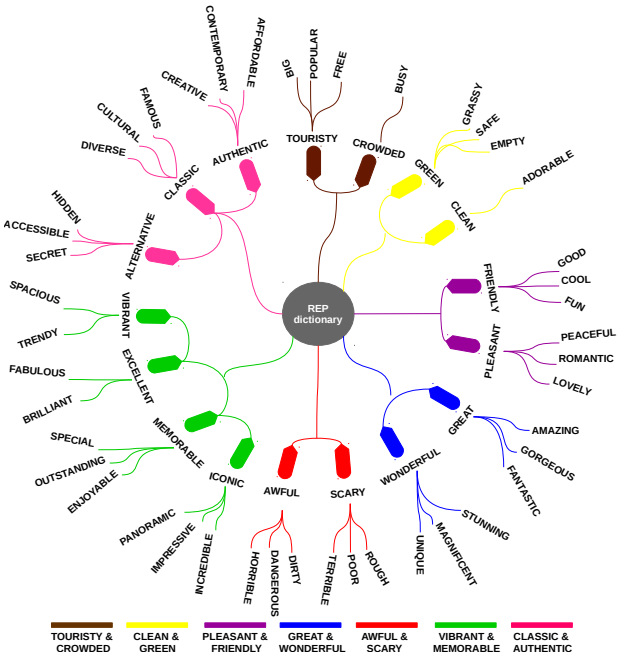

Figura 2. Ilustração do REP-dicionário.

Tabela 1. Palavras de cada categoria que compõem o REP-dicionário.

Rótulo da Categoria	Todas as Palavras da Categoria
Touristy & Crowded	*Big, busy, crowded, free, popular, pretty, touristy.*
Vibrant & Memorable	*Attractive, brilliant, enjoyable, excellent, fabulous, happy, iconic, impressive, incredible, memorable, outstanding, panoramic, rare, spacious, special, trendy, vibrant.*
Clean & Green	*Adorable, clean, empty, grassy, green, safe, shady.*
Pleasant & Friendly	*Convenient, cool, cute, different, friendly, fun, good, interesting, lovely, nice, peaceful, perfect, pleasant, quiet, relaxing, romantic, serene, tranquil.*
Awful & Scary	*Awful, cheap, dangerous, dirty, horrible, loud, poor, rough, scary, simple, terrible.*
Classic & Authentic	*Accessible, affordable, alternative, authentic, classic, contemporary, creative, cultural, diverse, famous, hidden, secret.*
Great & Wonderful	*Amazing, awesome, beautiful, calm, colorful, cozy, fantastic, gorgeous, great, magical, magnificent, scenic, spectacular, stunning, unique, wonderful.*

com base em uma investigação mais profunda das frases que as originaram.

A categoria *Touristy & Crowded* representa áreas com pontos turísticos. Além disso, esta categoria também representa áreas com elevado número de pessoas, que é comumente associada a áreas turísticas. Estudando a categoria *Vibrant & Memorable*, descobrimos que ela representa a percepção de áreas que podem impressionar seus visitantes, por exemplo, com belezas naturais e feitas pelo homem, que são ricas em características visuais e boas para tirar fotos, apreciar a paisagem e contemplação. A categoria *Clean & Green* representa a percepção de áreas com presença significativa da natureza, como parques e lagos, onde a paisagem verde e pouca (ou nenhuma) quantidade de lixo tendem a permitir que as pessoas aproveitem um ambiente agradável. É natural se questionar se a palavra "*shady*" pertence a este grupo. No entanto, descobrimos que essa palavra geralmente é usada para descrever a sombra de árvores. Estes são exemplos do nosso conjunto de dados: "*Perfection at rest is shady tree. Perfection in motion is line thereof.*" e "*Trees are very shady around the ages on the water. So ideal for summer sun breaks*".

