MINERAÇÃO DE REDES SOCIAIS PARA DETECÇÃO E PREVISÃO DE EVENTOS REAIS

JANAINA SANT'ANNA GOMIDE

MINERAÇÃO DE REDES SOCIAIS PARA DETECÇÃO E PREVISÃO DE EVENTOS REAIS

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

Orientador: Wagner Meira Junior
Co-orientador: Virgilio Augusto Fernandes Almeida

Belo Horizonte
Março de 2012

Gomide, Janaina Sant'Anna
G633m Mineração de redes sociais para detecção e previsão de eventos reais / Janaina Sant'Anna Gomide. — Belo Horizonte, 2012
xx, 85 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais - Departamento de Ciência da Computação.
Orientador: Wagner Meira Junior
Co-orientador: Virgilio Augusto Fernandes Almeida

1. Computação - Teses. 2. Redes sociais on-line - Teses. I. Orientador. II Coorientador. III. Título.

519.6*04.(043)

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

FOLHA DE APROVAÇÃO

Mineração de redes sociais para detecção e previsão de eventos reais

JANAÍNA SANT'ANNA GOMIDE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WAGNER MEIRA JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA - Co-orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO ALONSO VELOSO
Departamento de Ciência da Computação - UFMG

PROF. MAURO MARTINS TEIXEIRA
Departamento de Bioquímica e Imunologia - UFMG

Belo Horizonte, 02 de abril de 2012.

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais, Nize e Camilo, pelo amor, conselhos, apoio nas minhas decisões, por sempre me incentivarem e me darem condições para que eu realizasse mais essa conquista. Agradeço também a minha irmã, Camilinha, pela amizade e carinho. Obrigada Mila por estar sempre por perto e por ter me ajudado nas revisões do texto. Dedico essa vitória a vocês, mãe, pai e irmã, que são as pessoas que mais me incentivaram e apoiaram a fazer o mestrado.

Ao Júnior, amor da minha vida, agradeço por estar sempre comigo, pela sua compreensão e paciência. Obrigada pelo seu amor e carinho que me fazem sentir uma pessoa especial.

Agradeço aos meus amigos Thaty, Rodrigo e Douglas, que estiveram sempre juntos comigo, pela amizade, por não terem deixando que eu desanimasse, sempre dispostos a ajudar, me incentivando e ouvindo desabafos. Agradeço também à Glívia pelas palavras e apoio e pelas suas revisões. À minha amiga de infância, Flávia, pelos conselhos, amizade e pelos momentos de descontração.

Aos meus orientadores, Wagner Meira Jr. e Virgílio Almeida, por acreditarem em mim e me concederem a oportunidade de realizar o mestrado. Obrigada pela orientação. Considero um privilégio ter sido orientada pelos dois e ter aprendido com o conhecimento e experiência de cada um.

Agradeço aos colegas do laboratório e-Speed pelos momentos de descontração e companhia durante os dois anos de mestrado.

Finalmente, agradeço à CAPES pelo apoio financeiro e à Universidade Federal de Minas Gerais e ao Departamento de Ciência da Computação, pela minha formação e pelo provimento da infraestrutura e ambiente para desenvolvimento de pesquisas de qualidade.

Resumo

As redes sociais *online* fazem parte do cotidiano de milhões de pessoas do mundo inteiro. Cada vez mais pessoas utilizam essas redes para interagir, opinar e compartilhar conteúdos sobre os mais diversos tópicos, como diversão, clima, trabalho, família, trânsito e mesmo sua condição de saúde. Em suma, as redes sociais se tornaram mais um lugar social com significados próprios, evoluindo dinamicamente. Muitos acontecimentos são tardiamente percebidos e divulgados pelos meios de comunicação tradicionais, mas podem acontecer nas redes sociais em tempo real, sendo passíveis de serem detectados e de subsidiarem a construção de modelos de previsão. O objetivo dessa dissertação é utilizar o conteúdo disponível nas redes sociais para detectar a ocorrência e prever eventos da vida real. Para realizar essas tarefas, foi proposta uma metodologia que compreende desde a coleta das mensagens em redes sociais até a previsão de ocorrência de eventos, passando pela análise da correlação entre a natureza do conteúdo das mensagens e ocorrência de eventos, em termos de volume, tempo e espaço. A metodologia proposta foi aplicada a dois tipos de eventos reais: epidemia de dengue e enchentes. No caso da epidemia de dengue, observa-se uma alta correlação (74%) entre mensagens expressando experiência pessoal e a incidência da doença, o que permitiu construir um sistema de alerta da epidemia por localidade com acurácia maior que 90% para cidades com alta incidência. Além disso, foi possível obter resultados comparáveis para o segundo evento, sendo esses capazes de detectar a ocorrência de pontos de alagamento e de prever sua intensidade diariamente. Isso demonstra a aplicabilidade dessa proposta como complemento a mecanismos de vigilância tradicional, muitas vezes permitindo que ações sejam antecipadas e impactos sobre a população afetada sejam reduzidos.

Abstract

Online social networks are part of the everyday life of millions of people worldwide. More and more people use these networks to interact, provide feedback and share content about several topics such as entertainment, weather, work, family, traffic and even their health. In short, social networks have become a social place with its own meanings and evolving dynamicity. Many events are perceived and released later by the traditional media, but can occur in social networks in real time, being capable of being detected and subsidize the construction of predictive models. The objective of this dissertation is to use the data available on social networks to detect the occurrence and provide real-life events. To accomplish these tasks, we propose a methodology that extends from the collection of messages on social networks to predict the occurrence of events, through analysis of correlation between the nature of the message content and the occurrence of events in terms of volume, time and space. The proposed methodology was applied to two types of actual events: floods and dengue epidemics. In the case of the dengue epidemic, we found a high correlation (0.74) between messages expressing personal experience and the incidence of the disease, which allowed the building of an warning system of the epidemic by location with an accuracy greater than 90% for cities with high incidence . We also got comparable results for the second type of event, being able to detect the occurrence of flooding points and predict its intensity every day, demonstrating the applicability of our proposal to complement traditional surveillance mechanisms, often allowing anticipated actions and minimizing the impact on the affected population.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Desde a sua criação a *World Wide Web*, ou Web, impactou e modificou diversos aspectos no cotidiano das pessoas. O seu rápido crescimento nas últimas décadas fez dela a maior e mais conhecida fonte de dados publicamente acessível, Liu [2009]. Essa fonte de dados pode ser facilmente incrementada, acessada e pesquisada. Antes da Web, para encontrar uma informação era necessário consultar um especialista ou pesquisar em livros sobre o assunto. Entretanto, hoje em dia tudo está a poucos cliques de distância.

A Web é utilizada não apenas para encontrar a informação desejada, mas também para compartilhar informação e conhecimento e servir de canal para negócios. Além disso, a Web provê maneiras convenientes para as pessoas se comunicarem, expressarem opiniões sobre qualquer assunto e discutirem com outras pessoas de qualquer lugar do mundo por meio das redes sociais *online*. Essas redes fazem parte do dia a dia de milhões de pessoas e proporcionam um meio de comunicação que é mundialmente difundido. Cada vez mais pessoas utilizam as redes sociais *online* para interagir, opinar e compartilhar conteúdos sobre os mais diversos tópicos, que variam desde diversão, clima, trabalho, trânsito, até sua própria condição de saúde.

As rede sociais têm chamado atenção de diversos pesquisadores que visam correlacionar seu conteúdo com os acontecimentos da vida real. Isso acontece porque muitos eventos são tardiamente percebidos e divulgados pelos meios de comunicação tradicionais, enquanto nas redes sociais podem ser difundidos imediatamente, sendo passíveis de serem detectados e de subsidiarem a construção de modelos de previsão. Para exemplificar, em Sakaki et al. [2010] os autores relatam que quando ocorre um terremoto no Japão diversas mensagens são publicadas no mesmo instante no Twitter e esses relatos foram utilizados para criar um modelo que encontra o centro do terremoto e sua trajetória.

Desta forma, um problema fundamental é até que ponto as informações presentes nas redes sociais refletem fidedignamente eventos reais e podem ser utilizadas para prevê-los. A solução desse problema compreende responder três questões. A primeira questão diz respeito a como selecionar dentre os dados coletados, conteúdo relevante sobre o evento. Outra questão tem como objetivo verificar se há uma correlação entre características do evento na vida real, tais como magnitude e tendência, e sua repercussão nas redes sociais, considerando o volume e o conteúdo da mensagem, assim como sua localização espaço temporal. A terceira questão verifica o potencial de usar os dados das redes sociais para realizar uma detecção antecipada do evento na vida real, por exemplo através de um alerta.

O objetivo principal desta dissertação é propor uma metodologia para detecção antecipada de eventos reais a partir das redes sociais. A metodologia proposta para detecção e previsão do eventos da vida real a partir das redes sociais é composta por cinco etapas principais que vão desde a coleta das mensagens em redes sociais até a elaboração de um alerta. Após a coleta das mensagens sobre o evento é realizada a análise do conteúdo dos textos das mensagens para selecionar aquelas que ajudarão na previsão do evento. Para verificar a viabilidade do uso dessas mensagens como instrumento para previsão do evento é realizada a análise de correlação composta por três partes. A primeira correlaciona o volume de mensagens publicadas com o volume de ocorrências do evento, a segunda etapa agrupa regiões próximas com quantidade similar de ocorrências do evento e a terceira parte considera o intervalo do tempo de chegada entre as mensagens. Verificada a correlação, as redes sociais podem ser consideradas insumos para a previsão do evento. A previsão do evento é feita tanto em termos de volume de ocorrência quanto em termos da gravidade da situação de cada localização. Finalmente, é proposto um alerta para visualização dessas informações.

Os eventos reais, alvo dessa pesquisa, são algum acontecimento ou eventualidade que possuem certas particularidades. As características que esses eventos devem ter são: devem ser comentados nas redes sociais pelas pessoas que o vivenciaram para que haja mensagens a serem coletadas sobre o evento; ser de larga escala, ou seja, um grande número de pessoas devem estar envolvidas com o evento ou participar dele; influenciar no cotidiano das pessoas que por alguma razão são induzidas a postar sobre o acontecimento; e ter tanto localização no espaço quanto no tempo definidos. São exemplos desses eventos: grandes festas, lançamentos de filmes, eventos esportivos, doenças, campanhas políticas e terremotos. Certamente há eventos de impacto social que, por algum motivo, não são comentados nas redes sociais e esses não são aplicáveis nessa metodologia.

As mensagens publicadas nas redes sociais sobre o evento devem satisfazer

algumas premissas para serem consideradas fonte de dados em tempo real. Essas mensagens devem ser geradas espontaneamente, ser referenciadas no tempo e no espaço e expressar alguma opinião ou sentimento. As redes sociais também devem satisfazer algumas premissas, dentre elas ser altamente utilizadas pela sociedade. Além disso, as redes sociais devem disponibilizar as mensagens publicadas com seu texto, seu *timestamp*, o usuário que a escreveu e a localização geográfica declarada pelo usuário. Se todas as premissas anteriormente citadas forem atendidas, as redes sociais poderão ser consideradas como uma fonte de dados capaz de refletir os acontecimentos da vida real.

A efetividade da metodologia foi demonstrada aplicando-a em dois eventos reais distintos: epidemia de dengue e alagamentos. A dengue é uma doença febril aguda transmitida entre as pessoas pela picada do mosquito *Aedes aegypti*. Essa doença ocorre e dissemina-se especialmente nos países tropicais e subtropicais, onde as condições do meio ambiente favorecem o desenvolvimento e proliferação do seu vetor. Em 2011 foram registrados aproximadamente 730 mil casos da doença no Brasil. Para prever uma epidemia, o Ministério da Saúde monitora a quantidade dos vetores transmissores e caso haja uma grande quantidade de inseto em determinada região, concentram-se as campanhas e os esforços de prevenção nesses locais.No entanto, a presença de vetor não é um preditor de casos de doença. Além disso, uma vez iniciada a epidemia em determinada região, as autoridades públicas só tomam conhecimento da epidemia com um atraso de semanas, impedindo uma agilidade nos serviços de saúde para lidar com esta epidemia e deixando o sistema de saúde pública sobrecarregado. Neste contexto, o sistema baseado nas informações das redes sociais serviria para antever esta epidemia de forma mais rápida, permitindo um melhor planejamento por parte do governo. Dentre os resultados obtidos ao utilizar a metodologia proposta destaca-se a alta correlação (74%) encontrada entre as mensagens postadas e os casos de dengue notificados. É importante ressaltar que ao utilizar essas mensagens como insumo para previsão dos casos em cada município, metade desses possuem correlação superior a 60%, sendo que em cidades como Rio de Janeiro e Manaus os valores foram de 95% e 86% respectivamente.

O segundo grupo de eventos são os alagamentos e as enchentes que acontecem em diversos municípios do Brasil devido às fortes chuvas que costumam cair no verão. Como consequência, milhares de pessoas perdem seus bens, ficam desabrigadas e ficam sujeitas a desastres que podem causar vítimas fatais. Para monitorar os alagamentos e enchentes é utilizado um sistema que considera dados das chuvas e níveis de água nos rios, mas esse sistema não é disponível em todas as cidades do Brasil. Nesse contexto, utilizar as mensagens postadas nas redes sociais que se referem a esses eventos na

criação um sistema de alerta pode ajudar a informar mais pessoas rapidamente sobre a situação, diminuindo o número de vítimas. Dentre os resultados, é importante ressaltar a alta correlação (79%) obtida entre as mensagens coletadas e os pontos de alagamento e a previsão da situação de gravidade dos alagamentos, que foi correta em 81% dos dias.

A utilização dos dados provenientes das redes sociais pode ser vista como um complemento a mecanismos de vigilância tradicional, muitas vezes permitindo que ações sejam antecipadas e impactos sobre a população afetada sejam reduzidos. A metodologia proposta nessa dissertação assim como os resultados obtidos no contexto da dengue são utilizadas no Observatório da Dengue (http://www.observatorio.inweb.org.br/dengue/) com propósito de prever possíveis casos da doença e alertar sobre sua situação em cada cidade brasileira. Uma parceria foi firmada entre o Observatório da Dengue e o Ministério da Saúde com intuito de utilizar essa ferramenta como um sistema de vigilância complementar ao tradicional. O alerta desenvolvido nessa dissertação é disponibilizado ao Ministério da Saúde por meio de uma página web de acesso restrito que contém a avaliação da situação atual da incidência e da tendência da doença.

A metodologia proposta e parte dos resultados dessa dissertação foram publicados em Gomide et al. [2011] e Silva et al. [2011]. O artigo Gomide et al. [2011] apresentado no congresso *Web Science* 2011 foi reportagem no jornal alemão *Rhein Zeitung*[1], foi citado pela revista *NewScientist*[2] e pela revista brasileira Época[3]

A dissertação está organizada em quatro capítulos, além desta Introdução. O Capítulo 2 lista os trabalhos relacionados e explica como eles são complementados por este trabalho. No Capítulo 3 é apresentada a metodologia proposta. Em seguida, são apresentados os experimentos realizados com os dois eventos reais, epidemia de dengue e alagamentos no Capítulo 4. Finalmente, as conclusões do trabalho são apresentadas no Capítulo 5, bem como os trabalhos futuros.

[1]Link para matéria do jornal *Rhein Zeitung* sobre o trabalho Gomide et al. [2011]: (último acesso em 11/02/2012) http://www.rhein-zeitung.de/nachrichten/wissenschaft_artikel,-Twittern-bis-der-Arzt-kommt-Informatiker-entdecken-in-Brasilien-Denguefieber-Ausbrueche-_arid,263822.html

[2]Link para matéria da revista *NewScientist* que cita o trabalho Gomide et al. [2011]: (último acesso em 11/02/2012) http://www.newscientist.com/article/mg21128215.600-twitter-to-track-dengue-fever-outbreaks-in-brazil.html

[3]Link para matéria da revista Época que cita o Observatório da Dengue: (último acesso em 11/02/2012) http://revistaepoca.globo.com/Revista/Epoca/0,,EMI251340-15257,00-PROJETO+MONITORA+CASOS+DE+DENGUE+VIA+TWITTER.html

Capítulo 2

Trabalhos Relacionados

A Web tem chamado atenção de diversos pesquisadores devido à imensa quantidade de dados publicamente acessível e do seu caráter de tempo real. Três categorias de mineração de dados Web foram identificadas em Kosala & Blockeel [2000]. A primeira é a mineração do conteúdo Web que aplica técnicas de mineração de dados em conteúdos publicados na Internet tais como HTML, textos ou XML. A segunda é a mineração da estrutura da Web que opera na estrutura dos *hiperlinks*, a qual pode por exemplo, prover informações sobre o *page ranking* e melhorar os resultados de pesquisas. E, finalmente, a mineração do uso da Web que analisa o resultado das interações entre os servidores Web tais como logs, fluxos de cliques e transações em banco de dados.

Dentre as pesquisas realizadas na área de mineração de dados Web podemos citar a classificação de documentos para classificar conteúdo estruturado e semi-estruturado da Web na forma de *tags* HTML, como feito por exemplo em Weiss et al. [1996]. O trabalho realizado por Kumar et al. [1999] faz a identificação de comunidades Web utilizando dados de *hiperlinks*. Em Schafer et al. [2001] os autores notaram que sistemas de recomendação podem melhorar o comércio virtual ao tentar aumentar a venda cruzada de produtos relacionados. Por exemplo, a *Amazon.com* usa dados dos produtos do carrinho de compras para recomendar outros produtos. Esses são apenas alguns exemplos de trabalhos na área de mineração de dados Web.

Entretanto, a Web 2.0 não interliga apenas documentos ou páginas, mas também pessoas e organizações por meio das redes sociais *online*. A mineração de redes sociais visa extrair conhecimento a partir do conteúdo disponível nas redes sociais. Em Benevenuto et al. [2009] o comportamento do usuário foi caracterizado quanto à frequência com que esses se conectam e quanto aos tipos e sequências de atividades realizadas nas redes sociais. A influência das pessoas nas redes sociais é medida em Cha et al. [2010] ao comparar de forma detalhada três métricas: o grau de entrada,

os *retweets* e as menções dos usuários no Twitter. O artigo Guerra et al. [2011] usa endossos mútuos para aprender o viés entre os usuários e assim classificar opiniões manifestadas por esses nas mídias sociais em relação a um tópico.

Recentemente, alguns artigos demonstraram como os conteúdos disponíveis nas mídias sociais e na Web podem ser utilizados para detectar e prever eventos do mundo real. Em Tumasjan et al. [2010] a opinião sobre a eleição governamental alemã identificada nos *tweets* teve grande correlação com o resultado oficial das eleições. Similarmente, em Goel et al. [2010] o volume de consultas feitas no Yahoo! foi utilizado para prever a bilheteria da estreia dos filmes, as vendas de vídeo games e o *rank* das músicas na *Billboard Hot 100*. Asur & Huberman [2010] demonstraram como mensagens do Twitter podem ser usadas para prever bilheteria de filmes.

Além das utilizações citadas acima, a Internet também tem sido usada para monitorar surtos de doenças. Os primeiros trabalhos nessa direção utilizavam artigos de jornais que mencionavam a Influenza como fonte de informação sobre os surtos (Mawudeku & Blench [2006]; Brownstein et al. [2008]; Freifeld et al. [2008]).

Recentemente, o conteúdo disponível na Web vem sendo utilizado seja por meio da mineração das publicações relacionadas a doença em *blogs* (Corley et al. [2009]), seja com a análise dos registros nos sites de busca sobre consultas feitas relacionadas com a Influenza (Ginsberg et al. [2009]; Chan et al. [2011]; Althouse et al. [2011]), ou ainda por meio das mensagens postadas no Twitter (Culotta [2010]; Lampos & Cristianini [2010, 2011]; Lampos et al. [2010]; Chen et al. [2010]; Achrekar et al. [2011]).

Também são alvos de pesquisas que utilizam dados Web eventos que causam situações emergenciais, como terremotos e enchentes. Em Winerman [2009] o autor afirma que à medida que ocorre um evento que causa pânico, as pessoas buscam informações nas redes sociais. Esse artigo cita como exemplo a tragédia da Virgina Tech onde estudantes conseguiram formular uma lista completa de todos estudantes falecidos um dia antes das autoridades. O comportamento das pessoas nas redes sociais durante situações de emergência também tem sido tópico de pesquisa. Em Mendoza et al. [2010] e Starbird & Palen [2010] os autores determinaram como informações foram divulgadas em toda a rede por meio de *retweets* de notícias durante dois desastres naturais, a enchente do Rio Vermelho e incêndios em Oklahoma. As mensagens publicadas no Twitter também foram utilizadas para prever a ocorrência de terremotos em Sakaki et al. [2010] e Lampos & Cristianini [2011].

Esses artigos são diretamente relacionados ao trabalho proposto nessa dissertação, visto que utilizam as redes sociais como fonte de informação para detectar e/ou prever um acontecimento da vida real. Esses trabalhos se diferenciam quanto à metodologia utilizada para resolver esse problema. Os aspectos que os diferem são os dados

utilizados, a forma como esses são coletados e analisados, as possíveis técnicas para correlacionar esses com os dados reais e a maneira de realizar a previsão dos eventos.

A seguir é apresentada uma análise detalhada sobre a metodologia utilizada em cada um dos trabalhos relacionados. Os métodos analisados são: coleta dos dados, análise de conteúdo, análise de correlação, previsão usando redes sociais e sistemas de alerta. Cada seção corresponde a uma parte da metodologia proposta, sendo que nessas seções serão apresentados os métodos utilizados pelos artigos relacionados para realizarem essa tarefa e esses por sua vez serão contrastados com os métodos propostos nessa dissertação.

2.1 Coleta dos dados

Os dados utilizados pelos artigos relacionados se diferem pela fonte na qual são obtidos e pela maneira que são coletados.

Inicialmente, as fontes de dados *online* utilizadas eram notícias de jornais disponibilizadas pelo Google News, feed de notícias RSS ou e-mail do ProMED (Mawudeku & Blench [2006]; Brownstein et al. [2008]; Freifeld et al. [2008]). Em Corley et al. [2009] os dados analisados foram as publicações em blogs fornecidos pelo *Spinn3r*, que disponibilizou um total de 44 milhões de *posts* coletados de agosto a outubro de 2011.

Outros trabalhos utilizaram os registros das pesquisas realizadas pelos usuários das máquinas de busca. Em Goel et al. [2010] foram considerados os registros das consulta feitas no Yahoo! e na página music.yahoo.com. Já em Ginsberg et al. [2009] e Chan et al. [2011] foram utilizados os *logs* de consultas realizadas no Google. Além das consultas propriamente ditas, outros dados do Google também já foram considerados. Em Eysenbach [2006] os dados utilizados foram o número de clicks em propagandas feitas no Google, o Google Adsense, e em Althouse et al. [2011] foram usadas as estatísticas fornecidas pelo Google Insights sobre as consultas realizadas.

As redes sociais também já foram utilizadas para previsão dos acontecimentos da vida real. Na grande maioria dos artigos o Twitter foi utilizado para coletar as mensagens publicadas, sendo que há diferentes maneiras de coletar essas mensagens. Alguns autores (Asur & Huberman [2010], Sakaki et al. [2010], Tumasjan et al. [2010], Achrekar et al. [2011]) coletaram apenas mensagens que contenham termos relacionados ao evento que estão analisando. Outras pesquisas (Culotta [2010]; Ritterman et al. [2009]) coletaram todo o conteúdo do Twitter durante algumas semanas ou meses. E a outra forma de coleta é determinar as localidades de interesse e coletar todas as

publicações feitas a um raio de 10km como feito em Lampos & Cristianini [2010, 2011].

Segundo Culotta [2010] existem diversos motivos para considerar um modelo baseado em mensagens das redes sociais para prever um evento da vida real, ao invés de considerar registros de consultas em máquinas de busca. Primeiramente, as mensagens completas fornecem uma informação mais descritiva do que as consultas para caracterizar o evento. Além disso, os perfis dos usuários contém informações como localização, idade e sexo, o que possibilita um estudo estatístico mais detalhado permitindo que seja realizada uma análise demográfica.

Em [Ginsberg et al., 2009] os autores notaram que um evento não usual como *recall* de medicamentos poderia causar um alarme falso na previsão do surto de H1N1 ao considerar todas as consultas realizadas no Google. Ao utilizar as publicações do Twitter não é possível afirmar que o sistema é imune à falsos alertas, mas com um algoritmo de classificação de conteúdo é possível classificar o conteúdo das mensagens e eliminar parte das mensagens não relacionadas. Finalmente, o conteúdo do Twitter é publicamente disponível possibilitando a reprodutibilidade e o acompanhamento da pesquisa. Devido a esses motivos, foi escolhido utilizar as redes sociais, mais especificamente o Twitter, como fonte de dados *online*.

2.2 Análise de conteúdo

A análise de conteúdo tem como objetivo classificar o conteúdo das mensagens publicadas nas redes sociais e filtrar mensagens não relacionadas ao evento de interesse. Alguns artigos (Goel et al. [2010]; Achrekar et al. [2011]) simplesmente ignoram o conteúdo das mensagens e consideram todas as mensagens coletadas que contenham termos relacionados com o evento. Essa abordagem é bastante vulnerável, uma vez que está sujeita a considerar mensagens irônicas ou uma grande divulgação de um fato relacionado ao assunto de interesse.

Os artigos que coletaram todo o conteúdo do Twitter durante um certo período (Ritterman et al. [2009]; Culotta [2010]; Lampos & Cristianini [2010, 2011]) ou que utilizaram todas as consultas realizadas nas máquinas de busca (Ginsberg et al. [2009]; Chan et al. [2011]), consideram a porcentagem desses dados que contém termos relacionados ao evento. Em Culotta [2010] esse processo é feito em três etapas. Primeiro é feita uma seleção de palavras-chave e são utilizadas apenas algumas que reportam os sintomas da gripe H1N1 para selecionar os documentos. Depois, é realizada a geração de mais palavras-chave ao considerar as 5000 palavras mais frequentes nos documentos e assim é feito mais uma vez a seleção dos documentos. Finalmente, os documentos

não correlacionados são eliminados por meio de um classificador binário cujas classes são: positiva se reporta um sintoma e negativa, caso contrário. A validação cruzada foi utilizada para avaliar o classificador, cuja acurácia foi de 84%. Os autores Lampos & Cristianini [2010, 2011] fazem a extração automática das palavras-chave utilizando o LASSO (Tibshirani [1994]). Já em Ginsberg et al. [2009] e Chan et al. [2011] é feita uma avaliação de quantas palavras-chave devem ser utilizadas para melhor separar a parte das consultas relacionadas a influenza e a dengue, respectivamente. Após definida a quantidade de palavras-chave com a qual se obtém a maior correlação, são selecionadas as consultas que citam pelo menos uma delas.

