UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC

CURSO DE CIÊNCIA DA COMPUTAÇÃO

INOCIO FELIPE DA COSTA

MINERAÇÃO DE DADOS NA REDE SOCIAL TWITTER A RESPEITO DE CASOS DAS DOENÇAS DENGUE, ZIKA E CHIKUNGUNYA

CRICIÚMA

2016

INOCIO FELIPE DA COSTA

MINERAÇÃO DE DADOS NA REDE SOCIAL TWITTER A RESPEITO DE CASOS DAS DOENÇAS DENGUE, ZIKA E CHIKUNGUNYA

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Ciência da Computação, da Universidade do Extremo Sul Catarinense, UNESC.

Orientador: Prof. MSc. Paulo João Martins

CRICIÚMA

2016

INOCIO FELIPE COSTA

MINERAÇÃO DE DADOS NA REDE SOCIAL TWITTER A RESPEITO DE CASOS DAS DOENÇAS DENGUE, ZIKA E CHIKUNGUNYA

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Informática em Saúde.

Criciúma, 22 de junho de 2016.

BANCA EXAMINADORA

Prof. Paulo João Martins- MSc - (UNESC) - Orientador

Prof. Kristian Madeira- Dr - (UNESC)

Profª. Merisandra Côrtes de Mattos Garcia- Dra - (UNESC)

Dedico este trabalho primeiramente a Deus, que me iluminou e me deu forças em todos os momentos. Aos meus pais, Almiro e Marta pelo infinito amor dedicado e valores ensinados. A minha esposa Maiara por sempre estar ao meu lado, e a todos que de qualquer forma estiveram presente dando força e apoio.

AGRADECIMENTOS

Agradeço a Deus que me deu a vida e saúde para que esse sonho pudesse se tornar realidade.

Aos meus pais pelo apoio incondicional, me dando coragem para alcançar meus objetivos. Mãe, o seu amor, carinho e dedicação, me fez crescer e me tornar a pessoa que sou hoje. Pai, você é minha fortaleza, meu porto seguro, obrigada por não medir esforços para que eu chegasse ao final dessa caminhada.

A minha esposa pelo exemplo de bondade, dignidade e honestidade.

Ao meu professor e orientador Paulo que me auxiliou desde o começo.

Ao professor Kristian pela disponibilidade e paciência dedicada para a realização deste trabalho.

Aos professores pelos conhecimentos transmitidos durante todo o curso de Ciência da Computação.

Enfim a todos que, direta ou indiretamente, contribuíram para realização deste trabalho. Muito Obrigado!

RESUMO

A constante evolução das tecnologias de redes e comunicação motiva cada vez mais o acesso à internet, onde tem-se um espaço para a produção, circulação e manifestação de diferentes discussões acerca de vários assuntos. As redes sociais tem a capacidade de espalhar a informação com rapidez, e o conjunto dessa massa de dados pode ser explorado sobre a ocorrência de casos de doenças como a dengue, zica e chikungunya. Nesse capo pode-se revelar informações sobre o teor das publicações como por exemplo o conteúdo cômico, as campanhas, e mensagens que contem informação sobre a ocorrência das doença, entre outros tipos de manifestações. Para a categorização do textos que tem relação com um caso da doença utilizou-se a técnica máquina de suporte vetorial. Essa técnica é reconhecida em diversas aplicações na área de aprendizado de máquina. Também é aplicado com sucesso na classificação de texto. A técnica procura a separação máxima entre duas classes. Para a mineração de dados proveniente da rede social foi implementada a ferramenta Coletor de Dados, desenvolvido em linguagem Java com a utilização da API do Twitter, que permite a conexão para coleta em tempo real das mensagens. Para o armazenamento dos dados utilizou-se um banco de dados relacional. Após a fase de coleta e armazenamento realizou-se o pré-processamento. Essa fase tem a finalidade de reduzir o tamanho da massa de dados, tornando viável a geração da tabela valor atributo. Ainda na etapa, os dados são separadas em *tokens é* aplicado as funções *stemming e stop words*. Então aplicou-se o método que permite atribuir peso as palavras de acordo com sua frequência, o que determina o nível de importância, para o texto. Na sequência obteve-se a tabela de valor atributo para realizar os testes na ferramenta Weka. O teste foi realizado com as mensagens que contem menção as palavras: dengue, zica, e chikungunya, separadas em arquivos para realizar os teste de classificação. Os resultados do classificador aponta que a maioria das mensagens não contem em seu conteúdo relação com caso das doenças. A implementação SMO obteve mais indicadores com resultados superiores a LibSVM.

Palavras-chave: Redes Sociais. Dengue. Zica. Chikungunya. Máquina de suporte vetorial. Mineração de dados.

ABSTRACT

The constant development of networks and communication technologies motivates increasingly Internet access, which has a space for the production, circulation and manifestation of different discussions about various subjects. Social networks have the ability to spread information quickly, and the whole of this mass of data can be explored on the occurrence of diseases such as dengue, chikungunya and zica. Capo that can reveal information about the content of publications such as comic content, campaigns and messages containing information on the occurrence of the disease, among other events. For categorization of text that is related to a case of disease, we used the technical support vector machine. This technique is recognized in various applications in machine learning area. It is also successfully applied to the text classification. The technique seeks the maximum separation between two classes. For mining data from social network was implemented the Data Collector tool, developed in Java using the Twitter API, which allows connection to collect real-time messages. For storage of the data used a relational database. After the phase of collection and storage held preprocessing. This phase has the purpose of reducing the size of the mass of data, making possible the generation of attribute value table. Also in the step, the data are separated into tokens is applied and the stop words stemming function. Then he applied the method for assigning weight the words according to their frequency, which determines the level of importance for the text. Following obtained the attribute value table to perform the tests in Weka tool. The test was performed with the messages that mention the words it contains dengue, zica, chikungunya and separated files to perform classification test. The classifier results shows that most messages do not count in your content related to the case of diseases. The SMO implementation got more indicators with results exceeding LIBSVM.

Keywords: Social Networks. Dengue. Zica. Chikungunya. Vector support machine. Data mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANSI	American National Standards Institute
API	ApplicationProgramming Interface
ARPANET	Advanced Research Projects Agency
ARRF	Attribute-Relation File Format
BSD	Berkeley Software Distribution
COLT	Computational Learning Theory
CPCD	Centro de Prevenção e Controle de Doenças
DBI	Departamento de Bioquímica e Imunologia
GNU	General Public License
HSQLDB	HyperSQLDataBase
IBM	International Business Machines
ICE	Instituto de Ciências Exatas
IDE	Integrated Development Environment
IEEE	Institute of Electrical and Electronic Engineers
KDD	*Knowledge Discovery in Database*
LAC	*Lazy Associative Classification*
MSV	Maquinas de suporte vetorial
OPAS	Organização Pan-Americana de Saúde
RNA	Redes Neurais Artificiais
RSO	Redes Sociais Online
SBBD	Simpósio Brasileiro de Bancos de Dados
SGBD	Sistema de Gerenciamento de Banco de Dados
SMO	Sequential Minimal Optimization
SQL	Structured Query Language
SRS	Sites de Redes Sociais
SVM	Suport Vector Machine
TCP/IP	TransmissionControlProtocol/Internet Protocol
Tf-IDF	Term frequency – Inverse Document Frequency
UFMG	Universidade Federal de Minas Gerais
WEKA	Waikato Enviroment for Knowledge Analysis
WWW	World Wide Web

SUMÁRIO

1 INTRODUÇÃO

O aumento do número de pessoas com acesso à Internet e a variedade de dispositivos existentes contribui para a ampliação do acervo de diversas publicações (SANTOS et al, 2013).

Hoje em dia, tem-se discussões acerca de vários assuntos circulando na *Web*, expressando ideias na forma de elementos exibidos nas redes sociais (MACHADO et al, 2010). Conforme Leite (2010), esse é um campo com potencial a ser explorado. Com isso, tem-se a crença de realizar ações que possam antecipar um plano de ação mediante a obtenção de conhecimentos existentes nas informações de um meio. Acresce-se a isso, que este meio é atualizado constantemente, ao lado do crescente interesse a propósito do estudo de ambientes e fatores que apresente ideias pertinente a uma doença, por exemplo.

De acordo com Aranha e Passos (2006), a mineração de dados (*data mining*) conta com algoritmos apropriados e técnicas consagradas. Assim, observa-se a necessidade da aplicação de modelos computacionais direcionados à obtenção desse objetivo.

O conceito de mineração de dados é entendido como a aplicação de métodos para obtenção de conhecimento explorando grandes bases de dados (GOLDSCHMIDT; PASSOS, 2005). Também se pode entender como a descoberta de conhecimento por meio da estrutura e relação dos dados, investigando e trabalhando na área da inteligência artificial (CORDEIRO, 2004).

Segundo Santos et al (2013), existem diversos mecanismos para extrair o conhecimento de grandes massas de dados. Atualmente, a Web é um grande depósito de informações, sendo geradas a todo instante, das mais diversas formas, fontes e assuntos. São muitas pessoas transmitindo e recebendo informações por meio de *sites* de notícias, compartilhamentos, blogs, redes sociais, entre outros, que alteram o estilo e o objetivo, conforme a circunstância em que se encontra.

Contudo, ainda segundo Santos et al (2013), para minerar textos de redes sociais são encontrados alguns problemas, tais como bases não estruturadas. O texto que contêm opiniões é informal e, muitas vezes, apresentam palavras e expressões regionais e abreviações, além de expressões recentes ainda não reconhecidas pela linguagem culta.

A extração de conhecimento por meio de informações expressas em publicações na rede social tende a identificar a pretensão expressa por grupos na Web (MACHADO et al, 2010).

Diante do cenário apresentado, neste estudo será empregado um método, que será definido, para extrair o contexto das publicações sobre uma determinada doença, a ser caracterizada no decorrer da pesquisa, buscando resolver o problema de extrair o conhecimento acerca do núcleo radicado nessas discussões.

1.1 OBJETIVO GERAL

Minerar os dados públicos na rede social Twitter relacionados a ocorrência das doenças dengue, chikungunya e zika.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos desta pesquisa consistem em:

a) entender os conceitos oriundos da internet, redes sociais, e seus impactos na sociedade;
b) compreender as técnicas de coleta das expressões e informações da redes social;
c) realizar a coleta e pré-processamento dos textos e dados públicos da rede social;
d) aplicar o método *Support Vector Machine* (SVM) para identificar publicações relacionados com casos das doenças dengue, zika, chikungunya;
e) avaliar o desempenho do SVM, na classificação das mensagens coletadas da rede social, por meio das medidas estatísticas.

1.3 JUSTIFICATIVA

É importante considerar o uso de tecnologias na área da saúde. Segundo Leite (2010), é reconhecida a contribuição da tecnologia para obter conhecimento através das informações que auxiliam na identificação de parâmetros para a tomada de decisão. Conforme Antunes (2014), órgãos responsáveis pela saúde no Brasil já

demonstram a relevância de monitorar as mensagens compartilhadas a fim de definir uma estrutura de combate as doenças e a obtenção de indicadores que propiciam uma visão de determinada enfermidade em discussões por meio da análise de interações em redes sociais.