A categoria *Pleasant & Friendly* representa a percepção de áreas que podem permitir boas experiências aos visitantes, o que pode ser adequado para passear, por exemplo, com amigos e pessoas amadas. Pesquisando a categoria *Awful & Scary*, descobrimos que representa percepções de áreas que podem ter desencadeado experiências ruins para seus visitantes. Pode-se pensar que a palavra "*cheap*" está fora do contexto da categoria, no entanto, ela é comumente relacionado ao

tipo de lojas em torno do local em avaliação e, cujo contexto é semelhante ao contexto das palavras desta categoria, de acordo com o modelo Word2Vec. Estes são exemplos do nosso conjunto de dados: "*Don't like this place much just good to buy cheap souvenirs*" e "*Chinatown has gotten seedy and I irrelevant. Cheap, aggressive street vendors push their counterfeit wares. It's lost it's charm and allure. Streets have an awful stench.*".

Estudando a categoria *Classic & Authentic*, descobrimos que ela representa a percepção de áreas que podem fornecer experiências únicas aos seus visitantes. Por fim, a categoria *Great & Wonderful* tende a representar áreas que podem ter potencial de proporcionar experiências extraordinárias aos seus visitantes.

Após esta investigação, confirmamos que a nossa metodologia foi capaz de organizar as palavras em grupos que são semelhantes entre si para descrever áreas urbanas externas. De posse de grupos de palavras coerentes, ou seja, categorias, a tarefa de separar e classificar o conteúdo relevante de dados LBSN ruidosos torna-se mais viável.

3.4. Extração da Reputação

Nosso objetivo é extrair a reputação de áreas explorando uma grande coleção de dados ruidosos que contêm diferentes semânticas, ou seja, conjunto de dados de LBSN denotado por $\mathcal{D}_L$. Ao aplicar o REP-dicionário para capturar os documentos em $\mathcal{D}_L$ que correspondem ao alvo semântico (ou seja, a percepção das pessoas sobre áreas externas urbanas), possivelmente muitos casos falso positivos ocorrerão, uma vez que os qualificadores que compreendem o dicionário não se limitam a descrever lugares, mas também a descrever pessoas, coisas, etc. Para superar essa limitação, um passo-chave é o agrupamento de dados que têm similaridade espacial e semântica e, desconsiderando os documentos não relacionados à semântica investigada, ou seja, documentos com conteúdo não relacionado a áreas urbanas, ou percepções individuais, que podem não ser suficientes para representar a percepção de áreas externas.

Levando isso em consideração, propomos um algoritmo de agrupamento não supervisionado para agrupar documentos que são espacialmente e semanticamente relacionados. Esse agrupamento é um passo central para a identificação da reputação de áreas. O Algoritmo 1 resume os principais passos de nossa proposta[6]. Em seguida, descrevemos as principais etapas do algoritmo.

O algoritmo recebe como entrada: o modelo Word2Vec para áreas urbanas WV, descrito na seção anterior; um conjunto de dados de LBSN $\mathcal{D}_L$; um valor real ϵ representando uma distância; e um valor inteiro *minPts* representando um número de pontos. Primeiramente, o algoritmo remove qualquer ruído espacial do conjunto de dados, ou seja, filtra o conjunto de dados de todos os documentos cuja localização GPS é exatamente igual a um número grande de outros documentos. Esta etapa é importante para ajudar a evitar que dados inválidos sejam considerados, uma vez que a probabilidade desta situação ocorrer na prática é muito baixa para dados reais, sendo mais comum nos processos automáticos, como os usados pelos robôs web, como discutido em [Tasse et al. 2017]. Baseado no conjunto de dados $\mathcal{D}_L$, determinamos um limiar $th = 10$ para remover esse ruído, pois a maioria das localizações de GPS (cerca de 90%) têm no máximo 10 documentos associados a elas.

Em seguida, o algoritmo agrupa documentos de acordo com a similaridade espacial, usando o algoritmo *Density-Based Spatial Clustering of Application with Noise* (DBSCAN [Ester et al. 1996]) com $\epsilon = 100$ e *minPts* $= 3$. Testamos valores diferentes de $minPts$, e para todos os cenários considerados, ϵ não mudou significativamente. Usar DBSCAN para esta tarefa é interessante porque se baseia na densidade da vizinhança para identificar agrupamentos espaciais, sendo capaz de criar *clusters* com diferentes formatos e tamanhos, obtendo resultados satisfatórios

[6]Para favorecer o entendimento, o pseudocódigo não descreve as etapas de otimização aplicadas na prática.

