colloquium exactarum

MONITORAMENTO PARA PREVER EPIDEMIAS UTILIZANDO REDES SOCIAIS

MONITORING TO PREVENT EPIDEMICS USING SOCIAL NETWORKS

Bruno Neves Santos[1], Daniela Tereza Ascencio Russi[2], Francisco Assis da Silva[2], Mário Augusto Pazoti[2], Robson Augusto Siscoutto[2]

[1]Discente da Faculdade de Informática da UNOESTE. [2]Docente da Faculdade de Informática da UNOESTE

RESUMO - Atualmente as redes sociais têm atraído milhões de usuários, pois são ambientes favoráveis para o estudo de vários temas da computação (sistemas distribuídos, padrões de tráfego na internet, mineração de dados, sistemas de multimídias, entre outros), voltados a diversas áreas (epidemiologia, política, esportes, marketing). O conteúdo criado em redes sociais tornou-se um tema chave em pesquisas relacionadas ao tratamento e organização de grande quantidade de dados. Neste contexto, este trabalho propõe uma aplicação que é capaz de coletar dados compartilhados na rede social Twitter, processar esses dados e gerar informações de modo que obtenha resultados que possibilitam identificar prováveis concentrações de casos suspeitos de certa doença em determinado local e período, auxiliando a epidemiologia.
Palavras-chave: redes sociais; twitter; epidemiologia; mineração de dados.

ABSTRACT - Actually social networks have been attracting millions of users, because they are favorable environments for studying many computing themes (distributed systems, traffic patterns over the internet, data mining, multimedia systems, among others), directed to different areas (epidemiology, politics, sports, marketing). The concept created over the social networks became a key theme for researches related to processing and organization of huge amount of data. On this context, this research proposes an application that is able to collect data shared over the social network named Twitter, process this data and generate information in order to obtain results that permit identify probable concentrations of a certain disease on a determined place and period, assisting the epidemiology.
Keywords: social networks; twitter; epidemiology; data mining.

Recebido em: 21/08/2013
Revisado em: 27/09/2013
Aprovado em: 29/10/2013

1 INTRODUÇÃO

O monitoramento da difusão de uma epidemia em uma população é uma tarefa importante. Em geral, os dados recolhidos da população podem fornecer informações valiosas para autoridades de saúde sobre a localização, tempo e intensidade de uma epidemia, ou mesmo alertar as autoridades da existência de uma ameaça à saúde. No entanto, este procedimento é difícil e exige muitos recursos (LEITE, 2010).

Diversos métodos podem ser utilizados para estimar o número real de pessoas afetadas por uma determinada doença, mas a maioria desses métodos exige um investimento, infraestrutura e apresenta vários inconvenientes, tais como o atraso devido à agregação de informação e tempo de processamento (LAMPOS; CRISTIANINI, 2010).

Ginsberg et al. (2008) relatou o uso de dados de motores de busca para detectar agrupamentos geográficos com uma proporção elevada de consultas relacionadas com a saúde, particularmente no caso de "Gripe", demonstrando que informações sobre a difusão de uma doença, podem ser obtidas examinando o conteúdo de consultas de motores de busca.

A Web conta atualmente com um grande repositório de informações existente no mundo. Pessoas interagem todos os dias com uma enorme quantidade de dados e se perdem entre conteúdos diversos, sempre buscando encontrar o que realmente querem. A dificuldade está exatamente nesse ponto (como filtrar essas informações, e como recuperar apenas o conteúdo desejado). Existem diversas formas para dispor informações na Web: redes sociais, blogs, fóruns, entre outras. Todas as formas contendo suas especificidades, necessitando de abordagens distintas para a análise de seus conteúdos.

A rede social Twitter é um serviço que possibilita o compartilhamento de mensagens de até 140 caracteres e seus usuários podem seguir outros usuários e visualizar em sua página inicial as atualizações (mensagens compartilhadas) dessas pessoas na ordem que forem postadas. Em diversos trabalhos, já se identificou a importância de avaliar o sentimento das mensagens compartilhadas na rede social Twitter (ARAÚJO; GONÇALVES; BENEVENUTO, 2013; GONCALVES et al., 2013).

Identificar o sentido e classificar a emoção de um texto passou a ser foco em diversas pesquisas, sendo esta área de estudo conhecida como análise do sentimento. No tratamento de texto, são apresentados novos desafios, um deles é o fato de que o foco em extrair a opinião

expressa não é apenas descobrir sobre qual assunto se trata. Neste contexto, o Twitter tem se mostrado uma ótima plataforma, pois permite que usuários enviem mensagens curtas e frequentemente. Porém, novos desafios surgiram, por exemplo, o vocabulário utilizado, que é de caráter extremamente inconsistente e informal, o que dificulta o tratamento dos textos coletados é o fato de que estas características permitem que abreviações e variações na escrita de uma palavra representem um mesmo significado (JIANG et al., 2011).