Existem trabalhos que utilizam ferramentas de análise de sentimentos para avaliar o conteúdo das mensagens. Em Tumasjan et al. [2010] os sentimentos são extraídos automaticamente utilizando o LIWC2007 (Linguistic Inquiry and Word Count), um software que faz análise do texto para obter componentes emocionais, cognitivos e estruturais que usa um dicionário psicométrico. Esse software determina a taxa que certas cognições e emoções (p.ex., orientação do futuro, emoções positivas e negativas) estão presentes no texto. Outro exemplo de software que faz análise de sentimento é o LingPipe (www.alias-i.com/lingpipe) e foi utilizado em Asur & Huberman [2010] para classificar os *tweets* em positivo, negativo ou neutro. Para gerar um conjunto de treino foi utilizado o *Amazon Turk* (www.mturk.com) e a acurácia obtida nessa classificação foi de 98%.

Ao invés de analisar o sentimento das mensagens como feito nos artigos citados no parágrafo anterior, em Corley et al. [2009] os blogs foram classificados em três classes de acordo com suas publicações. As classes utilizadas foram: uma identificação própria de sintoma; a identificação de outra pessoa que tem sintoma (segunda mão); ou um artigo objetivo (ou opinião).

Uma outra maneira de analisar o conteúdo das mensagens é por meio do algoritmo *Support Vector Machine*, ou SVM, como feito em Sakaki et al. [2010]. Nesse trabalho foi utilizado o SVM para avaliar se o *tweet* está realmente se referindo a uma ocorrência de terremoto. As mensagens são classificadas em duas classes, positiva caso o *tweet* se refira a um terremoto que ocorreu no momento da publicação e negativa caso contrário. Três grupos de atributos são utilizados para construir o classificador, são eles: o número de palavras no tweet, o número de palavras chave no *tweet* e as palavras antes e depois da palavra chave para criar o seu modelo de classificação.

A análise de conteúdo realizada nessa dissertação se diferencia dos trabalhos citados acima nos seguintes aspectos. Primeiramente, as categorias de conteúdo são determinadas de acordo com a natureza do evento e a análise que se deseja realizar. Por isso propomos duas taxonomias diferenciadas, uma mais detalhada que contém cinco

classes e, outra que contém apenas duas classes. O outro aspecto é quanto ao algoritmo utilizado na classificação das mensagens. Para classificar automaticamente é necessário um algoritmo capaz de lidar com um grande volume de dados mesmo contando com um pequeno conjunto de treino e também de lidar com o desbalanceamento de classes. Um algoritmo que atende esses critérios é o *Lazy Associative Classification*, ou LAC, Veloso et al. [2006].

2.3 Análise de correlação

A análise de correlação entre os dados *online* e os eventos na vida real deve ser realizada com intuito de verificar a viabilidade de utilizar a Web para previsão dos mesmos. Entretanto, alguns trabalhos (Asur & Huberman [2010]; Sakaki et al. [2010]; Culotta [2010]; Goel et al. [2010]; Ritterman et al. [2009]; Achrekar et al. [2011]) não fazem nem uma caracterização mais detalhada sobre os dados coletados nem uma análise de correlação.

Em Tumasjan et al. [2010] os autores compararam a porcentagem de atenção recebida por cada partido no Twitter com os resultados da eleição do governo alemão. Além disso, analisaram se é possível inferir os laços ideológicos entre os partidos e as potenciais coalisões políticas depois da eleição por meio do conteúdo dos *tweets*.

Outros trabalhos (Lampos & Cristianini [2010, 2011]) fazem a correlação entre as séries temporais do volume de mensagens do Twitter e do índice de H1N1 e as taxas de chuva por região.

A análise de correlação possivelmente mais semelhante à realizada nesta dissertação é feita em Eysenbach [2006], onde os autores calcularam o coeficiente de correlação de Pearson entre o número de cliques nas palavras-chaves no Google com os dados epidemiológicos da flu no Canadá em um período de 33 semanas.

Nessa dissertação, a análise de correlação proposta calcula o coeficiente de correlação de Pearson entre a série temporal gerada pelas mensagens do Twitter e a série temporal dos registros oficiais sobre o evento. Esse coeficiente é calculado separadamente para cada localidade, visto que é importante considerar a localização geográfica para caracterizar o evento.

2.4 Previsão

A previsão de ocorrência do evento ou de sua intensidade pode ser realizada de diversas maneiras. Entretanto, alguns trabalhos relacionados (Tumasjan et al. [2010]; Corley

et al. [2009]) realizam apenas uma análise comparativa entre os dados oficiais do evento e os dados obtidos pela Web e não se preocupam com a previsão do evento.

Em Sakaki et al. [2010] são utilizados modelos probabilísticos para prever a localização e a trajetória dos terremotos no Japão. Foi utilizado um modelo temporal para aproximar o número de *tweets* de uma distribuição exponencial e obteve-se uma correlação de 87%. O modelo espacial utilizado para encontrar a localização e inferir a trajetória se baseia em métodos de estimação, como o filtro Kalman e o filtro de partícula.

Outra técnica que pode ser utilizada é a Média Móvel Autoregressiva (ARMA), a qual é apresentada no contexto de previsão de eventos em Chen et al. [2010]. Essa abordagem também foi utilizada em Achrekar et al. [2011] para prever a ocorrência da gripe H1N1 utilizando mensagens do Twitter.

A regressão linear foi utilizada na grande maioria dos trabalhos relacionados (Asur & Huberman [2010]; Culotta [2010]; Goel et al. [2010]; Ritterman et al. [2009]; Lampos & Cristianini [2010]; Ginsberg et al. [2009]; Chan et al. [2011]; Althouse et al. [2011]) e é a técnica que foi escolhida para a previsão de eventos implementada nessa dissertação. Em Asur & Huberman [2010] uma correlação superior a 90% foi gerada pela função de regressão linear, a qual utiliza as mensagens publicadas antes da estreia dos filmes para prever a bilheteria no final de semana de estreia.

Em Ginsberg et al. [2009] foi criado um modelo de regressão linear para prever a Influenza em nove regiões dos Estados Unidos. Um modelo para cada uma dessas regiões foi desenvolvido e a correlação média obtida foi de 0.9 (min = 0.8 e max = 0.96). A função de previsão gerada pela regressão linear foi criada utilizando 128 pontos (75% dos dados) e validada usando 42 pontos (25%). Para validar a previsão a nível de estado apenas Utah foi analisado, uma vez que não havia dados disponíveis sobre a Influenza para outros estados. Em Chan et al. [2011] os autores criaram um modelo de regressão linear para ajustar as séries temporais da fração de pesquisa do Google com o volume de casos oficiais da doença. Foi criada uma função para cada um dos nove países analisados e os dados utilizados foram do período entre 2003 e 2010. A correlação entre a função de previsão gerada e os dados oficiais foi de 0.82 a 0.99.

Há duas diferenças entre a previsão realizada nos trabalhos aqui apresentados e a previsão proposta nessa dissertação que devem ser ressaltadas. A primeira é quanto à especificidade da localização que é considerada em termos de cidades nos experimentos dessa dissertação e, na maioria das vezes, nem é diferenciada nos modelos criados nos trabalhos relacionados. A segunda grande diferença é a proposta da classificação da gravidade do evento em níveis de intensidade. Essa classificação não foi realizada em nenhum dos trabalhos encontrados.

2.5 Alerta

O alerta deve ser gerado para informar a população quando uma situação crítica do evento acontecer. A grande maioria dos trabalhos relacionados (Asur & Huberman [2010]; Culotta [2010]; Tumasjan et al. [2010]; Goel et al. [2010]; Ritterman et al. [2009]; Achrekar et al. [2011]; Corley et al. [2009]; Althouse et al. [2011]) não se preocupam em desenvolver um sistema de alerta.

Diferentes sistemas de alerta já foram desenvolvidos. Em Sakaki et al. [2010] foi implementada uma aplicação para reportar a ocorrência de terremotos no Japão. O sistema detecta o terremoto prontamente e envia e-mails para os usuários registrados. Nos testes realizados, essas notificações foram mais rápidas do que os anúncios espalhados pelo órgão oficial, o JMA.

Em Freifeld et al. [2008] uma página Web permite visualizar por meio do Google Maps as notícias publicadas sobre a Influenza. Outra ferramenta *online* implementada em Lampos et al. [2010], disponibiliza as taxas sobre essa doença que são inferidas baseadas no modelo proposto em Lampos & Cristianini [2010].

Os trabalhos apresentados em Ginsberg et al. [2009] e Chan et al. [2011] resultaram em ferramentas *online* que podem ser utilizadas para acompanhar a situação da Influenza e da dengue, respectivamente. O Google Flu Trends (http://www.google.org/flutrends/) disponibiliza um mapa mundial com a intensidade em que são realizadas pesquisas no Google sobre a Influenza em cada país e/ou estados. Há também o Google Dengue Trends (http://www.google.org/denguetrends/) que disponibiliza a frequência das consultas ao Google sobre a Dengue no Brasil e mais outros 9 países.

A principal diferença entre os trabalhos aqui apresentados e esta dissertação está no grau de especificidade da localização com a qual o evento é analisado e a periodicidade de atualização do status do alerta. O alerta proposto analisa os eventos semanalmente por cidade e não em termos de estados ou países como feito em Ginsberg et al. [2009]; Chan et al. [2011]. Além de disponibilizar a situação do evento prevista para o momento atual, foi criado um indicador da tendência do evento que visa informar se está aumentando, diminuindo ou mantendo o volume de ocorrências do evento.

Capítulo 3

Metodologia

A metodologia proposta nesse trabalho explora em profundidade cada etapa necessária para detecção e previsão do eventos da vida real a partir das redes sociais. Ao longo desse capítulo será explicado como selecionar conteúdo relevante sobre o evento dos dados coletados nas redes sociais, como verificar a correlação entre o evento na vida real e sua repercussão nas redes sociais, como usar as redes sociais como insumo para previsão do evento na vida real, como medir a tendência do evento e como utilizar os dados virtuais para criar um alerta sobre o evento. Parte da metodologia aqui proposta é apresentada em Gomide et al. [2011].

3.1 Visão Geral

A Figura 3.1 contém uma visão geral da metodologia. Os insumos utilizados são as redes sociais *online* e os dados oficiais. Os produtos gerados durante o desenvolvimento são os bancos de dados das mensagens georeferenciadas e das mensagens classificadas pelo conteúdo. Além disso, essa Figura mostra as cinco etapas da metodologia: (1) coleta das mensagens das redes sociais, (2) análise de conteúdo, (3) análise de correlação, (4) previsão e (5) alerta.

Definido o evento de interesse, a primeira etapa da metodologia é a coleta das mensagens que são publicadas nas redes sociais. Essas mensagens, relacionadas ao evento, servirão de insumo para a próxima etapa, cujo objetivo é classificar o conteúdo das publicações e selecionar as que são relevantes para as análises seguintes.

A terceira etapa da metodologia visa correlacionar os dados virtuais, sendo esses mensagens coletadas das redes sociais, com os dados reais, que são capazes de oferecer evidências sobre o evento. Essa correlação é feita tanto na dimensão temporal, quanto na dimensão espacial. Caso haja correlação entre os dados reais e virtuais, a próxima

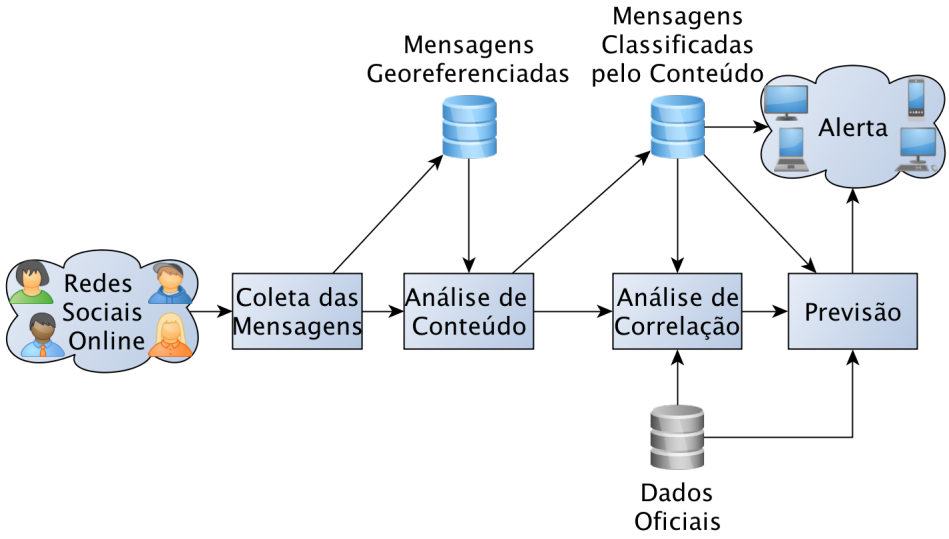


Figura 3.1: Visão geral da metodologia. As cinco etapas da metodologia são: coleta das mensagens, análise de conteúdo, análise de correlação, previsão e alerta.

etapa utiliza as mensagens das redes sociais como previsores do evento. O objetivo nessa fase é prever o número de ocorrências do evento e a situação de gravidade do mesmo.

A última parte do método utilizado é a elaboração do alerta que agrega todas as etapas anteriores e disponibiliza essa informação de forma visual.

Cada uma das etapas serão apresentadas em detalhe nas seções seguintes.

3.2 Coleta das Mensagens nas Redes Sociais Relacionadas ao Evento

As redes sociais que serão utilizadas devem ter algumas premissas para que seja possível utilizá-las como fonte de dados em tempo real. Essas redes devem ser altamente utilizadas pela sociedade, devem disponibilizar as mensagens publicadas com seu texto, seu *timestamp*, o usuário que a escreveu e a localização geográfica declarada pelo usuário. Uma das redes sociais que possui essas características é o Twitter.

O Twitter está entre uma das redes sociais mais utilizadas no Brasil juntamente com Orkut, Windows Live Profile e Facebook segundo uma pesquisa realizada em Agosto de 2010 pela comScore [2010]. Algumas dessas redes sociais disponibilizam seu conteúdo, enquanto outras mantém seus dados privados. Por exemplo, o Orkut e o Windows Live Profile não permitem coletar dados. Já o Facebook provê uma API (Interface de Programação de Aplicativos) (http://developers.facebook.com/) para

coletar seu conteúdo, mas a localização do usuário, informação fundamental para caracterização dos eventos, é disponível apenas sob autorização dele, o que inviabiliza a utilização dessa rede. Por fim, o Twitter é o único que disponibiliza seu conteúdo para coleta.

O Twitter fornece diversas API's (https://dev.twitter.com/docs) para tornar seu conteúdo disponível. Pode-se obter a rede de seguidores das pessoas, as mensagens publicadas (*tweets*) por usuários, por região geográfica, por data ou até mesmo por palavras específicas. Sem perda de generalidade e para facilitar a leitura e compreensão vamos discutir o restante da metodologia utilizando o Twitter, embora a metodologia proposta possa ser aplicada a quaisquer outras redes sociais que satisfizessem as mesmas premissas.

Para observar o evento de interesse foram coletadas as mensagens que contenham com menções às palavras relacionadas ao assunto. O primeiro passo para obtenção desses dados é a escolha dos termos adequados. Após definidos os termos, a coleta das mensagens publicadas no Twitter é iniciada, para que seja possível fazer o georeferenciamento das mesmas. As etapas necessárias desde a coleta das mensagens até a geração de um banco de dados com *tweets* georeferenciados, estão descritas na Figura 3.2.

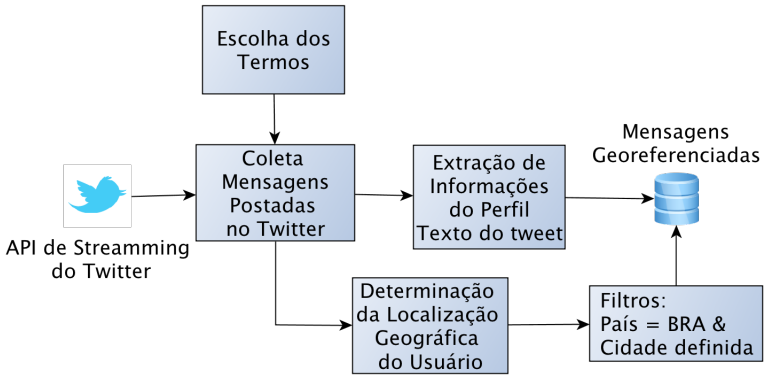


Figura 3.2: Diagrama contendo as fases da coleta das mensagens.

3.2.1 Escolha dos Termos

No intuito de obter as mensagens relacionadas ao evento, é criado um conjunto de termos que o caracterizam ou que são utilizados para referenciá-lo. As palavras devem abranger o assunto mesmo que para isso sejam coletadas publicações ambíguas ou não

relacionadas diretamente ao evento. As mensagens não relacionadas serão eliminadas posteriormente.

A escolha dos termos pode ser feita por meio de consulta à especialistas sobre o assunto ou utilizando os termos mais citados nas mensagens e reportagens previamente selecionadas.

3.2.2 Coleta das Mensagens Publicadas no Twitter

A API disponibilizada pelo Twitter para obter as mensagens relacionadas aos eventos é a Streaming API (https://dev.twitter.com/docs/streaming-api) que coleta em tempo real as publicações que possuem menções aos termos previamente escolhidas.

As mensagens, *tweets*, contém diversas informações, dentre elas as que nos interessam são: o identificador; o texto (limitado a 140 caracteres); o usuário; a localização do usuário; a data e horário que o *tweet* foi escrito (GMT+0). O identificador, o texto, o usuário, a data e o horário são obtidos diretamente pela API. Porém, a localização do usuário requer um último passo para ser propriamente definido. A determinação da localização é descrita a seguir.

3.2.3 Determinação da Localização Geográfica do Usuário

Conforme já mencionado, o local do evento é fundamental para sua caracterização. Nesta etapa, a localização do usuário é identificada procurando obter o maior nível de detalhe possível a partir das informações disponíveis nos *tweets* coletados.

Nos *tweets* podem haver várias informações para identificar a localização do usuário. As que utilizamos em ordem de prioridade pelas mais específicas são: o atributo *geo* que contém as coordenadas do *tweet* quando o usuário tem GPS em seu aparelho móvel; o atributo *place* através do qual o usuário declara o lugar, de uma lista de lugares cadastrados; ou o atributo *location* do objeto *user* que contem uma declaração em texto livre do possível lugar.

A informação contida no campo *location* é escrita em texto livre e pode conter locais inválidos como "Marte" ou "céu" inviabilizando a sua utilização conforme obtida pela API do Twitter. Nesse caso, a API Google Geocoding (http://code.google.com/apis/maps/documentation/geocoding/) permite filtrar locais inválidos e determinar exatamente a localização dos usuários que disponibilizam informações incompletas (p.ex., bh ao invés de Belo Horizonte).

Para exemplificar, considere que a informação declarada pelo usuário, no campo *location*, seja "bh". A requisição http feita para a API do Google Geocoding é

`http://maps.googleapis.com/maps/geo?q=bh` e a resposta obtida é a seguinte:


```
{  "name": "bh",
 "Status": {"code": 200,"request": "geocode"},
 "Placemark": [ {    "id": "p1",
   "address": "Belo Horizonte - Minas Gerais, Brazil",
   "AddressDetails": {
      "Accuracy" : 4,
      "Country" : { "AdministrativeArea" : {
        "AdministrativeAreaName" : "MG",
        "Locality" : {"LocalityName" : "Belo Horizonte"} },
        "CountryName" : "Brasil","CountryNameCode" : "BR"} },
  "ExtendedData": {
     "LatLonBox": {"north": -19.8351218,"south": -20.0029691,
       "east": -43.8105153,"west": -44.0666341} },
   "Point": {"coordinates": [ -43.9385747, -19.9190677, 0 ]}} ]
}
```


Sobre a resposta obtida pela API do Google Geocoding os seguintes campos merecem destaque. O campo *Accuracy* pertencente ao objeto *AddressDetails* fornece o nível de detalhe da localização obtida. Para localizações a nível de cidade, o valor do campo *Accuracy* é 4. Para localizações menos detalhadas (estado ou país) o valor desse campo é inferior a 4, e para localizações mais detalhadas (ruas, endereço completo) o valor é maior que quatro.

O objeto *Country* contém localização propriamente dita esquematizada na estrutura de cidade (*LocalityName*), estado (*AdministrativeAreaName*) e país (*CountryName*). A latitude e longitude estão no campo *coordinates* do objeto *Point*.

É importante ressaltar que a localização obtida é a declarada pelo usuário e pode não representar sua localização no momento da postagem. Por exemplo, um usuário que cuja localização declarada é Porto Alegre publicou uma mensagem sobre um evento que ocorreu durante sua viagem ao Rio de Janeiro.

Depois de obter a localização de cada mensagem, são selecionadas apenas aquelas de usuários do Brasil e com informação a nível de cidade.

3.3 Análise de Conteúdo

Durante análise e caracterização apresentadas nesta seção, as categorias de conteúdo são definidas, o algoritmo da análise de conteúdo é apresentado e as mensagens são classificadas.

Dentre as mensagens coletadas no Twitter, há aquelas que não são diretamente relacionadas ao evento. Por exemplo, considere que o evento de interesse seja terremoto, o *tweet* cujo texto é "Estou tremendo de medo da prova" não tem nenhuma relação com a ocorrência de um terremoto, apesar de conter o termo "tremendo" tipicamente usado para referenciar esse evento. No intuito de eliminar os *tweets* não relacionados ao evento é feita a análise de conteúdo do texto das mensagens.

Além de viabilizar a seleção apenas das mensagens que estejam diretamente relacionadas à ocorrência do evento, a análise de conteúdo também permite que a percepção do público sobre o assunto seja conhecida.

Devido ao grande número de mensagens, é inviável classificar todas as mensagens manualmente. Nesse sentido, um algoritmo de classificação é utilizado para estimar o conteúdo expresso no texto dos *tweets*.

A análise de conteúdo, Figura 3.3, se divide em duas fases: criação das categorias e a classificação do conteúdo. Na primeira fase, são definidas as categorias de conteúdo das mensagens. Na classificação do conteúdo, um conjunto de mensagens é classificado manualmente para ser utilizado como treino pelo classificador e, finalmente, todo o conjunto de mensagens é classificado.

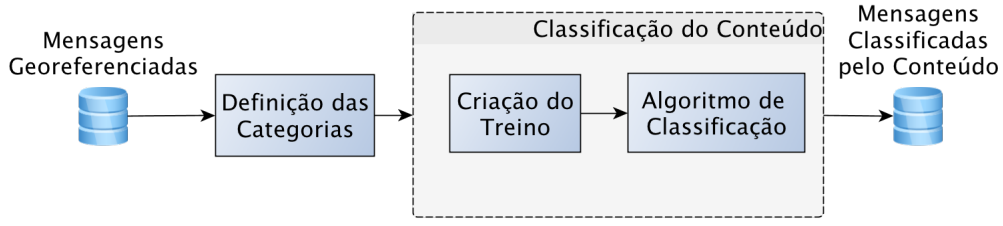


Figura 3.3: Diagrama contendo as fases da etapa de análise de conteúdo.

3.3.1 Definição das Categorias

As categorias de conteúdo devem fornecer a informação necessária para eliminar as mensagens que, apesar de conterem pelo menos alguma palavra-chave sobre o evento, não estão relacionadas com sua ocorrência. Além disso, as categorias podem ser definidas de forma que seja possível conhecer a percepção do usuário sobre o evento.

Nessa seção são apresentadas duas taxonomias para classificar mensagens que já foram utilizadas em trabalhos anteriores e possuem diferentes abordagens. A primeira taxonomia descrita em Chew & Eysenbach [2010] é composta por cinco classes e foi aplicada na classificação das mensagens publicadas no Twitter sobre a Influenza. A segunda, é composta por duas classes e foi utilizada em Sakaki et al. [2010] para classificar a ocorrência de terremotos. A escolha de qual classificação usar depende da natureza do evento e da análise que se deseja realizar. A seguir, as duas taxonomias serão explicadas detalhes.

A taxonomia composta por cinco categorias descrita em Chew & Eysenbach [2010] abrange diversos conteúdos que um texto pode ter. Os conteúdos nos quais as mensagens podem ser classificadas são: informação, experiência pessoal direta ou indireta, reações pessoais ou opiniões, piadas ou paródias e campanhas/propagandas. A descrição de cada classe se encontra na Tabela 3.1.

A grande maioria dos eventos pode ser classificada utilizando a taxonomia descrita na Tabela 3.1. Essa classificação é bastante detalhada e oferece várias visões sobre o evento. É possível observar qual a repercussão de campanhas públicas sobre o evento, conhecer a opinião das pessoas, mensurar a divulgação de informações do evento, perceber as críticas sobre o evento por meio de piadas ou ironias e, por fim, identificar as pessoas que vivenciaram esse acontecimento.

Alguns eventos podem ser analisados sob outro aspecto e por isso, uma segunda taxonomia para classificação de conteúdo também é discutida. Em Sakaki et al. [2010] é apresentada uma taxonomia composta por duas classes para verificar se um tweet relata a ocorrência de um terremoto em tempo real ou não. A descrição das duas classes está na Tabela 3.2. A primeira classe de conteúdo é para os *tweets* que descrevem uma situação do momento, algo que o usuário está vivendo no exato momento da publicação. A segunda classe é para qualquer outra descrição sobre o evento.

Essa classificação tem uma aplicabilidade mais voltada para eventos de caráter imediato como terremotos, enchentes ou engarrafamento. Os *tweets* que vão auxiliar na correlação e previsão desses eventos são os que descrevem uma situação do presente, no momento que a pessoa vivenciou, ou seja, uma descrição do evento em tempo real. Para esses eventos, uma taxonomia composta por duas classes é a mais apropriada.

Tabela 3.1: As categorias de conteúdo e sua descrição.

