Participatory Social Sensor: A Framework to Social Media Data Acquisition and Analysis

Ígor Araújo[1], Paulo H. L. Rettore[2] e Guilherme Maia[2]

[1]Departamento de Engenharia
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

igoral@ufmg.br,

{rettore,jgmm}@dcc.ufmg.br

Abstract. *Understanding urban mobility, people's feelings and social behaviors have been the focus of many researches and investments. Due to the costs to create a sensors' infrastructure in a given area, data about many aspects of a city become restricted to private companies, research groups and governments. In this scenario, considering that Location-Based Social Media (LBSM) may provide a new way to better comprehend the social behaviors with the user's viewpoint, we propose the use of LBSM as participatory sensing and design the Participatory Social Sensor (PSS), a friendly and open source framework to social media data acquisition and analysis. We develop the Twitter data acquisition and analysis process, aiming to guide new researchers in their application goals with a structured input setup, where they specify the spatial area, temporal interval, tags, and other parameters. The result is a set of visual analyses which describe the context overview, allowing researchers and students to conduct their projects, focusing only on the main research issues and help them to make project decisions.*

Resumo. *Compreender a mobilidade urbana, o sentimento das pessoas e os comportamentos sociais têm sido o foco de muitas pesquisas e investimentos. Devido ao custo de se criar uma infraestrutura de sensores em uma dada área, dados que descrevem vários aspectos das cidades têm se tornado restrito à empresas privadas, grupos de pesquisadores e governos. Nesse cenário, o uso de LBSM fornece uma nova maneira de compreender os comportamentos sociais a partir do ponto de vista do usuário. Diante disso, propomos o uso de LBSM como sensor participatório e desenvolvemos o PSS, um framework amigável e de código aberto para aquisição e análise de dados de mídia social. Desenvolvemos um processo de aquisição e análise de dados do Twitter, visando guiar pesquisadores em suas aplicações através de uma configuração de entrada estruturada, onde é especificada a área, intervalo temporal, tags e outros parâmetros. O resultado é um conjunto de análises visuais que descrevem uma visão geral do contexto, permitindo aos pesquisadores e estudantes conduzir seus projetos, focando somente nos problemas principais de pesquisa e os ajudando a tomar decisões.*

1. Introdução

O desenvolvimento das cidades e de toda uma sociedade depende da análise de dados para o planejamento de infraestruturas, gestão dos sistemas de transportes e compreensão dos aspectos sociais e econômicos. Nesse sentido, vários investimentos têm sido feitos como forma de compreender os comportamentos das comunidades, seja no trânsito ou em outros aspectos das vidas das pessoas. No entanto, o processo de coleta de dados em grandes cidades demanda grandes infraestruturas de sensores dos mais variados tipos, tornando o processo custoso e até mesmo inviável. Em outras palavras, esse processo depende de uma vasta cobertura de dados e comunicação, como dados de semáforos, dados de mídias sociais (como Twitter, Facebook, Instagram), dados de dispositivos móveis, dados de sensores veiculares, sensores meteorológicos, dentre outros. Grande parte desses dados, por sua vez, são restritos às companhias privadas, grupos de pesquisa, governos e em muitos casos, estão desatualizados. O monopólio desses dados é um fator preponderante que restringe a realização de estudos mais abrangentes, o que permitiria o desenvolvimento de novas soluções para essas cidades. Por esse motivo, a coleta de dados que caracteriza todo um contexto social e utiliza um processo de baixo custo se torna uma tarefa desafiadora mas necessária para o avanço da sociedade. A partir desses fatos, emerge o conceito de Sensoriamento Participativo – *Participatory Sensing*, que leva em consideração que as pessoas podem sensorear o ambiente ao seu redor e auxiliar na obtenção de determinado conhecimento.

Desse modo, propomos o Sensor Social Participativo – Participatory Social Sensor (PSS), um *framework* amigável de baixo custo para aquisição e análise de dados de Mídias Sociais Baseadas em Localização – Location-Based Social Media (LBSM). O PSS fornece para pesquisadores, desenvolvedores, educadores e alunos uma infraestrutura de código aberto pronta para coleta, tratamento e caracterização de dados de mídia social. A ferramenta tem como objetivos: i) auxiliar novos pesquisadores e alunos na compressão e análise de dados de mídias sociais; ii) direcionar o foco dos estudos aos dados e ao problema de pesquisa; iii) ser flexível o suficiente para dar suporte ao desenvolvimento de diversas aplicações; iv) ter código aberto para permitir modificações segundo os objetivos da aplicação; v) dar suporte ao ensino de disciplinas como ciência dos dados, mineração de dados, *big data* entre outras.