Conforme apresenta Aranha e Passos (2006), a mineração em bases não estruturadas, como os textos encontrados em redes sociais, integra diversas áreas, como a informática, a matemática, as línguas e ciências em geral. Partindo da mineração de dados e da utilização de métodos para explorar o conhecimento existente no conteúdo *Web,* a *Data Mining* ou mineração de dados levou ao surgimento do subgrupo *Text Mining* ou mineração de textos, que busca montar informações ocultas em meio a uma grande massa de dados sem formato contínuo, mas que podem se relacionar. Ainda conforme Aranha e Passos (2006), o conjunto de recursos mencionados organiza e propicia a descoberta de conhecimento oculto e ainda incompreensivo, mas existente, numa grande base de informações em forma de texto publicado em redes sociais. Este fator motiva a necessidade de realizar a análise que expresse e permite recuperar, padronizar e aplicar regras e técnicas específicas, capazes de trabalhar um volume acumulado de textos, tornando essa informação útil para possibilitar a definição de uma escolha (ARANHA; PASSOS, 2006).

Encontrar a associação entre as informações que circulam em uma rede social sobre assuntos relacionados a uma doença, pode confirmar a eficiência desses métodos como uma alternativa atualizada constantemente. Além disso, ao ser monitorada através de meios computacionais automatizados, proporciona um complemento no acompanhamento da evolução de casos de doenças (ANTUNES, 2014).

Com isso, acredita-se que a utilização de conhecimento extraído de informações contidas nas atualizações frequentes de interações em redes sociais poderá ser de grande valia à área da saúde, pois é um recurso a mais para obtenção de informações e para a tomada de decisão neste campo.

1.4 ESTRUTURA DO TRABALHO

O presente trabalho está dividido em seis capítulos, sendo esta introdução o primeiro, no qual é apresentada uma breve contextualização do tema, apresenta-se

os objetivos, além de se justificar o estudo e a forma pela qual o trabalho se encontra organizado.

O segundo capítulo trata das redes sociais, abordando sua definição, elementos e discorrendo sobre os sites com esta finalidade.

No terceiro capítulo é apresentado um breve apanhado teórico sobre a dengue, zika vírus e a febre chikungunya, devido ao fato de essas patologias serem consideradas no modelo proposto.

A mineração de dados provenientes das redes sociais é o tema do capítulo 4, bem como a s técnicas para classificar essas publicações, e o Support Vector Machine aplicado na tarefa de categorização das mensagens do Twitter.

O Capítulo 5 apresenta alguns trabalhos correlatos que obtiveram informações por meio das redes sociais, aplicando técnicas de aprendizado de máquina para classificação dos conteúdos existentes no texto coletado.

No Capítulo 6 são descritas as etapas do trabalho desenvolvido, a metodologia utilizada e os resultados obtidos por meio da classificação das publicações coletadas de rede social.

Finalizando, tem-se a conclusão deste trabalho acadêmico e as sugestões de trabalhos futuros.

2 REDES SOCIAIS

A chegada da Internet trouxe diversas mudanças para a sociedade. Ela tem suas origens por volta dos anos de 1960, durante a Guerra Fria, quando os Estados Unidos criam uma rede exclusivamente militar, com o objetivo de, no caso de um ataque russo, poder ter acesso a informações militares de qualquer lugar do país (LEITE; LEMOS, 2014).

Oficialmente, a rede pretendida pelos americanos foi criada em 1969 e denominada de ARPANET. Em princípio, a rede tinha 4 computadores distribuídos entre diferentes universidades. Dois anos mais tarde, já possuía cerca de 40 máquinas. Tal foi o crescimento deste sistema de comunicação em rede, que em seguida, dois pesquisadores criaram a arquitetura TCP/IP, utilizando os protocolos *Transmission Control Protocol* (TCP) *e Internet Protocol* (IP), que se tornou o padrão para comunicações dentro de redes de computadores (atualmente ainda usa-se estes protocolos) (TEIXEIRA, 2015).

As funções militares dissociaram-se da ARPANET e essa continuou a crescer e abrir-se para o mundo. Na década de 1980, a rede deixou de ter interesse militar e passou a ser interesse de outras agências que a viam com interesse científico e para fins acadêmicos (PAESANI, 2013).

Contudo, o mais importante elemento da verdadeira explosão da Internet foi o desenvolvimento do World Wide Web (WWW) ou simplesmente Web. O WWW nasceu em meados de 1980, no Laboratório Europeu de Física de altas energias, com sede em Genebra. É composto por hipertextos, ou seja, documentos cujo texto, imagem e sons são evidenciados de forma particular e podem ser relacionados com outros documentos (PAESANI, 2013).

A Internet trouxe uma revolução sem precedentes no mundo da computação e comunicações. Ela é, ao mesmo tempo, uma oportunidade com um alcance global, um mecanismo de propagação de informações e um meio de colaboração e interação entre indivíduos e seus computadores, independentemente da sua localização geográfica (TEIXEIRA, 2015).

Entre as inovações originadas pela Internet, incluem-se algumas expressivamente conhecidas e utilizadas, como as redes sociais.

Uma rede social é definida como um conjunto de dois elementos: atores e conexões, o primeiro é formado por pessoas, instituições ou grupos, enquanto o segundo são as interações ou laços sociais.

> Uma rede é uma metáfora para observar os padrões de conexões estabelecidas entre os diversos atores. A abordagem de rede tem seu foco na estrutura social, onde não é possível isolar os atores sociais e nem suas conexões (RECUERO, 2011. p 24).

Com base nessas considerações iniciais, neste capítulo apresenta-se aspectos sobre as redes sociais.

2.1 DEFINIÇÃO DE REDES SOCIAIS

O termo rede, vem do do latim *rete* sendo utilizado para definir uma estrutura que tem um padrão particular. Existem vários tipos de redes: computador, elétrica, social. As redes sociais podem ser definidas como estruturas onde muitas pessoas têm diferentes tipos de relacionamentos amigável (PAESANI, 2013).

O termo rede social é o nome dado aos diferentes sites que oferecem a possibilidade de contato com muitas pessoas para compartilhar conteúdo, interagir e criar comunidades de interesses semelhantes: trabalho, leituras, jogos, amizade, relacionamentos, entre outros (MARTINO, 2014).

Há mais de um século, a ideia de rede social era usada para mencionar as relações estabelecidas entre elementos de determinado sistema social. Porém, ouve-se falar mais recentemente deste conceito aplicado à Internet, querendo o mesmo significar uma estrutura constituída por pessoas ou organizações que compartilham interesses, motivações e objetivos comuns (PEREIRA; PEREIRA; PINTO, 2011).

As redes sociais tiveram o seu início em meados de 1995, quando Randy Conrads cria o website <classmates.com>, permitindo aos usuários recuperarem ou manterem contato com ex-colegas, de faculdade, universidade, trabalho, entre outros. Em seguida, outros sites passaram a promover redes de amigos em círculos online, quando o termo foi usado para descrever relacionamentos nas comunidades virtuais que apareciam (MARTINO, 2014).

Uma das primeiras redes sociais mais populares foi o Friendster, que surgiu em 2002 e foi criado para ajudar a encontrar amigos de amigos, e desde o início até maio de 2003, havia crescido para a quantidade de 300 mil usuários (MARTINO, 2014).

Em 2003, com a chegada de sites como o Tribe.net, MySpace, Ecademy, LinkedIn e Soflow, havia mais de 200 sites de redes sociais. A popularidade desses sites cresceu rapidamente e grandes empresas entraram no espaço das redes sociais. O Google lançou o Orkut em janeiro de 2004 para apoiar uma experiência que desenvolvida por um de seus empregados em seu tempo livre (TEIXEIRA, 2015).

Outra rede social muita conhecida, o Facebook, foi originalmente criado para suportar redes universitárias, em 2004. Os usuários do site foram obrigados a fornecer o endereço de e-mail associado com instituições de ensino. Essa rede posteriormente foi expandida para incluir os alunos do Ensino Médio, os profissionais e, finalmente, todos os potenciais utilizadores (TEIXEIRA, 2015).

Já o Twitter, termo em inglês que pode ser traduzido como piar ou chilrear, é o nome de uma rede de *microblogging*, que permite escrever mensagens na Internet que não excedam 140 caracteres. Essas entradas são conhecidas como *tweets*. Quando um usuário escreve uma mensagem em sua página no Twitter, ela é enviada automaticamente para todos os usuários que tenham escolhido a opção de recebê-las (seguidores). Esta mensagem também pode ser vista imediatamente no perfil do usuário (VIEIRA; MATOS; SLONGO, 2009).

A plataforma Twitter surgiu em outubro de 2006, em San Francisco, nos Estados Unidos, como uma rede social com algumas características autênticas. Entre eles, pode-se citar que permite um serviço absolutamente livre, sem anúncios publicitários e fácil de usar (TEIXEIRA, 2015).

Com a popularidade das redes sociais online e o acesso as novas tecnologias, a comunicação por meio da utilização de dispositivos confortáveis e com custo acessível, encarregaram-se da alta abrangência e necessidade de uso desses equipamentos no dia a dia. São pessoas comuns, empresas, governos e instituições, que utilizam essas ferramentas para chegar a praticamente qualquer lugar onde exista uma sociedade conectada à Internet. Pode iniciar de forma individual ou até por meio de grupos, que também podem replicar essas informações, tornando a comunicação ilimitada para quem está conectado e dispondo de uma acesso a uma rede social online (CASTELLS; CARDOSO, 2005).

Por muitas vezes, as redes sociais são as protagonistas de informações, como por exemplo, para a propagação de notícias importantes como campanha presidencial, catástrofes naturais, entre outros. Fenômenos diferentes atrai a atenção de pessoas em todo mundo a todo o momento. Esses fenômenos, por vezes muito

diferentes, apresentam as formas de organização, identidade, conversão e mobilização social: “o advento da Comunicação Mediada pelo Computador”, que admite aos indivíduos comunicarem-se, permitindo que redes fossem criadas e expressas nesses espaços. Por isso, as redes sociais não conectam apenas computadores, mas pessoas (RECUERO, 2011).

2.2 ELEMENTOS DAS REDES SOCIAIS

O principal elemento da rede social são os atores que são representados pelos nós. Trata-se das pessoas envolvidas na rede que se analisa, ou seja, eles operam de forma a moldar os mecanismos sociais, por meio da interação e da composição de laços sociais. No entanto, conforme apresenta Recuero (2011), os atores são formados de maneira um pouco diferenciada por causa do distanciamento entre os envolvidos na interação social, principal característica da comunicação mediada por computador. Assim, neste caso, a mencionada autora ainda afirma que os atores sociais trabalham com construções do ciberespaço. Um ator pode ser representado por diversas maneiras, entre elas, pode-se citar, o Facebook, Twitter e Instagram.

As redes sociais compõem uma das estratégias subjacentes utilizadas pela sociedade para o compartilhamento da informação e do conhecimento, o indivíduo vai apresentando e ampliando sua rede, que é muito peculiar. Ele se reúne com seus semelhantes e estabelece relações de trabalho, amizade, enfim, relações de interesses, que se desenvolvem e se modificam conforme a sua trajetória. Assim, cria sua inserção na realidade social, mediante as relações entre atores que as interagem. Durante toda a vida, as pessoas desenvolvem relações na sociedade em que estão inseridas. Primeiro no domínio familiar, na escola, na comunidade em que vivem e no trabalho. É da natureza humana estar conectado a outras pessoas e estruturar a sociedade em rede. Portanto, as relações que as pessoas desenvolvem e mantêm é que fortalecem a esfera social (TOMAEL; ALCARÁ; DI CHIARA, 2006).