A relevância desse trabalho se destina a identificação de problemas epidemiológicos em um curto espaço de tempo. Com a identificação prévia de pequenas concentrações de focos de certa doença em uma determinada região, acredita-se que seja possível combater a proliferação dessa doença atacando os locais onde estão ocorrendo muitos focos. Dessa forma, isso tende a melhorar a eficiência do combate a doença e uma melhor utilização e aproveitamento dos recursos utilizados nesses combates, sendo possível controlar esses focos e evitar epidemias.

As demais seções deste trabalho estão organizadas da seguinte maneira: na Seção 2 é descrita suscintamente redes sociais; na Seção 3 é apresentado o Twitter que é uma rede social gratuita, onde seus usuários enviam e recebem pequenos textos; na Seção 4 é apresentado o classificador Bayesiano; na Seção 5 são apresentados os conceitos Web Mining que se refere à extração ou mineração do conhecimento através de grandes quantidades de dados e Text Mining que tem como principal objetivo analisar o conhecimento implícito de grandes quantidades de textos escritos em linguagem natural; na Seção 6 é apresentada a análise do sentimento que é uma área da mineração de textos muito utilizada, e que lida com as opiniões expressas em textos; na Seção 7 é apresentada a metodologia desenvolvida para realização do trabalho; por fim, na Seção 8 são apresentadas as conclusões e considerações finais do trabalho.

2 REDES SOCIAIS

A Web vem experimentando uma nova onda de aplicações associada à proliferação das redes sociais e o crescimento de sua popularidade. Várias redes sociais surgiram, incluindo redes profissionais (Linkedin), redes de amigos (Myspace, Facebook, Orkut), e redes para o compartilhamento de conteúdos específicos, tais como mensagens curtas (Twitter), diários e blogs, fotos (Flickr) e vídeos (Youtube) (BENEVENUTO; ALMEIDA; SILVA, 2011).

Segundo Benevenuto, Almeida e Silva (2011) redes sociais *on-line* têm atraído milhões de usuários. O conteúdo criado e disseminado via interações sociais, passou na frente do e-mail como a atividade *on-line* mais popular. Tanta popularidade está associada a uma funcionalidade comum de todas as redes sociais *on-line* que permite que usuários criem e compartilhem conteúdo nesses ambientes. O conteúdo das redes sociais pode variar de simples mensagens de texto comunicando eventos do dia-a-dia até mesmo a conteúdo multimídia, como fotos e vídeos.

Apesar de tanta popularidade e da enorme quantidade de conteúdo disponível, o estudo de redes sociais ainda está em seu início, já que esses ambientes estão experimentando novas tendências e enfrentando novos problemas e desafios.

3 TWITTER

O Twitter é uma rede social gratuita onde os usuários enviam e recebem pequenos textos. Cada usuário tem uma lista de seus seguidores e outra lista de usuários que ele segue. Mensagens de texto de até 140 caracteres (os "tweets") são escritos e enviados (ou "twitados") para que todas as pessoas que seguem o usuário recebam essa mensagem. Simultaneamente, o usuário está constantemente recebendo atualizações na lista de mensagens dos usuários que ele segue enviam.

Diversos aplicativos foram desenvolvidos para interagir com o Twitter, entre eles estão os complementos para navegadores de internet, os softwares embarcados em celulares, e integração com jogos de videogame. O usuário também pode utilizar um simples SMS para "twittar", entretanto esse serviço pode ser cobrado pela operadora telefônica. A explicação para o limite de 140 caracteres do Twitter vem justamente do fato de que uma mensagem SMS contém no máximo esse número de caracteres e o Twitter foi inicialmente desenvolvido com essa ideia.

Atualmente, o Twitter ainda não utilizou propagandas ou anúncios no site, o que poderia ser uma tendência natural de um site gratuito, com tantos usuários. Entretanto, existem empresas que patrocinam pessoas influentes (com muitos seguidores) para "twitarem" mensagens em favor dessas empresas e de seus produtos.

Segundo Cheng, Evans e Singh (2011), o Twitter possui uma grande quantidade de usuários que são usuários assíduos e postam várias mensagens por dia. São trocadas cerca de três milhões de mensagens por dia. Em maio de 2011, o Twitter alcançou 300 milhões de usuários e esse número vem crescendo, embora seja em ritmo menor ao

alcançado em 2009, quando ocorreu a maior taxa de crescimento.

4 CLASSIFICADOR BAYESIANO

Os classificadores Bayesianos são capazes de predizer a probabilidade de um determinado registro fazer parte de uma classe. Esses classificadores disponibilizam uma poderosa técnica de classificação supervisionada, considerando que todos os atributos de entrada sejam independentes entre si e tenham a mesma importância (FONSECA, 2007).

A análise de dados bayesiana define modelos práticos para realizar inferências a partir de dados utilizando modelos de probabilidade para valores observados e valores que não foram observados, que se deseja aprender. A característica essencial deste tipo de análise de dados é a utilização da probabilidade para quantificar a incerteza na inferência (HRUSCHKA JÚNIOR, 2003).