Conteúdo	*Descrição*
Informação	*Tweets* contendo notícias, atualizações ou informações sobre o evento. Pode ser o título ou resumo de uma reportagem.
Experiência Pessoal	Usuário mencionando uma experiência direta (pessoal) ou indireta (por exemplo, amigo,familiares ou colegas de trabalho) com o evento ou com efeitos socais ou econômicos causados por esse.
Opinião	Publicações com a opinião do usuário sobre o evento, situação, reportagem ou expressando a necessidade de saber mais informação. Geralmente um comentário.
Piada ou Ironia	*Tweets* contendo piadas ou uma opinião bem-humorada sobre o evento que não se refira a uma experiência pessoal.
Campanha ou Propaganda	*Tweets* contendo um anúncio ou sobre o evento no sentido de motivar as pessoas para tomar atitudes que ajudem a evitá-lo. Usuários que reproduzem textos mencionados em campanhas públicas feitas para alertar sobre o evento ou para prevenir.

Tabela 3.2: As categorias de conteúdo e sua descrição.

Conteúdo	*Descrição*
Evento em tempo real	*Tweets* contendo descrição de algo que está acontecendo no exato momento em que foi publicado. Ou seja, o evento sendo reportado pelas pessoas em tempo real. Na maioria das vezes com verbo no presente.
Outros	*Tweets* contendo qualquer outra informação sobre as enchentes ou alagamentos como por exemplo, notícias de algum jornal ou comentários de alagamentos que ocorreram em outro dia.

A escolha de qual classificação usar depende da análise que se deseja realizar com as mensagens sobre o evento. No intuito de obter uma visão detalhada sobre o conteúdo dos *tweets*, a discriminação dos possíveis conteúdos feita por meio da classificação com cinco classes é mais indicada. Na correlação e previsão de eventos, os *tweets* que serão mais representativos são os classificados como experiência pessoal, esses descrevem a vivência da própria pessoa que publicou a mensagem ou de algum conhecido. Porém, na correlação e previsão de eventos de caráter imediato a segunda classificação é

fundamental para selecionar apenas os *tweets* que descrevem uma situação vivenciada no exato momento da publicação da mensagem e não sobre algo passado.

3.3.2 Classificação do Conteúdo

A classificação do conteúdo das mensagens é composta por duas etapas. A primeira etapa é a criação de um conjunto de mensagens previamente classificadas, ou treino, que consiste em exemplos formados pelo par atributos da mensagem e sua classe. A segunda etapa é a execução do algoritmo que realiza a tarefa de classificação. Essa tarefa, também chamada de aprendizado supervisionado, analisa os dados de treinamento e os utiliza para construir uma função de inferência cujo valor de saída é a classificação para qualquer mensagem de entrada a partir de seus atributos, Liu [2009].

3.3.2.1 Criação do Treino

No intuito de criar o conjunto de treino as mensagens coletadas devem ser selecionados aleatoriamente para serem classificadas manualmente. O treino é gerado apenas uma vez antes da execução do algoritmo, o qual é executado sempre quando uma nova mensagem é publicada.

Qualquer pessoa pode ler o texto dos *tweets* e classificá-los de acordo com seu conteúdo, desde que seja previamente instruída quanto às classes de conteúdo e o que essas representam.

As mensagens são selecionadas para serem rotuladas manualmente pois não é viável rotular todas as mensagens coletadas devido ao grande número obtido.

Para estimar a qualidade do classificador, foi utilizada a técnica de Validação Cruzada (Liu & Özsu [2009]) com 5 partições do conjunto de treino.

No método de Validação Cruzada denominado $k - fold$, ou k partições, os dados são particionados de forma aleatória em k subconjuntos mutualmente exclusivos do mesmo tamanho, Zaki & Meira Jr. [2012]. Um subconjunto é removido e os $k - 1$ restantes são utilizados para criar um novo modelo de regressão. O novo modelo é usado para prever os valores dos dados do subconjunto removido. Esse processo é realizado k vezes de forma que, a cada vez, um subconjunto diferente dos k subconjuntos é selecionado para teste. Ao final das k iterações, calcula-se a acurácia sobre os erros encontrados, obtendo uma medida confiável sobre a capacidade do modelo de representar o processo de previsão dos dados.

3.3.2.2 Algoritmo de Classificação

Com o intuito de classificar as mensagens automaticamente é necessário um algoritmo capaz de lidar com um grande volume de dados mesmo contando com um pequeno conjunto de treino e também de lidar com o desbalanceamento de classes. Um algoritmo que atende esses critérios é o *Lazy Associative Classification*, ou LAC (Veloso et al. [2006]).

O LAC gera uma função de mapeamento representada por um conjunto de regras de associação. Tais regras são geradas a partir de um conjunto de padrões frequentes extraídos da base de treinamento.

Entretanto, um classificador associativo pode gerar um número muito grande de regras, muitas delas desnecessárias durante a classificação, por não serem aplicáveis a nenhuma instância de teste.

O LAC, classificador associativo sob demanda, gera regras específicas para cada instância de teste. Essa estratégia obtém uma projeção da base de treinamento somente com instâncias que possuem pelo menos um atributo em comum com a instância de teste. A partir desta projeção e do conjunto de atributos da instância de teste, as regras são induzidas e ordenadas, e a melhor regra do conjunto é utilizada para a classificação. Pelo fato das regras serem induzidas a partir do conjunto de atributos da instância de teste, todas as regras geradas serão aplicáveis (Veloso et al. [2006]).

Na classificação do texto das mensagens, os atributos são as palavras (ou *tokens*) do texto publicado e as classes são as categorias de conteúdo.

Para cada mensagem, o LAC gera a probabilidade dessa pertencer a cada das classes definidas. A classe com maior porcentagem é a classe do conteúdo prevista para a mensagem. Dessa forma, todas as mensagens têm seu conteúdo classificado automaticamente.

3.4 Análise de Correlação

Durante a análise de correlação apresentada nesta seção, investigamos se os dados obtidos nas redes sociais servem como uma fonte de dados representativos sobre o evento. Caso a correlação entre os dados das redes sociais e os dados reais sobre o evento seja verificada, as mensagens publicadas podem ser utilizadas tanto como previsores do evento quanto na elaboração do alerta.

A análise de correlação deve ser feita sempre levando em consideração quatro dimensões: volume, conteúdo, localização e tempo. O volume representa a quantidade de mensagens do Twitter que contém em seu texto pelo menos um dos termos

relacionados ao evento. O conteúdo se refere à percepção e ao sentimento do público expressados no texto das mensagens publicadas. A localização é a informação geográfica declarada pelo usuário que escreveu a mensagem ou o local da ocorrência do evento. A última dimensão, o tempo, é referente a quando os *tweets* foram enviados ou quando ocorreu o evento.

A Figura 3.4 contém uma visão geral da análise de correlação. A primeira análise considera o volume das ocorrências ao longo do tempo e correlaciona as séries temporais obtidas por meio das mensagens das redes sociais e pelas ocorrências oficiais do evento. Essa correlação é mensurada considerando também o deslocamento ao longo do tempo para que seja possível observar se há um atraso ou avanço da repercussão do evento nas redes sociais em relação a ocorrência do evento na vida real.

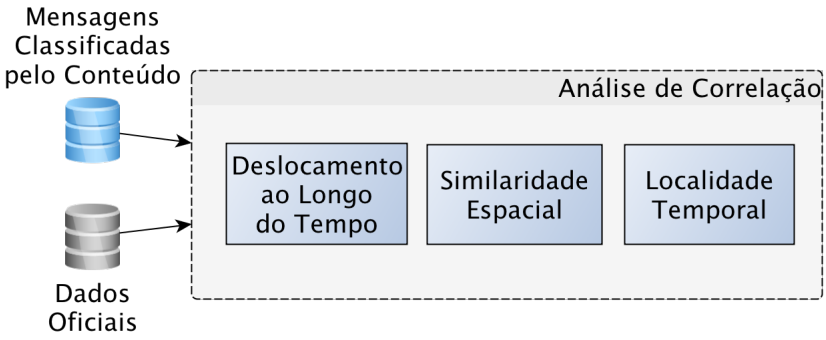


Figura 3.4: Diagrama contendo as partes da análise da correlação.

A segunda análise tem o intuito de encontrar regiões próximas com índices similares de ocorrência do evento e compara as regiões encontradas nos dados das redes sociais e nos dados oficiais.

A última análise considera o intervalo de tempo entre a chegada das mensagens. É esperado que em um período crítico para o evento haja uma maior concentração de mensagens enviadas no Twitter.

3.4.1 Deslocamento ao Longo do Tempo

Nesta análise é mensurada a similaridade entre o volume das ocorrências do evento e o volume das mensagens relacionadas provenientes do Twitter. A correlação linear entre essas duas variáveis é calculada para verificar como elas se comportam. A hipótese é que quando o volume de ocorrências sobre o evento aumenta ou o evento é grave, há

também um aumento da repercussão no Twitter representado pelo aumento do número de mensagens publicadas.

Serão criadas, para cada localização, duas séries temporais : $T = t_1 \dots t_n$ para os dados do Twitter, e $O = o_1 \dots o_n$ para os dados oficiais onde n é o tamanho das séries. Para mensurar a correlação existente entre as duas séries será calculado o coeficiente de correlação de Pearson. A fórmula desse coeficiente é a seguinte:

$$r = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2}\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}}$$

sendo, $\bar{t}$ e $\bar{o}$ são as médias das séries T e O, respectivamente.

Esse coeficiente, representado por r, quantifica o grau de correlação entre duas variáveis e assume valores entre -1 e 1. O valor de r igual a zero significa que não há uma relação linear entre as duas variáveis. O valor 1 indica uma correlação perfeita positiva e o valor -1 também indica uma correlação perfeita, porém inversa, ou seja, quando uma variável aumenta, a outra diminui. Quanto mais próximo de 1 ou -1, mais forte é a associação linear entre as duas variáveis.

Além disso, deseja analisar o deslocamento dessa correlação ao longo do tempo com intuito de observar se a repercussão do evento no Twitter acontece ao mesmo tempo que na vida real, se há algum atraso ou avanço. A correlação cruzada, Brourke [1996], é a correlação entre duas séries considerando um atraso $d = 0, 1, 2, ...n - 1$ no tempo de uma das séries. A correlação cruzada r_d em um atraso d no tempo é definida como:

$$r_d = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(o_{(i-d)} - \bar{o})}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2}\sqrt{\sum_{i=1}^{n}(o_{(i-d)} - \bar{o})^2}}$$

O resultado da correlação cruzada, r_d, é interpretado da mesma forma que r. O desvio entre as duas séries, d, pode ser variado de 0 até $n - 1$, sendo n o tamanho da série.

3.4.2 Localidade Temporal

O intervalo de tempo entre a chegada das mensagens é uma medida que permite analisar se, quando o evento ocorre, as mensagens são publicadas todas juntas ou se são enviadas ao longo do tempo. É esperado que a publicação das mensagens em um período crítico para o evento exiba forte localizade de referência, ou seja, são enviadas com maior frequência em um mesmo intervalo de tempo do que em períodos normais.

Para essa análise, criamos o *Event Index* (EI), ou índice do evento, uma medida derivada do intervalo entre o tempo de chegada das mensagens no Twitter. Há

um EI para cada localização em um certo intervalo de tempo. O EI é o valor da área abaixo da curva do gráfico representado na Figura 3.5. A curva é a função de distribuição acumulada (CDF) do intervalo entre o tempo de chegada (IAT) das mensagens publicadas no Twitter.

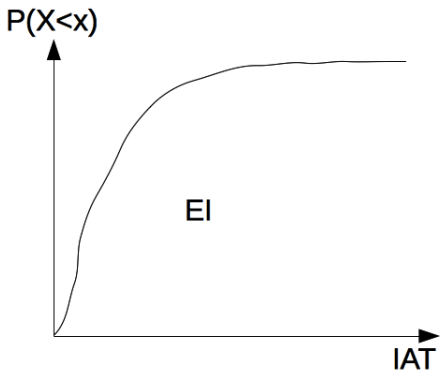


Figura 3.5: Cálculo do Event Index(EI).

A curva da CDF é gerada da seguinte forma. Primeiro, as mensagens são ordenadas por ordem de envio, ou seja, pelo horário em que foram publicadas. O intervalo entre o tempo de chegada de cada mensagem é calculado, esses valores são ordenados e armazenados em um vetor, chamado IAT. A função $P(X < x)$ corresponde à probabilidade de que a variável aleatória X assuma um valor inferior ou igual a determinado x. Nesse contexto, os valores de x são os valores em IAT.

Quanto maior o valor de EI menor o intervalo entre o tempo de chegada das mensagens no Twitter, ou seja, mais mensagens foram publicadas em um intervalo pequeno de tempo. O valor de EI deve ser comparado com o número de ocorrências do evento, o número de mensagens postadas e a situação oficial do evento. Além disso, o valor de EI deve ser comparado entre períodos em que não ocorreu o evento e períodos críticos do evento. Dessa forma, é possível observar se os valores de EI são maiores em períodos críticos para o evento e se durante esse período há uma maior concentração no envio das mensagens.

3.4.3 Similaridade Espacial

Esta seção descreve a análise da similaridade espacial que tem o intuito de encontrar locais próximos com níveis similares de ocorrência do evento em um dado espaço de tempo. Essas regiões similares serão encontradas utilizando um algoritmo de agrupamento que será executado para cada período de tempo levando em consideração o volume de ocorrências do evento em cada local.

O algoritmo ST-DBSCAN, Birant & Kut [2007], é uma técnica de agrupamento baseada em densidade. Esse algoritmo é uma extensão do DBSCAN, Ester et al. [1996], que possui as seguintes vantagens: não requer a priori a especificação do número de grupos que devem ser gerados e tem a habilidade de descobrir agrupamentos cuja forma é arbitrária.

O ST-DBSCAN determina os agrupamentos de acordo com informação não-espacial, espacial e temporal. Nesse contexto, a informação não-espacial consiste no volume de ocorrências do evento. Informação espacial consiste na localização do evento e a temporal corresponde ao período de tempo (mês, semana, dia ou horário) em que o evento foi observado.

Cada localização é representada por um ponto com latitude e longitude. Para um agrupamento ser formado, é necessário que um número mínimo de locais, ou pontos (*MinPts*) sejam próximos um do outro (distância entre os locais deve ser menor que *Eps1*) e tenham níveis de ocorrência similares (diferença entre o volume deve ser menor que *Eps2*). Para encontrar os valores dos parâmetros *Eps1* e *Eps2* foi utilizada a heurística descrita em Ester et al. [1996]. O valor do *MinPts* depende da natureza do evento e deve ser analisado separadamente em cada caso.

Antes de explicar o funcionamento do algoritmo, dois conceitos serão definidos. Um *objeto núcleo* é um ponto cuja vizinhança, definida por uma circunferência de raio *Eps1*, tem pelo menos o número mínimo de pontos (*MinPts*) com uma diferença máxima de *Eps2* entre seus valores não-espaciais. Um *objeto borda* é um ponto que não é *núcleo* mas é alcançável por qualquer *objeto núcleo*.

O algoritmo é explicado resumidamente a seguir. Para cada ponto p existente, se esse ponto ainda não tiver sido associado a nenhum agrupamento, então procura por todos os seus vizinhos, considerando tanto *Eps1* e *Eps2*, do ponto p. Se o número de vizinhos for menor que *MinPts* então marca p como ruído. Caso contrário, um novo agrupamento é criado e o ponto p e seus vizinhos $q_{1...n}$ são assinalados como pertencentes a esse novo grupo. Para cada vizinho q encontrado, procura seus respectivos vizinhos $o_{1...n}$. Dentre os vizinhos encontrados, aqueles que não forem ruído ou que ainda não estiverem em um grupo, são atribuídos a esse novo agrupamento. Uma descrição mais detalhada está em Birant & Kut [2007].

Depois que os agrupamentos foram criados, verifica-se a correlação entre os grupos gerados utilizando a base com os dados oficiais e os gerados utilizando a base com mensagens do Twitter. É desejável que os locais (pontos) que estão em um determinado grupo da base oficial também estejam juntos na base de mensagens do Twitter. A correlação entre os agrupamentos é medida pelo *Rand Index* (Rand [1971]). Dado um conjunto de n locais, $S = L_1, ..., L_n$, e duas partições de S para comparar, $X = x_1, ..., x_n$

e $Y = y_1, ..., y_n$, é definido:

- a, o número de pares de elementos de S que estão no mesmo conjunto em X e no mesmo conjunto em Y
- b, o número de pares de elementos de S que estão em diferentes conjuntos em X e em diferentes conjuntos em Y
- c, o número de pares de elementos de S que estão no mesmo conjunto em X e em diferentes conjuntos em Y
- d, o número de pares de elementos de S que estão em diferentes conjuntos em X e no mesmo conjunto em Y

O *RandIndex*, R, é:

$$R = \frac{a+b}{a+b+c+d}$$

.

Intuitivamente, $a+b$ são os números de agrupamentos que concordaram entre X e Y e $c+d$ são o número de desacordo entre X e Y. O valor de R varia entre 0 e 1, sendo que 0 indica que não tem nenhuma correspondência entre os agrupamentos gerados e 1 indica que os agrupamentos gerados são exatamente os mesmos em ambas as bases.

Após realizadas as três análises de correlação entre os dados do Twitter e os dados oficiais, caso seja verificada a correlação, o Twitter pode ser considerado uma boa fonte de informação sobre o evento e pode ser usado como insumo para prever a ocorrência do mesmo.

3.5 Redes Sociais como Previsores

Ao utilizar as redes sociais como previsores deseja-se prever o número de ocorrências do evento e prever a situação de gravidade do evento apenas utilizando as mensagens publicadas no Twitter. A regressão linear é utilizada para criar uma função que utiliza o número de mensagens do Twitter para prever o número de ocorrências do evento. A partir desse número previsto, classifica-se a situação de gravidade do evento em cada região.

As fases da previsão do evento, as informações necessárias e os resultados gerados podem ser visualizados na Figura 3.6.

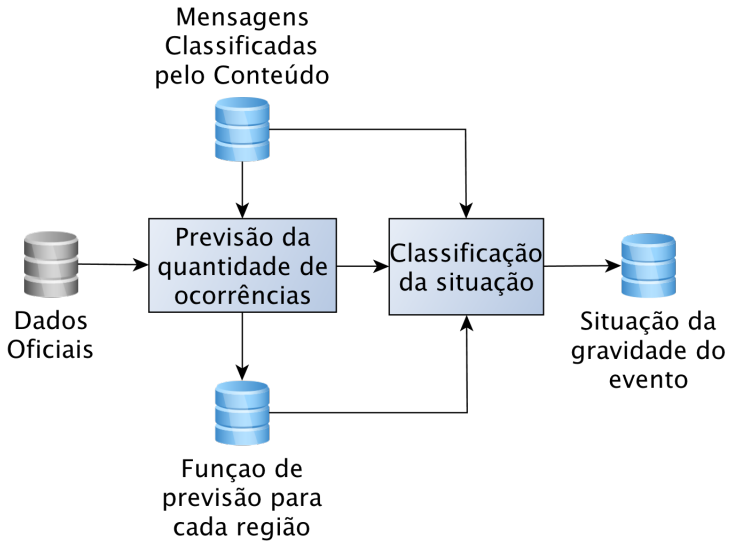


Figura 3.6: Diagrama contendo as fases da previsão do evento.

3.5.1 Previsão da quantidade de ocorrências do evento

Esta seção descreve como inferir a quantidade de ocorrências do evento para uma determinada localização. Para tal, o volume de mensagens do Twitter sobre o evento é utilizado para ajustar um modelo de regressão linear que deve se aproximar do número de ocorrências oficiais.

A regressão linear modela a relação entre duas variáveis pelo do ajuste de uma equação linear para os dados observados. Uma variável, y, é chamada de variável dependente ou variável resposta, e outra variável, x, é chamada de variável independente ou explanatória. Na regressão linear, temos a hipótese de que o valor de y depende do valor de x e expressamos matematicamente esta relação por meio de uma equação, Wang & Jain [2003]. A variável y é o número de ocorrências do evento e x é o número de mensagens publicadas no Twitter.

Assumindo que a associação entre x e y é linear, ou seja, descrita adequadamente por uma reta, essa pode ser descrita com a fórmula:

$$y = a + bx$$

sendo, a o coeficiente linear (valor que y assume quando x for zero) e b o coeficiente angular (inclinação da reta que mede o aumento ou redução de y para cada aumento

de uma unidade em x).

A regressão é usada para duas finalidades. A principal é prever o valor de y, ou seja, o número de ocorrências do evento a partir do valor de x que é do número de mensagens no Twitter. Depois de desenvolver um modelo, se um valor qualquer de x é dado sem o valor de y, o modelo ajustado (equação linear) pode ser utilizado para fazer a previsão do valor de y, basta substituir o valor de x no modelo para encontrar o valor de y. A outra finalidade é estimar o quanto x influencia ou modifica y. Para tal verifica-se o valor de b na equação. Para cada variação de uma unidade de x o valor de y aumenta ou diminui o equivalente a b unidades.

Além disso, por meio da regressão é possível verificar se a associação entre essas variáveis pode ser explicada pelo acaso. Essa questão é respondida realizando-se um teste t para verificar se o coeficiente angular, b, é diferente de zero. Se for zero, a reta não tem inclinação alguma, então x não interfere em y.

Outra informação que deseja-se obter é o percentual de variação de y explicado pela variação de x. Essa resposta é dada pelo coeficiente de determinação, R^2, gerado como resultado da regressão linear.

No intuito de avaliar o modelo de regressão criado foi utilizada a validação cruzada, Liu & Özsu [2009], técnica para avaliar como os resultados de uma análise estatística generalizam um conjunto de dados independentes.

O software R[1] foi utilizado para realizar a regressão linear e a validação cruzada.

3.5.2 Classificação da situação do evento

Para finalizar a previsão do evento, é realizada a classificação da situação desse em classes que representam a gravidade de sua ocorrência. Essas classes serão definidas de acordo com a natureza de cada evento. Por exemplo, algumas doenças tem sua incidência classificada utilizando três classes (baixa, média e alta) enquanto outros eventos podem ser classificados utilizando apenas duas classes caracterizando a ocorrência ou não do evento.

Na seção 3.5.1 foi descrito como criar uma função de regressão considerando o número de *tweets* de experiência pessoal para prever o número de ocorrências do evento. Essa função foi utilizada para prever o número de ocorrências para cada local em um determinado período de tempo e, a partir dessa previsão, a situação do evento foi classificada em uma das classes de intensidade.

[1]Link para acesso ao pacote de software R: http://www.r-project.org (último acesso em 11/02/2012)

A classificação criada utilizando as mensagens do Twitter foi comparada com a classificação gerada com os dados oficiais. Para validar o classificador desenvolvido, verificamos a matriz de confusão que é um resumo do desempenho do classificador, Zaki & Meira Jr. [2012]. As métricas derivadas da matriz de confusão são:

- Taxa de erro: número de previsões erradas dividido pelo número total de previsões.
- Taxa de falso positivo: porcentagem de falso positivo para cada classe.
- Taxa de verdadeiro positivo: porcentagem de verdadeiro positivo para cada classe.
- Acurácia: número de previsões corretas dividido pelo número total de previsões.
- Precisão: mede a habilidade da predição em classificar os positivos. É número de verdadeiros positivos dividido pelo número de positivos.

Essas métricas foram utilizadas para avaliar o desempenho da classificação e avaliar se o método proposto é capaz de classificar a gravidade da situação do evento para cada uma das regiões consideradas.

3.6 Alerta

O último item da metodologia é a elaboração de um sistema de alerta que mostre visualmente a situação do evento em uma determinada região geográfica utilizando como insumo os dados do Twitter.

Em um sistema de alerta sobre o evento é indispensável mostrar simultaneamente dois indicadores: a situação atual e a tendência do evento. A situação atual indica o que está ocorrendo no exato momento, como está a repercussão do evento no Twitter. A tendência indica se essa repercussão tem aumentado ou diminuído nas últimas semanas.

Utilizar somente um dos dois indicadores não torna o sistema de alerta confiável. Por exemplo, considere que a situação atual do evento em determinada localização está dentro de uma faixa aceitável. Dessa forma, na visualização da situação atual não haverá nenhum destaque, ou seja, nada que mereça um alerta. No entanto, se essa mesma localização apresentou um aumento de três vezes no valor em relação à última semana, algo está fora da normalidade para gerar tal variação. Logo, devemos chamar atenção para essa localização por meio do alerta no indicador de tendência.

Assim como a situação atual, a tendência não pode ser considerada separadamente. Por exemplo, considere que a tendência de ocorrência do evento

aumentou nas últimas semanas, o que indica que o sistema estará em alerta. Entretanto, se esse número parar de crescer, a tendência é de estabilidade e não deve-se mostrar um alerta mesmo se o número de ocorrências continuar alto. Logo, o sistema de alerta deve conter ambos os indicadores para fornecer uma informação mais completa sobre a situação do evento.

3.6.1 Avaliação da situação atual

A situação atual visualizada no sistema de alerta se baseia no volume de *tweets* cujo conteúdo foi classificado como sendo um relato de experiência pessoal ou uma descrição sobre o evento em tempo real. Para calculá-la, utiliza-se uma função de regressão linear (seção 3.5.1) específica para cada região, o volume de mensagens publicadas com o conteúdo desejado e, em alguns casos, a população da região.

A função de regressão linear é utilizada para prever o número de ocorrências do evento a partir do volume de *tweets*. A situação atual do evento é representada pelo valor previsto da quantidade de ocorrências do evento.

Essa informação é visualizada em uma escala de cores que varia do branco passando pelo amarelo até o vermelho. A cor branca representa a situação de normalidade e a cor vermelha, alerta máximo. Essa escala é gerada comparando-se o volume de ocorrências previsto com a classificação da situação na região. Por exemplo, para algumas doenças há uma classificação em baixa, média e alta incidência. Já para outros eventos há apenas duas classificações.