O restante deste artigo está organizado da seguinte forma: na Seção 2 são discutidos os trabalhos relacionados. A Seção 3 apresenta a arquitetura do PSS. A Seção 4 descreve os resultados das análises em um determinado contexto. Na Seção 5, apresentamos os resultados do uso da ferramenta e um estudo de caso. A Seção 6 descreve como a ferramenta será demonstrada no evento. Por fim, concluímos na Seção 7.

2. Trabalhos Relacionados

Com a difusão da Internet e com as LBSMs como parte do cotidiano das pessoas, uma grande quantidade de dados são gerados diariamente. Dessa forma, vários estudos foram desenvolvidos utilizando-se desses dados para criar aplicações e serviços, como por exemplo, no contexto de trânsito [Xu et al. 2018]. Santos et al. [Santos et al. 2018] apresentam uma metodologia para melhorar a compreensão do tráfego urbano. Ainda no cenário de trânsito, [Donahue et al. 2018] fazem uso das mídias sociais para compreender a dinâmica de visitas que os motoristas fazem em Twin Cities, Minnesota, EUA e ainda gerar informações que facilitam a gestão dos espaços verdes e parques da região.

Já no trabalho desenvolvido por [Gaurav et al. 2013], foi investigado o poder da mídia social em prever os candidatos vencedores em eleições que aconteceram na América Latina. Para tal, foi desenvolvida uma técnica para análise de *tweets* com base na análise de termos que remetem aos candidatos desses países. No fim, a abordagem mostrou-se eficiente na predição dessas eleições.

[Karami et al. 2018] utilizaram dados de mídia social como forma de caracterização de doenças (diabetes, obesidade) e comportamentos sociais (dietas, exercício), desenvolvendo uma abordagem cujo objetivo é caracterizar a opinião da população em relação a esses assuntos de saúde como forma de controle e combate à doenças relacionadas.

Diferente das abordagens descritas acima, propomos uma ferramenta amigável e de código aberto de propósito geral para aquisição e análise de dados do Twitter. Tal ferramenta permite que pesquisadores, estudantes e professores desenvolvam aplicações tendo como base um *framework* para coleta, análise e caracterização de dados de mídia social. Dessa forma, por meio de um arquivo de configuração, é possível gerar resultados visuais que possam guiar os usuários em seus respectivos objetivos, seja ele pesquisa, ensino ou aprendizado.

3. Arquitetura do PSS

PSS é uma ferramenta de código aberto que tem como objetivo auxiliar alunos, professores e pesquisadores (usuários) a desenvolver suas investigações com o foco em questões de pesquisa. Ou seja, os processos de coleta, tratamento e caracterização dos dados de mídia social são realizados pela ferramenta, permitindo maior flexibilidade e poder de decisão aos usuários. A estrutura da ferramenta é apresentada na Figura 1. O processo de aquisição e análise de dados inicia-se quando o usuário insere os parâmetros desejados no arquivo de configuração. Os dados são então coletados e armazenados em arquivos de texto para a etapa de tratamento. Nesta etapa, os *tweets*, que estão em arquivos JSON, são transformados em arquivos CSV e servem de entrada para a etapa de análise de dados. Por fim, é realizada a caracterização dos dados e geradas as visualizações para melhor compreensão do contexto de interesse. Devido à limitação de espaço, foi criada uma página Web contendo a descrição completa da ferramenta, funcionalidades, código fonte, guia de instalação e um tutorial em vídeo.

3.1. Entrada

Para iniciar o uso do PSS, é necessário preencher apenas o arquivo de configuração, especificando o tempo de coleta de dados, o período de análise, a região geográfica de interesse (por meio de um *bounding box*), as palavras (*tags*) de referência para o rastreio dos *tweets* (*tracks*), fuso horário da região de coleta, além do nome do lugar de onde serão coletados os dados.