Os classificadores Bayesianos podem ser encontrados em muitos trabalhos na literatura, para diversos propósitos, como no trabalho de Curotto (2003), que trabalhou com mineração de dados utilizando um classificador bayesiano integrado em um sistema de banco de dados. Ainda utilizando banco de dados, em um trabalho posterior (CUROTTO, 2006) o autor realizou a classificação bayesiana em dados multirelacional nos bancos SQL SERVER 2000 e SQL SERVER 2005, utilizando para isso uma técnica de desnormalização. Em Ceci, Appice e Malerba (2003) encontra-se um trabalho que trata da integração de um sistema de mineração de dados multirelacional a um sistema de banco de dados, baseado na classificação bayesiana.

Em Fonseca (2007) encontra-se uma explanação sobre o teorema de Bayes, que pode ser traduzido por Witten (2005), como:

$$P(H|E) = P(E|H).\ P(H) / P(E) \qquad (1)$$

onde *H* é a hipótese a ser testada e *E* é a evidência associada com a hipótese.

Da visão de classificação a hipótese é a variável dependente e representa a classe predita. A evidência é determinada pelos valores dos atributos de entrada. P(E|H) é a probabilidade condicional associada com *H*, a evidência verdadeira fornecida. P(H) é a probabilidade a priori, que denota a probabilidade da hipótese antes da apresentação de qualquer evidência.

O teorema de Bayes mostra uma maneira de calcular a probabilidade a posteriori a partir das probabilidades a priori, sendo que a probabilidade condicional e a priori são facilmente computadas a partir dos dados de treinamento.

5 *WEB MINING* E *TEXT MINING*

Atualmente, ninguém questiona o valor da Web como fonte de informações. Em todo o mundo tem-se uma grande quantidade de páginas espalhadas, contendo os mais variados tipos de informação e em várias línguas, disponíveis em dados não estruturados, como imagens, sons, vídeos, textos, etc. Sabendo-se o quão custoso é realizar a tarefa de se extrair essas informações através de meios não computacionais, foi desenvolvido um conceito derivado da expressão *Data Minging*, o conceito de *Web Mining*.

Segundo Han e Kamber (2006), *Data Mining* (ou mineração de dados), se refere à extração ou mineração do conhecimento através de grandes quantidades de dados. Uma definição semelhante seria que a mineração de dados designa uma área de trabalho e investigação, pertencente à Inteligência Artificial, que tem como objetivo a descoberta de conhecimento, de estrutura e relações dos conteúdos das Bases de Dados (CORDEIRO, 2004). No contexto *World Wide Web*, utiliza-se o conceito *Web Mining* (ou mineração da Web).

A mineração na Web possui algumas subdivisões que são mostradas na Figura 1 (PASSOS; ARANHA, 2006):

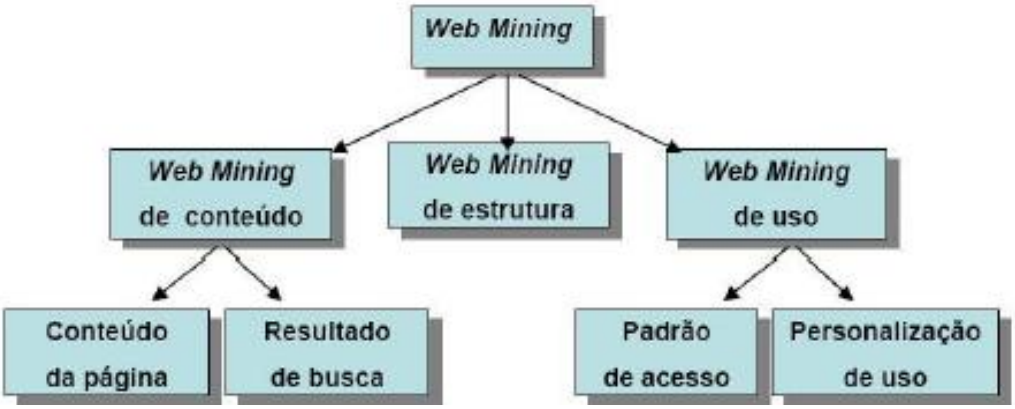

Figura 1. Mineração na Web.

Fonte: (PASSOS; ARANHA, 2006).