Para cada localização são estabelecidos limites inferiores e superiores para a escala de cores. O limite superior (LS) é o valor no qual começa a situação crítica, alarmante, do evento. O limite inferior (LI) é o menor valor possível para o evento, ou seja, como se não houvesse nenhuma ocorrência desse. Quanto mais próximo do LS, mais próxima do vermelho será a cor representada na visualização. Caso exceda o LS, será utilizado o vermelho absoluto. Da mesma forma, caso seja menor que o LI será utilizado o branco absoluto. Para valores intermediários, entre LI e LS, será utilizada uma escala em degradê variando sobre a porcentagem entre o valor mínimo e o valor máximo. Quando há uma terceira classe, é estabelecido também o limite intermediário (LM) e a escala em degradê fica da seguinte forma: para valores entre LI e LM será utilizada uma escala variando do branco ao amarelo sobre a porcentagem entre o valor mínimo e o valor médio. E para valores entre LM e LS será utilizada um degradê variando do amarelo passando pelo laranja até o vermelho.

3.6.2 Avaliação da tendência

A tendência fornece um indicativo se o número de *tweets* em determinada localidade continua constante, se tem aumentado ou diminuído em relação a dois períodos de tempo anteriores. Esse período depende da natureza do evento. Por exemplo, eventos como epidemias de doenças podem ser analisados semanalmente e eventos mais imediatos, como queimadas e terremotos, diariamente.

A tendência é representada pelo o Z-score (Larsen & Marx [1986]). Esse *score* é derivado a partir da média do número de *tweets* do período atual (x), da média de *tweets* de dois períodos anteriores (μ) e do desvio padrão do número de *tweets* durante esse período (σ). A fórmula do Z-score é:

$$(x - \mu)/\sigma$$

O valor do Z-score é a diferença do total de *tweets* do período atual e da média de *tweets* do período anterior em unidades de desvio padrão. Se o número de *tweets* atual não varia em relação ao período anterior, o valor do Z-score é zero. Se houver uma diminuição do número de *tweets*, o valor é negativo, e se aumentar, o valor é positivo.

Essa informação pode ser visualizada em uma escala de cores que varia do branco, passando pelo amarelo, até o vermelho. As localizações que tiveram uma diminuição no número de *tweets* (Z-score negativo) terão cor branca; as que apresentaram um aumento (Z-score positivo) terão cor vermelha e os que ficaram com valor constante (Z-score nulo) terão cor amarela.

O limite superior (LS) e o limite inferior (LI) da escala de cores são iguais para todas as localizações visto que a semântica do Z-score é a mesma. O LS é 2 e o LI é -1. Dessa forma, caso a diferença da média do número de *tweets* do período atual (x) e da média do período anterior (μ) exceda cinco vezes o desvio padrão (σ), a cidade será colorida de vermelho e, se for menor do que um desvio padrão, sua cor será branca. Quanto mais próximo do LS, mais próxima do vermelho será a cor representada no mapa. Caso exceda o LS, será utilizado o vermelho absoluto. Para valor de Z-score igual a zero, a cidade será colorida de amarelo. Para valores entre 0 e LS, será utilizado uma escala em degradê variando do amarelo, passando pelo laranja até chegar ao vermelho cujo valor é uma porcentagem entre 0 e LS. Para valores entre LI e 0, será utilizada uma escala em degradê variando do branco ao amarelo sobre a porcentagem entre LI e 0.

3.6.3 Síntese

A metodologia que foi apresentada nesse capítulo compreende todas as fases para, a partir dos dados disponíveis nas redes sociais, detectar e prever eventos da vida real. Os eventos que a metodologia abrange são aqueles comentados nas redes sociais pelas pessoas que o vivenciaram e que possuem um grande número de pessoas envolvidas. Além disso, o evento deve ter localização no espaço e no tempo definidos. Alguns exemplos desses eventos são lançamento de filmes, jogos, epidemias, terremotos ou engarrafamento.

Capítulo 4

Experimentos e Resultados

Neste capítulo são apresentados e discutidos os resultados obtidos ao aplicar a metodologia proposta em dois cenários distintos. O primeiro cenário é a dengue, doença que atinge centenas de milhares de pessoas no Brasil todos os anos. O segundo cenário são os alagamentos e as enchentes que causam grande prejuízo à população.

4.1 Dengue

A dengue é uma doença febril aguda causada pelo vírus da Dengue, um arbovírus da família *Flaviviridae*, que inclui quatro sorotipos distintos: 1, 2, 3 e 4. Quando uma pessoa apresenta infecção por um desses agentes, ela fica protegida para uma nova contaminação pelo mesmo subtipo. Nenhum sorotipo é mais perigoso que outro, mas quando um novo tipo entra em circulação, há um grande risco de epidemia, pois poucos indivíduos são imunes a ele. Além disso, a ocorrência de epidemias anteriores causadas por outros sorotipos aumenta o risco de casos graves, [CDC, 2012].

Dengue é transmitida para humanos pela picada do mosquito *Aedes aegypti*, que se desenvolve em áreas tropicais e subtropicais. Os sintomas da infecção geralmente começam depois de 4-7 dias da picada do mosquito e duram tipicamente 3-7 dias, [WHO, 2012]. Os sintomas variam de pessoa para pessoa, algumas podem nunca manifestar sintomas significativos, mas outras podem sentir dor de cabeça, dores musculares, dor nos olhos e cansaço, dentre outros.

Em grande parte das regiões dos trópicos e subtrópicos, a dengue é endêmica, ou seja, ocorre todo ano, geralmente durante a época na qual a população do mosquito *Aedes* está alta. Essa doença afeta mais de 100 países em desenvolvimento e subdesenvolvidos. A Organização Mundial de Saúde (OMS) estima que cerca de 2,5

bilhões de pessoas correm o risco de infecção e cerca de 50 a 100 milhões de infecções ocorrem globalmente a cada ano.

No Brasil, segundo o informe epidemiológico sobre o balanço da dengue realizado pelo Ministério da Saúde[1], o número de casos de dengue notificados em 2011 foi de aproximadamente 730 mil. Em torno de 54% dos casos se concentraram em quatro estados: Rio de Janeiro, São Paulo, Amazonas e Ceará. Na região norte, os municípios de Manaus e Rio Branco apresentaram os maiores números de casos notificados e foram responsáveis por 62% dos casos na região. Na região nordeste, o município de Fortaleza se destaca pelo grande número de casos notificados. Na região sudeste, os estados do Rio de Janeiro e Espírito Santo apresentaram aumento quando comparado ao mesmo período de 2010. Os estados de Minas Gerais e São Paulo tiveram redução do número de casos em relação a 2010. No entanto, a situação encontrada em Ribeirão Preto merece destaque pelo enorme número de notificações. Na região Sul, o município de Londrina teve destaque pelo grande número de casos. Todos os estados da região centro-oeste apresentaram redução no número de casos em 2011, quando comparado a 2010.

A previsão de epidemias de dengue é de grande importância para o Ministério da Saúde e para as autoridades de saúde pública. A vigilância epidemiológica juntamente com medidas de controle adequadas são os pilares para a prevenção de epidemias da doença, especialmente porque vacinas anda não são disponíveis, Runge-Ranzinger et al. [2008]. Dessa forma, os sistemas de vigilância devem ser capazes de detectar esses eventos para prover indicadores confiáveis que orientem as medidas de controle.

A metodologia proposta nessa dissertação pode ser aplicada no contexto da dengue com intuito de colaborar no combate à dengue por meio da disponibilização de alertas e ferramentas que possam orientar ações de combate e prevenção à doença. Essa colaboração tem como objetivo atender a demanda existente na vigilância da doença que irá utilizar uma nova fonte de dados, as redes sociais *online*, como provedor de informações sobre a situação atual da dengue.

Os resultados obtidos nessa dissertação no contexto da dengue fazem parte do Observatório da Dengue (http://www.observatorio.inweb.org.br/dengue/), um sistema de vigilância epidemiológica ativa a partir de dados internet, desenvolvido em parceria com o Instituto Nacional de Ciência e Tecnologia em dengue (INCT em dengue). O sistema permite visualizar as informações coletadas de diversas formas e provê estimativas acerca da incidência de dengue em determinada região.

O Observatório da Dengue firmou uma parceria com o Ministério da Saúde para que os dados obtidos por meio das redes sociais fizesse parte da campanha nacional

[1]Link para balanço sobre a dengue em 2011: (último acesso em 11/01/2012) http://portal.saude.gov.br/portal/arquivos/pdf/informe_dengue_2011_37_39.pdfs

de combate à dengue[2]. O Observatório da Dengue é utilizado como um sistema complementar ao sistema de vigilância tradicional e disponibiliza o alerta, desenvolvido nessa dissertação, com a avaliação da situação da incidência e da tendência da doença. Além disso, disponibiliza relatórios semanais com informações separadas por região geográfica e municípios com população acima de 100 mil habitantes[3].

Os resultados alcançados são descritos em Gomide et al. [2011] e Silva et al. [2011].

4.1.1 Base de dados

Nesta seção, serão descritas em detalhe as duas bases de dados sobre a Dengue. A primeira base de dados contém os dados oficiais sobre a dengue no Brasil. O Ministério da Saúde disponibilizou todos os casos de dengue notificados no Brasil durante 2010 e 2011 até final de setembro.

A segunda base de dados é composta pelas mensagens publicadas no Twitter que se referem a dengue. A coleta dessas mensagens começou dia 21/11/2010 e continua sendo realizada visto esses dados são utilizados pelo Observatório da Dengue.

Como a coleta das mensagens no Twitter teve início dia 21/11/2010 e os dados oficiais sobre os casos de dengue vão até final de setembro de 2011, os experimentos foram realizados considerando o período de 21/11/2010 até 30/09/2011.

4.1.1.1 Ministério da Saúde

Por meio de uma parceria entre o Ministério da Saúde e a Universidade Federal de Minas Gerais (UFMG), os dados oficiais sobre a dengue no Brasil foram disponibilizados. O Ministério da Saúde disponibilizou um banco de dados contendo os dados de cada notificação da doença durante 2010 e 2011.

A base de dados contém os seguintes atributos: a data da notificação, a data dos primeiros sintomas, o município de notificação e a classificação final do caso. A classificação final do caso, segundo Ministério da Saúde, é dividida em cinco classes: dengue clássico, dengue com complicações, febre hemorrágica, síndrome do choque e descartado. Nos experimentos considera-se que um caso de dengue ocorreu, se for classificado em uma das quatro primeiras classes, caso contrário é descartado.

[2]Link para lançamento da campanha contra dengue pelo Ministério da Saúde (último acesso: 11/02/2012) http://portalsaude.saude.gov.br/portalsaude/noticia/3563/162/ministerio-lanca-campanha-de-combate-a-dengue.html

[3]Link para a notícia sobre o Observatório da Dengue publicada no *blog* do Ministério da Saúde: (último acesso 11/02/2012) http://www.blog.saude.gov.br/monitoramento-das-redes-sociais-auxiliara-luta-contra-a-dengue/

Durante o período de 21/11/2010 e 30/09/2011 foram notificados 736.281 casos de dengue no Brasil. A Figura 4.1 contém o número de casos de dengue notificados no Brasil ao longo desse período.

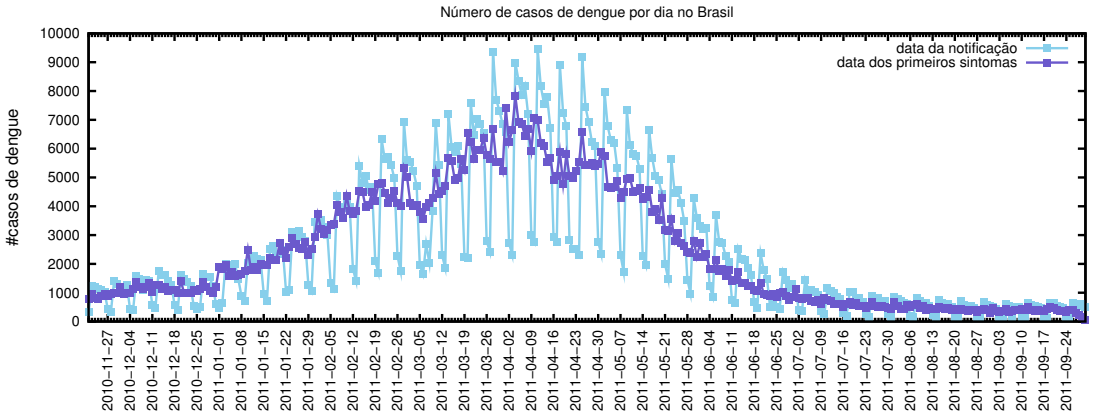


Figura 4.1: Número de casos de dengue por dia notificados no Brasil durante 21/11/2010 e 30/09/2011. A linha azul clara representa o número de casos pela data de notificação e a linha azul escura pela data dos primeiros sintomas.

O nível de intensidade da dengue é classificado de acordo com sua incidência. Para o cálculo da incidência, divide-se o número de notificações pelo quantitativo populacional do município e multiplica-se este valor por 100 mil. O Ministério da Saúde considera três níveis de incidência de dengue: baixa (menos de 100 casos/100 mil habitantes), média (de 100 a 300 casos/100 mil habitantes) e alta (mais de 300 casos/100 mil habitantes). Essa classificação será utilizada em nossas análises.

4.1.1.2 Twitter

A coleta das mensagens no Twitter teve início dia 21/11/2010 e como a base de dados do Ministério da Saúde vai até o dia 30/09/2011, o período que vamos considerar nas análises será do dia 21/11/2010 até o dia 30/09/2011. Durante todo esse período, houve uma falha na coleta entre os dias 23/12/2010 e 04/01/2011, período que será desconsiderado nas análises.

Os termos escolhidos para coleta das mensagens relacionadas com a dengue são: dengue e *aedes*.

A Tabela 4.1 apresenta o número de *tweets* e usuários coletados, e a parte desses dados que são do Brasil e apresentam informação de localização a nível de cidade. Aproximadamente metade dos *tweets* sobre dengue são de usuários do Brasil e mais da metade das mensagens do Brasil possuem localização a nível de cidade. Quase 90% dos usuários brasileiros declararam sua localização a nível de cidade.

Tabela 4.1: Número de *tweets* e usuários presentes na base de dados sobre a Dengue do Twitter. Período da coleta foi de 21/11/2010 até 06/01/2012.

#*tweets*	925.727
#*tweets* do Brasil	460.816 (49,78%)
#*tweets* do Brasil com cidade	296.096 (31,98%)
#usuários	470.798
#usuários do Brasil	219.632(46,65%)
#usuários do Brasil com cidade	196.152(41,66%)

A função densidade de probabilidade (PDF) do número de *tweets* por usuário é mostrada na Figura 4.2. A grande maioria dos usuários postaram apenas uma mensagem sobre a dengue.

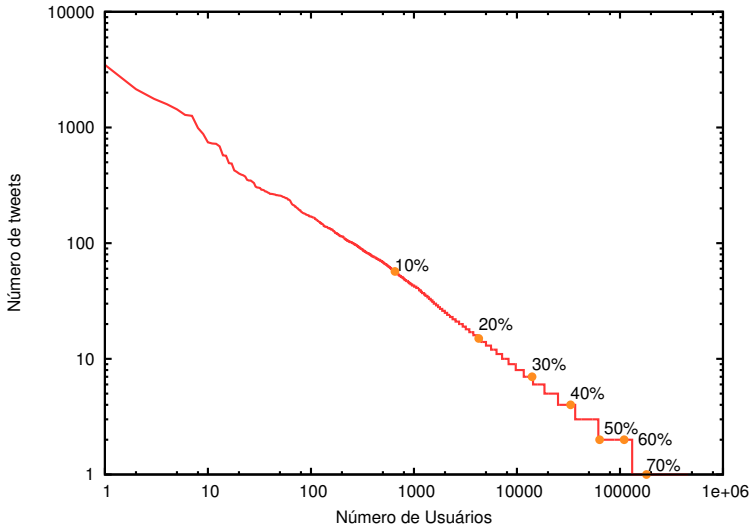


Figura 4.2: Número de *tweets* por usuario em escala logarítmica.

Durante o período de 21/11/2010 a 30/09/2011 foram coletadas 296.096 mensagens de 196.152 usuários diferentes do Brasil. A Figura 4.3 contém o número de *tweets* sobre dengue no Brasil ao longo desse período. Vale ressaltar que durante o período de janeiro a abril houve um maior número de mensagens publicadas no Twitter sobre a dengue e é nesse período que a maior parte dos casos de dengue foram notificados.

Há *tweets* de 3.424 cidades do Brasil, entretanto, aproximadamente 60% dos *tweets* são de apenas 26 cidades. As dez cidades com maior número de *tweets* são: Rio de Janeiro, São Paulo, Manaus, Natal, Fortaleza, Belo Horizonte, Brasília, Curitiba, Recife e Belém. A função densidade de probabilidade (PDF) do número de *tweets* por

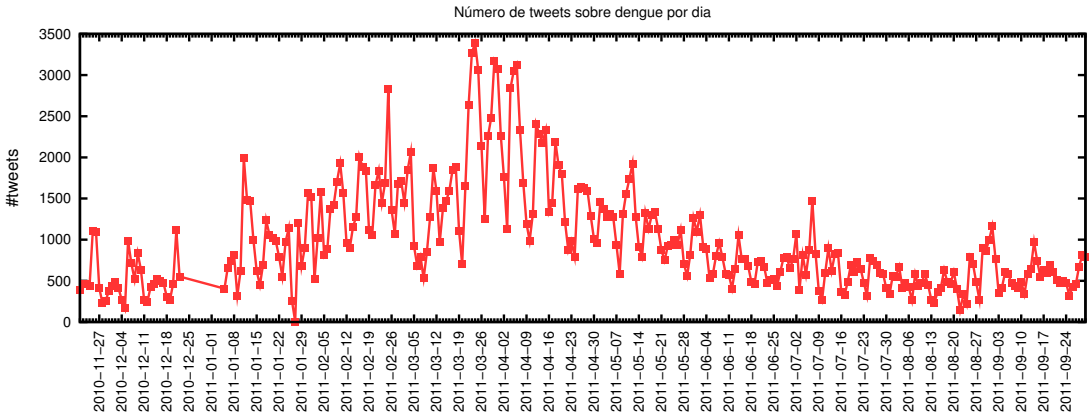


Figura 4.3: Número total de *tweets* coletados com localização a nível de cidade durante todo período de coleta.

cidade é mostrado na Figura 4.4.

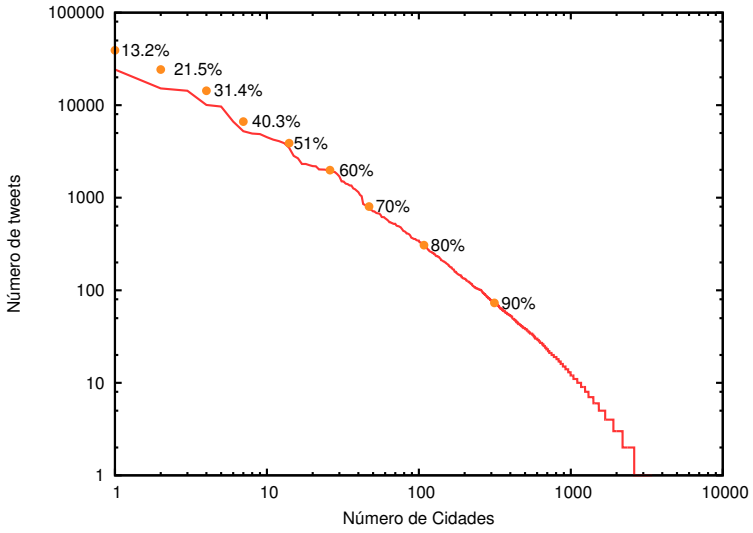


Figura 4.4: Número de *tweets* por cidade em escala logarítmica.

O número de mensagens e usuários da base de dados do Twitter e o número de casos de dengue da base de dados do Ministério da Saúde separados por estado do Brasil se encontram na Tabela 4.2.

4.1.2 Análise de Conteúdo

Nesta seção, é realizada a análise de conteúdo do texto das mensagens relacionadas à dengue. Primeiramente, é feita a caracterização dessas mensagens e as diversas classes de conteúdo são exemplificadas por meio de alguns *tweets*. Depois, descreve-se como

Tabela 4.2: Número de mensagens e usuários da base de dados sobre a Dengue do Twitter e número de casos de dengue notificados da base do Ministério da Saúde.

Estados do Brasil	*#tweets*	*#usuários*	*#casos de dengue*
Norte			
AC	2.565	836	119.137
AM	18.551	4.428	135.715
PA	6.032	3.046	102.150
RO	1.103	639	156.081
RR	442	194	34.607
TO	2.182	599	122.527
Nordeste			
AL	2.878	1.252	125.513
CE	14.312	5.222	108.041
MA	2.897	1.306	148.058
RN	20.257	6.540	37.786
SE	3.374	1.271	54.536
PB	8.090	3.720	17.502
PI	4.702	1.146	54.132
Centro-Sul			
DF	7.160	3.668	6.316
GO	6.857	3.335	72.522
MS	3.875	2.037	16.033
MT	2.533	1.442	10.894
Sudeste			
ES	6.936	3.275	118.574
MG	25.645	11.061	63.560
RJ	53.509	21.181	185.566
SP	70.356	39.202	223.811
Sul			
PR	18.620	9.236	126.731
RS	12.271	7.845	1.748
SC	6.719	4.422	8.156

o treino é criado e avalia-se o desempenho do classificador. E, por último, o resultado da classificação do conteúdo das mensagens é analisado.

As mensagens postadas no Twitter sobre dengue são classificadas de acordo com o conteúdo do seu texto. O LAC, classificador associativo sob demanda, formará as regras utilizando como atributo as palavras (ou tokens) da mensagem. Antes de ser classificado, cada *tweet* teve seu texto processado da seguinte forma: remoção da acentuação; remoção dos caracteres RT, que classificam a mensagem como um *retweet*; remoção da menção às páginas web (p.ex., http); remoção da menção aos usuários;

as letras maiúsculas foram substituídas por letras minúsculas; remoção de todos os caracteres alfa-numéricos, tais como vírgulas e pontos.

O número de mensagens, número de atributos (palavras ou tokens da mensagem), tamanho do vocabulário (número de tokens diferentes) e a média do número de atributos por mensagens são apresentados na Tabela 4.3.

Tabela 4.3: Características das mensagens postadas no Twitter sobre dengue.

Número de mensagens	925.727
Número de atributos (tokens do tweet)	11.803.826
Tamanho do vocabulário (tokens diferentes)	301.405
Média do número de atributos por mensagem	12,75 (min=1, max=39)

O treino criado contém 2204 mensagens que foram classificadas com a ajuda de 15 alunos do curso de Ciência da Computação da UFMG. Foram apresentadas aos alunos as cinco classes de conteúdo nas quais as mensagens seriam classificadas. O significado de cada conteúdo foi explicado por meio da descrição presente na Tabela 3.1. E, além disso, três exemplos de *tweets* para cada classe foram mostrados para que eles pudessem compreender melhor o significado de cada uma delas. A porcentagem de cada classe de conteúdo presente no treino está na Figura 4.5. Alguns exemplos de *tweets* de cada classe estão na Tabela 4.4.

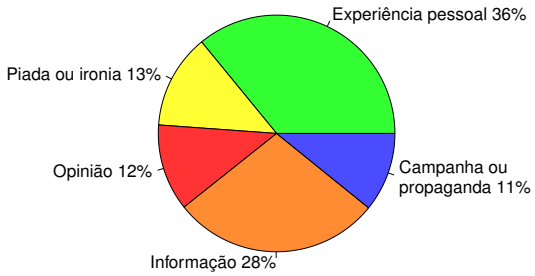


Figura 4.5: Porcentagem dos *tweets* classificados em cada classe de conteúdo no treino.

Para avaliar o classificador de conteúdo das mensagens foi realizada uma validação cruzada com 5 partições do conjunto de mensagens classificadas manualmente. Os valores para precisão, taxa de verdadeiro positivo e acurácia para cada uma das cinco classes estão na Tabela 4.5. O classificador para a classe experiência pessoal classificou corretamente 92% dos *tweets*. Embora para as classes opinião, piada e

Tabela 4.4: As categorias de conteúdo e exemplos de *tweets*.

Experiência Pessoal
• É, estou com suspeita de dengue... agora é esperar e voltar ao hospital em breve. :(
• Bom dia!! Acho que estou com dengue...
• ainda estou com dengue, mas já melhorei bastante! :D
Opinião
• po velho vc n ve os avisos de como previnir a dengue da globo?
• Eu não concordo que exista essa queda da epidemia de #Dengue
• um absurdo... na rua da minha avó em Sepetiba, RJ. Tem um foco de dengue numa casa velha...já foram na pref milhares d vezes
Informação
• RT @g1: Brasil já registra quase o dobro do número de casos de dengue em relação ao ano passado http://tinyurl.com/27ymker
• Dê uma olhada nesse vídeo – Secretário da Saúde do Estado fala sobre campanha contra a dengue ... http://bit.ly/csHTMB
• Correio do Povo: Mobilização reduz casos de dengue, diz Sesau http://ow.ly/1cgAap
Piada ou Ironia
• Estou tão carente que deixei um vaso com água parada aqui em casa só para ter companhia do mosquito da dengue.
• Porque a loira balança o copo antes de tomar água? R: Porque água parada é dengue.
• O Ministério da Saúde adverte: seque-se bem após o banho, água parada em pneus dá dengue.
Campanha ou Propaganda
• Agora é Guerra! Todos contra a Dengue. Fazemos a nossa parte. Faça a sua.
• vamos se unir contra a dengue....nada de deixar água parada em pneus..
• nunca deixe água parada em qualquer recipiente já que o mosquito da dengue anda solto e vamos evitar essa doença

campanha o classificador tenha previsto corretamente uma pequena parte dos *tweets*, aproximadamente 90% dos casos que foram previstos como sendo dessas classes estavam corretos. Para a classe informação o classificador acertou aproximadamente 44% das

previsões e mais de 60% dos *tweets* previstos para essa classe estavam corretos.