Enquanto os atores representam os nós da rede em questão, as conexões de uma rede social podem ser percebidas de diversas maneiras. De um certo modo, as conexões em uma rede social são constituídas dos laços sociais que são formados através da ligação entre os atores, ou seja, são as conexões o principal foco de estudo das redes sociais, pois é sua variação que altera as estruturas desses grupos. Essas

influências, na Internet, são percebidas graças à possibilidade de manter os rastros sociais, dos indivíduos os quais continuam ali. Nas redes sociais, cada usuário pode construir uma imagem que o representa na Internet, por meio de um perfil de uma rede social online. A interação entre participantes é feita por meio de comentários, quando cada um é livre para opinar e comentar sobre o conteúdo disponibilizado nessas páginas. Por exemplo, um comentário em uma rede social online, continua ali até que alguém o delete ou a rede social online saia do ar (RECUERO, 2011).

2.3 SITES DE REDES SOCIAIS

Os Sites de Redes Sociais (SRS) são um dos aspectos mais populares para a compreensão das redes sociais na Internet. Esses sites não são exatamente um elemento novo, mas uma consequência da assimilação das ferramentas de comunicação mediada pelo computador pelos atores sociais. Os SRS são os espaços utilizados para expressão das redes sociais na Internet (RECUERO, 2011).

A forma como a informação é apresentada no SRS dependem de uma variedade de perspectivas e "construção" de um objetivo comum entre as comunidades que estão organizadas em torno deles. Tipicamente, em um ambiente social, as pessoas compartilham suas histórias e experiências com os outros, naturalmente (MIRAGEM, 2013).

O aparecimento das redes sociais pode ser visto como um ponto na abertura de novas possibilidades para a produção, circulação e apropriação da informação. As redes sociais dependem exclusivamente da interação entre pessoas, pois é a partir do debate e conexão entre eles que os conteúdos propriamente ditos são levantados e compartilhados. Para isso, usa-se a tecnologia da informação como recurso, definindo-se como ferramentas online projetadas para permitir a interação social a partir do compartilhamento e da criação colaborativa da informação nos mais diversos formatos. Estas atividades podem ser textos, imagens, vídeos e áudios. Tanto a influência mútua como a forma pela qual a informação será proporcionada vai depender das perspectivas e visões de mundo do usuário ou do grupo que compartilhou o conteúdo (VIEIRA; MATOS; SLONGO, 2009).

Tabela 1. Características das redes sociais.

Rede Social	Característica
Facebook	Lançada em 2004, a rede social permite criar perfis pessoais ou páginas como empresa, figura pública, marca, comunidade, entre outros. Permite também elaborar enquetes, usar aplicativos, entre outros serviços.
Flickr	Lançado em 2004, é a maior comunidade de fotografia do mundo, com amplo armazenamento gratuito. Além de permitir novas maneiras de organizar fotos e vídeos.
Instagram	Uma rede social online lançada em 2010, permite que seus usuários façam compartilhamento de fotos e vídeos escolhendo um filtro para mudar sua aparência. Podendo compartilhar com o Facebook, Twitter, Tumblr e outros.
LinkedIn	Lançado oficialmente em 2003, o objetivo do site é reunir profissionais, através de listas de contatos (pessoas ou empresas), promovendo a interatividade entre profissionais e a prática de networking.
Tumblr	Lançado em 2007 é feito de blogs, ou seja, blogging que permite aos usuários publicarem textos, imagens, vídeo, links, citações, áudios e diálogos.
Twitter	Serviço de microblogging criado em 2006. Permite publicar textos de até 140 caracteres, fotos e links.
Youtube	Lançado em 2005, o conteúdo disponibilizado é exclusivo para o compartilhamento de vídeos com a possibilidade dos usuários fazerem comentários a respeito do que foi publicado

Fonte: Do autor.

As redes sociais são baseadas em sua maioria na teoria conhecida como seis graus de separação, que basicamente diz que todas as pessoas no mundo entram em contato com outra, com um máximo de seis pessoas como intermediárias na cadeia. Por exemplo: a pessoa conhece alguém que por sua vez conhece alguém, que por sua vez conhece alguém e assim por diante, formando uma cadeia de seis ou menos pessoas envolvidas. Isso faz com que o número de pessoas conhecidas na cadeia cresça exponencialmente à medida que o número de ligações aumente, para formar uma rede que conecta todos (MIRAGEM, 2013).

As redes sociais trazem um sentimento de imparcialidade aos assuntos. É comum falar sobre as redes em termos de igualdade, e dizer que elas consistem em relações entre pares que padronizam e unificam esses relacionamentos. As redes sociais existem apenas em situações de assimetria, caso contrário, nenhuma rede seria necessária, ou seja, pares simétricos podem se comunicar, mas os pares assimétricos devem se envolver. A rede social possui um potencial grandioso, elas

podem derrubar governos, construir novos impérios a partir das cinzas do antigo, e por meio de sua própria conectividade, podem se disseminar rapidamente em novos espaços. Muitas vezes, as redes sociais são descritas como fora de controle (SILVEIRA, 2010).

Contudo, mesmo em que pese fatores negativos, as rede sociais no atual contexto têm a importância ímpar de disseminar informações sobre temas de grande relevância para a sociedade. A exemplo, pode-se citar o caso da dengue, patologia esta que vem assustando muitas regiões do mundo, sobretudo no Brasil, onde se registra um surto de outras enfermidades a essa associadas. Sobre este tema, discorre-se no capítulo a seguir, por ser objeto de estudo deste trabalho.

3 DENGUE, ZICA E CHIKUNGUNYA

A dengue é uma moléstia muito comum na América Latina, na América do Sul e no Caribe. É transmitida pelo mosquito *Aedes aegypti*, que habita os grandes aglomerados urbanos e ataca o homem durante o dia, ao sugar seu sangue, transmitindo o vírus da doença (BRASIL, 2013).

A dengue é uma doença infecciosa não contagiosa, de origem viral, caracterizada por uma patologia febril aguda com duração limitada (2-7 dias), com mal-estar grave, acompanhada de erupções cutâneas, hemorragia, entre outros sintomas (PESSANHA, 2009).

A palavra dengue é de origem espanhola e diz respeito ao fato das pessoas acometidas adquirirem estranhos trejeitos motivados pela febre alta e pela tremura corporal, tornando-se, dessa forma, "dengosas". Para os nativos caribenhos a palavra dengue significa "quebra ossos", o que representaria também o estremecimento e a contorção que acometem o doente em razão da febre alta. A dengue é transmitida pelo mosquito *Aedes aegypti* (SILVA; SILVA, 2009).

A origem do vírus dengue tem sido objeto de intensa discussão. Alguns autores atribuíram a origem africana relacionada ao comércio de escravos, que teria levado o vírus pelo mundo afora. Outros autores propuseram que o vírus teria se originado na Malásia, a partir de um ciclo silvestre da floresta, envolvendo macacos. Outros acreditam que o vírus *Aedes aegypti* é proveniente do Egito, sendo encontrado no rio Nilo, de onde derivou o seu nome (SILVA; SILVA, 2009).

O *Aedes aegypti* vive e deposita seus ovos em volta e no interior das casas, em recipientes utilizados para o armazenamento de água para as necessidades domésticas como vasos e potes, em pneus velhos e outros objetos que podem funcionar como reservatório de água. Sua capacidade de voo é de cerca de 100 metros, de modo que o mosquito que pica é o mesmo que cresceu nos arredores (BRASIL, 2013).

O mosquito *Aedes aegypti* mede menos de um centímetro, tem aparência inofensiva, cor café ou preta e listras brancas no corpo e nas pernas. Costuma picar, transmitindo a dengue, nas primeiras horas da manhã e nas últimas da tarde, evitando o sol forte, mas, mesmo nas horas quentes, ele pode atacar à sombra, dentro ou fora de casa. O indivíduo não percebe a picada, pois não dói e nem coça no momento (BRASIL, 2013).

O *Aedes aegypti* (figura 1) tem uma coloração castanha escura, o tórax possui tegumento coberto por escamas castanhas escuras e outras branco-prateadas. O abdômen é escuro, pode possuir manchas branco-prateadas formando anéis, e as pernas traseiras possuem faixas brancas dando a ilusão de se tratar de listras (PESSANHA, 2009).

Figura 1 – Mosquito Aedes aegypti

Fonte: Pessanha (2009).

Além da dengue, o mosquito *Aedes aegypti* transmite outras duas doenças distintas: as infecções causadas pelo zika vírus e a febre chikungunya.

O zika é uma patologia causada cujos sintomas são semelhantes aos da dengue (febre alta, dor de cabeça, dor nas articulações e dores musculares, três ou sete dias depois de ser infectado pelo mosquito). Embora a maioria dos pacientes tende a se sentir melhor dentro de dias ou semanas, algumas pessoas podem desenvolver dor nas articulações e rigidez de forma intermitente durante meses (SILVA; SILVA, 2009).

As complicações são raras, mas foram descritos surtos na Polinésia e mais recentemente no Brasil, encontrando-se cada vez mais numerosas as evidências de uma ligação entre o vírus e microcefalia eM bebês de mulheres que foram infectadas (BRASIL, 2016).

A febre chikungunya (do africano “andar curvado”), por sua vez, é caracterizada por início súbito de febre, geralmente acompanhada de dor nas articulações. Além disso, muitas vezes ocorrem dores musculares, dores de cabeça,

náuseas, fadiga e erupção cutânea (BRASIL, 2016).

Apesar de serem enfermidades totalmente distintas, o fator chave das três patologias é a transmissão pelo mesmo vetor (*Aedes aegypti*). Porém, as características de cada uma mostram diferenças substanciais: os pacientes com dengue apresentam dores musculares geralmente mais altas e mais fortes, já na febre chikungunya, é observada uma febre superior e dores mais intensas nas articulações (mãos, pés, joelhos, coluna), enquanto com zika, os acometidos não apresentam características particulares, mas na maioria dos casos, relata-se erupções na pele, e diarreia e conjuntivite em alguns pacientes (BRASIL, 2013).

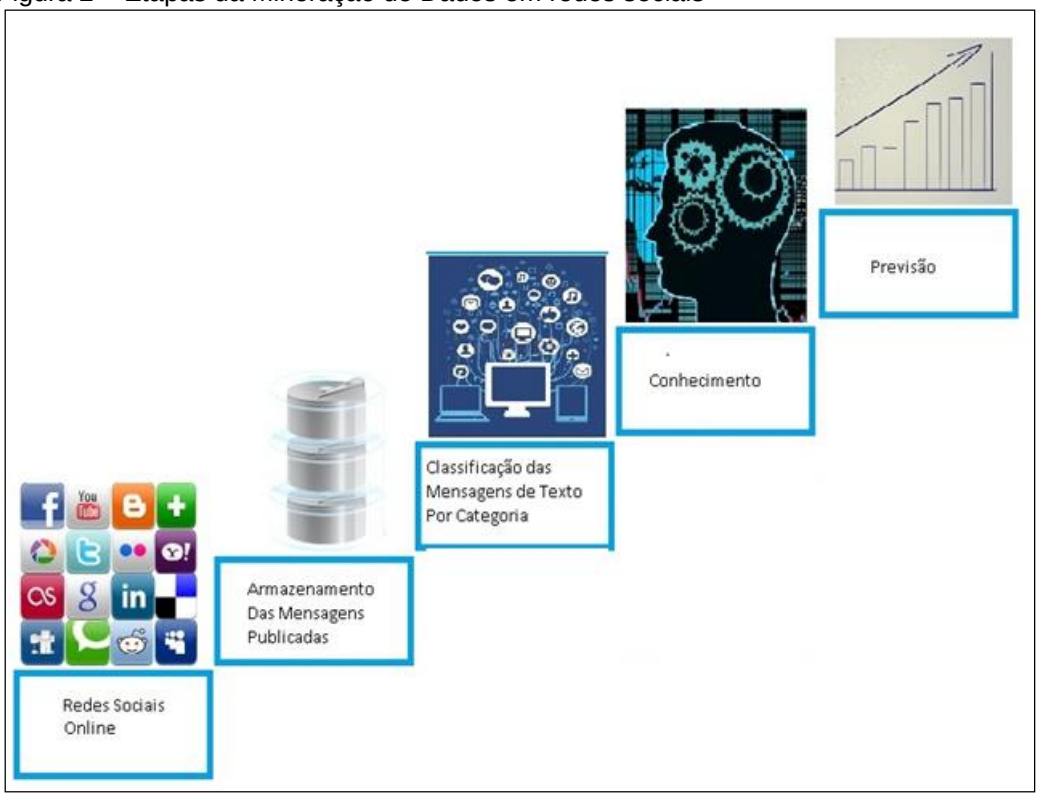
Não há nenhum tratamento específico ou vacina disponível para prevenir estas doenças. Segundo Brasil (2016), uma das formas de prevenção é compartilhar informações com os cidadãos, incentivando programas de mobilização social, que contribuam para ampliar as medidas de prevenção e minimizar o risco de disseminação do vírus.