Contudo, também é possível especificar os parâmetros para gerar as visualizações dos dados coletados, como por exemplo: o tipo de análise que se deseja fazer (temporal, espacial ou espaço-temporal). Além disso, o usuário pode especificar se deseja realizar análises como: análise de emoções, sumarização do dados textuais, análise dos usuários mais frequentes, gerar os *traces* desses usuários e o respectivo intervalo de tempo da análise.

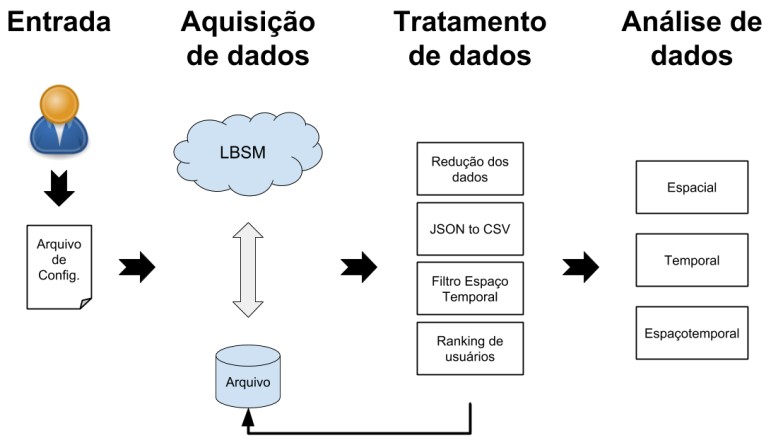


Figura 1: Arquitetura do PSS.

3.2. Aquisição de Dados

A primeira versão do PSS conta com o processo de aquisição de dados usando a API do *Twitter*, *Tweepy*, através do *bounding box* da região, *tags* de pesquisa como forma de redirecionamento dos assuntos dos *tweets* coletados e o tempo (horas) que o usuário deseja realizar a coleta. Dessa forma, os *tweets* coletados em formato JSON são armazenados em arquivos de texto e nomeados com a data da coleta. Para otimizar os custos computacionais, foi conduzido um processo de redução de parâmetros que busca eliminar todas as informações que não serão utilizadas nas etapas posteriores, como por exemplo, imagem padrão da conta, cor da borda da barra lateral, cores da imagem de plano de fundo do perfil, cor do link do perfil, id do *tweet* respondido, URL para imagens da página principal e demais descrições do perfil dos usuários. É importante ressaltar que o processo de aquisição de dados foi desenvolvido para suportar a implementação de novas fontes de dados, tornando a infraestrutura robusta para futuras melhorias.

3.3. Tratamento de Dados

Nesta etapa são eliminados dos arquivos JSONs as informações irrelevantes e os dados resultantes são agrupados em um único arquivo CSV, contendo as seguintes características: id do *tweet*, latitude, longitude, texto, *timestamp* e o nome do usuário. Além disso, são selecionados apenas os *tweets* que possuem informações de localização.

Como os *tweets* coletados podem estar fora da área especificada pelo *bounding-box* (inconsistência do *stream* da API do Twitter), faz-se uma nova filtragem espacial para certificar que apenas dados da região de interesse serão analisados. Por fim, são verificados os usuários mais ativos na plataforma, ou seja, usuários com maior número de *tweets* publicados dentro da região espaço-temporal especificada, identificando as contas mais relevantes da região. Ao identificar essas contas, é realizada uma pesquisa utilizando a API REST do Twitter, com o objetivo de coletar o histórico dos últimos 3000 *tweets* de cada conta, que serão posteriormente analisados nas etapas de identificação dos trajetos dos usuários.

3.4. Análise de Dados

Nessa etapa é realizada a caracterização dos *tweets* coletados. Essas análises abrangem as dimensões: Espacial, Temporal e Espaço-Temporal. O objetivo desta etapa é caracterizar

os dados coletados de forma a guiar o usuário do *framework* PSS nas etapas de tomada de decisão de acordo com seus objetivos individuais.

3.4.1. Espacial

Nesta seção, o interesse é analisar o comportamento dos *tweets* no espaço, a partir da visualização dos dados no mapa e sua densidade. O objetivo é compreender a i) Distribuição dos *tweets* no espaço; ii) Densidade dos *tweets* no espaço, permitindo identificar as regiões que possuem maior concentração de dados e direcionar o foco da investigação. Por exemplo, ao perceber que uma dada região tem uma densidade muito grande de *tweets*, pode-se iniciar outra coleta, nessa região, conseguindo assim mais dados para desenvolver eventuais pesquisas e aplicações.