Tabela 4.5: Resultados da validação cruzada com 5 partições na tarefa de classificação do conteúdo das mensagens. Obs.:V.P. é Verdadeiro Positivo

Classe	*Métrica*	*Valor*
Experiência Pessoal	Precisão	0.7412 (min=0.7079, max=0.8012)
	Taxa de V.P.	0.5918 (min=0.5535, max=0.6734)
	Acurácia	**0.9258** (min=0.8922, max=0.9567)
Opinião	Precisão	0.8669 (min=0.8637, max=0.8693)
	Taxa de V.P.	0.2484 (min=0.1875, max=0.2857)
	Acurácia	0.0516 (min=0.0091, max=0.1257)
Informação	Precisão	0.6644 (min=0.6476, max=0.6729)
	Taxa de V.P.	0.6474 (min=0.6053, max=0.6831)
	Acurácia	0.4431 (min=0.4200, max=0.4655)
Piada ou Ironia	Precisão	0.8723 (min=0.8575, max=0.8815)
	Taxa de V.P.	0.5726 (min=0.3272, max=0.7619)
	Acurácia	0.1345 (min=0.0818, max=0.2126)
Campanha ou Propaganda	Precisão	0.8939 (min=0.8868, max=0.9007)
	Taxa de V.P.	0.5603 (min=0.4158, max=0.6603)
	Acurácia	0.1823 (min=0.1391, max=0.2375)

A Figura 4.6 apresenta o número de *tweets* de cada conteúdo sobre a dengue semanalmente durante todo o período. Os *tweets* de experiência pessoal começam a aumentar significativamente no início do ano e são significativos até abril, o que coincide com a época da epidemia de dengue. O volume de *tweets* de informação também aumenta nessa época. Já os *tweets* de ironia, opinião e campanha não tem grandes variações ao longo do período.

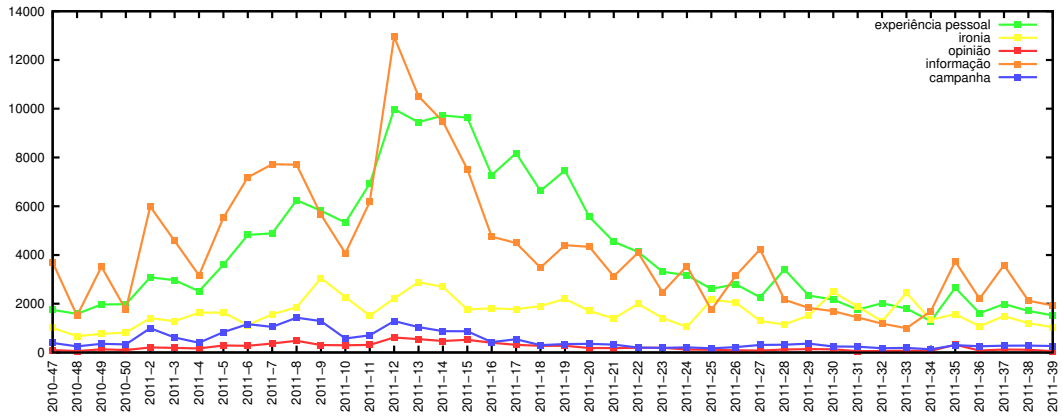


Figura 4.6: Número de *tweets* de cada classe de conteúdo por semana durante todo o período.

4.1.3 Análise de Correlação

As análises realizadas nesta seção correlacionam as mensagens sobre a dengue obtidas pelo Twitter com as informações fornecidas pelo Ministério da Saúde sobre essa doença. Os resultados para as três análises de correlação realizadas são descritos a seguir.

4.1.3.1 Deslocamento ao Longo do Tempo

Nestes experimentos é mensurada a correlação entre o volume dos casos de dengue notificados e o volume das mensagens do Twitter. Primeiro, serão consideradas todas as cidades com mais de 100 mil habitantes[4] e, depois, é criado um limiar para o número de *tweets* que cada município deve ter. Por último, escolhemos 12 cidades para mostrar o resultado da correlação ao longo do tempo.

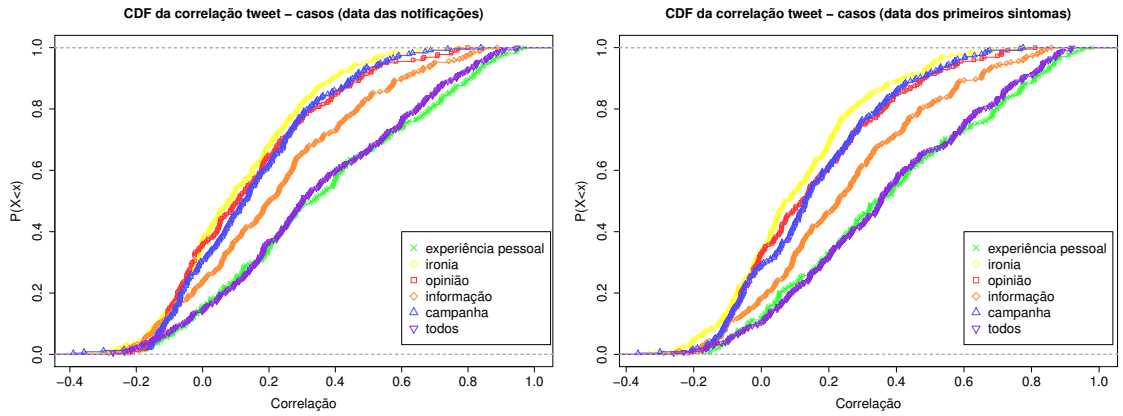
A correlação é calculada considerando o volume semanal durante todo o período de 21/11/2010 até 30/09/2011. Entretanto, como houve falha na coleta do Twitter durante o período de 23/12/2010 a 04/01/2011, as duas últimas semanas de dezembro de 2010 e a primeira semana de janeiro de 2011 serão desconsideradas do cálculo, que irá avaliar 42 semanas.

Foram criadas duas séries temporais para cada município a partir da base de dados do Ministério da Saúde. Uma delas considera a data do caso de dengue como sendo a data de notificação da doença, e a outra, considera a data do caso como sendo a data dos primeiros sintomas reportado pelo paciente.

Com a base de dados do Twitter criamos seis séries temporais para cada município. Uma delas considera o volume total de *tweets* e as outras cinco são para as cinco categorias de conteúdo.

A Figura 4.7 contém a função densidade acumulada (CDF) da correlação entre as séries do Twitter e as séries do Ministério da Saúde. As correlações das séries do Twitter com a série formada pela data de notificação dos casos e com a data dos primeiros sintomas são ilustradas nas Figuras 4.7a e 4.7b respectivamente. As correlações que utilizam os *tweets* de ironia, opinião e campanha são os que apresentam a menor correlação com os dados do Ministério da Saúde. Já os *tweets* de informação apresentam uma correlação um pouco maior mas, as melhores correlações são quando consideramos todos os *tweets* ou os *tweets* de experiência pessoal. Aproximadamente 40% dos municípios apresentam correlação maior que 50% e 20% possuem correlação maior que 70% quando utilizamos todos os *tweets* ou apenas os de experiência pessoal.

[4]Link para lista dos municípios com mais de 100 mil habitantes: (último acesso em 11/01/2012) http://pt.wikipedia.org/wiki/Anexo:Lista_de_munic%C3%ADpios_do_Brasil_acima_de_cem_mil_habitantes



(a) (b)

Figura 4.7: CDF da correlação de *tweets* das cinco categorias de conteúdo e considerando todos os *tweets* com o número de casos de dengue por data de notificação (a) e por data dos primeiros sintomas (b).

Apesar das melhores correlações ocorrerem quando se considera os *tweets* classificados como sendo de experiência pessoal ou todo o volume de *tweets*, algumas cidades apresentaram alta correlação, e outras cidades apresentaram uma correlação muito baixa. Ao investigar o porquê dessa diferença observa-se que há vários municípios que não possuem *tweets* ou que possuem poucas mensagens. A falta de *tweets* pode indicar ausência de dengue ou pode ser falta de divulgação dessa informação nas redes sociais. Por isso, é importante realizar a correlação considerando um limiar mínimo do número de publicações. Nessa próxima correlação serão considerados apenas municípios cujo total de *tweets* de experiência pessoal é no mínimo o equivalente a um *tweet* por dia. Após essa seleção, restaram apenas 47 cidades das 285 que estávamos analisando.

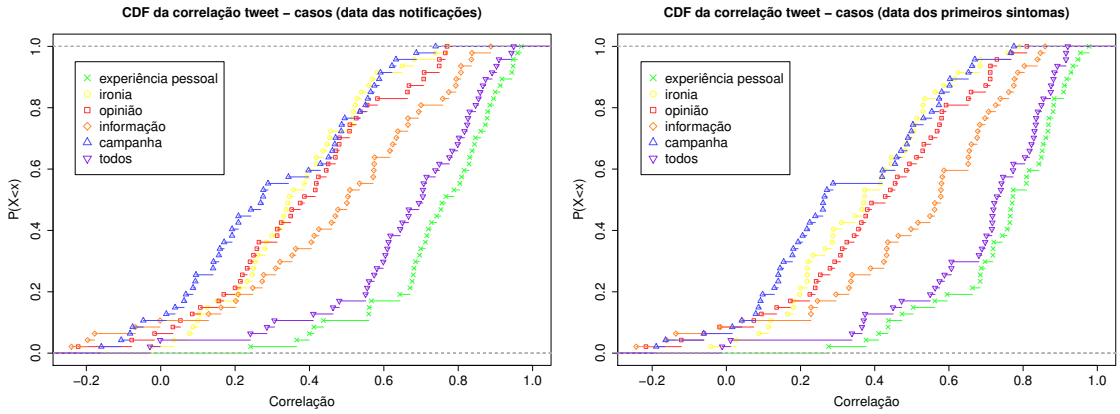
A média e o desvio padrão das correlações entre os dados do Twitter e do Ministério da Saúde para as cidades para as quais foi registrado no mínimo o equivalente a um *tweet* de experiência pessoal por dia estão na Tabela 4.6. Na média, a melhor correlação obtida é entre os *tweets* de experiência pessoal e o número de casos por dia quanto relatou-se os primeiros sintomas.

A Figura 4.8 ilustra a CDF da correlação para as cidades que tiveram no mínimo o equivalente a uma mensagem publicada por dia. As melhores correlações são obtidas quando são utilizados os de experiência pessoal. Além disso, tanto quando a data das notificações, Figura 4.8a, ou a data dos primeiros sintomas, Figura 4.8b, é utilizada, apenas 20% dos municípios possuem correlação menor ou igual a 60%. E, aproximadamente 60% dos municípios possuem correlação maior que 70%, ou seja, há uma forte correlação entre o número de *tweets* de experiência pessoal e o número de

Tabela 4.6: Média e desvio padrão para as correlações realizadas considerando o limiar de um *tweet* de experiência pessoal por dia, considerando o total de dias.

tweets	*Casos - Média (desvio padrão)*	
	data de notificação	*data dos primeiros sintomas*
Todos	0,6599 (0,2229)	0,6851 (0,2152)
Exp. Pessoal	0,7454 (0,1699)	**0,7485 (0,1704)**
Informação	0,4612 (0,2882)	0,5023 (0,2727)
Opiniã	0,3751 (0,2354)	0,4039 (0,2421)
Campanha	0,2987 (0,2387)	0,3242 (0,2423)
Ironia	0,3624 (0,1866)	0,3751 (0,1945)

casos de dengue notificados pelo Ministério da Saúde.



(a) (b)

Figura 4.8: CDF da correlação de *tweets* das cinco categorias de conteúdo e considerando todos os *tweets* com o número de casos de dengue por data de notificação (a) e por data dos primeiros sintomas (b) dos municípios cujo total de *tweets* é no mínimo o equivalente a um *tweet* de experiência pessoal por dia.

O próximo passo é comparar as correlações e verificar qual gera o melhor resultado. As comparações são realizadas entre quatro correlações, as que utilizam todos os *tweets* ou apenas os *tweets* de experiência pessoal e as que utilizam os casos oficiais de dengue pela data da notificação ou pela data dos primeiros sintomas. São realizadas um total de seis comparações resultantes da combinação das quatro séries duas a duas. Com o intuito de verificar qual é a melhor combinação de séries, é calculada a diferença dos valores obtidos para cada cidade em cada uma dessas correlações e é gerado um intervalo de confiança de 99% para essa diferença. Na Tabela 4.7 há o intervalo de confiança para cada uma dessas diferenças. Conclui-se que, com 99%

de confiança, as correlações que utilizam os *tweets* de experiência pessoal são sempre melhores do que as correlações que utilizam todos os *tweets* e não há diferença ao considerar a data de notificação ou a data dos primeiros sintomas na correlação.

Tabela 4.7: Intervalo de confiança de 99% das comparações entre as correlações.

	tweets - sintomas	E.P. - notificação	*tweets* - notificação
E.P. - sintomas	$[0,0161; 0,11065]$ ↑	$[-0,0211; 0,0273]$	$[0,0292; 0,14805]$ ↑
tweets - sintomas	-	$[-0,1066; -0,0140]$ ↓	$[-0,0011; 0,0516]$
E.P. - notificação	-	-	$[0,0379; 0,1331]$ ↑

Concluímos que a melhor correlação entre Twitter e dados oficiais do Ministério da Saúde é obtida quando se considera os *tweets* de experiência pessoal e a data dos primeiros sintomas. Doze cidades foram escolhidas para uma análise mais detalhada. Dentre as cidades escolhidas, 9 são capitais: Belém, Belo Horizonte, Brasília,Fortaleza, Manaus, Natal, Rio Branco, Rio de Janeiro, São Paulo; e 3 são cidades do interior: Londrina, Niterói, Ribeirão Preto. O resultado da correlação, assim como o número de casos de dengue por 100 mil habitantes e o número de *tweets* de experiência pessoal para cada uma dessas cidades estão na Tabela 4.8. Há uma forte correlação entre os dados do Twitter e do Ministério da Saúde para todos os municípios apresentados, uma cidade que merece destaque é o Rio de Janeiro que apresentou uma correlação de 97,79%.

Tabela 4.8: Número de casos de dengue por 100 mil habitantes, volume de *tweets* de experiência pessoal (e.p.) e a correlação para as doze cidades escolhidas.

Cidade	*#casos/100mil hab.*	*#tweets e.p.*	*Correlação (r) para d=0*
Belém	171,32	1.930	0,7641
Belo Horizonte	92,38	2.192	0,7183
Brasília	142,37	2.156	0,8489
Fortaleza	1.445,13	3.731	0,9340
Londrina	1.456,15	328	0,8089
Manaus	3.034,66	5.602	0,9545
Natal	1.297,02	5.186	0,8835
Niterói	1.155,79	747	0,9119
Ribeirão Preto	4.796,85	2.439	0,9474
Rio Branco	6.392,67	1.016	0,8090
Rio de Janeiro	1.197,09	16.035	0,9779
São Paulo	58,53	8.971	0,8807

A próxima análise é a correlação considerando um desvio (d) de quatro semanas. Dessa forma, a série temporal do Twitter é deslocada para gerar um atraso de até quatro semanas e um adiantamento de até quatro semanas, resultando num total de nove séries temporais. Essas séries temporais deslocadas são correlacionadas com a série oficial com o número de casos da dengue e verifica-se quando o melhor valor é obtido.

Na Figura 4.9 há um gráfico para cada cidade. Os valores no eixo X menores que zero representam o Twitter defasado em relação às notificações reportadas pelo Ministério da Saúde, e os valores maiores que zero, o Twitter adiantado. As cidades de Brasília, Fortaleza, Manaus, Niterói, Ribeirão Preto, Rio de Janeiro e São Paulo, apresentaram maior correlação para o valor de $d = 0$, ou seja, o volume de mensagens no Twitter correlaciona com o volume de casos do Ministério da Saúde na mesma semana, sem nenhum atraso ou adiantamento. Já nas cidades de Londrina, Rio Branco e Natal, a série do Twitter apresentam um atraso de uma semana em relação aos dados oficiais. Em Belém, a correlação entre os dados é alta para $d = 0$ mas é maior ainda se consideramos três semanas de atraso do Twitter. A única cidade na qual o Twitter foi adiantado em relação ao Ministério da Saúde foi em Belo Horizonte cuja maior correlação foi obtida para $d = 2$, ou seja, os dados do Twitter se adiantaram em relação a duas semanas em relação às notificações do Ministério da Saúde.

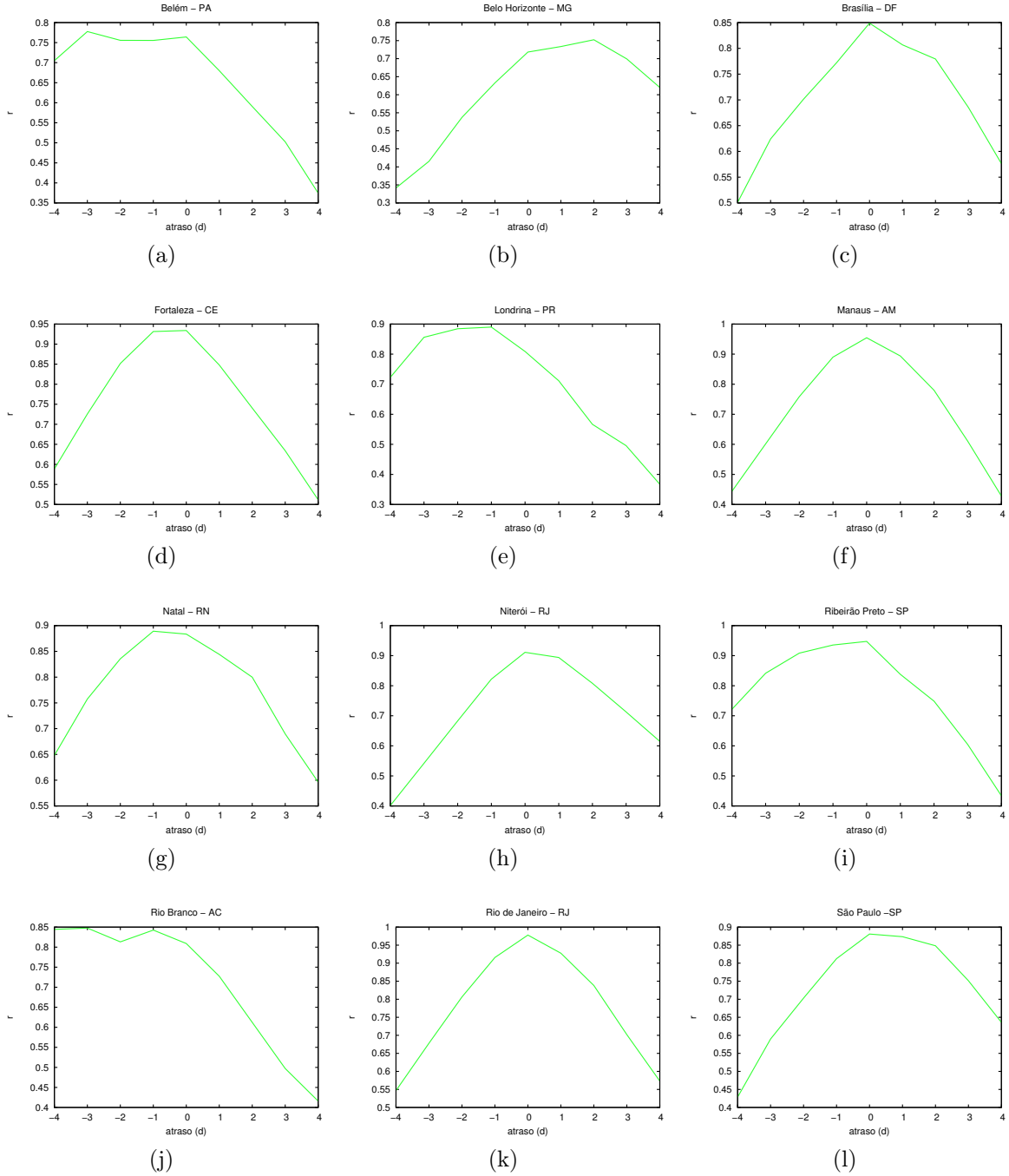
Essas correlações, considerando o desvio de semanas, dependem da cultura das pessoas do município em relação a usar as redes sociais, e do rigor do processo de notificações por parte da secretaria de saúde desse município que pode ser rigorosa ou não quanto a notificação dos casos de dengue. Mas de uma forma geral, conclui-se que o volume de *tweets* no Twitter se correlaciona com o volume de casos de dengue ao considerar a mesma semana.

4.1.3.2 Localidade temporal

Nesta seção são apresentados os resultados da análise da localidade temporal. Para cada data e cada município é definido um *Event Index*, conforme descrito na seção 3.4.2.

Ao analisar o *Event Index* tem-se o intuito de verificar se, durante as semanas cuja incidência de dengue foi alta, há uma maior concentração das mensagens se comparado com um período de baixa incidência. Em outras palavras, verificar se o *Event Index* é maior durante os períodos críticos para a dengue do que em períodos de baixa incidência da doença.

A Figura 4.10 ilustra o histograma do *Event Index* para cada uma das três classes



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figura 4.9: Correlação entre *tweets* de experiência pessoal e casos de dengue considerando a data dos primeiros sintomas com um desvio de 4 semanas.

de incidência para os municípios de Manaus e do Rio de Janeiro. Observe que para média e alta incidência, Figuras 4.10b, 4.10c, 4.10e e 4.10f, os valores de *Event Index* são maiores do que para época de baixa incidência da dengue, Figuras 4.10a e 4.10d.

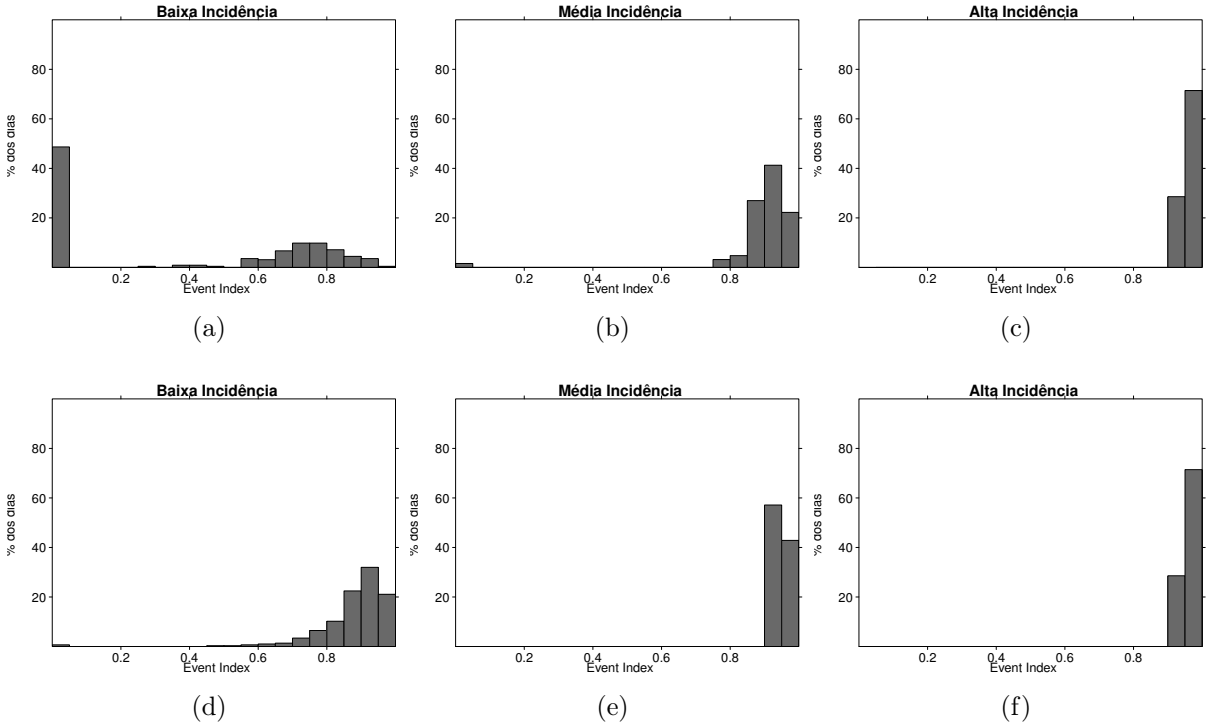


Figura 4.10: Histograma do *Event Index* para Manaus e Rio de Janeiro. Em (a), (b) e (c) histograma cidade de Manaus em períodos de baixa, média e alta incidência de dengue, respectivamente. E em (d), (e) e (f) para a cidade do Rio de Janeiro.

O valor do *Event Index* é comparado com o número de *tweets* e com o número de casos de dengue na Figura 4.11. Observe que os dias pertencentes às semanas de alta incidência de dengue sempre possuem alto valor de *Event Index*. Além disso, nos dias com mesmo volume de *tweets* é possível diferenciar as classes de incidência de dengue ao considerar o valor do *Event Index*, cujo valor é maior para semanas com maior incidência.

4.1.3.3 Similaridade Espacial

Nesta seção são ilustrados experimentos e resultados da análise espacial. Primeiramente, foi criado um arquivo para cada uma das 42 semanas com o volume de *tweets* de experiência pessoal e de casos notificados pela data dos primeiros sintomas. Foram definidos os parâmetros de entrada do algoritmo de agrupamento

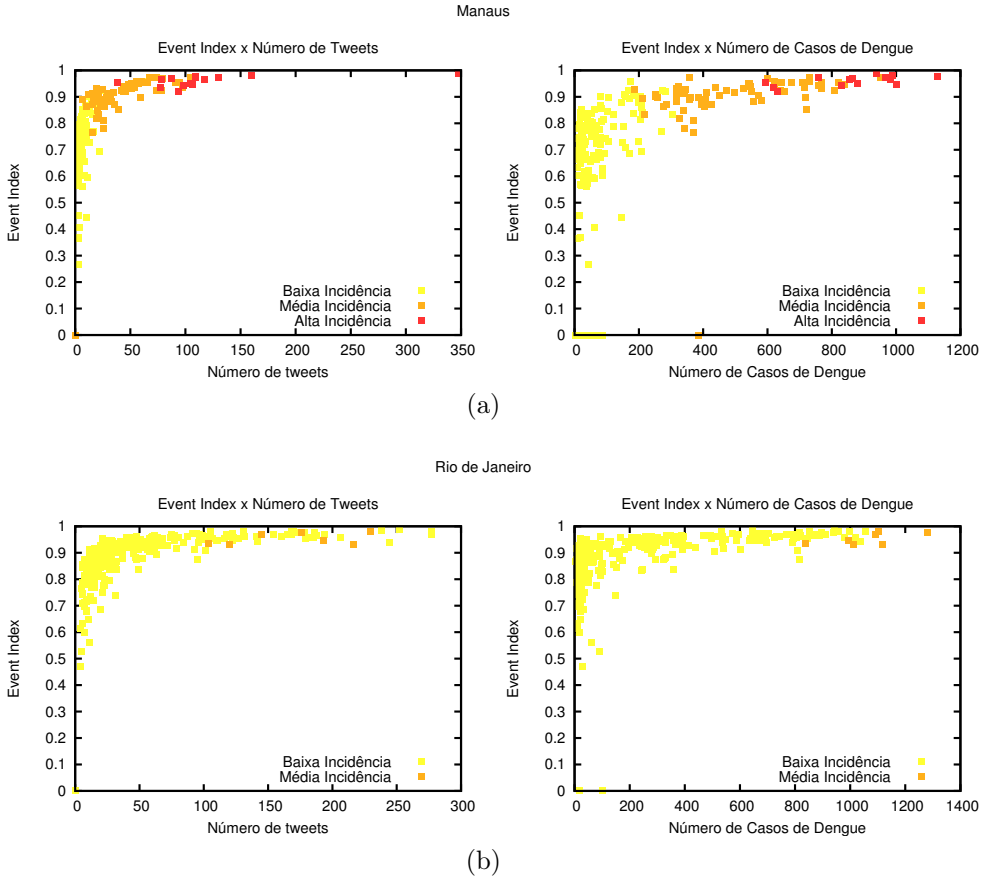


Figura 4.11: Gráficos do Event Index x Número de *tweets* e Event Index x Número de casos de dengue para as cidades de Manaus (a) e Rio de Janeiro (b).

espaço-temporal. Os dados do Twitter e do Ministério da Saúde foram agrupados e avalia-se a correlação entre os agrupamentos obtidos.