Algoritmo 1: Agrupamento de documentos com similaridade espacial e semântica.

entrada: WV, $\mathcal{D}_L$, ϵ, *minPts*.
saída : Documentos com significante similaridade espacial e semântica.

```
// clean remove todos os documentos com ruído espacial de D_L
D'_L ← clean(D_L);
C ← DSCAN com ε e minPts em D'_L;
foreach C ∈ C do
    // emmpyVec cria um vetor vazio de tamanho |C|
    colorDoc ← emmpyVec();
    // emptyMat cria uma matriz quadrada de tamanho |C|
    simDocMatrix ← emptyMat();
    foreach doc_i ∈ C do
        foreach s ∈ doc_i do
            // wordsModel retorna um vetor com todas as palavras de s em WV
            words ← wordsModel(s, WV);
            if size(words) == 0 then
                continua;
            end
            vector_i ← meanWordVec(words, WV);
            foreach doc_j ∈ C do
                foreach s ∈ doc_j do
                    words ← wordsModel(s, WV);
                    vector_j ← meanWordVec(words, WV);
                end
                if size(words) == 0 then
                    continua;
                end
                simSentList.append(cosseno(vector_i, vector_j));
            end
        end
        simDocMatrix[i, j] ← mean(simSentList);
    end
    for i ← 0; i < |C|; i++ do
        score ← mean(simDocMatrix[i, z], ∀z ≠ i);
        if score ≤ 0 then
            colorDoc[i] ← preto;
        end
        else
            colorDoc[i] ← vermelho;
        end
    end
end
```

mesmo na presença de ruídos. Este processo de agrupamento resulta em um conjunto de *clusters* $\mathcal{C}$.

Depois disso, o algoritmo executa o passo de similaridade semântica. Para cada *cluster* $C \in \mathcal{C}$, consideramos documentos distintos $doc_i, doc_j \in C$, e calculamos a similaridade do cosseno entre todas as sentenças em doc_i com as de doc_j. Para isso, identificamos todas as palavras de cada $s \in doc_i$ que estão presentes em WV. Para cada uma dessas palavras o modelo retorna o vetor de palavras (descrito na seção anterior), que representa a similaridade de uma determinada palavra para todas outras do modelo. Por exemplo, se s possui duas palavras no modelo, teremos dois vetores de palavras. Fazemos a mesma coisa com sentenças $s' \in doc_j$. Se s', por ventura, contém somente uma palavra presente em WV, então terá apenas um vetor de palavras. Quando uma sentença possui mais do que um vetor de palavras, um único vetor é gerado com base na média entre esses vetores, sendo este considerado na comparação. Observe que, eventualmente, nenhum vetor de palavras é encontrado, ou seja, nenhuma palavra de s ou s' pertence a WV. Nesse caso, ignoramos essa comparação específica[7]. Um valor entre -1 e 1 é devolvido pela similaridade do cosseno.

Como resultado, todas as sentenças $s \in doc_x$ têm um valor de similaridade com outras sentenças de documentos vizinhos (i.e., em um mesmo *cluster*), valores que são armazenados em *simSentList*. Consideramos que o valor médio de *simSentList* representa proximidade p do doc_x entre outros membros no *cluster* C. Após essa computação, usamos uma matriz quadrada *simDocMatrix*, de tamanho $|C|$, para armazenar os valores p de todos os documentos contra todos

[7]Esses passos são também considerados em várias implementações do Word2Vec disponíveis publicamente.

os documentos do *cluster* C.

Finalmente, calculamos a distância média de cosseno, *score*, para cada documento explorando *simDocMatrix*, representando, assim, a semelhança semântica de um documento com outros documentos no mesmo *cluster*. Os documentos com *score* maior que 0 são coloridos em vermelho e os documentos restantes são coloridos em preto. Como resultado, todos os documentos que têm similaridade espacial são identificados como tendo também similaridade semântica significativa.