Segundo Passos e Aranha (2006) *Web Mining* de conteúdo (ou mineração do conteúdo da Web) trabalha com o conteúdo das páginas, extraem informações através da análise dos textos, imagens, etc. Nesta área, também se localizam os motores de busca, utilizados para encontrar páginas que o usuário deseja através de critérios. *Web Minging* de uso (ou mineração do uso da Web) trabalha com a forma como os usuários

navegam, investigando a conduta dessas pessoas na Web. É examinado a sequência de navegação de uma pessoa, por exemplo, para examinar qual parte de um site precisam ser representadas do ponto de vista da tendência dos seus *links*, uma vez que se examine que certa informação não é muito acessada. Diversas aplicações podem ser estimadas usando a abordagem de *Web Usage Mining*, como a sugestão de produtos que uma pessoa possa gostar em um site de vendas, de acordo com a análise de sua navegação pelo site da empresa. *Web Mining* de estrutura (ou mineração da estrutura da Web) analisa a estrutura dos *hiperlinks* presentes nas páginas Web, examinando o relacionamento das mesmas. Verificam-se características como: qual página é a que contém mais ligações com outras? Qual é a mais referenciada em certo contexto? Por fim, tem-se a *Text Mining* ou mineração de textos, que tem como principal objetivo analisar conhecimento implícito de grandes quantidades de textos escritos em linguagem natural. A definição do *Text Mining* é similar àquela dada à mineração de dados, sendo a diferença o modo usado para minerar: na *Data Mining*, a mineração é feita em uma base de dados.

A quantidade de documentos em um trabalho de mineração de textos é chamada de corpus. Em geral, o corpus analisado é composto por uma quantidade de documentos, que são primeiramente coletados usando algum método definido pelo usuário. Neste trabalho utiliza-se o conteúdo da página que pode ser visualizado no nó à esquerda da árvore exibida na Figura 1, na subárea da *Web Mining* conhecida como *Web Content Mining*, visto que será examinado o conteúdo presente na Web, mais especificamente, textos presentes na rede social Twitter.

6 ANÁLISE DO SENTIMENTO

Uma área da mineração de textos muito utilizada é a que lida com as opiniões expressas em textos. Pode-se analisar o que está sendo expresso por alguém sobre determinado filme, livro, ou qualquer outro tópico através de análises computacionais, tal tratamento é conhecido como análise de sentimento.

Há algum tempo, pessoas procuram as opiniões de outras como modo de direcionarem suas compras e suas considerações sobre um assunto. Com a chegada da Web como fonte de informações, grande parte dos usuários tem procurado nela textos que disponibilizem esse tipo de informação desejada (opiniões sobre algo). Pang e Lee (2008) relataram que a maioria das pessoas que realizam pesquisa sobre produtos, conseguem informações

importantes e decisivas na hora de optar pela compra dos mesmos.

Entende-se que o número de textos expostos na Web é incomensurável e que na maioria das vezes não se acha o que deseja. Tendo em vista a crescente procura por métodos que ajudem na análise de grandes quantidades de texto, a necessidade de softwares que façam o estudo do sentimento tem aumentado, porém, a construção de uma aplicação que realize a análise de sentimento de maneira correta e rápida não é nem um pouco trivial, e várias pesquisas na área ainda encontram-se em andamento.

O número de dificuldades existentes na hora de fazer um processo de análise de sentimento é grande. Considerando alguns aspectos: fazer com que o computador tenha capacidade de interpretar um determinado documento, e a emoção contida nele. O trabalho não é fácil nem para seres humanos, sendo que diferentes visões e opiniões influenciam no modo como cada um interpreta um texto, como por exemplo, a frase: (essa gripe ainda vai me matar) esta é uma frase que pode ter interpretações diferentes.

Existem algumas dificuldades para se realizar a interpretação, tais como, a necessidade de filtrar textos que possuem conteúdo opinativo daqueles que são objetivos. A descrição dos resultados encontrados também forma um problema, caso se deseja informações extras além da quantidade que foi classificada como pertencente a uma classe de opiniões positivas ou negativas (tarefa dentro da mineração de opinião conhecida como *sentiment polarity*). Pode-se desejar visualizar partes exibindo onde estão ocorrendo divergências de opinião entre os usuários e/ou partes mostrando as partes chave do sentimento do usuário a cerca do assunto sendo analisado.

Um tipo de aplicação importante é a utilização na área dos negócios. Uma empresa quer visualizar feedback sobre seus serviços/produtos, ter conhecimento de como está o reconhecimento de um lançamento, analisar opiniões relacionadas à concorrência, entre outros objetivos. Através dos dados encontrados, podem-se realizar novas estratégias de marketing, propor inovações nos produtos ou processos da empresa, etc.

7 METODOLOGIA APLICADA

Neste trabalho, a ferramenta desenvolvida para coletar e armazenar mensagens compartilhada no Twitter foi implementada no Visual Studio 2010, usando a linguagem de programação C#. Para o desenvolvimento foi utilizado a Twitter API, que é um serviço proporcionado pelo Twitter

que permite aos desenvolvedores acessarem dados e funções do Twitter diretamente via cliente *Desktop*, via dispositivo móvel ou mesmo *on-line*. Com a biblioteca Twtterizer para a realização da coleta das mensagens, foram definidos parâmetros de pesquisa como: descrição que significa o assunto que foi abordado; data inicial e data final que informa o período da pesquisa; localidade onde é definido um ponto de pesquisa e o raio para se encontrarem quais locais serão atingidos; e palavras chave pelas quais a ferramenta vai procurar dentro das mensagens.