Os valores semanais tanto para o volume de *tweets* de experiência pessoal quanto para o número de casos notificados são transformados em taxas de incidência, ou seja, o volume por 100 mil habitantes para formar o agrupamento. Como nas outras análises, consideramos os municípios com mais de 100 mil habitantes.

Há três parâmetros de entrada do algoritmo: o número mínimo de pontos (*MinPts*), a distância geográfica máxima entre dois pontos (*Eps1*) e a diferença máxima entre as taxas de incidência da doença (*Eps2*). O valor de *MinPts* será 2, pois dois municípios já formam uma região. O valor de *Eps1* foi encontrado pela heurística descrita em Ester et al. [1996] e seu valor está entre 1.0 e 3.5. O único valor que varia da base do Twitter para a base do Ministério da Saúde é o *Eps2*. Para a base do Twitter o valor de *Eps2* encontrado pela heurística está entre 0.1 e 0.3 de diferença entre o número de *tweets* por 100 mil habitantes. Já para a diferença entre as taxas de incidência do número de casos o valor está entre 0.5 e 1.5.

O algoritmo ST-DBScan foi executado para todos esses parâmetros. Para medir a correlação entre os agrupamentos formados, foi calculado o *Rand Index* para a combinação de todas as configurações dos parâmetros. Os valores do *Rand Index* para essas combinações estão na Figura 4.12.

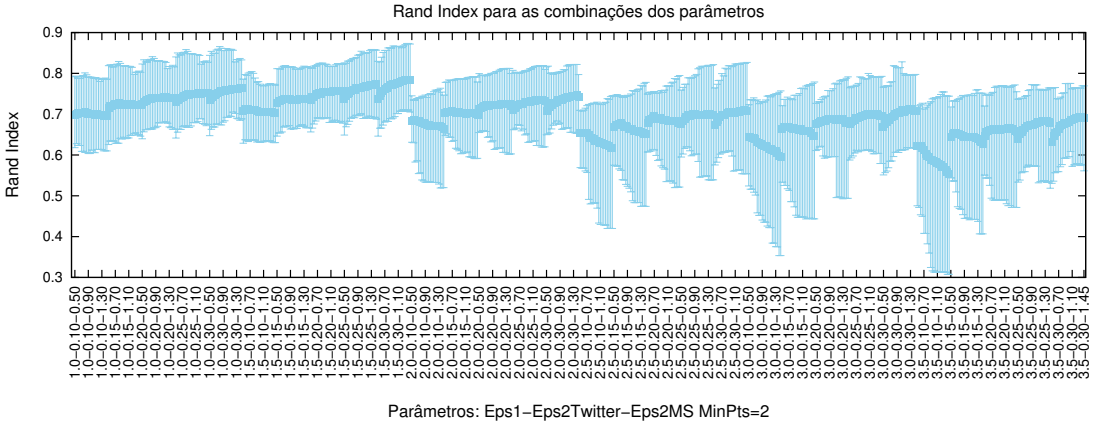


Figura 4.12: Valores do *Rand Index* para todas as combinações de parâmetros. Os valores dos parâmetros estão na seguinte ordem: valor do Eps1, valor do Eps2 para a incidência de *tweets* e valor do Eps2 para a incidência de casos notificados. O valor de MinPts é 2.

A configuração de parâmetros que gerou o maior *Rand Index* na média das semanas é: *Eps1*=1,5; *Eps2Twitter*=0,3; *Eps2MS*=1,45; *MinPts*=2. As características dos agrupamentos obtidos são apresentados na Tabela 4.9.

Tabela 4.9: Características dos agrupamentos formados com a configuração cuja correlação gerou maior valor médio do *Rand Index*. Apresentamos a média do valor para todas as semanas, o valor mínimo e o valor máximo.

Rand Index	0,7844 (min= 0,7053, max=0,8718)	
	Twitter	*Ministério da Saúde*
#agrupamentos	30,61 (min=21, max=40)	25,07 (min=16, max=35)
#pontos ruído	96,57 (min=61, max=141)	93,83 (min=42, max=144)
#pontos nos clusters	187,43 (min=143, max=223)	190,17 (min=140, max=242)

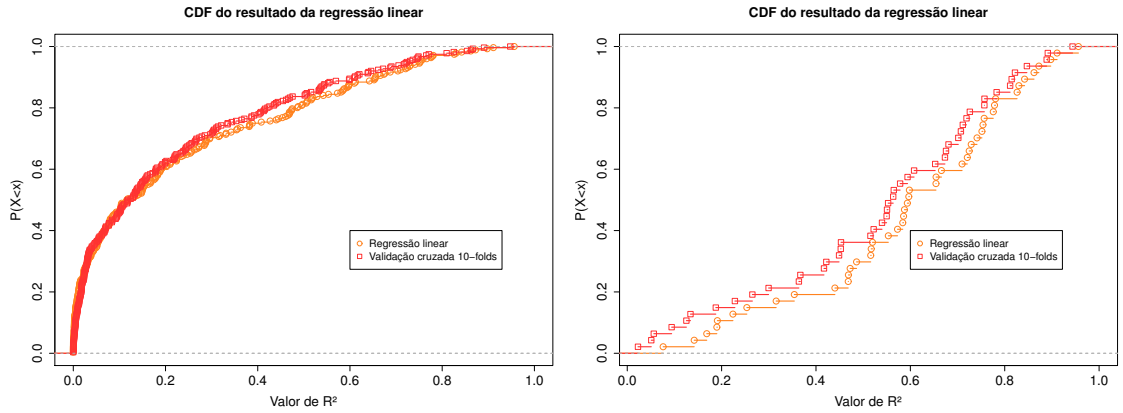
4.1.4 Prevendo a Dengue

As mensagens publicadas no Twitter podem servir de instrumento para a previsão do número de casos de dengue. Nessa seção, os resultados da previsão do volume de casos de dengue é mostrado e, além disso, cada município é classificado de acordo com o nível (baixo, médio ou alto) de incidência da doença.

4.1.4.1 Inferir a quantidade de casos de dengue

Para prever a quantidade de casos de dengue, foi gerado um modelo de regressão linear para cada município. Esse modelo de regressão linear considera duas variáveis: t, o número semanal de *tweets* classificados como sendo de experiência pessoal, e o, o número semanal de casos de dengue notificados por data dos primeiros sintomas. Essas variáveis foram escolhidas em razão de terem apresentado maior correlação como apresentado na seção 4.1.3.1.

A regressão foi realizada para todas as cidades com mais de cem mil habitantes e também para as cidades com um suporte mínimo de uma média de um *tweet* de experiência pessoal por dia. As CDFs das correlações obtidas por meio das regressões lineares se encontram na Figura 4.13. Em ambos os gráficos, as curvas da correlação para a regressão com todos os pontos e para validação utilizando os 10 partições estão bem próximas. Quando todos os 285 municípios são considerados, Figura 4.13a, observa-se que apenas 20% deles possuem correlação maior que 60%. Entretanto, ao descartar as cidades com nenhum ou poucos *tweets* e impor um limiar para o número de *tweets*, a correlação melhora significativamente. Na Figura 4.13b é possível observar que aproximadamente 20% dos municípios possuem correlação inferior a 40% e metade dos municípios possuem correlação superior a 60%.



(a) (b)

Figura 4.13: CDF da do resultado da regressão linear (a) e resultado da validação cruzada com 10 partições (b).

As três cidades que apresentaram maior correlação foram Rio de Janeiro, Manaus e Ribeirão Preto. Os resultados da regressão linear e da validação cruzada se encontram na Tabela 4.10. O valor de R^2 para a validação cruzada com 10 partições e ao considerar todas as datas são parecidos.

Tabela 4.10: Resultado da regressão linear. Na função de previsão, o é número de casos previstos e t é número de *tweets* de experiência pessoal

Cidade	*Função de previsão*	R^2	*p-value*	R^2 *da validação cruzada com 10-partições*
Manaus	$o = 93,5834 + 8,3029 \times t$	0,9110	$<$ 2,2e-16	0,8629
Ribeirão Preto	$o = 174,765 + 15,075 \times t$	0,8976	$<$ 2,2e-16	0,8908
Rio de Janeiro	$o = 402,31 + 5,826 \times t$	0,9562	$<$ 2,2e-16	0,9477

O número de *tweets* de experiência pessoal, o número de casos de dengue notificados e o número de casos de dengue previsto utilizando a função criada pela regressão linear ao longo do período de epidemia da dengue do ano de 2011 se encontram na Figura 4.14.

4.1.4.2 Classificando a intensidade da incidência de dengue

A regressão linear é utilizada para criar uma função que infere o número de casos de dengue semanalmente para determinado município a partir do número de *tweets* de experiência pessoal. Nessa seção detalhamos como o valor previsto é utilizado para classificar a situação de cada município, o que será feito com o mesmo critério do Ministério da Saúde. A análise foi feita considerando os municípios com mais de 100 mil habitantes.

Para avaliar a situação de cada município, o Ministério da Saúde considera a incidência dos casos de dengue. Ela é alta quando há mais de 300 casos por 100 mil habitantes; média entre 100 e 300 e baixa entre 0 e 100 casos por 100 mil habitantes.

Com o número previsto para cada cidade em uma determinada semana, calcula-se o valor para 100 mil habitantes e verifica-se em qual classe se encontra. A classificação obtida utilizando os dados do Twitter é comparada com a incidência calculada pelo Ministério da Saúde para avaliar a previsão realizada.

O número de cidades que apresentam pelo menos uma semana classificada em cada classe de incidência de dengue se encontra na Tabela 4.11. Todas as 285 cidades consideradas possuem pelo menos uma semana com baixa incidência de dengue. E, apenas cinco cidades apresentaram alguma de suas semanas com alta incidência de dengue de acordo com o Ministério da Saúde, são elas: Foz do Iguaçu, Rio das Ostras, Manaus, Ribeirão Preto e Rio Branco. As três últimas cidades tiveram alguma semana com alta incidência segundo a previsão com o número de *tweets*.

Na Figura 4.15 se encontram 3 CDFs do resultado da classificação da incidência

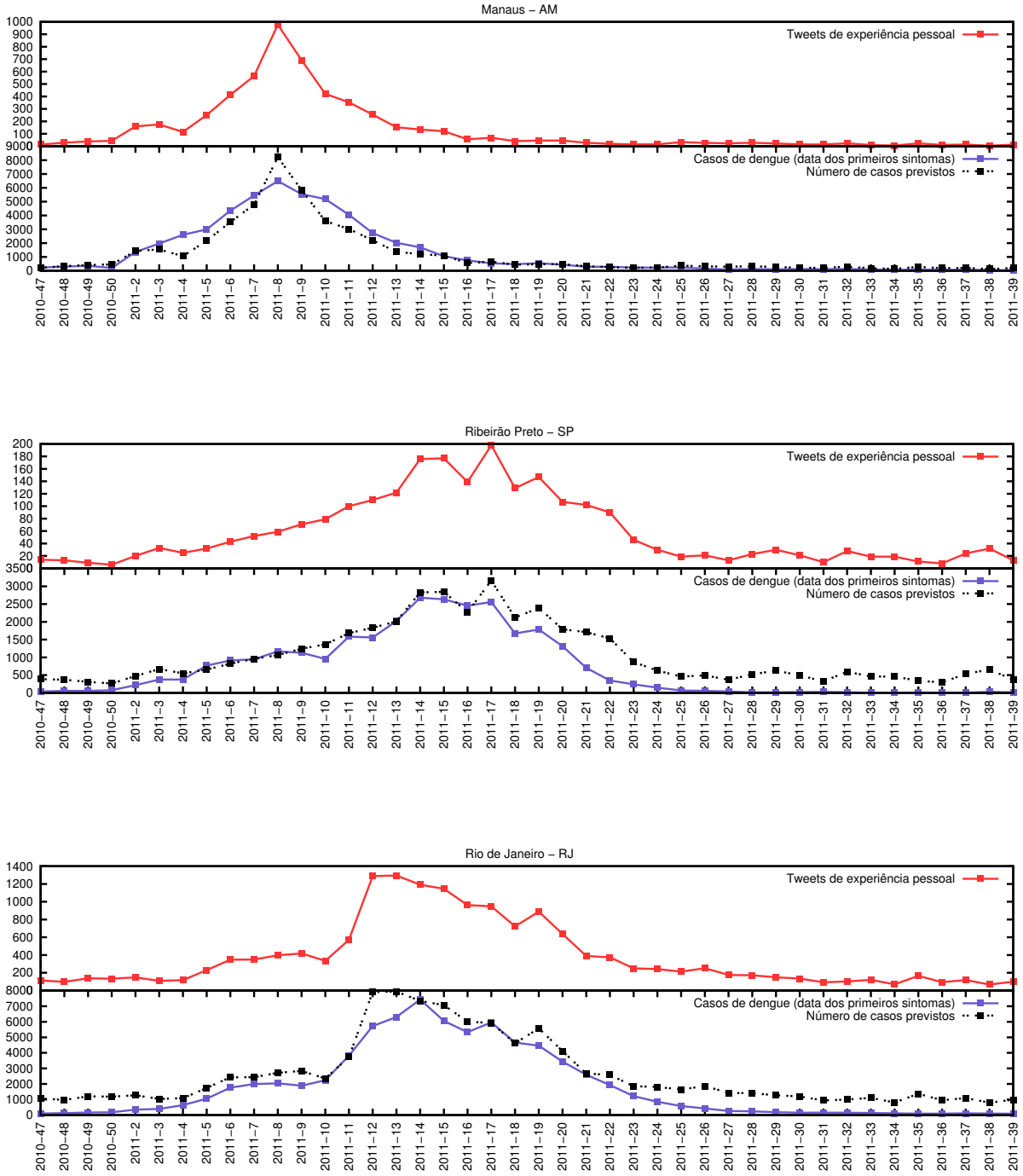


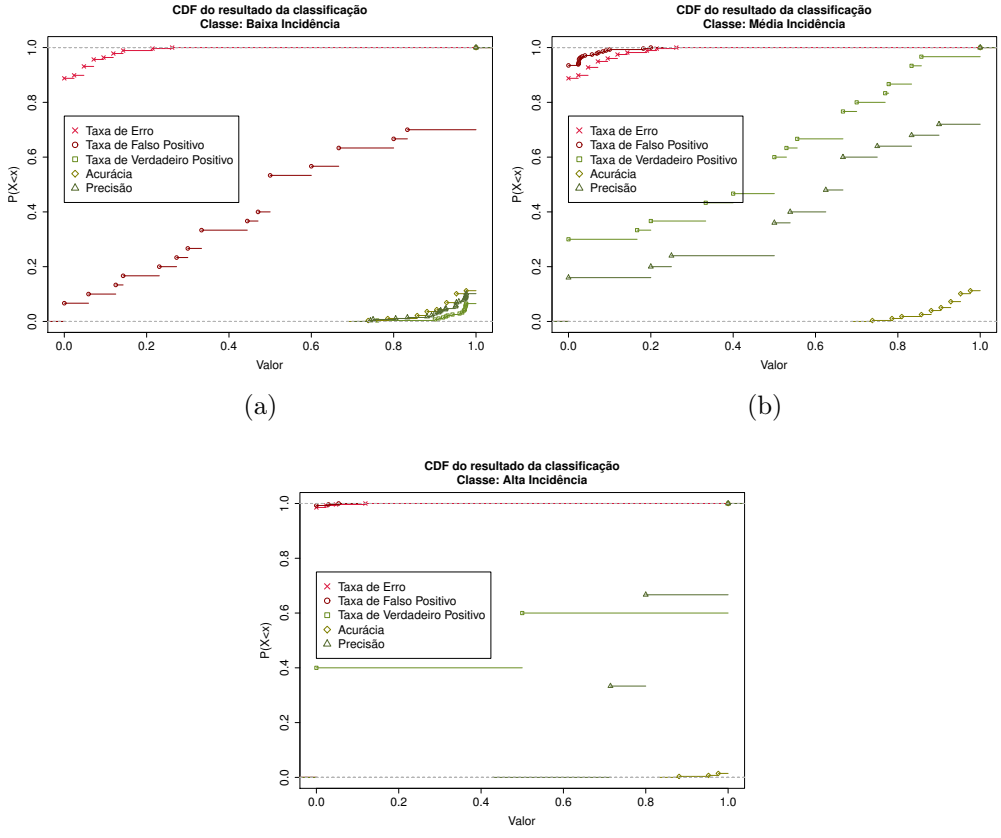
Figura 4.14: Volume de *tweets* de experiência pessoal, número de casos de dengue notificados por data dos primeiros sintomas e valor do número de casos previstos utilizando a regressão linear.

da dengue. Todas as cidades apresentaram taxa de erro menor que 30% nas semanas cuja classe foi baixa incidência, Figura 4.15a, e as taxas de verdadeiro positivo, acurácia

Tabela 4.11: Quantidade de cidades que possuem alguma semana classificada em cada uma das três classes de incidência.

Classe de incidência casos/100mil hab.	*#Cidades* *Ministério da Saúde*	*Previsão*
Baixa (0 - 100)	285	285
Média (100 - 300)	30	25
Alta (mais de 300)	5	3

e previsão foram quase todas acima de 90%. Para a média incidência, Figura 4.15b, as taxas de erro e falso positivo em todas as cidades foram menores do que 30% e a acurácia em apenas 20% das cidades foi menor do que 100%. Para a classe de alta incidência, 4.15c, as taxas de erro e falso positivos para aproximadamente 100% das cidades foram menor que 10% e a taxa de acurácia foi alta.



(a) (b)

(c)

Figura 4.15: CDF do resultado da classificação da incidência da dengue para os municípios. Em (a) os resultados para baixa incidência, em (b) para a incidência média e em (c) para alta incidência.

4.1.5 Alerta contra dengue

O alerta reuni praticamente todas as partes anteriores: volume de *tweets* de experiência pessoal, volume de casos de dengue, função de regressão linear. Além disso, agrega um novo conceito, o da tendência dos casos de dengue.

Para cada um dos municípios com mais de 100 mil habitantes, são apresentados semanalmente dois indicadores: a situação atual (ou incidência) e a tendência da dengue. A situação atual indica o que está ocorrendo no exato momento e é baseada no volume de *tweets* de experiência pessoal. A tendência indica se esse volume tem aumentado ou diminuído nas últimas semanas.

A seguir, é descrito o cálculo da incidência relativa de dengue e da tendência relativa de dengue. A visualização dos resultados para todas as cidades com mais de 100 mil habitantes pode ser acessada http://homepages.dcc.ufmg.br/~janaina/dissertacao/sistemaAlerta/

4.1.5.1 Avaliação da incidência relativa de dengue

A incidência relativa de dengue é calculada semanalmente por município. No seu cálculo é utilizado o volume de *tweets* de experiência pessoal, a função gerada pela regressão linear e a população do município.

A função de previsão gerada pela regressão linear é utilizada para gerar o número de casos de dengue previsto a partir do volume de *tweets*. O valor da incidência relativa é o valor previsto da quantidade de casos de dengue.

Para visualizar a intensidade da incidência de dengue é utilizada uma escala de cores. Essa escala é gerada comparando-se o número de casos de dengue previsto com a classificação do Ministério da Saúde. O Ministério da Saúde classifica a situação da dengue de acordo com o número de casos por 100 mil habitantes. A incidência é baixa para 0 a 100 casos por 100 mil habitantes por semana, média para 100 a 300 e alta acima de 300.

Para cada município são estabelecidos limites para o número de casos de acordo com a população. O limite inferior (LI) para o número de casos é 0. O limite superior (LS) é 300 casos por 100 mil habitantes. Quanto mais próximo do LS, mais próxima do vermelho será a cor representada na visualização. Caso exceda o LS, será utilizado o vermelho absoluto. Da mesma forma, caso seja menor que o LI será utilizado o branco absoluto.

Há também o limite intermediário (LM) que é o valor correspondente a 100 casos por 100 mil habitantes. Para valores entre LI e LM será utilizada uma escala em degradê variando do branco ao amarelo sobre a porcentagem entre o valor mínimo e

o valor médio. E para valores entre LM e LS será utilizada um degradê variando do amarelo passando pelo laranja até o vermelho. Na Figura 4.16 está essa escala de cores com os limites para as faixas de cores.

Figura 4.16: Escala de cores para alerta sobre a incidência relativa de dengue.

A visualização da incidência relativa da dengue no alerta pode ser conforme mostrada na Figura 4.17. Observe que no início do ano a incidência estava bem alta, durante a epidemia, e depois foi diminuindo.

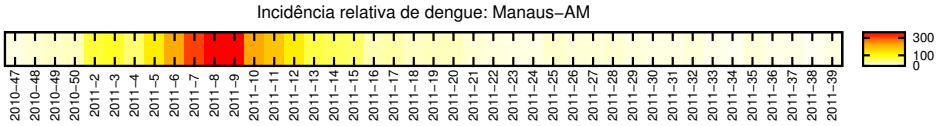


Figura 4.17: Visualização do sistema de alerta: Incidência relativa da dengue para Manaus.

4.1.5.2 Avaliação da tendência relativa de dengue

A tendência relativa de dengue é calculada semanalmente por município. No seu cálculo é utilizado o volume de *tweets* de experiência pessoal de cada dia da semana atual e das duas semanas anteriores.

Utilizamos o Z-score, apresentado na seção 3.6.2, para estimar a tendência. A média é calculada utilizando a média dos 14 dias das semanas anteriores e o valor atual é a média dos 7 dias da semana atual.

Para visualizar a intensidade da tendência de dengue foi utilizada uma escala de cores. Essa escala é gerada comparando o valor do Z-score com dois limites pré-definidos. O limite inferior (LI) é o valor -1, ou seja, houve a diminuição de 1 desvio padrão em relação às duas semanas anteriores. O limite superior (LS) é o valor 2, ou seja, houve o aumento de 2 desvios padrões em relação às semanas anteriores. Quanto mais próximo do LS, mais próxima do vermelho será a cor representada na visualização. Caso exceda o LS, será utilizado o vermelho absoluto. Da mesma forma, caso seja menor que o LI será utilizado o branco absoluto. Na Figura 4.18 está essa escala de cores com os limites para as faixas de cores.

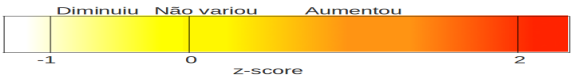


Figura 4.18: Escala de cores para alerta sobre a tendência relativa de dengue.

A visualização da tendência relativa da dengue no alerta para a cidade de Manaus é mostrada na Figura 4.19. Observe que já nas semanas 5 e 6 a tendência para os casos de dengue em Manaus era de aumento o que só foi percebido na incidência de dengue na semana 7.

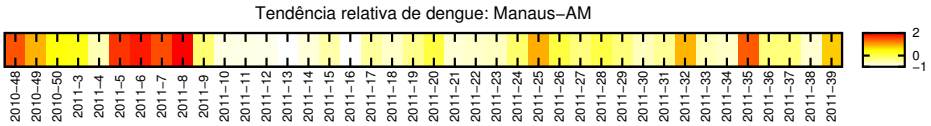


Figura 4.19: Visualização do sistema de alerta: Tendência relativa da dengue para Manaus.

4.2 Alagamentos e Enchentes

Todos os anos, durante as chuvas torrenciais que costumam cair no verão, diversos municípios do Brasil sofrem com os alagamentos e as enchentes. Em função desses desastres milhares de moradores das regiões atingidas pelas chuvas perdem seus bens, ficam desabrigados e, mais grave, esses desastres causam vítimas fatais. Em 2012, várias cidades de Minas Gerais sofreram com as chuvas e mais de 50 mil pessoas ficaram desabrigadas[5] e, em 2011, a região serrana do Rio de Janeiro sofreu o maior desastre climático da história do país devido as enchentes que provocaram mais de 900 vítimas fatais[6].

O monitoramento das chuvas e alagamentos é possível por meio de um sistema que cruza dados meteorológicos (chuvas) e hidrológicos (níveis de água nos rios e córregos) coletados por meio de uma rede de estações telemétricas e de um radar meteorológico[7].

[5]Link para reportagem sobre as chuvas em Minas Gerais: http://glo.bo/xcRevQ (último acesso em 11/02/2012)

[6]Link para reportagem sobre enchente da região serrana do Rio de Janeiro: (último acesso em 11/02/2012) http://pt.wikipedia.org/wiki/Enchentes_e_deslizamentos_de_terra_no_Rio_de_Janeiro_em_2011

[7]Link para reportagem sobre sistema de monitoramento de enchente de São Paulo: (último acesso em 11/02/2012) http://exame.abril.com.br/economia/meio-ambiente-e-energia/noticias/sp-tem-novo-sistema-monitoramento-enchentes-604040

Em São Paulo, há um órgão responsável por monitorar e prever enchentes, o Centro de Gerenciamento de Emergências (CGE), que é equipado com um radar meteorológico capaz de fazer a previsão do tempo com até 15 dias de antecedência. Esse centro de gerenciamento consegue antever a chuva e emitir boletins para os principais órgãos envolvidos com a emergência na cidade. O centro indica em seu sistema a situação de cada região de São Paulo utilizando os estados de observação, atenção (quando começa um alagamento), alerta (alagamento, mais enchente) e alerta máximo (decretado apenas com autorização do prefeito).