Considerando apenas os documentos do conjunto de dados LBSN com similaridade espacial-semântica, ou seja, que foram coloridos em vermelho pelo algoritmo de agrupamento apresentado, utilizamos o REP-dicionário para rotulá-los. Desta forma, um dado documento com pelo menos uma palavra correspondente ao REP-dicionário pode ser rotulada com uma das sete categorias do dicionário. Caso duas ou mais palavras estejam presentes no dicionário, o documento é rotulado com as categorias correspondentes, sendo possível atribuir múltiplas categorias ao mesmo documento. Isso ajuda a refletir a heterogeneidade da reputação, onde a mesma área pode ter reputações distintas de acordo com a opinião das pessoas. Alternativamente, o documento pode não ter palavras relacionadas à reputação, o que implica que, apesar de ter um conteúdo semântico relacionado com a área urbanas, ele não está relacionado à reputação e, portanto, é desconsiderado.

4. Reputação de Áreas Urbana Externas de Chicago

Para testar o arcabouço proposto, usamos um conjunto de dados coletado a partir do Twitter, para mapear a reputação das áreas urbanas externas de acordo com a percepção das pessoas compartilhada nas mídias sociais (Seção 4.1). Além disso, validamos nossos resultados realizando uma pesquisa *online* para reunir a percepção das pessoas sobre algumas áreas específicas de Chicago (Seção 4.2).

4.1. Mapeamento da Reputação

Considerando Chicago como cenário de avaliação de nosso arcabouço, coletamos dados conforme descrito na Seção 3.1. No total, 39.348 documentos compõem o conjunto de dados usado para entender como os usuários avaliam as áreas externas urbanas. Ao analisar as frases nesses documentos, o modelo Word2Vec gerou um vocabulário com 2.187 palavras únicas, onde o tamanho do *corpus* é 55.717, o tamanho da janela é $ws = 8$, a frequência mínima da palavra no corpus é $minCount = 20$ e o número de características é $m = 300$.

Também coletamos documentos do Twitter, de janeiro a maio de 2017, para construir nosso conjunto de dados LBSN. Ao todo, 71.450 foram coletados e após remover o ruído espacial, restaram 22.806 documentos. O número restante de documentos após o agrupamento espacial com DBSCAN é 20.618, e 14.136 desses documentos também possuem similaridades semânticas significativas de acordo com o Algoritmo 1.

Com isso, estamos prontos para gerar os mapas de reputação das áreas urbanas externas, mapas que chamamos de REP-Map. Devido à limitação de espaço, neste estudo, apresentamos e analisamos apenas algumas áreas da cidade. No entanto, o protótipo interativo da nossa ferramenta REP-Map estará disponível na página Web do projeto, onde será possível verificar os resultados para outras áreas. Ao usar mapas de calor, o nível da reputação na cidade para cada categoria é destacado de acordo com o número de percepções coletivas observadas em certas regiões. No REP-Map, as cores escuras e suavizadas indicam baixa intensidade de reputação, mas o oposto, ou seja, cores claras e com brilho forte, correspondem a alta incidência de reputação.

A Figura 3 mostra os REP-Maps para cada categoria de reputação separadamente em quatro regiões diferentes de Chicago: *Downtown*, *Near South Side*, partes dos bairros *North Center*