Para realizar a busca por mensagens do Twitter, foi construída uma consulta, sendo passados os parâmetros da pesquisa desejada. Os parâmetros utilizados na consulta foram: “Gripe”, “Dengue” e “Vomito”. Sendo assim a consulta ficou da seguinte maneira: “Select Gripe OR Febre OR Vomito”. Após isso, foram definidos os parâmetros de localidade e período da pesquisa, que foram: “Localidade: Presidente Prudente”, “Raio: 50 km”, “Data inicial: 11/11/2013”, “Data Final: 17/11/2012”. A pesquisa retornou todas as mensagens compartilhadas no Twitter que possuíam um ou mais termos passados por parâmetros. As mensagens foram postadas em uma data dentro do período estipulado e a localidade do usuário que se encontra dentro de um raio de 50 km da cidade de Presidente Prudente. Na Tabela 1 encontra-se uma amostra de como a API do Twitter retornou essa pesquisa. Por motivo de privacidade o ID e Login dos usuários não foram exibidos.

Tabela 1. Dados recebidos pela API Twitter.

ID Mensagem	Idioma:	Localização	Data:	Mensagem:
1	pt	Prudente	11 nov 2012 14:15:00	Essa gripe não vai passa nunca?
2	pt	P. Pte	12 nov 2012 20:15:00	Acho que estou com dengue.
3	pt	Indiana	15 nov 2012 10:10:52	To bem, só gripe e febre pra ajuda.

A Figura 2 exibe o modelo de dados utilizados para armazenar as mensagens coletadas.

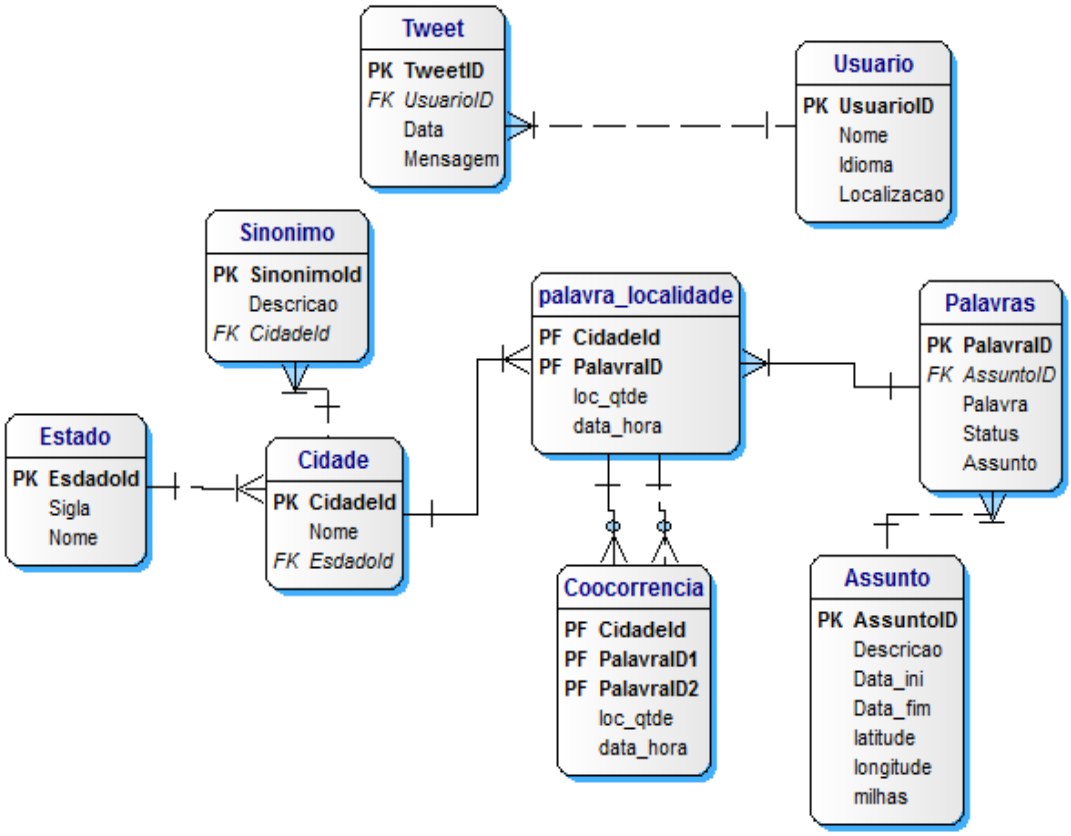

Figura 2. Modelo da Base de Dados.