As situações de emergência que acontecem na vida de milhares de pessoas em diversos locais do mundo têm repercussão nas redes sociais. As pessoas postam mensagens nas redes sociais sobre tal acontecimento como uma forma de alerta. Por exemplo, terremotos no Japão [Sakaki et al., 2010] e no Chile [Mendoza et al., 2010], inundação e queimada [Vieweg et al., 2010] são exemplos de situações de perigo que tiveram repercussão nas redes sociais.

Ao aplicar a metodologia neste estudo de caso é possível analisar e prever alagamentos e, além disso, disponibilizar um mecanismo de alerta sobre esses acontecimentos. A seguir vamos descrever as bases de dados utilizadas e os resultados obtidos para cada etapa da metodologia.

4.2.1 Base de dados

Nesta seção as bases de dados sobre alagamentos serão descritas em detalhe. A base de dados do Twitter contém as mensagens publicadas que se referem aos alagamentos, pontos de alagamento e enchentes. A coleta dessas mensagens foi realizada para o período de 20/10/2010 até 11/05/2011.

Os dados oficiais utilizados pelas autoridades para fazer o monitoramento das enchentes são o volume de chuva, níveis de água nos rios e córregos e o número de pontos de alagamento. Desses dados oficiais os que são disponíveis para pesquisa são o volume de chuva e número de pontos de alagamento.

Eventos como enchentes e alagamentos devem ser alertados em tempo real, mas como as informações oficiais são fornecidas diariamente, vamos considerar o período da análise como sendo dias.

4.2.1.1 Dados oficiais

Para analisar a ocorrência de alagamento e enchente serão utilizados dois tipos de dados oficiais disponíveis, o volume de chuva e o número de pontos de alagamento. O volume

de chuva diário para as capitais do Brasil que é disponibilizado *online* pelo Instituto Nacional de Meteorologia (INMET)[8].

No entanto, o índice pluviométrico não é suficiente para caracterizar a ocorrência de alagamento ou enchente. Um alto índice de chuva em determinado dia pode provocar um grande transtorno na cidade causando diversos pontos de alagamento mas pode também significar uma chuva fraca ao longo do dia ou em determinados pontos da cidade e que não causou nenhum alagamento. Por isso é necessário considerar também o número de pontos de alagamento.

A única cidade do Brasil que monitora o número de pontos de alagamento é a cidade de São Paulo. O Centro de Gerenciamento de Emergências (CGE) provê um sistema *online*[9] que fornece o número diário de pontos de alagamento para São Paulo.

As análises serão realizadas apenas para o município de São Paulo para o período de 20/10/2010 até 11/05/2011. O número de pontos de alagamento e o volume de chuva para São Paulo durante esse período estão na Figura 4.20.

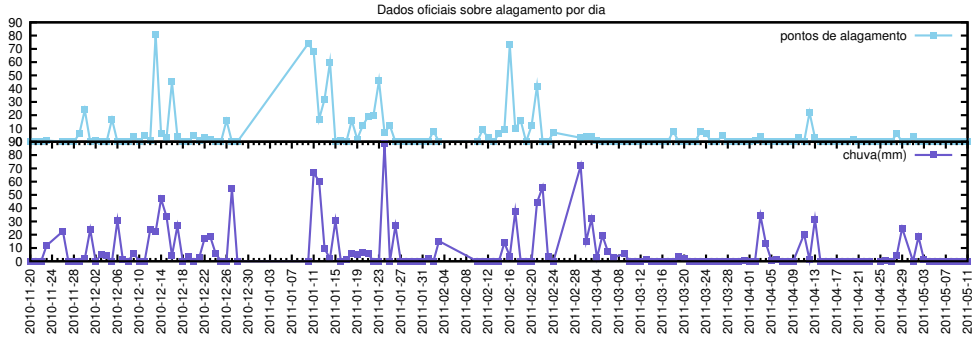


Figura 4.20: Número de pontos de alagamento e MM de chuva.

A função densidade de probabilidade (PDF) do número de pontos de alagamento pode ser vista na Figura 4.21. Em 62% dos dias não houve nenhum ponto de alagamento registrado e aproximadamente 90% dos dias teve menos de 16 pontos de alagamento.

4.2.1.2 Twitter

A coleta de mensagens postadas no Twitter sobre alagamentos e enchentes foi realizada para o período de 20/10/2010 até 11/05/2011. Durante todo esse período, houve

[8]Link para o índice pluviométrico diário das capitais do Brasil: http://www.inmet.gov.br/sim/cond_reg/tempoCapitais.php?data= (último acesso em 11/01/2012)

[9]Link para o número de pontos de alagamento diário para cidade de São Paulo: http://www.cgesp.org/pontosdealagamento_dia.php (último acesso em 11/01/2012)

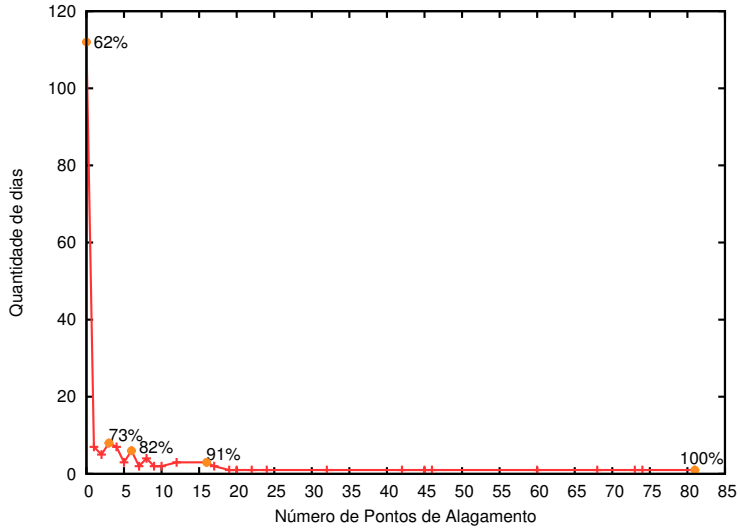


Figura 4.21: A função densidade de probabilidade (PDF) do número de pontos de alagamento.

falha na coleta entre os dias 23/12/2010 e 10/01/2011 e entre os dias 19/04/2011 e 28/04/2011, esses períodos serão desconsiderados nas análises.

Os termos escolhidos para coleta das mensagens relacionadas com os pontos de alagamento são: enchente, enchentes, alagamento, alagado, alagada, inundação, inundado, inundada, inundação.

Na Tabela 4.12 há o número de *tweets* e usuários coletados, e a parte desses dados que são do Brasil e apresentam informação de localização a nível de cidade. Aproximadamente 70% dos *tweets* sobre alagamento são do Brasil e quase 10% são do município de São Paulo. Aproximadamente 65% dos usuários que postaram mensagem são brasileiros e por volta de 8% se declaram como sendo da cidade de São Paulo.

Tabela 4.12: Número de *tweets* e usuários presentes na base de dados sobre alagamento do Twitter. Período da coleta foi de 20/10/2010 até 11/05/2011.

#*tweets*	626.202
#*tweets* do Brasil	428.447 (68,42%)
#*tweets* de São Paulo	51.694 (8,26%)
#usuários	362.327
#usuários do Brasil	228.857(63,16%)
#usuários de São Paulo	26.329(7,67%)

A função densidade de probabilidade (PDF) do número de *tweets* por usuário é mostrada na Figura 4.22. Aproximadamente 80% dos usuários postaram cinco ou menos *tweets* e 41% dos usuários postaram apenas uma mensagem.

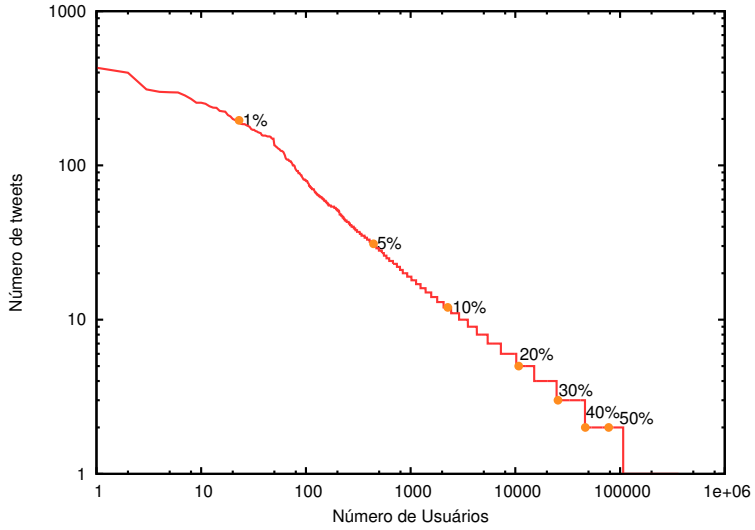


Figura 4.22: Número de *tweets* por usuario em escala logarítmica.

Durante o período de 20/10/2010 até 11/05/2011 foram coletadas 428.447 mensagens de 228.857 usuários diferentes do Brasil. A Figura 4.23 contém o número de *tweets* postados no Brasil ao longo desse período. O grande número de *tweets* entre os dias 11 e 15 de janeiro são reflexo da repercussão da catástrofe que ocorreu na região serrana do Rio de Janeiro devido as enchentes[10].

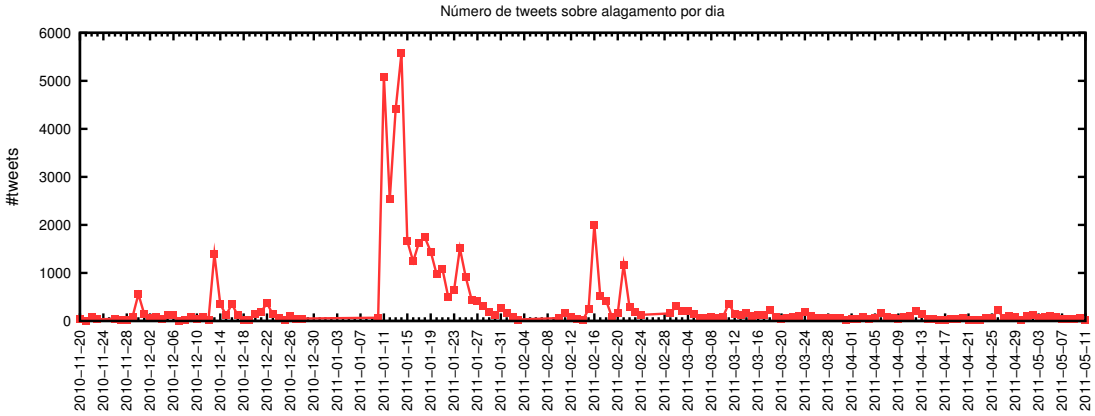


Figura 4.23: Número total de *tweets* coletados com localização a nível de cidade durante todo período de coleta.

[10]Link para reportagem sobre enchente da região serrana do Rio de Janeiro: (último acesso em 11/01/2012) http://pt.wikipedia.org/wiki/Enchentes_e_deslizamentos_de_terra_no_Rio_de_Janeiro_em_2011

4.2.2 Análise de Conteúdo

Nessa seção são apresentados os resultados da análise de conteúdo que foi feita nas mensagens sobre os alagamentos e as enchentes. Inicialmente, algumas características dessas mensagens serão apresentadas e os tipos de conteúdo serão exemplificados. A seguir, descreve-se como treino é criado e avalia-se o desempenho do classificador. E, por último, o resultado da classificação do conteúdo de todas as mensagens é discutido e algumas características desses dados são apresentados.

As mensagens postadas no Twitter sobre os alagamentos e as enchentes terão seu conteúdo classificado. O LAC, classificador associativo sob demanda, formará as regras utilizando como atributo as palavras (ou tokens) da mensagem. Antes de ser classificado, cada *tweet* teve seu texto processado da seguinte forma: remoção da acentuação; remoção dos caracteres RT, que classificam a mensagem como um *retweet*; remoção da menção às páginas web (p.ex., http); remoção da menção aos usuários; as letras maiúsculas foram substituídas por letras minúsculas; remoção de todos os caracteres alfa-numéricos, tais como vírgulas e pontos.

O número de mensagens, número de atributos (palavras ou tokens da mensagem), tamanho do vocabulário (número de tokens diferentes) e a média do número de atributos por mensagens são apresentados na Tabela 4.13.

Tabela 4.13: Características das mensagens postadas no Twitter sobre alagamentos.

Número de mensagens	428
Número de atributos (tokens do *tweet*)	3210
Tamanho do vocabulário (tokens diferentes)	2352
Média do número de atributos por mensagem	7,5 (min=1, max=22)

Esse cenário de aplicação tem um sentido mais imediato, ou seja, não é de interesse saber se já ocorreu alguma enchente ou se ontem ocorreram pontos de alagamento. Os *tweets* que vão auxiliar na análise de correlação e previsão serão os *tweets* que descrevem uma situação do presente, do momento atual que a pessoa está vivenciando. Devido a esse motivo, utiliza-se a classificação composta por duas classes descritas na Tabela 3.2.

O treino criado contém 428 mensagens que foram classificadas com a ajuda de 2 alunos do curso de Ciência da Computação da UFMG. Foi apresentado aos alunos as classes de conteúdo nas quais as mensagens seriam classificadas e mostramos três exemplos de *tweets* para cada classe para que eles pudessem compreender melhor o significado de cada uma delas. A porcentagem de cada classe de conteúdo presente no

treino é 285 para presente e 143 para outros. Alguns exemplos de *tweets* de cada classe estão na Tabela 4.14.

Tabela 4.14: As categorias de conteúdo e exemplos de *tweets*.

Evento em tempo real
• Aqui na Artur de Azevedo, esquina com a Mateus Grow já está tudo alagado. A água tá chegando na calçada. Chuva em São Paulo.
• a João Paulo continua alagada, já quebraram o vidro do onibus, e as pessoas estão saindo por coordas.. E a policia? Chego e foi embora.
• Presa em SP... Tudo alagado! 9 de julho, estados unidos, brasil... Tudo sem condições! http://yfrog.com/h2y76tij
Outros
• CPI das enchentes interroga empresas prestadoras de serviços relacionados à manutenção urbana.
• vendo as enchentes na região serrana no rio no globonews. que os deuses sejam misericordiosos, a coisa tá feia demais :(
• Prefeitura e Estado precisam de planejamento e obras de longo prazo para combater as enchentes em São Paulo http://bit.ly/egxUWS

Para avaliar o classificador de conteúdo das mensagens foi feita uma validação cruzada com 5 partições no conjunto de mensagens classificadas manualmente. Os valores para precisão, taxa de verdadeiro positivo e acurácia estão na Tabela 4.15. Aproximadamente 75% das previsões foram corretas e quase 80% das previsões de alagamento estavam corretas.

Tabela 4.15: Resultados da validação cruzada com 10 partições na tarefa de classificação do conteúdo das mensagens.

Métrica	*Valor*
Precisão	0.7878 (min=0.7391, max=0.8475)
Taxa de verdadeiro positivo	0.4392 (min=0.3950, max=0.5135)
Acurácia	0.7698 (min=0.7616, max=0.7878)

De todos dos *tweets* que referenciam a cidade de São Paulo, o número de *tweets* classificados como sendo da classe evento em tempo real são 2037 (3,94%). A Figura 4.24 mostra o número de *tweets* da classe evento em tempo real. Os picos no número de *tweets* são referentes a dias com muitos pontos de alagamento na cidade, ver Figura

4.20, ou nos dias da enchente da região serrana do Rio de Janeiro que repercutiu por todo o Brasil.

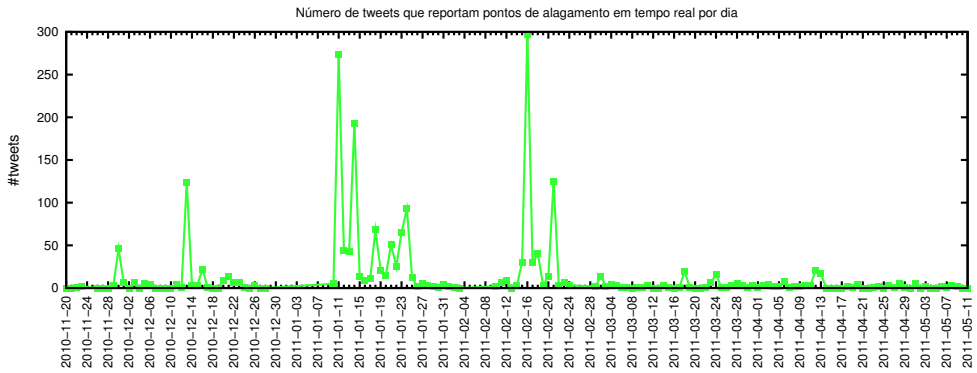


Figura 4.24: Número total de *tweets* que reportam algum ponto de alagamento em tempo real com localização a nível de cidade durante todo período de coleta.

4.2.3 Análise de Correlação

As análises realizadas nesta seção sempre correlacionam as mensagens publicadas no Twitter sobre os alagamentos com as informações fornecidas pelo CGE sobre esses acontecimentos. Como é considerado apenas o município de São Paulo nos experimentos e não há informações sobre as regiões desse município, não será feita a análise da similaridade espacial visto que não há regiões para ser agrupadas. Os resultados para as outras duas análises de correlação realizadas são descritos a seguir.

4.2.3.1 Deslocamento ao Longo do Tempo

Nesses experimentos é mensurada a correlação entre o volume de chuva e o número de pontos de alagamento com o volume das mensagens do Twitter. A correlação é calculada considerando o volume diário durante todo o período de 20/10/2010 até 11/05/2011 excluindo os dias que houve falha na coleta, especificamente o período entre os dias 23/12/2010 e 10/01/2011 e entre os dias 19/04/2011 e 28/04/2011. No total serão 179 dias considerados.

As séries temporais formadas pelo volume de chuva e o número de pontos de alagamento para o município de São Paulo são considerados os dados oficiais sobre o evento. Essas séries oficiais serão comparadas com duas outras séries. Uma delas considera todas mensagens sobre as enchentes e alagamentos postadas no Twitter e a outra contém o volume apenas das mensagens sobre o evento em tempo real.

A Tabela 4.16 contém o resultado da correlação entre as quatro séries. Observe que o melhor resultado, aproximadamente 80% foi obtido ao correlacionar o volume de *tweets* que descrevem uma situação sobre as enchentes em tempo real com o número de pontos de alagamento.

Tabela 4.16: Correlação de Pearson

	chuva(mm)	*pontos de alagamento*
Todos tweets	0,3297	0,6276
tweets sobre o evento em tempo real	0,3461	**0,7950**

O próximo passo é mostrar como é a correlação considerando um desvio (d) de dias. A correlação cruzada foi feita com um desvio de sete dias, ou seja, com o deslocamento em relação a semana anterior e a semana posterior. Como a melhor correlação obtida foi utilizando o volume de *tweets* sobre o evento em tempo real e o número de pontos de alagamentos serão utilizadas essas duas séries como as séries dos dados do Twitter e dos dados oficiais, respectivamente.

Na Figura 4.25 há um gráfico com o resultado da correlação cruzada. Os valores no eixo X menores que zero representam o Twitter defasado em relação ao número de pontos de alagamento e os valores maiores que zero, adiantado. Observe que a melhor correlação foi obtida sem nenhum atraso das mensagens em relação aos pontos de alagamento, comprovando que houve repercussão da ocorrência desses pontos de alagamento na rede social no exato momento da ocorrência desses.

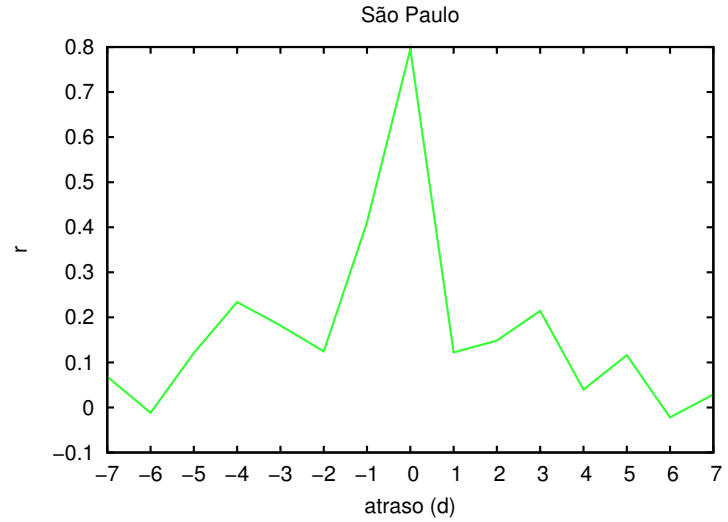


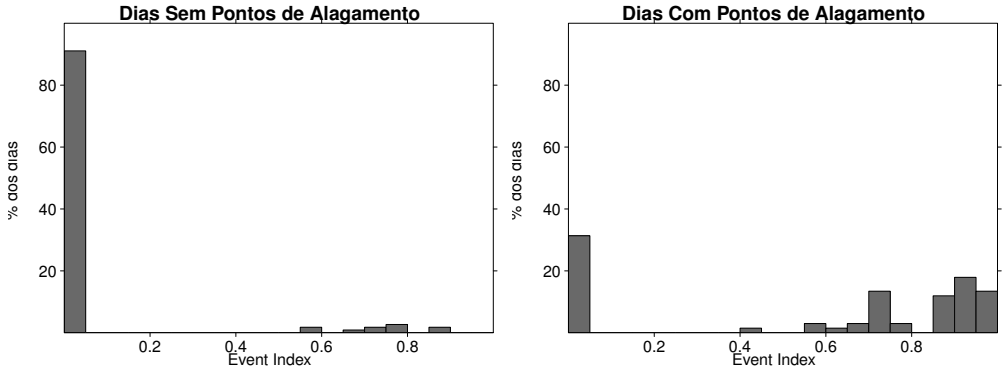
Figura 4.25: Correlação cruzada entre *tweets* sobre o evento em tempo real e pontos de alagamento com um desvio de 7 dias.

4.2.3.2 Localidade temporal

Nessa seção é analisada a localidade temporal dos *tweets* sobre alagamentos e enchentes. Para cada dia é definido um *Event Index*, conforme descrito em 3.4.2.

O intuito de analisar o *Event Index* é verificar se, durante os dias que foram registrados pontos de alagamento, houve uma maior concentração das publicações se comparado com um período que não teve nenhum ponto de alagamento. Dessa forma é possível verificar se o *Event Index* é maior durante os períodos de alagamento.

O histograma do *Event Index* para cada dia com e sem ponto de alagamento é mostrado na Figura 4.26. Observe que para os dias com alagamento, Figura 4.26b, os valores de *Event Index* são maiores do que para dias sem, Figura 4.26a. Isso significa que nos dias com alagamento as mensagens chegam juntas provavelmente no horário crítico, ou seja, no horário em que houve maior concentração do número de pontos de alagamentos na cidade.



(a) (b)

Figura 4.26: Histograma do *Event Index* para o município de São Paulo em dias que não teve ponto de alagamento (a) e em dias que houveram pontos de alagamento (b).

O valor do *Event Index* será analisado em relação ao número de *tweets* sobre o evento em tempo real e o número de pontos de alagamento. Na Figura 4.27 essa comparação é feita. Quanto maior o número de *tweets*, maior o valor do *Event Index* e, na maioria dos dias sem pontos de alagamento, esse valor é inferior a 0.8, exceto nos dias da tragédia da região serrana do Rio de Janeiro que teve grande repercussão. Além disso, apesar do número de *tweets* sobre o momento atual ser pouco em alguns dias com pontos de alagamento, o valor desse índice foi alto na maioria desses dias. A maior parte dos dias sem alagamento, o valor do *Event Index* é menor do que nos dias com alagamento. Além disso, quanto mais pontos de alagamento maior o valor do *Event Index*.

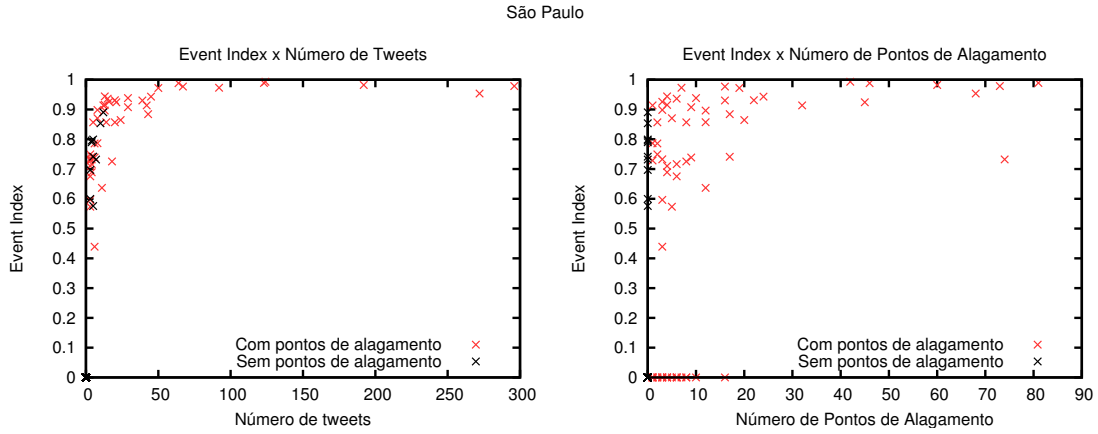


Figura 4.27: Gráficos do Event Index x Número de *tweets* do presente e Event Index x Número de pontos de alagamento para o município de São Paulo.

4.2.4 Prevendo pontos de alagamento

As mensagens publicadas no Twitter podem servir de instrumento para a previsão do número de pontos de alagamento. Nessa seção, os resultados da previsão do número de pontos de alagamento é mostrado e, além disso, a situação do município de São Paulo é classificada de acordo a ocorrência desses pontos.