3.4.2. Temporal

Na dimensão temporal, levam-se em consideração as características dos *tweets* ao longo do tempo, ou seja, como esses dados se comportam ao longo das horas do dia, dias da semana e meses do ano. Nesse processo são realizadas as seguinte análises: i) Histograma dos *tweets* ao longo das horas do dia; ii) Densidade dos *tweets* ao longo das horas do dia; iii) Histograma dos *tweets* ao longo dos dias da semana; iv) Densidade dos *tweets* ao longo dos dias da semana; v) Histograma dos *tweets* ao longo dos meses do ano; vi) Densidade dos *tweets* ao longo dos meses do ano.

3.4.3. Espaço-Temporal

Esta etapa realiza a análise dos *tweets* nas dimensões espacial e temporal. Para tal, foram geradas as seguintes análises de dados: i) *Tweets* plotados ao longo das horas do dia; ii) Densidade de *tweets* plotados ao longo das horas do dia; iii) Densidade de *tweets* plotados ao longo dos dias da semana iv) *Tweets* plotados ao longo dos meses do ano v) Densidade de *tweets* plotados ao longo dos meses do ano vi) Análise de sentimentos da região vii) Análise de sentimentos da região ao longo das horas do dia viii) Análise de sentimentos da região ao longo dos dias da semana ix) Análise de sentimentos da região ao longo dos meses do ano x) Histórico dos usuários mais frequentes xi) Trace dos usuários mais frequentes xii) Descrição textual dos usuários por sentimento e horas do dia.

Em (i) conseguimos ver a distribuição espacial dos dados e como eles variam ao longo do tempo. Fica nítido, por exemplo, que nas horas iniciais do dia (01:00h às 06:00h), os *tweets* são mais esparsos; Já em (ii e iii), podemos ver onde essa concentração é mais densa, considerando as horas do dia e dias da semana, respectivamente. Em seguida, realizamos essas análises para os meses do ano (iv e v), sendo possível observar o comportamento desses *tweets* em uma escala maior. Também é feita a análise de sentimentos (positivos, neutros ou negativos) dos *tweets* em uma dada região, sendo que a análise (vi) considera o sentimento agregado, (vii) o sentimento por hora, por dias da semana em (viii) e meses do ano em (ix). Assim, podemos analisar regiões onde os sentimentos são ruins, dado fatores externos (como engarrafamentos, acidentes, desvios),

bem como bons sentimentos (pessoas em momentos de lazer, como em parques, casas de show).

A próxima etapa (x), constitui na análise dos usuários mais frequentes. Ou seja, foi realizado o ranking dos *n* primeiros usuários que mais criaram *tweets*. Com esses usuários selecionados, foi possível observar o comportamento individual deles no espaço-tempo. Na análise (xi), observamos como essas contas se movimentam no espaço-tempo, ou seja, criamos os traces de *tweets* dos usuários e exportamos essa análise para um arquivo de CSV com todos os traces dos usuários. Com esses dados, podemos analisar a cobertura e, principalmente, a mobilidade dessas contas no espaço-tempo. Por fim, (xii) considera os textos dos usuários para criar uma sumarização que descreva a região, seja por sentimento ou horas do dia.

4. Resultados

Para as análises contida nos gráficos gerados, foi utilizado uma base constituída de 158.413 *tweets* adquiridos entre os dias 14-09-2018 e 06-11-2018. Além disso, foram selecionadas *tags* de pesquisa que remetem a incidentes de trânsito *congestion, accident, construction, planned event, road hazard, disabled vehicle, traffic, jam, car, weather*. Os *tweets* analisados possuíam geolocalização e estavam localizados em Manhattan - Nova York.

Alguns resultados do processamento e caracterização dos dados do Twitter são apresentados na Figura 2. Devido à limitação de espaço, são apresentadas a análise temporal, que contém a frequência de *tweets* ao longo das horas do dia, a densidade no espaço, o sentimento agregado em toda a região de observação e por fim, o exemplo de um trace de *tweets* de um determinado usuário.