Os dados foram distribuídos e armazenados nas tabelas do banco de dados da seguinte maneira:

- **Usuário**: dados referentes ao usuário que postou a mensagem, armazenando o Id do usuário, nome, idioma e sua localização;
- **Tweet**: dados referente às mensagens coletadas, Id da mensagem, usuário responsável, Data de envio e conteúdo da mensagem;
- **Sinônimos** dados referentes à localidade do usuário, que muitas vezes estavam de forma abreviada, sendo assim feito uma referência de qual cidade cada sinônimo pertencia, foram tratados como o exemplo da cidade de Presidente Prudente. Para referenciar a cidade podem ser usados: “P. Pte”, “Pres. Prudente” e “P.P”;
- **Cidade** e **Estado:** dados de estado e cidades que foram utilizados para identificar qual cidade um sinônimo se referenciava;
- **Assunto** e **Palavras:** dados utilizados como parâmetros nas buscas realizadas;
- **Palavras:** dado utilizado como parâmetros para se realizar a consulta;
- **Palavra_Localidade:** dados gerados após a realização de uma coleta, contendo Id

da cidade de onde se originou as mensagens, id das palavras contidas na mensagens, quantidade de ocorrência das mesmas e o período ocorrente;

- **Coocorrência:** dados de coocorência de palavras é utilizada para informar onde e quando mais de uma palavra está sendo mencionada em uma mesma mensagem, informando também a data e a quantidade de menções.

Na Tabela 2 é apresentado um exemplo partindo de uma coleta utilizando o assunto “dengue” e os termos “gripe”, “febre” e “dor de cabeça”.

Tabela 2. Exemplo de coocorência de termos em uma mesma mensagem.

MENSAGENS	COOCORÊNCIA
Morrendo de dor de cabeça! Essa gripe vai me matar...	Coocorência dos termos gripe e dor de cabeça
Muita febre! #dengue	Coocorência dos termos febre e dengue

Com essas informações é possível observar quando uma determinada palavra é mensurada, e se existe coocorência dos termos abordados, ou seja, quando em uma mesma mensagem são encontrados mais de um termo.

Com essa ferramenta foram coletas mensagens compartilhadas no Twitter, durante o período de 11/11/2012 ao dia 17/11/2012. As mensagens coletadas foram as que possuíam os termos “dengue”, “gripe”, “febre” e “dor de cabeça”, sendo as mensagens foram armazenadas no banco de dados apresentado na Figura 2, e com isso foram geradas informações sobre coocorência dos termos em determinados locais e períodos.

Também foram empregados algoritmos de aprendizado de máquina, a fim de analisar o sentimento e tentar interpretar o conteúdo de cada mensagem. Para isso foi utilizados algoritmos *open-source*, de autoria de Rafael José Alencar de Almeida, todos disponíveis no repositório https://github.com/rafjaa/analisador-cyberbullying-twitter. Esses algoritmos são todos implementados na linguagem de programação Python, que é de uma linguagem *open-source* e possui um amplo suporte para se trabalhar com extração e processamento de dados.

Com a utilização desses algoritmos coletou-se uma amostra de mensagens compartilhadas na rede social Twitter, referentes ao assunto dengue, e aos termos considerados sintomas da mesma. Essas mensagens foram processadas e classificadas

de forma automatizadas, dividindo-as nas categorias “positivo”, “negativo” e “neutro”.

Também foram coletadas mensagens compartilhadas no Twitter no período de 11/10/2012 ao dia 17/11/2012. A coletada de dados foi feita utilizando consultas via protocolo HTTP GET à URL https://search.twitter.com/search.json?q=termo, o parâmetro “termo” se refere ao item consultado. Com as respostas obtidas foram gerados arquivos do tipo JSON (*JavaScript Object Notation*), que recebeu uma lista de mensagens referentes a pesquisa, juntamente com suas demais informações como: horário de envio da mensagem, localidade e usuário responsável pela mesma.

Após a etapa de coleta, as mensagens foram processadas a fim de se extrair apenas o conteúdo relevante para a pesquisa: texto, e data de envio. No texto da mensagem foram retirados os termos que continham menos de 3 caracteres, e também o próprio termo da busca. A Tabela 3 mostra um exemplo de mensagem processada.

Tabela 3. Exemplo de mensagem processada.

TEXTO ORIGINAL	Eu acho que to com dengue! Febre ta me matando.
TEXTO PROCESSADO	Acho que com dengue febre matando

Após o processamento, foi realizada uma classificação automatizada das mensagens coletadas. Para isso foi utilizado um filtro Bayesiano, que é um método de aprendizagem de máquina supervisionado, que requer uma etapa de treinamento, com a classificação de mensagens manualmente em duas categorias predefinidas. Esse algoritmo permite a classificação de um texto em categorias, utilizando análises estatísticas da ocorrência de suas palavras em outros textos pré-classificados.

SEGARAN (2008) relatou que devido à abordagem de classificação “ingênua” do filtro Bayesiano, que não leva em consideração a ordem e o relacionamento entre as palavras avaliadas. O seu uso necessita de um baixo custo computacional em relação a outras técnicas de classificação como, por exemplo, rede neural e SVM (*Support Vector Machine*), o que torna altamente viável para a classificação de uma grande amostra de tweets.