Diante disso, a seguir apresentam-se aspectos sobre mineração de dados e as redes sociais, para, na sequência, verificar-se como este processo pode atuar em casos de saúde pública, especificamente em relação as doenças causadas pelo *Aedes aegypti*.

4 MINERAÇÃO DE DADOS E AS REDES SOCIAIS

A extração de informações úteis no volume de dados publicado nas redes sociais online por meio das técnicas de mineração de dados é alvo da atenção de pesquisas e aplicações práticas. O emprego de tecnologias preparadas para operar esse volume de dados com a intenção de encontrar informações com a aspiração de tornar os dados sem valor em algo produtivo. Sem o tratamento dos métodos de mineração de dados, parte do conhecimento que é publicado nas Redes Sociais Online (RSO) permaneceria escondido no meio das inúmeras publicações que ocorrem nas RSO (PINHEIRO, 2008) (figura 2).

Figura 2 – Etapas da Mineração de Dados em redes sociais



Fonte: Adaptado de Pinheiro (2008).

Em determinadas ocasiões, a mineração de dados é confundida com recuperação de dados. É importante esclarecer que a seleção de registros utilizando um sistema de gerenciamento de bancos de dados ou a procura utilizando os mecanismos de busca da Internet pertencem à área de recuperação de dados. Já a mineração de dados está relacionada à descoberta de conhecimento em base de dados, do inglês *Knowledge Discovery in Database* (KDD), que é a ação genérica de transformar uma grande massa de dados, em princípio sem sentido, em um

conhecimento relevante (TAN; STEINBACH; KUMAR, 2009). Diante do cenário é visto a necessidade da aplicação da mineração de texto. A qual a mineração em dados não estruturados como por exemplo as mensagens provenientes da rede social consiste em automatizar a coleta de dados. A fase inicial é de coleta de dados. Em seguida, vem a fase de pré-processamento e indexação, a qual rotula entidades no texto e emprega o conhecimento de linguística computacional para tratar cada palavra de acordo com a classe gramatical a qual pertence, induzindo os limites da sua classificação. Outra fase é a aplicação de algoritmos e, por fim, a análise dos resultados. A tecnologia de mineração de textos ganhou ainda mais destaque com a atenção dispensada nos últimos anos as redes sociais online, tendo em vista o volume de dados publicados diariamente na Internet (PINHEIRO, 2008).

A fase de pré-processamento exemplificado na figura 3 como a fase de preparação dos dados é importante porque limpa os dados que serão minerados e também processos, como a redução e categorização da importância das variáveis (PINHEIRO, 2008).

Figura 3 – O plano para minerar texto comparado à realidade.

Fonte: Adaptado de Miller (2005).

A fase de preparação dos dados ocupa o maior tempo na mineração de texto. Sendo essa etapa relevante para obtenção de informações seguras. O pré-processamento é a etapa que por meio de funções trata o texto para para obter as informações corretas (MILLER, 2005, tradução nossa) e então preparar esse dado

não estruturado para a classificação por meio de métodos de aprendizado de máquina.

4.1 DEFINIÇÃO DOS MODELOS

Os modelos de classificação fazem a caracterização de diferentes objetos e classes (TAN; STEINBACH; KUMAR, 2009). A modelagem preditiva busca antecipar a média condicional do alvo por meio dos valores fornecidos na entrada, ou seja, estimar a probabilidade condicional de cada classe, por meio dos valores de entrada. De acordo com as probabilidades posteriores, cada ocorrência será classificada pela maior possibilidade (PINHEIRO, 2008).

Para que uma função alvo f relacione-se aos conjuntos de atributos y para cada um dos rótulos z pré-determinado, é necessário a aprendizagem dos modelos de classificação (função alvo) (TAN; STEINBACH; KUMAR, 2009).

Variáveis de frequência com a estimação de pesos dos termos são utilizados nos modelos preditivos, onde é especificado a variável e a frequência dos valores, com entradas semelhantes a situação do treino, podendo ocorrer do resultado situar-se próximo de algum espaço, assim, criando uma correlação linear a redundância das variáveis semelhantes. A desvantagem desse processo é que provoca inconstâncias na estimação dos parâmetros. Por meio de um treino inicial, com uma pequena amostra, o processo do modelo preditivo é aplicado a uma população de dados disponível. Para cada decisão, existe um peso para uma variável ou uma constante numérica. As medidas das variáveis são representações após o tratamento dos casos específicos. Essas consequências não dependem dos valores alvo dos dados iniciais do processo de treino. Um aspecto importante que deve ser observado no modelo preditivo é a generalização. Para isso, determina-se quanto será acurado no modelo para os casos não previstos durante o treinamento. Os modelos de mineração de texto, a exemplo de outros métodos de estimação não lineares flexíveis, são sensíveis ao ajuste a menor ou maior (TAN; STEINBACH; KUMAR, 2009).

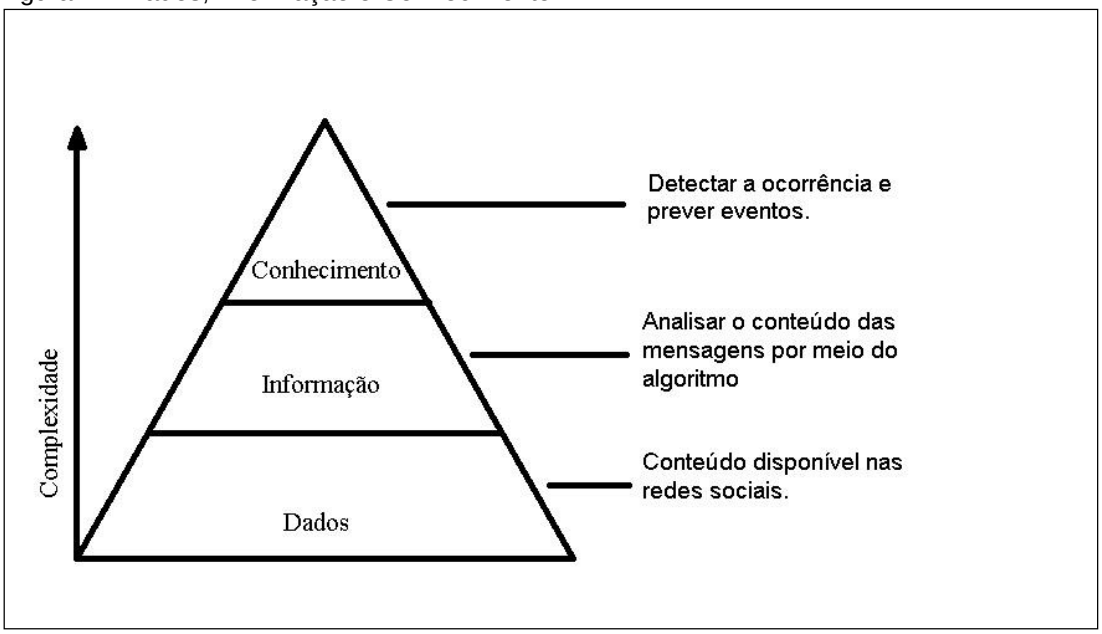
Existem formas distintas de se avaliar o erro de generalização, mas em comum a importância está em determinar o erro de treinamento e saber avaliar que o erro de treinamento não tem relação com o erro de generalização que é a taxa de erro nos dados de teste, sendo que o valor pode ser bem diferente (TAN; STEINBACH;

KUMAR, 2009).

4.1.1 Técnicas para classificar as publicações da rede social

Conforme Golschmidt e Passos (2005), o classificador baseado em regras é um processo para dispor os dados utilizando um conjunto de regras “se...então”, sendo as regras da amostra simuladas de forma separada. Por meio de algoritmos computadorizados, é possível separar, classificar e agrupar os dados permitindo a análise da informação. Na figura 4 tem-se após a etapa de coleta dos dados, é seguida a etapa de processamento da informação, diferenciando as etapas da hierarquia de dados, informação e conhecimento.

Figura 4 – Dados, Informação e Conhecimento.



Fonte: Adaptado de Goldschmidt e Passos (2005).

A mineração nas RSO é um processo de conhecimento em que o usuário interage com a coleção de textos previamente coletada da rede social, visando extrair informações uteis por meio da identificação de padrões. O texto é explorado por meio dos sistemas de mineração de texto, que por sua vez são espelhados na mineração de dados de tal maneira que possuem semelhanças com o pré-processamento e os algoritmos para descoberta do conhecimento, além de ferramentas que auxiliam na visualização dos resultados. Mas existe uma diferença no tratamento de dados não estruturados e dados estruturados. Quando o dados está estruturado, espera-se que

os dados já tenham sido armazenados de forma estruturada, diferente dos dados em forma de texto que demanda aplicação de operações que identificam e tratem a linguagem natural, transformando o texto que são dados não estruturados em um formato intermediário, a tabela atributo-valor (FELDMAN; SANGER, 2007, tradução nossa). Após a etapa de pré-processamento a base de texto coletada está preparada para ser aplicada nas técnicas de aprendizado de máquina, nesse trabalho o método de maquinas de suporte vetorial será utilizado na fase para classificar as mensagens em duas categorias.

4.1.2 Máquinas de suporte vetorial

Support Vector Machines (SVM) é uma técnica de classificação fundamentada na teoria de aprendizagem com recursos da estatística (TAN; STEINBACH; KUMAR, 2009). É vasta a área de aplicação dessa técnica com êxito, como exemplo a classificação de imagens, o reconhecimento de áudio e o diagnóstico de doenças com análise e previsões na área de bioinformática (LUTSA et al, 2010; VAPNIK, 1998). Também é encontrado o emprego da técnica em áreas como na classificação automática de proteínas, o método para a predição de estruturas terciarias a partir de suas estruturas primarias (BISOGNIN, 2007). Outra aplicação a área da agropecuária, na qual, a partir de algumas medições, é possível prever a exata evolução do peso dos animais (ALONSO; VILLA; BAHAMONDE, 2015). E no processamento de imagens, com a segmentação de cores (WANGA; WANGA; BU, 2011), e até na previsão do risco de acidentes, com a avaliação dos dados do trânsito em tempo real (ABDEL-ATY, 2013). É utilizado, ainda, na identificação do público-alvo nas redes sociais, como um instrumento de customização de marketing, auxiliando no entendimento dos clientes que publicam e comentam a respeito de uma marca (LO; CHIONG; CORNFORTH, 2015).

Diante do amplo alcance e aplicação a técnica tem despertado a atenção, firmando-se o aproveitamento do SVM na tarefa de classificação dos dados não estruturados, como o modelo dos textos contidos nas redes sociais online, que é base para esse trabalho.