4.2.4.1 Inferir a quantidade de pontos de alagamento

Para prever o número de pontos de alagamento, foi gerado um modelo de regressão linear. Esse modelo de regressão linear considera duas variáveis: t, o número diário de *tweets* classificados como sendo sobre o evento em tempo real, e o, o número diário de pontos de alagamento. Essas variáveis foram escolhidas em razão de terem gerado a melhor correlação como apresentado na seção 4.2.3.1.

Os resultados da regressão linear e da validação cruzada se encontram na Tabela 4.17. O valor de p é extremamente baixo e podemos concluir que o acaso para previsão dos valores é uma explicação pouco provável. O valor de R^2 para a validação cruzada com 10 partições é de 63%.

Tabela 4.17: Resultado da regressão linear. Na função de previsão, o é número de casos previstos e t é número de *tweets* sobre o evento em tempo real.

Cidade	*Função de previsão*	R^2	*p-value*	R^2 *da validação cruzada com 10-partições*
São Paulo	$o = 1,872301 + 0,29104 \times t$	0,63	$<$ 2,2e-16	0,61

O número de *tweets* sobre o evento em tempo real, o número de pontos de alagamento e o valor previsto do número de pontos de alagamento utilizando a função criada pela regressão linear ao longo do período são apresentados na Figura 4.28.

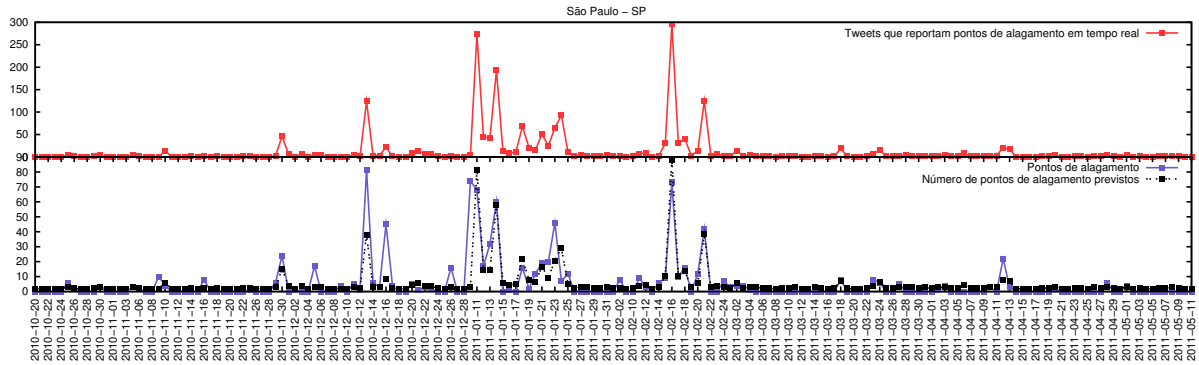


Figura 4.28: Número de *tweets* sobre o evento em tempo real, o número de pontos de alagamento e o valor previsto do número de pontos de alagamento utilizando a regressão linear.

4.2.4.2 Classificando a ocorrência de alagamentos

A regressão linear foi utilizada para criar uma função que infere o número de pontos de alagamento diariamente para a cidade de São Paulo a partir do número de *tweets* sobre o evento em tempo real. Nessa seção, o valor previsto será utilizado para classificar a situação do município.

A classificação da ocorrência dos alagamentos é feita considerando se houve ou não pontos de alagamento. São utilizadas duas classes, não ocorreu alagamento para os dias que não houve nenhum registro de ponto de alagamento e ocorreu alagamento para quando foi registrado um ou mais pontos de alagamento.

Dado o número diário de pontos de alagamentos previsto pela função de regressão linear (seção 4.2.4.1), é definido um limiar para determinar se houve ou não pontos de alagamento. Esse limiar foi variado utilizando valores entre o menor número de pontos de alagamento previsto (1,872301) e o maior (88,3112). A curva ROC gerada se encontra na Figura 4.29 e a área abaixo da curva foi de 0,8847.

Para realizar a classificação é necessário primeiro definir um valor para o limiar que irá determinar se houve ou não pontos de alagamento. Esse limiar deve considerar o compromisso entre a precisão e a revocação do classificador pois, se o classificador classificar sempre como verdadeiro, então tem revocação perfeita, mas baixa precisão.

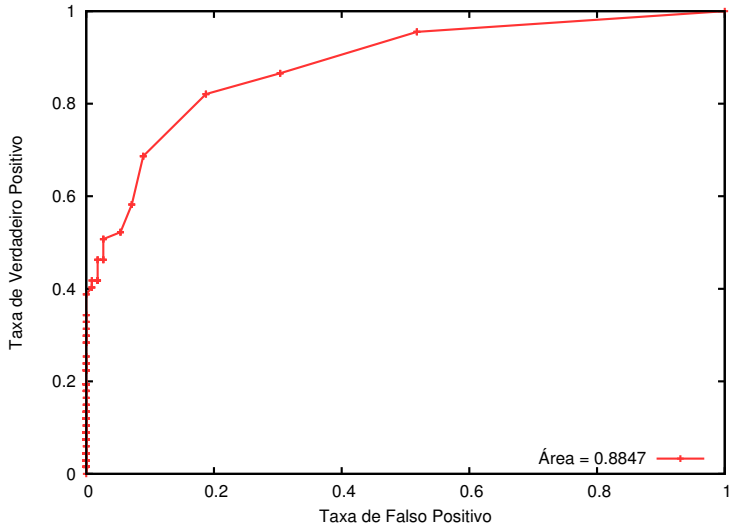


Figura 4.29: Curva ROC gerada variando o limiar do número de pontos de alagamento previsto para determinar a ocorrência de alagamentos.

Entretanto, se classificar como falso todos os exemplos, a precisão será perfeita, mas a revocação será baixa. Por essa razão é utilizado o limiar no qual a precisão e a revocação são iguais, esse ponto é chamado de *Break-even* (Liu [2009]).

A Figura 4.30 ilustra o valor da previsão e da revocação para os possíveis limiares. O valor no qual essas taxas mais se aproximam é aproximadamente 2,5 e esse será o valor do limiar. Dessa forma, caso o valor previsto para o número de pontos de alagamento for maior que 2,5, será definido que houve alagamento e, caso contrário, não houve alagamento.

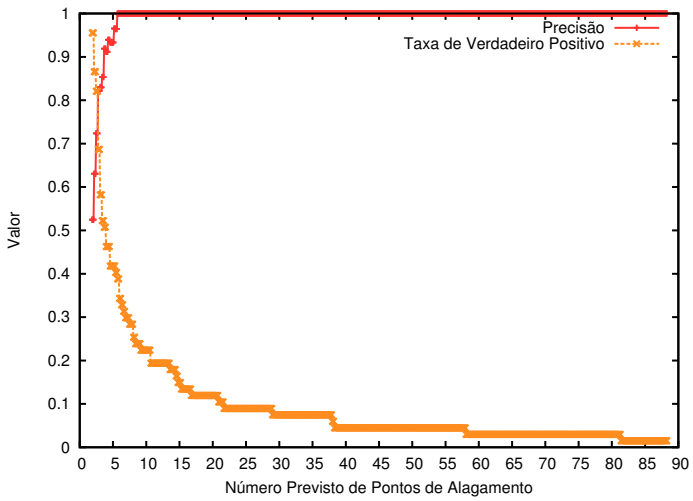


Figura 4.30: Valor da precisão e da recovação (taxa de verdadeiro positivo) para os possíveis limiares.

O resultado da classificação prevista foi comparado com a classificação obtida com os dados oficiais. Os números de dias que são classificados em cada uma das classes se encontram na Tabela 4.18.

Tabela 4.18: Quantidade de dias que são classificados em cada uma das classes.

Classe	*Pontos de Alagamento*	*Previsão*
Não houve alagamento	112	103
Houve alagamento	67	76

Na Tabela 4.19 é apresentado o resultado da classificação. A taxa de erro e a taxa de falso positivo foram aproximadamente 18%, 81% das previsões estavam corretas e 82% das previsões para os dias com pontos de alagamento foram corretamente identificadas.

Tabela 4.19: Resultado da classificação da situação do alagamento para o município de São Paulo.

Métrica	*Valor*
Taxa de Erro	0.1843
Taxa de Falso Positivo	0.1875
Taxa de Verdadeiro Positivo	0.8209
Acurácia	0.8156
Precisão	0.7236

4.2.5 Alerta para pontos de alagamento

O alerta reúne todas as métodos apresentados anteriormente: volume de *tweets* sobre os alagamentos em tempo real, número de pontos de alagamento, função de regressão linear. Além disso, agrega o conceito da tendência da ocorrência dos pontos de alagamento.

Dois indicadores são mostrados diariamente: a situação atual e a tendência dos alagamentos. A situação atual indica o que está ocorrendo no exato momento e será baseada no volume de *tweets* sobre os alagamentos em tempo real. A tendência indica se esse volume tem aumentado ou diminuído nas últimos dias.

A seguir vamos descrever como foi feito o cálculo da situação atual relativa dos alagamentos e da tendência relativa.

4.2.5.1 Avaliação da situação atual dos pontos de alagamento

A situação atual dos pontos de alagamento é calculada diariamente. No seu cálculo é utilizado o volume de *tweets* que descrevem o evento em tempo real e a função gerada pela regressão linear. A função de previsão gerada pela regressão linear é utilizada para gerar o número de alagamentos previsto a partir do volume de *tweets*.

Para visualizar a intensidade da situação atual dos alagamentos criamos uma escala de cores. Essa escala é gerada pela comparação do número de pontos de alagamento previsto com dois limiares. O limite inferior (LI) para o número de pontos de alagamento é 2,5, esse valor definido na seção 4.2.4.2 como limiar para a ocorrência de pontos de alagamento. O limite superior (LS) é 30, pois acima desse valor é considerada uma situação alarmante para a quantidade de pontos de alagamento. Quanto mais próximo do LS, mais próxima do vermelho será a cor representada na visualização. Caso exceda o LS, será utilizado o vermelho absoluto. Da mesma forma, caso seja menor que o LI será utilizado o branco absoluto.

A visualização da situação atual relativa dos pontos de alagamento no sistema de alerta pode ser conforme mostrada na Figura 4.31.

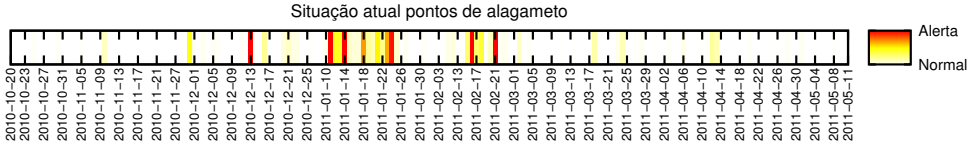


Figura 4.31: Visualização do sistema de alerta: Situação atual dos pontos de alagamento.

4.2.5.2 Avaliação da tendência

A tendência relativa dos alagamentos é calculada diariamente. No seu cálculo é utilizado o volume de *tweets* que descrevem o evento em tempo real de cada dia da semana atual e das duas semanas anteriores.

Utilizamos a fórmula do Z-score, apresentada na seção 3.6.2, para calcular a tendência. A média é calculada utilizando a média dos dois dias anteriores e o valor atual é o valor do dia.

Para visualizar a intensidade da tendência de dengue criamos uma escala de cores. Essa escala é gerada pela comparação do valor do Z-score com dois limites pré-definidos. O limite inferior (LI) é o valor -1, ou seja, houve a diminuição de 1 desvio padrão em relação às duas semanas anteriores. O limite superior (LS) é o valor 2, ou seja, houve o

aumento de 2 desvios padrões em relação às semanas anteriores. Quanto mais próximo do LS, mais próxima do vermelho será a cor representada na visualização. Caso exceda o LS, será utilizado o vermelho absoluto. Da mesma forma, caso seja menor que o LI será utilizado o branco absoluto.

O valor 0 significa que não houve alteração da semana atual para as duas anteriores e para representá-lo será utilizada a cor amarela. O valor de Z-score igual a 0 é o limite intermediário (LM). Para valores entre LI e LM será utilizada uma escala em degradê variando do branco ao amarelo sobre a porcentagem entre o valor mínimo e o valor médio. E para valores entre LM e LS será utilizada um degradê variando do amarelo passando pelo laranja até o vermelho.

A visualização da tendência relativa dos alagamentos no alerta pode ser conforme mostrada na Figura 4.32.

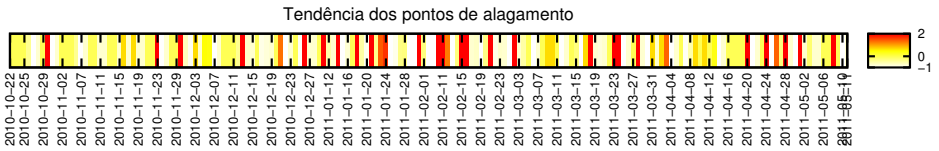


Figura 4.32: Visualização do sistema de alerta: Tendência dos pontos de alagamento.

Capítulo 5

Conclusões e Trabalhos Futuros

Nesta dissertação foi apresentada uma metodologia para detecção e previsão de eventos reais utilizando a mineração de redes sociais *online*. O processo inicia-se com a coleta das mensagens publicadas sobre o evento de interesse no Twitter. A partir dos *tweets* coletados, as informações são extraídas e é determinada a localização geográfica do usuário. O texto de cada mensagem é classificado pelo seu conteúdo utilizando o algoritmo LAC (Classificação Associativa sob Demanda) com o intuito de selecionar apenas as mensagens relevantes para detecção e previsão do evento.

Para verificar a viabilidade do uso das redes sociais como instrumento para previsão dos eventos reais, é realizada a análise de correlação entre as mensagens classificadas pelo conteúdo e os dados oficiais sobre o evento. São propostas três análises de correlação. A primeira utiliza a correlação cruzada para observar se há um atraso da repercussão do evento no Twitter. A segunda é a análise da similaridade espacial, na qual os locais próximos e com níveis similares de ocorrências do evento são agrupados, sendo que os grupos formados com os dados das redes sociais são correlacionados com os agrupamentos dos dados oficiais. Por fim, é feita uma análise que utiliza o intervalo de tempo entre a chegada das mensagens para verificar se em períodos críticos para o evento há uma maior tendência de chegar mais mensagens em um período menor de tempo.

Verificada a correlação entre os dados virtuais e os dados reais, as redes sociais podem ser consideradas insumo para previsão dos eventos reais. Primeiramente é feita a previsão do volume de ocorrências do evento por meio de uma função de regressão linear gerada para cada região. E, a partir do número previsto, classifica-se a situação de gravidade da região.

A última etapa da metodologia é a elaboração do sistema de alerta sobre o evento. Esse sistema propõe a visualização dos dados previstos para a situação atual do evento

e para a tendência do evento.

A metodologia proposta foi aplicada a dois tipos de eventos reais: epidemia de dengue e enchentes. No caso da epidemia de dengue, observa-se que o classificador de conteúdo para a classe experiência pessoal classificou corretamente 93% das mensagens. Ao utilizar essas mensagens obteve-se uma alta correlação (74%) entre mensagens expressando experiência pessoal e a incidência da doença, sendo que cidades como Rio de Janeiro e Manaus apresentaram correlação de 98% e 95% respectivamente. A similaridade espacial média foi 78%. Na previsão do volume de casos da dengue utilizando a função de regressão linear, apenas 20% dos municípios apresentaram correlação menor que 40% e metade dos municípios possuem correlação superior a 60%. As cidades com maior correlação foram Rio de Janeiro (95%), Ribeirão Preto (89%) e Manaus (86%). Na previsão da gravidade da situação da doença, quase todas as cidades com semanas classificadas como baixa incidência tiveram acurácia e previsão acima de 90%. Para média incidência a acurácia em apenas 20% das cidades foi menor que 100% e para alta incidência, a taxa de verdadeiro positivo maior que 90%.

Para o segundo evento, alagamentos e enchentes, observa-se que o classificador do conteúdo acertou em 75% das mensagens. Essas mensagens foram utilizadas na correlação com o número de pontos de alagamento e a correlação foi de 79%. O *Event Index* apresentou maiores valores nos dias em que houve pontos de alagamento, o que comprova a localidade de referência temporal do tempo de chegada das mensagens. A previsão dos pontos de alagamento gerada pela função de regressão linear teve uma correlação de 61% com os dados oficiais. A previsão para a gravidade da situação foi correta em 81% dos dias.

É importante ressaltar que esses resultados demonstram a aplicabilidade dessa proposta como complemento a mecanismos de vigilância tradicional, muitas vezes permitindo que ações sejam antecipadas e impactos sobre a população afetada sejam reduzidos.

A metodologia proposta nessa dissertação assim como os resultados obtidos no contexto da dengue são utilizadas no Observatório da Dengue com propósito de acompanhar o que é dito pelos usuários das redes sociais para prever possíveis casos da doença e alertar sobre sua situação em cada cidade brasileira. Recentemente, o Observatório da Dengue firmou uma parceria com o Ministério da Saúde com intuito de utilizar essa ferramenta como um sistema complementar ao sistema de vigilância tradicional. O alerta desenvolvido nessa dissertação é disponibilizado por meio de uma página web de acesso restrito que contém a avaliação da situação atual da incidência e da tendência da doença.

Considera-se como principais contribuições do trabalho a proposição de uma

metodologia que realiza a detecção e previsão de eventos de impacto utilizando como insumo as mensagens postadas nas redes sociais, bem como a metodologia para classificar o conteúdo das mensagens postadas e a elaboração do sistema de alerta que disponibiliza as informações de forma visual (Gomide et al. [2011], Silva et al. [2011]). Além disso, não é do nosso conhecimento a realização de outros experimentos com dados da Dengue e dos alagamentos no Brasil.

A relevância dessa pesquisa é demonstrada pela contribuição da metodologia proposta e dos seus resultados experimentais obtidos. Entretanto, é preciso lembrar que as análises realizadas possuem algumas limitações. Primeiramente, o banco de dados do número de casos da dengue disponibilizado pelo Ministério da Saúde não contém todos os casos de dengue que ocorreram no Brasil, mas apenas os casos notificados pelos médicos e reportados pelo governo. Outra limitação é que muitas mensagens postadas são descartadas devido à falta de informação sobre a localização do usuário e, como consequência, algumas cidades não são analisadas pelo fato de haver poucos ou nenhum *tweet*. Além disso, a localização do usuário não é obtida pela mensagem postada e sim pelo usuário que a postou. Dessa forma, a análise que foi realizada ignora a mobilidade do usuário. Finalmente, a faixa etária dos usuários das redes sociais é em sua maioria de 18 a 35 anos e não reflete toda a população atingida em ambos os cenários utilizados nos experimentos.

Como trabalhos futuros, há várias frentes de continuidade dessa pesquisa. A primeira é melhorar a qualidade da análise de sentimento, por exemplo, por meio da utilização de algoritmos que considerem os relacionamentos entre os usuários. Além disso, é possível melhorar a análise de conteúdo por meio da atualização constante do conjunto de treino para incorporar novos termos que estão sendo utilizados para referenciar o evento. A segunda melhoria do trabalho é analisar a viabilidade de utilizar outras fontes de informações *online* disponíveis tais como *blogs*, notícias e o *Google Insights*. Outra possível melhoria é considerar um nível mais detalhado de localização do usuário, por exemplo, as regiões dos municípios que podem ser obtidas por meio das coordenadas de GPS disponibilizadas apenas pelos usuários que postam por *Smartphone*. Considera-se, inclusive, que essas possíveis novas fontes de informação possam ser agregadas com as redes sociais criando um novo modelo de detecção e previsão de eventos. Contudo, pretende-se associar informações sobre perfil demográfico do uso da internet na criação dos modelos de previsão, assim como associar a penetração das redes sociais em cada região considerando a idade dos usuários para tornar o modelo de previsão mais refinado.

Referências Bibliográficas

Achrekar, H.; Gandhe, A.; Lazarus, R.; Ssu-Hsin Yu & Liu, B. (2011). Predicting flu trends using Twitter data. Em *IEEE INFOCOM 2011 - IEEE Conference on Computer Communications Workshops*, pp. 702--707. IEEE.

Althouse, B. M.; Ng, Y. Y. & Cummings, D. A. T. (2011). Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis*, 5(8):e1258.

Asur, S. & Huberman, B. A. (2010). Predicting the future with social media. Em *Proceding of IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 492–499. IEEE.

Benevenuto, F.; Rodrigues, T.; Cha, M. & Almeida, V. (2009). Characterizing user behavior in online social networks. Em *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pp. 49--62, New York, NY, USA. ACM.

Birant, D. & Kut, A. (2007). St-dbscan: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*, 60:208--221.

Brourke, P. (1996). Cross Correlation. `http://paulbourke.net/miscellaneous/correlate/`.

Brownstein, J. S.; Freifeld, C. C.; Reis, B. Y. & Mandl, K. D. (2008). Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Med*, 5(7):e151.

CDC (2012). Centers for Disease Control. `http://www.cdc.gov/dengue/`.

Cha, M.; Haddadi, H.; Benevenuto, F. & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. Em *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA.

Chan, E. H.; Sahai, V.; Conrad, C. & Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis*, 5(5):e1206.

Chen, L.; Achrekar, H.; Liu, B. & Lazarus, R. (2010). Vision: towards real time epidemic vigilance through online social networks. Em *ACM Workshop on Mobile Cloud Computing Services: Social Networks and Beyond*, pp. 1--5. ACM.

Chew, C. & Eysenbach, G. (2010). Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11):e14118.

comScore (2010). Orkut continua liderando o mercado de redes sociais no brasil, e a audiencia do facebook quintuplica. `http://tinyurl.com/346u9na`.

Corley, C.; Mikler, A. R.; Singh, K. P. & Cook, D. J. (2009). Monitoring influenza trends through mining social media. Em *Procedings of International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, pp. 340–346. CSREA Press.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. Em *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pp. 115--122, New York, NY, USA. ACM.

Ester, M.; Kriegel, H.-P.; Sander, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Em *International Conference on Knowledge Discovery and Data Mining*, pp. 226–231. AAAI Press.

Eysenbach, G. (2006). Infodemiology:tracking flu-related searches on the web for syndromic surveillance. Em *AMIA Annu Symp Proc.*, pp. 244--248.

Freifeld, C. C.; Mandl, K. D.; Reis, B. Y. & Brownstein, J. S. (2008). Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association (JAMIA)*, 15(2):150–157.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012--4.

Goel, S.; Hofman, J. M.; Lahaie, S.; Pennock, D. M. & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486--17490.

Gomide, J.; Veloso, A.; Jr., W. M.; Almeida, V.; Benevenuto, F.; Ferraz, F. & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. Em *ACM SIGWEB Web Science Conference (WebSci)*.

Guerra, P. H. C.; Veloso, A.; Meira, Jr, W. & Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. Em *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Diego, CA.

Kosala, R. & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations*, 2(1):1–15.

Kumar, R.; Raghavan, P.; Rajagopalan, S. & Tomkins, A. (1999). Trawling the web for emerging cyber-communities. Em *Proceedings of the eighth international conference on World Wide Web*, WWW '99, pp. 1481--1493, New York, NY, USA. Elsevier North-Holland, Inc.

Lampos, V. & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. Em *2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*, pp. 411--416. IEEE Press.

Lampos, V. & Cristianini, N. (2011). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

Lampos, V.; De Bie, T. & Cristianini, N. (2010). Flu detector - tracking epidemics on twitter. *Machine Learning and Knowledge*, 6323:599--602.

Larsen, R. & Marx, M. (1986). *An introduction to mathematical statistics and its applications*. Prentice-Hall.

Liu, B. (2009). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 1st ed. 2007. corr. 2nd printing edição.

Liu, L. & Özsu, M. T., editores (2009). *Encyclopedia of Database Systems*. Springer US.

Mawudeku, A. & Blench, M. (2006). Global public health intelligence network (gphin). Em *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*.

Mendoza, M.; Poblete, B. & Castillo, C. (2010). Twitter under crisis: Can we trust what we rt? Em *1st Workshop on Social Media Analytics (SOMA '10)*. ACM Press.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846--850.

Ritterman, J.; Osborne, M. & Klein, E. (2009). Using prediction markets and Twitter to predict a swine flu pandemic. Em *Proceedings of the 1st International Workshop on Mining Social Media*.

Runge-Ranzinger, S.; Horstick, O.; Marx, M. & Kroeger, A. (2008). What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? *Tropical Medicine International Health*, 13(8):1022--1041.

Sakaki, T.; Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. Em *Proceedings of the 19th international conference on World wide web*, WWW '10, pp. 851--860, New York, NY, USA. ACM.

Schafer, J. B.; Konstan, J. A. & Riedl, J. (2001). E-commerce recommendation applications. *Data Min. Knowl. Discov.*, 5(1-2):115--153.

Silva, I. S.; Gomide, J.; Barbosa, G.; Veloso, A.; Santos, W.; Ferreira, R. & Jr., W. M. (2011). Observatório da dengue: Surveillance based on twitter sentiment stream analysis. Em *Simpósio Brasileiro de Banco de Dados (SBBD)*.

Starbird, K. & Palen, L. (2010). Pass it on?: Retweeting in mass emergencies. Em *Information Systems for Crisis Response and Management Conference*, Seattle, WA, USA.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267--288.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G. & Welpe, I. M. (2010). Predicting elections with twitter : What 140 characters reveal about political sentiment. *Word Journal Of The International Linguistic Association*, pp. 178--185.

Veloso, A.; Meira Jr., W. & Zaki, M. J. (2006). Lazy associative classification. Em *International Conference on Data Mining*, pp. 645--654. IEEE Computer Society.

Vieweg, S.; Hughes, A. L.; Starbird, K. & Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness.

Em *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pp. 1079--1088, New York, NY, USA. ACM.

Wang, G. & Jain, C. (2003). *Regression analysis: modeling & forecasting*. Graceway Pub.

Weiss, R.; Velez, B.; Sheldon, M. A.; Namprempre, C.; Szilagyi, P.; Duda, A. & Gifford, D. K. (1996). Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. Em *Proceedings of the Seventh ACM Conference on Hypertext*, pp. 180--193.

WHO (2012). World Health Organization. `http://www.who.int/tdr/diseases/default.htm`.

Winerman, L. (2009). Crisis Communication. *Nature*, 457:376--378.

Zaki, M. & Meira Jr., W. (2012). *Fundamentals of Data Mining Algorithms*. Cambridge University Press.