Para realizar o treinamento do filtro Bayesiano, foram coletadas 120 mensagens, as quais foram classificadas manualmente dentre as três categorias conforme descritas na Tabela 4, as quais foram utilizadas no treinamento do classificador Bayesiano.

Tabela 4. Exemplo de mensagens classificadas manualmente.

MENSAGEM COLETADA	CLASSIFICAÇÃO MANUAL
Morrendo de febre!	Positivo
Já estou melhor, gripe nunca mais.	Negativo
Gripe infinita!	Neutro

Foram coletadas mais de 7000 mensagens, as mesmas foram processadas, e classificadas automaticamente com a utilização do algoritmo Bayesiano. A Figura 3 demostra uma visualização do resultado gerado pelo filtro, distribuindo as mensagens por termo e dias da semana.

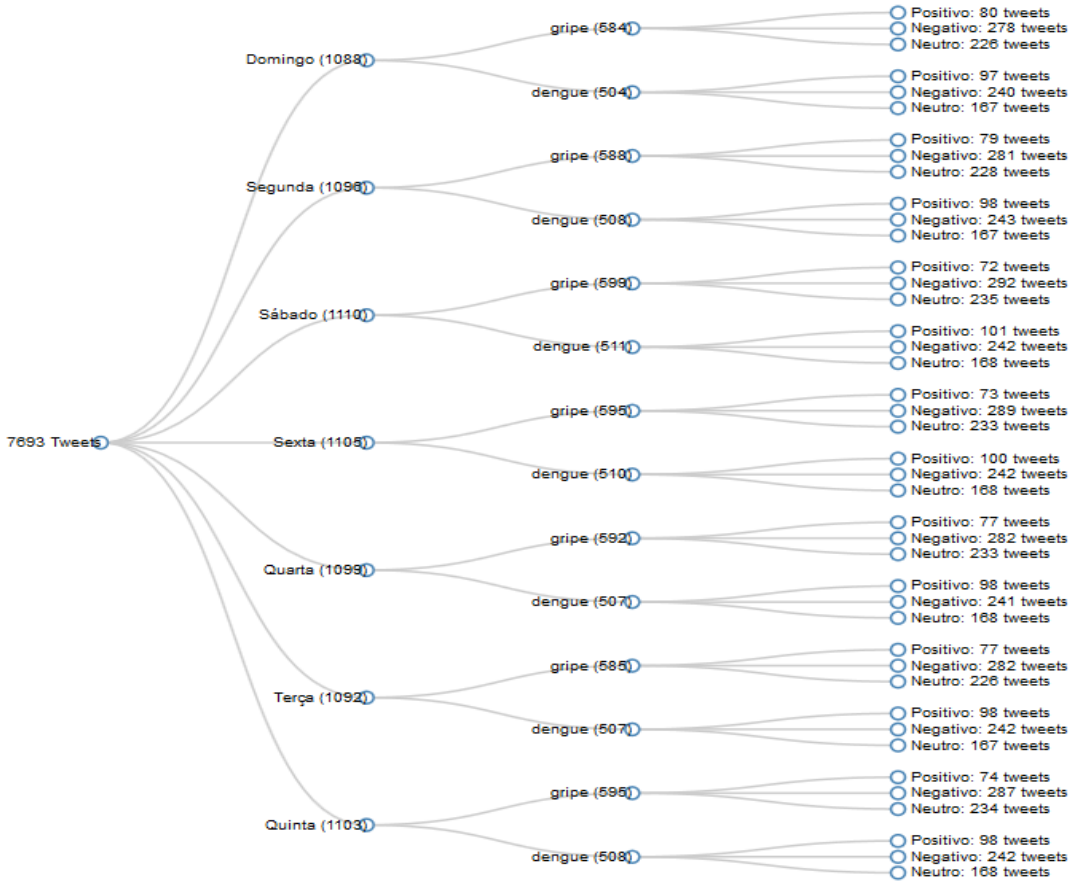

Figura 3. Visualização da classificação gerada.

Pode-se observar, na Figura 3, que a maioria das mensagens foi classificada como neutra. Isso ocorreu devido ao fato de que o arquivo de treino utilizado possuía poucos registros cadastrados, outro fator determinante nesse resultado foi à quantidade de ironias e sarcasmos contido nas mensagens, o que dificultou a classificação, fazendo com que o algoritmo classificasse de forma inadequada algumas mensagens.

8 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi abordada a análise de sentimento, mostrando uma das formas de como a técnica de mineração de opinião pode ser utilizada na classificação de conteúdos compartilhados no Twitter, técnica essa que classifica os textos utilizando pontuações (*scores*), com o auxílio do algoritmo de filtragem Bayesiano. Entretanto existem alguns problemas a serem resolvidos, como detecção de ironias, correção de textos escritos incorretamente e falsas opiniões.