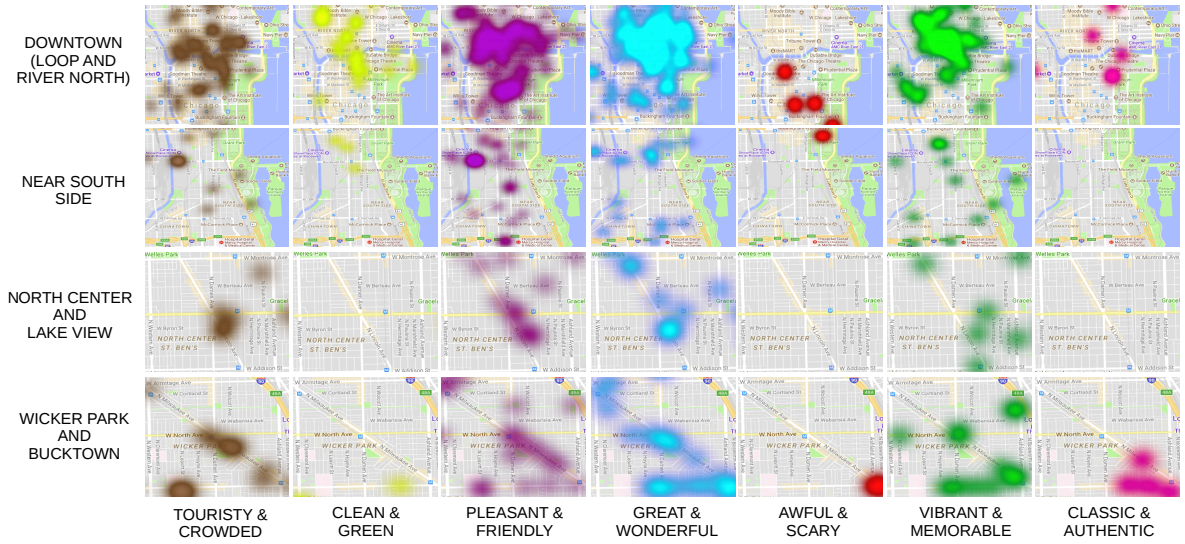

Figura 3. REP-Map em alguns bairros de Chicago.

e *Lake View*, e partes dos bairros *Wicker Park* e *Bucktown*. Quanto ao centro de Chicago, não é uma surpresa que quase todas as categorias possuem *clusters* de alta intensidade, abrangendo os bairros *Loop* e *River North*, conhecidos como centros comerciais e financeiros da cidade, com muitos restaurantes, bares, atrações turísticas e locais para eventos diversos, atraindo um grande número de visitantes e residentes. Embora a categoria *Awful & Scary* tenha acontecido de forma isolada nesta área, note que há poucas sobreposições entre os *clusters* desta categoria com os demais, sugerindo que tais áreas podem ter algum problema que afeta negativamente as experiências dos usuários, ao ponto de encorajá-los a compartilhar sua opinião nas mídias sociais.

O bairro *Near South Side* é considerado predominantemente *Pleasant & Friendly*, onde há um *cluster* roxo de alta intensidade e outros menos intensos, mas ainda significativos, espalhados pela área estudada. A segunda e terceira reputações mais populares são *Great & Wonderful* e *Vibrant & Memorable*, respectivamente. Outra categoria significativa nessa área é *Touristy & Crowded*, com *clusters* identificados em áreas comerciais ou pontos turísticos.

As áreas estudadas dos bairros *North Center* e *Lake View* consistem principalmente em áreas residenciais e comerciais, com poucos parques, locais culturais e pontos turísticos para visitar, resultando na ausência de *clusters* das categorias *Clean & Green* e *Classic & Authentic*. Como podemos ver, os *clusters Touristy & Crowded* são exibidos próximos a escolas e áreas comerciais, representando lugares com grande concentração de pessoas. Ao longo da avenida *N Lincoln Ave.*, podemos ver *clusters* de alta intensidade das categorias *Pleasant & Friendly* e *Great & Wonderful*, indicando, possivelmente, que fazer compras e/ou comer nessas áreas, são atividades que usuários acham agradáveis. A ausência de *clusters Awful & Scary* e a existência de *clusters Vibrant & Memorable* localizados perto de residências, sugere que é um bom lugar para viver.

Estudando a região do bairro *Wicker Park*, que é conhecido como um centro de comércio e cultura da cidade, podemos ver várias sobreposições entre a reputação de todas as categorias ao longo da avenida *N Milwaukee Ave.* e da rua *W Divison St.*, principalmente perto de restaurantes e bares, onde é esperado surgir reputações heterogêneas devido à grande variedade de sons, elementos visuais, odores, etc, presentes nessas áreas, o que pode, potencialmente, desencadear percepções distintas nas pessoas. Além disso, também podemos ver que alguns *clusters Pleasant & Friendly* e *Vibrant & Memorable* são mapeados em alguns parques, como *Churchill Field Park* e *Walsh Park*, indicando que nessas áreas muitos usuários se divertirem.

4.2. Avaliação com Usuários

Realizamos uma pesquisa *online* de 22 de Novembro à 14 de Dezembro de 2017, com o intuito de obter a opinião dos voluntários sobre as áreas urbanas externas avaliadas na seção anterior, com exceção do centro da cidade, para validar nossos resultados. A região do centro foi removida da validação, porque, como discutimos, é comum a ocorrência de todas as reputações nessas áreas. Nós recrutamos voluntários que vivem ou conhecem Chicago e pedimos que respondessem um questionário.