Para iniciar as tarefas com a vastidão de textos nas mídias sociais, é preciso encontrar uma pequena parte do todo. A função objetiva utiliza como solução uma otimização convexa com algoritmos competentes. O SVM controla a capacidade,

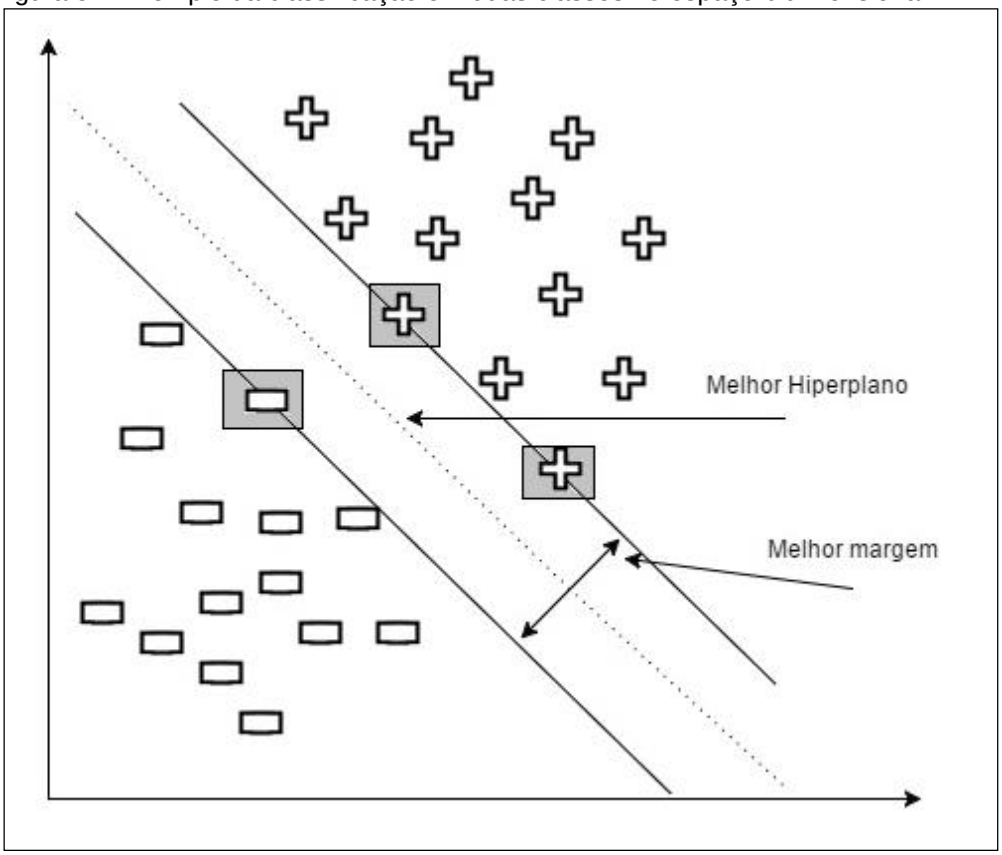
aumentando o limite de decisão. É preciso fornecer os parâmetros, um para um atributo de dados com sua categoria e introduzir variáveis testadas para o valor do atributo, com sua categoria de dados (TAN; STEINBACH; KUMAR, 2009).

Para descobrir se uma mensagem postada no Twitter tem relação com um caso da doença dengue ou é apenas um comentário ou campanha publicitária que não tem analogia com a ocorrência da doença, o método SVM apresentará o resultado da separação entre as mensagens de ocorrência da enfermidade e as mensagens que não têm relação com um caso de dengue. Segundo Vapnik (1998), o SVM é aplicado em diferentes áreas inclusive para categorização de texto.

4.1.3 O Support Vector Machine aplicado na tarefa de categorização das mensagens do Twitter

O Support Vector Machine (SVM) tem recebido atenção por parte de pesquisadores na área acadêmica principalmente na tarefa de mineração de dados. A técnica possui funções de classificação em categorias, regressão e funções de *ranking* (RankSVM). Existem duas propriedades especiais do SVM. A primeira é a alta generalização, porque maximiza a margem; segunda é a propriedade de suporte para aprendizado das funções não lineares incorporado ao *kernel trick*. O SVM utiliza as características da geometria para extrair e calcular o melhor hiperplano e separar os dados do treinamento (STITSON et al, 1996). O melhor hiperplano entre os grupos de dados obtém uma margem maximalizada conforme a figura 5. A margem é a distância do hiperplano até o ponto de dados mais próximo. O procedimento não generaliza o problema, ele apenas minimiza o erro, tentando encontrar a melhor margem, que é a que tem a maior separação entre a linha que classifica e os pontos dos dados (DILRUKSHI, ZOYSA; CALDERA, 2013).

Figura 5 – Exemplo da classificação em duas classes no espaço bidimensional.



*Os vetores de suporte estão destacados com quadrado cinza.
Fonte: Adaptado de Cortes e Vapnik (1995).

O SVM foi desenvolvido por Vapnik e ampara a teoria de aprendizado estatístico, estabelecendo os princípios necessários para a obtenção de uma generalização de qualidade. A ideia do SVM principiou-se em 1992, por Boser, Guyon, e Vapnik na conferencia COLT - *Computational Learning Theory*, o *Support Vector Machines* está ligado com o aprendizado de máquina supervisionado, equipado com métodos de classificação e regressão. Pertencem ao grupo geral de classificador linear e constitui um instrumento de previsão embasado pela teoria de aprendizado estatístico, para maximizar a precisão da previsão (VAPNIK, 1998).

SVM foram desenvolvidos inicialmente para a classificação e depois estendidos para a regressão. A técnica consiste num classificador binário em que a saída da função é positiva ou negativa, mas a classificação multiclasse pode ser implementada por meio da combinação de múltiplos classificadores binários, fazendo o acoplamento de pares. SVM binários são classificadores que discriminam pontos de duas categorias de dados. Cada ponto de dados é representado por um vetor de n dimensões e pertence a apenas uma classe. Existe mais de uma linha que separa o grupo de dados no hiperplano. Com a finalidade de conseguir a separação máxima

entre as duas classes, o SVM escolhe o hiperplano que tem a maior margem, que é o somatório da distância mais curta entre o ponto de separação até o ponto de dados mais próximo de ambas as categorias. O método faz o mapeamento de espaço de entrada para o ajustamento mais adequado do problema de classificação. Na prática é importante escolher o hiperplano que garanta a maior distância, formando uma estrada que separa os dados evitando as classificações erradas (VAPNIK, 1998, tradução nossa).

Entre as várias etapas da implementação do algoritmo SVM, o *mapping* é que transforma o espaço dos atributos dos dados num espaço multidimensional. O objetivo do *mapping* é permitir a separação linear dos dados. A segunda etapa, *features*, é a definição do hiperplano, que é construído a partir de alguns dos atributos dos dados, após a etapa de *mapping*. No processo *feature selectionI* ocorre a seleção desses atributos (VAPNIK, 1998, tradução nossa).

5 TRABALHOS CORRELATOS

A tarefa de mineração com bases nos textos publicados na internet tem apresentado um campo de pesquisa com resultados admiráveis. Alvo de estudos atuais que buscam a obtenção de modelos probabilísticos e através da classificação de qualquer evento que seja de relevância para a obtenção de conhecimento, pode ser utilizado nas publicações das redes sociais, assim como os parâmetros de espaço e tempo (SAKAKI, 2010).

Na área acadêmica, é crescente o número de pesquisas desenvolvidas e essas apresentam excelentes resultados. Os resultados, em alguns casos, são utilizados em aplicações práticas, bem como o modelo proposto no Observatório da Dengue (http://www.observatorio.inweb.org.br/dengue), criado pela equipe do Departamento de Ciência da Computação da UFMG e do Departamento de Bioquímica e Imunologia – ICB/UFMG, vinculado ao Instituto Nacional de Ciência e Tecnologia em Dengue e ao Instituto Nacional de Ciência e Tecnologia para a Web, em parceria com Ministério da Saúde e a Organização Pan-Americana de Saúde (OPAS). No Observatório da Dengue, que foi desenvolvido como um sistema de vigilância online baseado nos dados publicados em redes sociais, blogs entre outras fontes da internet, os textos, coletados em tempo real, oferecem informações com o propósito de prever focos dos casos da doença e informar sobre a situação nas cidades brasileiras (GOMIDE et al, 2011).

5.1 PESQUISAS DE OPINIÃO EM REDES SOCIAIS

Esta dissertação foi apresentada ao Instituto de Ciências Exatas (ICE) da Universidade Federal de Minas Gerais (UFMG) - Departamento de Ciência da Computação, por Renato Miranda Filho, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação, na cidade de Belo Horizonte, no estado de Minas Gerais, no ano de 2012. No trabalho, o autor cria uma estrutura teórica e ferramental que permite realizar pesquisas de opinião por meio de dados coletados em redes sociais e validar sua eficácia diante dos métodos tradicionalmente utilizados, como entrevistas e enquetes realizadas por instituições renomadas. Para identificar

usuários *Spammers*[1], foi utilizado métodos supervisionados, o classificador SVM não linear, utilizando a ferramenta *easy*, equipada com o pacote libSVM e também foi utilizado a ferramenta Weka. Do mesmo modo os resultados mostram que as pesquisas em redes sociais podem, sim, ser um complemento aos métodos tradicionais (MIRANDA FILHO, 2012).

5.2 UM SISTEMA PARA VIGILÂNCIA EPIDEMIOLÓGICA UTILIZANDO AS REDES SOCIAIS

Este artigo desenvolvido por D. Brito, J. Gomide, W. Santos, W. M. Junior, A. V. e V. Almeida e foi apresentado no Simpósio Brasileiro de Bancos de Dados (SBBD), no ano de 2012. Segundo os autores, para detecção de eventos por meio das redes sociais, faz-se necessário artifícios dinâmicos, levando em conta o contexto da doença dengue, que ocorre em regiões de ambiente tropical, onde encontra se o transmissor, que é o mosquito *Aedes Aegypti* infectado. Quando ocorre um surto, a quantidade de casos é conhecida, porém com um atraso, adiando as providências que necessitam ser tomadas pelas autoridades. Para tornar o planejamento mais ágil para controle da infestação, o trabalho propôs o emprego do Classificador Associativo sob Demanda, do inglês *Lazy Associative Classification* (LAC), que confronta regras partindo de um treino prévio, feito manualmente. O classificador rotula em cinco categorias: experiência pessoal, paródia ou ironia, opinião, informação e campanha. Baseando-se nas informações expressas no Twitter é analisada a tendência estimada da incidência do surto de dengue nas cidades com mais de cem mil habitantes (BRITO et al, 2012).

5.3 PREVENDO A TENDÊNCIA DA GRIPE USANDO DADOS DO TWITTER

Este artigo escrito por H. Achrekar, A. Gandhe, R. Lazarus; S. Yu e B. Liu, publicado em Institute of Electrical and Electronics Engineers (IEEE) INFOCOM 2011 - IEEE Conference on Computer Communications Workshops, pp. 702--707. IEEE no ano de 2011. Com o desafio de prever a ocorrência da gripe H1N1 utilizando mensagens coletadas do Twitter nos anos de 2009 e 2010, utilizando as mensagens

[1] Spammers são usuários que enviam spam, ou seja, são pessoas que enviam diversas mensagens para outros usuários.

postadas, para prever e rastrear as ocorrências de gripe H1N1. A dificuldade atual do monitoramento dos Centros de Prevenção e Controle de Doenças (CPCD) é a demora de uma a duas semanas entre o tempo que o paciente é diagnosticado e o momento em que os dados são disponibilizados em relatório com as informações do número de casos da doença.