4.1. Descrição

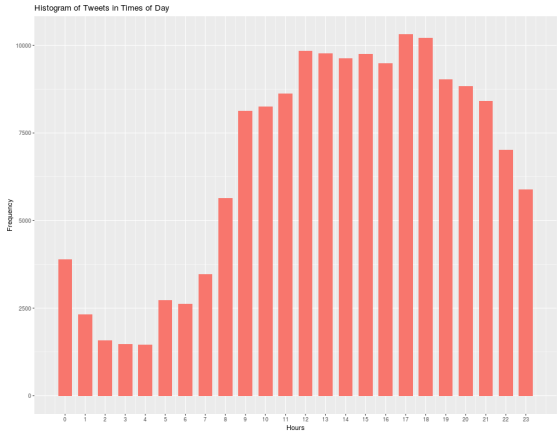
Outra análise realizada pelo PSS considera os textos dos usuários para criar uma sumarização que descreva a região, seja por sentimento ou horas do dia. A seguir, é apresentada uma sumarização por sentimentos de -6, 0 e 6, como exemplo.

> **The summary for emotion scale:-6** This idiot, illiterate ignorant #UNFIT ILLEGITIMATE crook liar campaigning for the GOP'S NOMINATION but they are h… **The summary for emotion scale:0** "New York, New York: a city so nice they named it twice"#thatwindwasnojoke #newyork, New York New York New York #maisenza is #everywhere Ph by filippotartaglia03 & rosimarty, New York New York, New York #NewYork #BigApple #ShotOniPhone, New York New York, New York, you put me in an empire state of mind, New York **The summary for emotion scale:6** Happy Birthday my Baby best friend my everything happy 12th birthday I love you to infinity an beyond… CHEERS to this little one who I love so much! Happy 27th birthday Jennaaaaa!! Have fun today my libra baby! Hope… A very happy birthday to my best friend and favorite travel companion, GING GING! I love growing and going through…

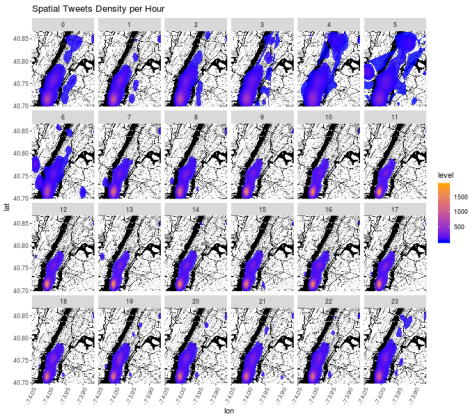
Para tal análise, foi utilizado o pacote *syuzhet* em R, que retorna valores relativo ao quão positivo e quão negativo é o texto analisado, no caso, o *tweet*. Com essas duas informações, foi realizado a soma do parâmetro positivo menos o negativo, chegando a uma escala que representa se seu texto possui emoções boas, valores positivos na escala, ou ruins, valores negativos na escala.

5. Estudo de Caso: T-Incident

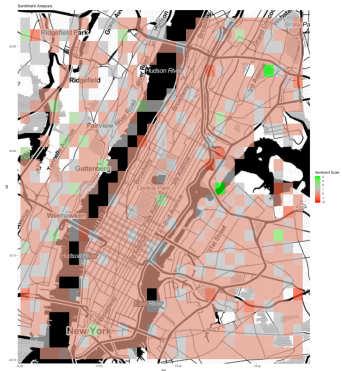
Utilizando o *framework* PSS foi possível coletar e analisar dados de Nova York e produzir uma aplicação descrita no seguinte estudo de caso: "Serviço de Detecção e Enriquecimento de Eventos Rodoviários Baseado em Fusão de Dados Heterogêneos para VANETs", artigo aceito para publicação no SBRC 2019.



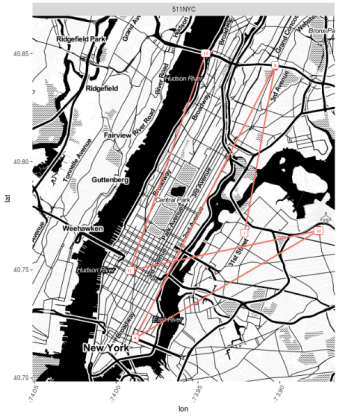
(a) Frequência de *tweets* por hora.



(b) Densidade de *tweets* por região e hora.



(c) Sentimento geral da região.



(d) Trace de usuário.

Figura 2: Exemplos de análises temporal, espacial e espaço-temporal dos dados.