O questionário contém três figuras dos bairros avaliados de Chicago, sem a exibição dos *clusters* de reputação. Solicitamos aos participantes que escolhessem um ou mais conjuntos de palavras que, segundo sua opinião, melhor caracterizam as áreas urbanas externas exibidas nas figuras. Os conjuntos de palavras são os mesmos que compõem as categorias de reputação do REP-dicionário (veja Tabela 1). Também, incluímos dois conjuntos extras de palavras qualificadoras de áreas urbanas externas, mas não pertencentes ao REP-dicionário. Além disso, oferecemos uma opção vazia, a fim de não impor a escolha de pelo menos um dos conjuntos informados. Para cada figura, também oferecemos a possibilidade para os participantes informarem, através de texto livre, suas impressões pessoais sobre a área externa urbana exibida. Além disso, solicitamos informações básicas sobre os participantes, como o grau de escolaridade, gênero, faixa etária e o nível de conhecimento sobre as áreas urbanas da cidade (baixo, médio e alto).

No total, recrutamos 18 participantes para Chicago. A maioria, sete participantes, conhece as principais áreas e algumas menos populares (nível de conhecimento médio). Cinco participantes declararam ter um nível de conhecimento alto sobre a cidade, incluindo vários lugares menos populares, e outros cinco só conhecem algumas áreas populares (nível de conhecimento baixo). Apenas um participante não conhecia a cidade. Mulheres representam a maioria (10 mulheres e oito homens), de um grupo principalmente adulto (13 com faixa etária de 31 a 40, 1 de 41 a 50 e o restante até 30 anos) e com alto nível de escolaridade (nove possuem um curso superior completo, quatro possuem o título de mestre, três possuem doutorado ou pós-graduação avançada e dois estavam cursando um curso superior).

Em seguida, para evitar ruído nos dados, desconsideramos a opinião de seis participantes cujo conhecimento sobre a cidade não era suficiente (nível declarado baixo). Para os 12 participantes que possuem nível de conhecimento médio ou alto sobre as áreas urbanas de Chicago, o bairro *Near South Side* é *Vibrant & Memorable* (na opinião de 58,3% deles), *Pleasant & Friendly* (33,3%), *Great & Wonderful* (33,3%), e *Touristy & Crowded* (33,3%). A similaridade das opiniões dos participantes e o REP-map gerado para a mesma área é muito boa. De acordo com o REP-map, a reputação mais intensa é *Pleasant & Friendly*, no entanto, a mensagem principal ainda foi capturada.

Da mesma forma, a reputação da região do bairro *North Center*, na opinião de participantes com um conhecimento médio ou alto sobre Chicago, é predominantemente *Pleasant & Friendly* (50%), seguido de *Vibrant & Memorable* (33,3%) e *Touristy & Crowded* (25%). A semelhança do REP-Map com este resultado também é impressionante, onde *Pleasant & Friendly* corresponde a 33,3% das percepções, *Touristy & Crowded* e *Vibrant & Memorable* cerca de 18,5%. Outra categoria significativa descoberta pelo REP-Map é *Great & Wonderful* (29,6%), completando o espectro de reputação refletido pelo bairro *North Center*. É importante mencionar que, os demais participantes, com pouco conhecimento da cidade, tiveram dificuldade em classificar a reputação desta área. A maioria deles informou que nenhuma das opções na pesquisa foi boa para descrever a área.

Uma participante (com alto nível de conhecimento da área, do sexo feminino, de 31 a 40 anos e com diploma de bacharel), nos dá uma dica interessante desta área que pode ajudar a

justificar os resultados do REP-Map:

> *"It is an area of the city with low crime and a relatively higher average income than most neighborhoods on the South and West sides of the city. The streets, parks, and yards are well kept."*

O bairro *Wicker Park* foi o cenário com maior heterogeneidade de reputação entre os pesquisados. De acordo com a opinião do grupo com bom conhecimento da cidade, esta região é *Pleasant & Friendly* (41,67%), *Touristy & Crowded* (33,3%), *Great & Wonderful* (25%), *Vibrant & Memorable* e *Classic & Authentic* (ambos com 16,67%). Com o REP-Map, também foi identificada uma alta heterogeneidade na reputação desta área: *Great & Wonderful* (28,7%), *Pleasant & Friendly* (24,1%), *Touristy & Crowded* (19,6%), *Vibrant & Memorable* (18,7%) e *Classic & Authentic* (6%). Embora há uma ligeira diferença de ordem, os resultados mostram uma forte correlação entre a percepção das pessoas e o algoritmo REP-Map. Novamente, os participantes restantes tiveram dificuldade em classificar a reputação desta área, e eles optaram principalmente por uma resposta vazia (nenhum conjunto de palavras).