Foi possível averiguar que o Twitter fornece uma avaliação mais próxima das ocorrências de gripe, de tal modo que oferece uma alternativa significativa para as equipes da área da saúde atuarem no controle das epidemias de gripe.

O método constitui de conjuntos diferentes de dados que compõe o conjunto de textos explorados no Twitter inicialmente e o conjuntos de *retweetes*, com os dados recolhidos foi calculado a correlação entre esses conjuntos de dados e os dados do CPCD, Concluinram que o modelo apresentado no experimento considera modelos estatísticos que o torna capaz de prevê com antecedência as informações, alguns exemplos são, o número de sintomas da gripe, e o número de casos esperados para a próxima semana (Achrekar et al, 2011) .

5.4 UM MONITOR DE DENGUE COM MENSAGENS COLETADAS NO TWITTER

Este artigo escrito por J. Gomide, A. Veloso, W. M. Junior, V. Almeida, F. Benevenuto, F. Ferraz e M. Teixeira e publicado em Proceedings of the 3rd International Web Science Conference no ano de 2011, teve por objetivo monitorar a propagação de uma doença e entender como ela se espalha, por meio do número de registros nas redes sociais, no período em determinada localização. A ferramenta proposta fornece essa informação de forma rápida e sensível às variações, sendo esse o desafio a ser superado pelas agências de saúde. Normalmente os setores públicos levam semanas para ter esses resultados, acarretando na demora para adotar medidas que combata a proliferação da doença. As mídias sociais online como o Twitter são uma fonte contínua de informações, pois ali as pessoas se comunicam sobre os mais diversos assuntos, entre esses, relacionados à dengue. A pesquisa indica uma vigilância sobre os casos de incidência da dengue, considerando quanto, quando e onde, assim, dimensionando, os casos por meio das redes sociais (GOMIDE et al, 2011).

Para realizar esse sistema de vigilância ativo da dengue, utilizou-se quatros parâmetros: Volume, que representa a quantidade de *tweets* mencionando a palavra

dengue; localização, que é a informação geográfica associada a um *tweet* que menciona o assunto; tempo, que se refere a quando foi publicado; e, por último, o sentimento contido na publicação, que classifica as publicações em cinco categorias: Experiência Pessoal, quando o autor expressa um caso de dengue; Sarcástico, quando é irônico ou uma piada envolvendo o tema; Parecer, que são opiniões sobre o fato relacionado; Recursos, que são informações; e Marketing, que são campanhas de prevenção. Assim, concluindo que o Twitter mostra potencial para emitir relatórios com o resumo semanal da situação da doença em cada cidade (GOMIDE et al, 2011).

6 O CONHECIMENTO DAS OPINIÕES PÚBICAS NAS REDES SOCIAIS

Para a viabilidade desta pesquisa foram realizados os procedimentos que compreendem: a coleta das publicações na rede social, processamento e preparação do texto, para, em seguida, proceder-se a classificação desse conteúdo por meio do método de aprendizagem de máquina SVM. E por meio desse, confirmar o funcionamento do método para classificar o texto das opiniões contidas nas redes sociais em duas categorias. Essas, consistem em: publicações de texto que apresenta relação com uma ocorrência da doença; e as que não contêm relação com uma ocorrência real da doença.

Seguindo esse processo, optou-se por utilizar uma base de dados colhida na rede social Twitter das mensagens publicadas em tempo real. Para realizar a coleta foi desenvolvido o *software* Coletor de Dados, que permitirá a recolher as mensagens e armazenar em banco de dados relacional. O *software* Coletor de Dados permite que o usuário informe a palavra-chave, assim, a aplicação somente coleta as mensagens de texto que contenha esse termo.

A próxima etapa é filtrar e preparar as mensagens coletadas. A realização de tal procedimento será por meio do *software* Weka, utilizaou-se as implementaç~eos SMO e LibSVM que estão disponíveis *software* Weka. Então realiza-se a demonstração dos indicadores obtidos com o método SVM para separação das mensagens em duas categorias que constitui conforme a tabela 2 mensagens que têm a relação com um caso da doença e as que não têm relação com um caso de dengue.

Tabela 2. Mensagens a respeito da dengue

Classe: Não tem relação com um caso da doença.	**Classe: Tem relação com um caso da doença.**
"sorte q meu whatsapp não entra água, se não ia tá tendo foco de dengue já, de tão parado"	"Dps q eu sarar dessa dengue, vou passar repelente até na unha do dedin do pé"
"RT @EricaErijs123: hj eu to igual o mosquito da dengue, cheio de casos por aí"	"Vou ir no medico amanhã dnv , minha febre voltou agora , concerteza e dengue ! Tnc ."
"Até a dengue tem casos e eu não"	"Os particulares estão na mesma situação por causa da epidemia de Dengue,não aguento mais sentir dor ☹"
"Piauí mantém redução de casos de dengue, diz Sesapi"	"essa dengue podia sarar logo"
"BH hj tá mais parado que a água da dengue"	"A dengue me pegou, tentar escapar, não consegui "
"aqui ta tão parado que to com medo de dá dengue kkkkkkk"	"Com dengue, Dinho Ouro Preto posta foto em hospital de SP: Que roubada"
"criando meu próprio foco de dengue pra infectar as inimiga"	"Estou com dengue, minhas férias começaram daquele jeito"
"Pra mim todo mosquito que existe no mundo é da dengue"	"Minha tia tá com dengue"
"Cá estou eu assistindo uma palestra sobre dengue mas pensando no festival de londrina e quão louca eu vou ficar ?"	"não desejo dengue pra ninguém, é horrível, ta tudo muito difícil pra mim"

Fonte: Do Autor.

Conforme a separação demostrada na tabela 2, pretende-se que por meio da aplicação do método SVM classificar e chegar ao número de mensagens coletadas que realmente têm relação com um caso da doença.

6.2 METODOLOGIA

Para a realização deste trabalho acadêmico foram empregadas as seguintes etapas metodológicas: levantamento bibliográfico; coleta de dados publicados na rede social Twitter; armazenamento dos dados em banco de dados local; geração do arquivo texto, filtragem e processamento do texto, conversão do

arquivo para usar no *software* Weka, demonstração do método de aprendizagem de máquina SVM para realizar a classificação das mensagens publicadas na rede social; validação dos resultados obtidos.

No decorrer da etapa de levantamento bibliográfico, a atenção principal foi o entendimento dos temas de redes sociais, breves dados sobre a doença em questão e mineração de dados em redes sociais, assim compondo a escrita da fundamentação teórica.

6.2.1 coleta e armazenamento das mensagens

A base de dados utilizada na presente pesquisa contém o texto das publicações. Essa por sua vez necessita ter as seguintes premissas: um grande número de usuários, e ser constantemente atualizada com mensagens de texto, também é preciso permitir a coleta das publicações em tempo real. Entre as redes sociais mais utilizadas no Brasil está o Twitter, que, conforme já referido, é uma rede social onde as pessoas publicam mensagens de texto curtas sendo constantemente utilizada para expressar situações do dia a dia. Além disso possui *Application Programming Interface* (API), que permite coletar as mensagens em tempo real[2]. Para o desenvolvimento da aplicação Coletor De Dados utilizou-se a linguagem de programação Java e a *Integrated Development Environment* (IDE), NetBeans[2] que é um ambiente de desenvolvimento integrado para essa linguagem. A API é disponibilizada pelo Twitter e possui a documentação[3] completa para o desenvolvimento na linguagem de programação Java, bem como em outras linguagens.

Para o armazenamento das mensagens, optou-se por utilizar o Sistema de Gerenciamento de Banco de Dados (SGBD), *HyperSQL DataBase*[4] (HSQLDB).esse SGBD é gratuito e possui portabilidade. Apesar de ser um banco leve possui recursos exigidos para o bom funcionamento da aplicação coletor de dados.

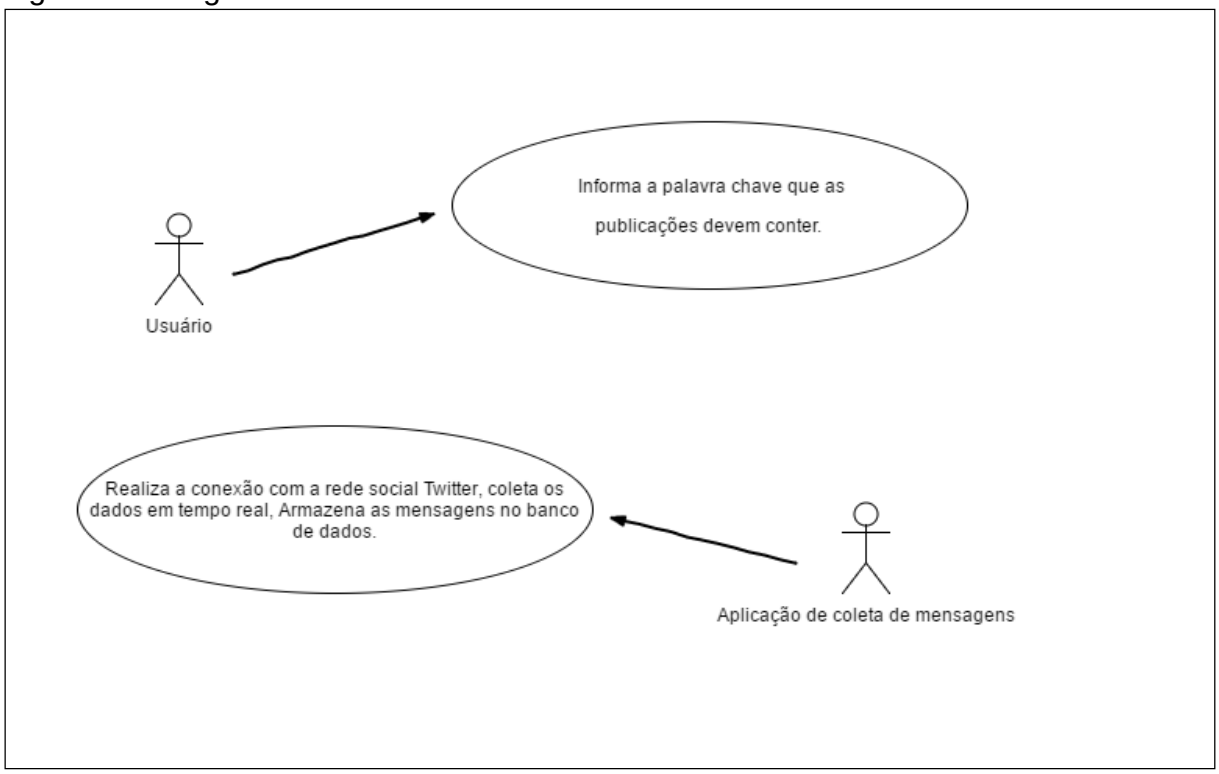
---

[2] Disponível para download no seguinte endereço eletrônico: http://twitter4j.org/javadoc/index.html
[3] Disponível para download no seguinte endereço eletrônico: http://netbeans.org.
[4] Disponível para download no seguinte endereço eletrônico: http://hsqldb.org.

6.2.2 Desenvolvimento da aplicação Coletor de Dados

O desenvolvimento do *software* foi utilizado o Diagrama de Casos de Uso, que tem o objetivo demonstrar de forma objetiva a utilização do *software*. Um diagrama de caso de uso descreve um cenário que mostra as funcionalidades da aplicação do ponto de vista do usuário (GUEDES, 2008). O usuário deve ver no diagrama de Casos de Uso as principais funcionalidades de seu sistema (figura 6).