O Twitter Incident (T-Incident) é uma arquitetura robusta de baixo custo para detecção e enriquecimento de eventos rodoviários baseado na fusão de dados heterogêneos. Com o uso do PSS, foi possível identificar os potenciais de uso da mídia social na região. Em seguida, com a coleta de dados do Twitter utilizamos as funções de extração de características para identificar grupos de palavras relevantes para permitir a detecção e descrição de eventos rodoviário. Além disso, implementamos os estágios de agrupamento espaço-temporal e de aprendizado, uma vez que a programação do PSS é estruturada e modularizada, viabilizando o processo de criar interfaces e novas funções. A Figura 3 apresenta os estágios da abordagem T-Incident. Como resultado da metodologia, foi fornecido um serviço apurado de detecção e descrição de incidentes rodoviários com acurácia acima de 90% para as métricas *F1 score*, *Recall* e *Precisão*.

6. Demonstração

Como forma de demonstração, será utilizado no dia da apresentação um computador que apresentará em tempo real a coleta, tratamento, caracterização e visualização das informações. Além disso, serão coletados previamente dados para uma análise mais completa, levando-se em conta diferentes contextos sociais, intervalos de tempos e regiões de análise. Como forma de instrução, será feito slides para guiar as pessoas que acompa-

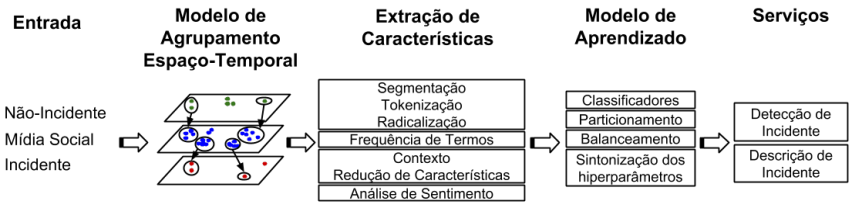


Figura 3: Arquitetura do T-Incident.

nham a explicação. Além disso, a descrição completa da ferramenta, tutorial em vídeo[1], código-fonte[2] e o guia de instalação da ferramenta estão disponíveis para a comunidade na página da ferramenta[3].

7. Conclusão

Este trabalho apresentou o PSS, um *framework* amigável, de baixo custo e código aberto para coleta, tratamento e análise de dados de LBSM.

Com a solução proposta, desenvolvedores, pesquisadores, educadores e estudantes podem conduzir seus trabalhos e se preocupar apenas com os dados da mídia social e não com toda a etapa de coleta e tratamento desses dados. Além disso, a etapa de caracterização dos dados de mídia social realizada pelo PSS pode guiar os usuários durante a tomada de decisão em seus projetos. Outro benefício no uso do PSS é em auxiliar o ensino que envolva o uso de grandes quantidades de dados, como mineração de dados, ciência dos dados, ciência da informação e aprendizado de máquina, tornando sistemático e padronizado o processo de coleta, tratamento e análise.

Como trabalhos futuros, pretende-se integrar outras bases de dados de mídias sociais; desenvolver uma interface Web; melhorar a interação com o usuário, tornando o uso ainda mais fácil e intuitivo, permitindo também que usuários de outras áreas tirem proveito da ferramenta.

Referências

Donahue, M. L., Keeler, B. L., Wood, S. A., Fisher, D. M., Hamstead, Z. A., and McPhearson, T. (2018). Using social media to understand drivers of urban park visitation in the twin cities, mn. *Landscape and Urban Planning*, 175:1–10.

Gaurav, M., Srivastava, A., Kumar, A., and Miller, S. (2013). Leveraging candidate popularity on twitter to predict election outcome. In *Proceedings of the 7th workshop on social network mining and analysis*, page 7. ACM.

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., and Shaw Jr, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, 38(1):1–6.

Santos, B. P., Rettore, P. H., Ramos, H. S., Vieira, L. F. M., and A.F. Loureiro, A. (2018). Enriching traffic information with a spatiotemporal model based on social media. In *ISCC*, Natal, Brazil.

Xu, S., Li, S., and Wen, R. (2018). Sensing and detecting traffic events using geosocial media data: A review. *Computers, Environment and Urban Systems*, (June).

[1] `https://youtu.be/FC2QWlg8x60`
[2] `https://github.com/ufmg-pss/ufmg-pss.github.io`
[3] `https://ufmg-pss.github.io/`