A reputação geral refletida pelos bairros capturadas com a nossa metodologia parece adequada para representar a percepção das pessoas que conhece bem Chicago. Isso pode ajudar usuários com pouco conhecimento sobre a cidade, incluindo turistas, em melhor compreendê-la, ajudando-os a melhor explorar as áreas urbanas.

5. Considerações Finais

O presente trabalho explora a utilização de mídias sociais para extrair e mapear a reputação de áreas urbanas externas. Para isso, definimos um arcabouço para auxiliar no processo de aprendizagem e mapeamento da opinião coletiva em relação a reputação de áreas urbanas externas a partir de uma grande coleção de dados ruidosos. Para validar nosso arcabouço, realizamos um estudo com usuários voluntários. É notável a correlação entre os resultados alcançados com nosso arcabouço e a opinião dos voluntários, em todos os bairros avaliados. Conseguindo, assim, atingir o objetivo deste estudo: identificar a reputação refletida em áreas urbanas externas de forma escalável, e assim, fornecer mecanismos importantes para ajudar as pessoas a compreenderem melhor as semânticas existentes em diferentes áreas da cidade. Como trabalho futuro, pretende-se evoluir este arcabouço para incorporar outras fontes de dados, tais como, Flickr, Instagram e Facebook, na avaliação da reputação urbana ao ar livre em diferentes cidades.

Agradecimentos

Os autores agradecem o apoio financeiro concedido pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2015/07538-1 e 2015/24494-8), CAPES (PDSE 88881.132016/2016-01), CNPq (processo 401802/2016-7) e CNPq-URBCOMP (processo 403260/2016-7).

Referências

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Flaes, J. B., Rudinac, S., and Worring, M. (2016). What multimedia sentiment analysis says about city liveability. In *European Conference on Information Retrieval*, pages 824–829. Springer.

Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38. ACM.

Henshaw, V. (2013). *Urban smellscapes: Understanding and designing city smell environments*. Routledge.

Jiang, S., Qian, X., Mei, T., and Fu, Y. (2016). Personalized travel sequence recommendation on multi-source big social media. *IEEE Transactions on Big Data*, 2(1):43–56.

Kim, J., Cha, M., and Sandholm, T. (2014). Socroutes: safe routes based on tweet sentiments. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 179–182. ACM.

Marsden, P. V. and Lin, N. (1982). *Social structure and network analysis*, volume 57. Sage.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Naik, N., Philipoom, J., Raskar, R., and Hidalgo, C. (2014). Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 779–785.

Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. (2012). Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 965–968. ACM.

Quercia, D., O'Hare, N. K., and Cramer, H. (2014a). Aesthetic capital: what makes london look beautiful, quiet, and happy? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 945–955. ACM.

Quercia, D., Schifanella, R., and Aiello, L. M. (2014b). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 116–125. ACM.

Quercia, D., Schifanella, R., Aiello, L. M., and McLean, K. (2015). Smelly maps: the digital life of urban smellscapes. *arXiv preprint arXiv:1505.06851*.

Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2013). Uma Fotografia do Instagram: Caracterização e Aplicação. In *Proc. of XXXII SBRC*, Brasília, DF.

Steiger, E., Resch, B., and Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, 30(9):1694–1716.

Tasse, D., Liu, Z., Sciuto, A., and Hong, J. I. (2017). State of the geotags: Motivations and recent changes. In *ICWSM*, pages 250–259.

Weigelt, K. and Camerer, C. (1988). Reputation and corporate strategy: A review of recent theory and applications. *Strategic management journal*, 9(5):443–454.

Yang, D., Zhang, D., Yu, Z., and Wang, Z. (2013). A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 119–128. ACM.

Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., and Xiong, H. (2015). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725.