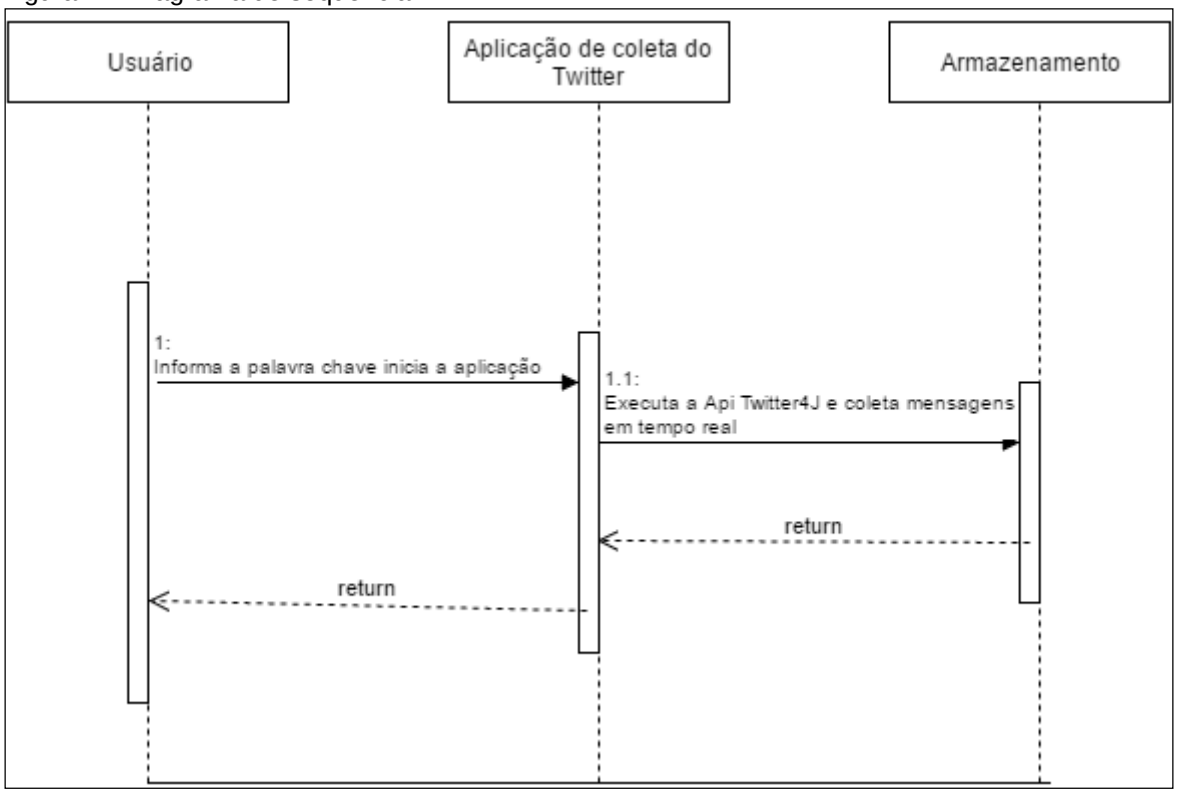
Figura 6 – Diagrama de Casos e Usos.



Fonte: Do autor.

Também esboçou-se o diagrama de sequência que, segundo Guedes (2008) tem o objetivo de mostrar como as interações entre os objetos e as mensagens trocadas no decorrer do tempo para a realização de uma tarefa no contexto (GUEDES, 2008). Na aplicação o usuário deve informar um termo e clicar no botão confirmar. Então o software realiza por meio da API Streaming uma conexão com a rede social Twitter e obtém em tempo real as mensagens que contém o termo indicado pelo usuário. A cada mensagem publicada o sistema armazena no banco de dados e o usuário é informado do conteúdo da mensagem, o que confirma o sucesso da realização da coleta (Figura 7).

Figura 7 – Diagrama de sequência.

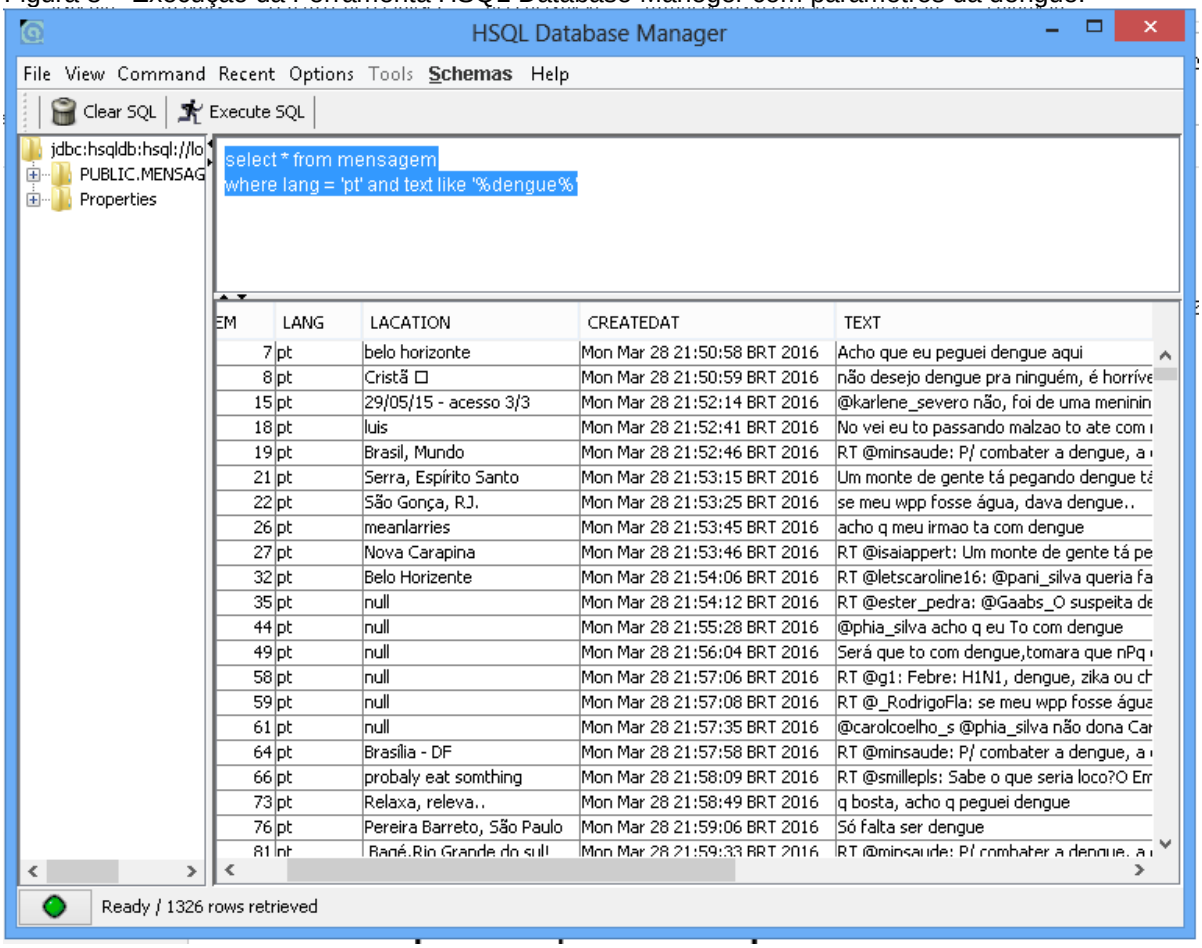


Fonte: Do autor.

O SGBD é onde as mensagens serão gardadas *HyperSQL DataBase* (HSQLDB) é um banco de dados relacional que utiliza a linguagem de consulta padrão o *Structured Query Language* (SQL), o HSQLDB foi desenvolvido na linguagem de programação java, tem como características possuir um pequeno servidor que executa várias *threads* simultaneamente, é transacional tendo alta taxa de atualização que proporciona a possibilidade de trabalhar com um amplo volume de dados, suas tabelas podem ser armazenadas em disco ou memória, além de possuir a ferramenta que permite a operação por meio de linhas de comando em SQL. O HSQLDB suporta os recursos de SQL padronizado conforme o comitê americano, *American National Standards Institute* (ANSI), na versão SQL-92 [3]. Por meio do *HSQL Database Maneger*, conforme ilustrado na figura 8, onde é realizado os comandos SQL. Ele é um *software* livre para uso e distribuição, com base na licença Berkeley Software Distribution (BSD), compatível com as licenças de código aberto,

[3] O padrão SQL foi implementado pala International Business Machines (IBM) e definido como padrão pela primeira vez em 1986 na versão SQL-86 pelo ANSI, um ano depois a ISO também reconheceu o modelo como padrão. Com novos recursos especificado na versão SQL-92 é esse é um padrão amplamente utilizado pelos SGBD.

Figura 8 – Execução da Ferramenta HSQL Database Maneger com parâmetros da dengue.



Fonte: Do Autor.

O modelo implementado visa manter da melhor forma as informações importantes para armazenar as mensagens de texto da RSO, utilizando o banco de dados e recursos de linguagem padrão SQL para criação das tabelas, inserção dos registros e a realização consultas. Na implementação, foi adotada a nomenclatura da tabela de – mensagem - A figura 9 demonstra a entidade mensagem e os atributos que formam a base de dados.

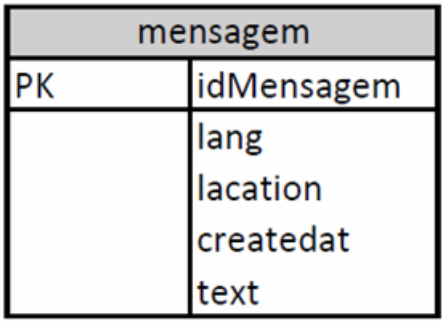
Figura 9 – Campos e requisitos da tabela que armazena as mensagens coletadas.

Campo	Tamanho	Tipo	Descrição
idMensagem	9	inteiro	Número de identificação da mensagem
lang	2	string	Linguagem do texto
lacation	50	string	Localização fornecida pelo usuário do Twitter
createdat		data/hora	Data e hora em que foi publicada a mensagem
text	150	texto	O texto da mensagem

Fonte: do Autor.

Para armazenar as mensagens da RSO foi determinado que os campos que são relevantes para esse trabalho a figura 10 demostra as entidades.

Figura 10 – Modelo entidade da tabela.



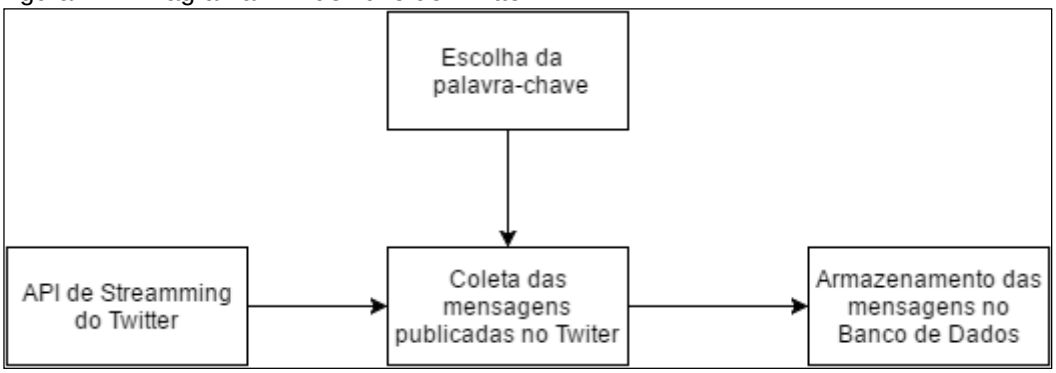
Fonte: Do autor.