Várias dificuldades foram encontradas no processo de análise do sentimento como: distinguir se um texto é uma opinião ou um fato, identificar se num fato existem opiniões embutidas, detectar sarcasmo e ironias para evitar resultados contrários aos que realmente eram o que os autores queriam expressar, identificar, dentro de um mesmo texto, a que objetos cada opinião se referem, quando pronomes são usados para referenciar os objetos do texto podem dificultar a identificação deles, quando um texto é escrito com ortografia ou sintaxe errada dificulta bastante a interpretação.

Um trabalho de grande valia, um aperfeiçoamento do arquivo de treino, ou a utilização de novos algoritmos de classificação de textos devem aperfeiçoar os resultados.

Com os resultados obtidos pode-se observar que, para a obtenção de melhores resultados, seria necessário o auxílio de profissionais da área de saúde para que se possa alcançar uma forma de comparar os resultados utilizando as informações disponíveis.

REFERÊNCIAS

ARAÚJO, M.; GONÇALVES, P.; BENEVENUTO, F. Métodos para análise de sentimentos no Twitter. In: SIMPÓSIO BRASILEIRO DE SISTEMAS MULTIMÍDIA E WEB (WEBMEDIA). **Proceesings...** Salvador, Brasil. November, 2013.

BENEVENUTO, F.; ALMEIDA, J.M.; SILVA, A.S. **Coleta e análise de grandes bases de dados de redes sociais online**. Disponível em: http://homepages.dcc.ufmg.br/~fabricio/download/jai2012.pdf. Acesso em: Outubro de 2011.

CECI, M.; APPICE, A.; MALERBA, D. Mr-SBC: a Multi-Relational Naive Bayes Classifier. In: Knowledge Discovery in Databases (PKDD). 2003.

CHENG, A.; EVANS M.; SINGH, H. **Inside Twitter**: a in-depth look inside the Twitter World. Disponível em: <http://www.sysomos.com/insidetwitter>. Acesso em: jun. 2011.

CORDEIRO, J.P.C. **Extracção de elementos relevantes em texto/páginas da World Wide Web**. 2004. 174f. Dissertação (Mestrado em Inteligência Artificial e Computação) - Departamento de Ciências de Computadores, Faculdade de Ciências da Universidade do Porto.

CUROTTO, C.L. **Integração de recursos de data mining com gerenciadores de bancos**

de dados relacionais. 2003. 216f. Tese (Doutorado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro.

CUROTTO, C.L.; EBECKEN N.F.F. **Multi-relational data mining in Microsoft® SQL Server™.** Rio de Janeiro: COPPE-UFRJ, 2006.

FONSECA, M.P.A. **Classificação Bayesiana de grandes massas de dados em ambientes ROLAP**. 2007. 111f. Tese (Doutorado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro.

GINSBERG, J. et al. Detecting influenza epidemics using search engine query data. **Nature**, v.457, n.7232, p.1012-1014, 2008.

GONCALVES, P. et al. Comparing and combining sentiment analysis methods. In: ACM CONFERENCE ON SOCIAL NETOWRKS (COSN'13). **Proceedings…** Boston, USA. Oct 2013. http://dx.doi.org/10.1145/2512938.2512951

HAN, J.; KAMBER, M. **Data mining**: concepts and techiniques. 2. ed. San Francisco, CA: Morgan Kaufmann, 2006.

HRUSCHKA JÚNIOR, E.R. **Imputação Bayesiana no contexto da mineração de dados**. 2003. 106 f. Tese (Doutorado em Ciências em Engenharia Civil) - Universidade Federal do Rio de Janeiro.

JIANG, L. et al. Target-dependent Twitter sentiment classification. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: Human Language Technologies – v.1., 49. **Proceedings…** Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. p. 151–160.

LAMPOS, T.; CRISTIANINI, N. **Tracking the flu pandemic by monitoring the Social Web**. Bristol, UK: Intelligent Systems Laboratory University of Bristol, 2010.

LEITE, E.M. Análise da correlação entre dengue e indicadores sociais a partir do sig. **HYGEIA**, v.6, n.11, 44-59, 2010. Disponível em: <www.hygeia.ig.ufu.br/>.

PANG, B.; Lee, L. Opinion mining and sentiment analysis. **Foundation and Trends in Information Retrieval**, v.2, n.(1-2), p.1-135, 2008. http://dx.doi.org/10.1561/1500000011

PASSOS, E.; ARANHA, C. A tecnologia de mineração de textos. **RESI-Revista Elerônica de Sistemas de Informação**, n.2, 2006.

SEGARAN, T. **Programando a inteligência coletiva**. Rio de Janeiro: Alta Books, 2008.

WITTEN, I.H. **Data mining**: practical machine learning tools and techniques. 2. ed. San Francisco, USA: Morgan Kaufmann Publishers, 2005.