Além do campo que guarda o conteúdo das publicações da rede social, também optou-se por guardar outras informações utes para o trabalho tendo em vista que para garantir as informações da pesquisa é necessário ter os parâmetros: a linguagem em que a mensagem foi escrita o local onde foi publicado essa mensagem, a data e hora que tem elevada importância para comparação com dados do boletim epidemiológico da secretaria de vigilância em saúde do ministério da saúde.

6.2.4 Twitter Streaming API

A API de fluxo do Twitter permite a integração com a aplicações desenvolvidas em Java para monitorar dados públicos desta rede, a Twitter4j é uma biblioteca que permite a integração com a aplicações desenvolvidas em Java. Para se utilizar, a mesma deve ser incluída no arquivo twitter4j-core-4.0.4.jar no *classpath* da aplicação, assim o IDE Netbens reconhece as classes da api que permite conexão e coleta dos dados em tempo real da rede social Twitter, como pode ser visto na ilustração da figura 11, onde a aplicação recebe do usuário a palavra chave para filtrar as mensagens do Twitter e conecta-se por meio da API de Streamming efetuando o recolhimento das publicações e armazena no bando de dados.

Figura 11 – Diagrama API de fluxo do Twitter.



Fonte: Do autor.

Para realizar a conexão com a API de *Streamming* do Twitter é necessário os parâmetros *AuthConsumerKey*, *AuthConsumerSecret*, *AuthAccessToken*, *AuthAccessTokenSecret*, que são fornecidos no site da rede social mediante o cadastro no qual é provido o ingresso na rede, com o devido acesso, encontra-se a área destinada a desenvolvedores, e nesta se obtém as informações das aplicações e os códigos utilizados na autenticação.

6.3 PRÉ-PROCESSAMENTO DE TEXTOS

O pré-processamento é uma das fases da mineração de textos. O texto contido nas redes sociais precisa ser estruturado na forma de tabela de atributo-valor e assim, possibilitar a utilização de textos em algoritmos de aprendizado de máquina. A partir do número de palavras, independente de regras, são contados e esse valor determina o peso de cada atributo.

Tabela 3. Exemplo da tabela de Atributo Valor.

Mensagem	pacient	morad	mosquit	grip	...
t1	0,33	0,5	0,4	0,5	...
t2	0,2	0,3	0,33	0,3	...
t3	0	0,4	0	0,1	...
t4	0,4	0,6	1	0,2	...
t5	1	0,1	0,22	0	...
t6	0,6	0	0,33	1	...
tn	...	...	...	...	...

Fonte: Do autor.

Na tabela, cada postagem do Twitter, indicada pela letra t, contém um peso para cada atributo, que são as palavras da mensagem. Por essa razão, antes de gerar

a tabela valor-atributo, é preciso atenção especial na fase de pré-processamento, a fim de tratar e reduzir a grande quantidade de palavras existentes no vocabulário, evitando que a tabela contenha um número elevado de atributos. A seguir, serão descritas as funções utilizadas na fase de pré-processamento de texto.

6.3.1 Funções de pré-processamento de texto

O *Word Tokenization* é um processo que limita o texto, separando em palavras ou frases dependendo da implementação. Mas para utilização em textos curtos, como o da rede social Twitter, optou-se por organizar em palavras tendo como delimitador os espaços. Também escolheu-se por remover os sinais de pontuação, caracteres especiais e os números, uma vez que esses dados isoladamente fornecem pouca relevância para a finalidade de pré-processamento no texto da rede social.

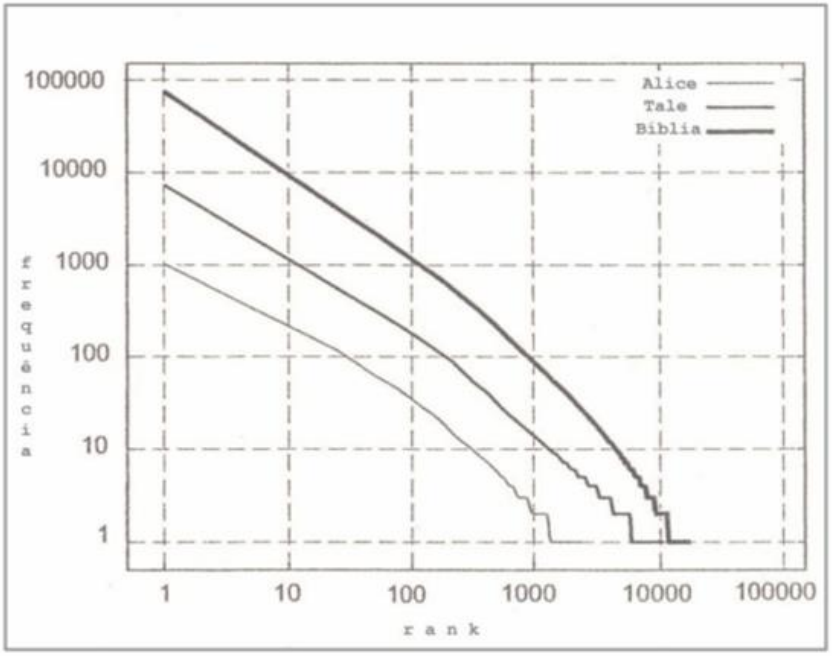
Outra função, a *stemming,* está diretamente ligada ao idioma a qual o texto foi publicado na rede social. Para realizar a normalização linguística, é preciso encontrar o radical de dada palavra do texto e, assim, as variações da palavra são reduzidas à forma comum, reduzindo bem a quantidade de *tokens.* O algoritmo mais utilizado para *stemming* é o Porter, que foi desenvolvido para a língua inglesa (PORTER, 2006). Para a língua portuguesa, existem algoritmos que foram adaptados do algoritmo de Porter, e devido às palavras com significados diferentes, ocorrem erros chamados de *over-stemming*, sendo que em alguns casos, os *tokens* não serão reduzidos ao mesmo radical morfológico. Para este trabalho foram consideradas apenas mensagens postadas no idioma português. Todas as mensagens que possuem o termo escolhido são coletadas e armazenadas, mas na geração do arquivo texto é aplicado o filtro para mensagens escritas na língua portuguesa.

No *Stop Words* é gerado uma lista de palavras que não são significativas para o algoritmo de aprendizado quando consideradas isoladamente, como por exemplo, artigos, preposições, pronomes e conjunções. Essas palavras devem ser removidas do texto *e*, consequentemente, reduzindo o número de *tokens*.

A distribuição de termos conforme Konchady (2006), é utilizada para determinar as palavras com pouca representação no conjunto de mensagens coletadas da RSO. Termos que têm pouca frequência não têm representação para análise dos textos, assim como as palavras que ocorrem com muita frequência que também não têm grande valia para o conjunto de mensagens, uma vez que estão

presentes na maioria das publicações. George Kingsley Zipf propôs baseado na experiência com termos e a frequência que esses aparecem no texto pesquisado, os métodos que reduz o esforço computacional para realizar a mineração de texto, que resumidamente consiste na contagem e determinação de pesos tratando em especial as palavras que raramente são usadas e as que são usadas constantemente por que essas apresentam menor relevância para os algoritmos que classificam os textos em categorias essa fórmula ficou reconhecida como a Lei de Zipf. Com ajustes nas pontas dos valores, é possível obter a relação entre a frequência da ocorrência das palavras e a quantidade que ocorrem no conjunto de textos (KONCHADY, 2006). A figura 12 exemplifica a escala fundamentada na lei de *Zipf*.

Figura 12 – Distribuição de termos segundo a lei de Zipf*



*****Exemplos: A Bíblia, Alice e Tale.
Fonte: Konchady (2006).

O método *Term frequency – Inverse Document Frequency* (Tf-IDF) consiste na frequência de uma palavra para determinar o nível de importância para o texto, com o intuito de diminuir o peso de palavras frequentes, e um fator que acresce ao longo do conjunto de mensagens coletadas. A fórmula para o TF-IDF onde w é a palavra e d é a mensagem que contém a palavra (FELDMAN; SANGER, 2007, tradução nossa).

$$\boldsymbol{TF - IDF\_Peso}\ (\boldsymbol{w}, \boldsymbol{d}) = \boldsymbol{TermFreq}(\boldsymbol{w}, \boldsymbol{d}).\log(\frac{\boldsymbol{N}}{\boldsymbol{DocFreq}(\boldsymbol{w})})$$

Onde, TermFreq(w, d) é a frequência da palavra nas mensagens, N é o

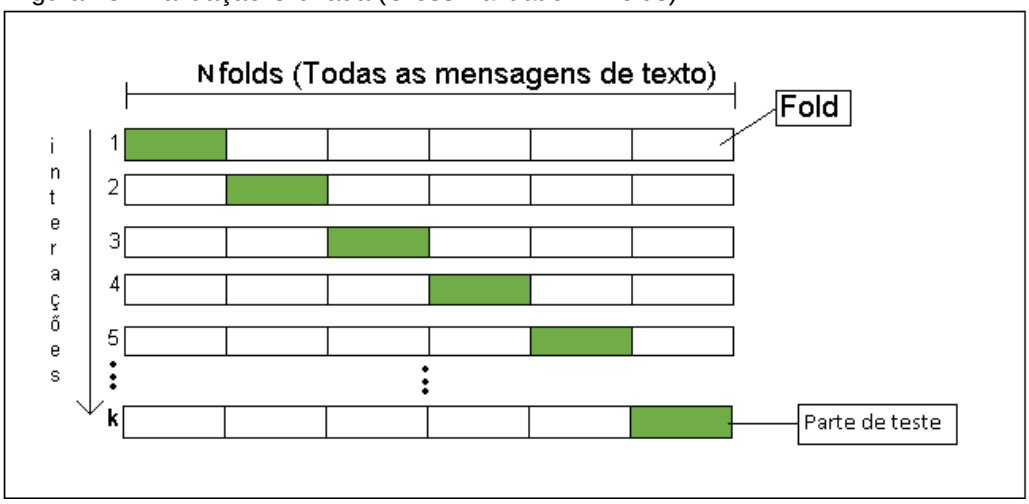
número total de mensagens, e DocFreq(w) é o número de mensagens que contém a palavra w.

6.3.2 Base de treino

Em métodos supervisionados é necessário criar a base de treino, é uma base de conhecimento com um modelo treinado, antes de executar o método de classificação é necessário gerar a base de treino por uma pessoa que identifique e diferencie para classificar as mensagens de acordo com o conteúdo nas classes determinadas, nesse caso será utilizado duas classes sendo elas “pos” para quando a mensagem ter o conteúdo relacionado com um caso da doença e “neg” quando a mensagem não tem relação com um caso ocorrido da doença.

Devido à grande quantidade de mensagens coletadas não é viável classificar manualmente todas as mensagens, para determinar a quantidade de mensagens de treino utilizou-se a técnica de validação cruzada. Nessa técnica a amostragem é feita de forma a garantir que cada classe seja devidamente representada em ambos os conjuntos de treino. Com o procedimento os dados são selecionados aleatoriamente para treino, e o restante é usado para teste. As taxas de erro das diferentes interações são calculadas para se obter uma taxa de erro global (WITTEN; FRANK; HALL, 2011, tradução nossa).

Figura 13 - Validação Cruzada (*Cross Validation* n-folds)



Fonte: do autor.

A base de dados é dividida aleatoriamente em n partes e as classes em cada dobra são representados com a mesma proporção dos dados originais. Em

seguida, o modelo de classificação é treinado em n-1 partes, e da parte restante é usado para testar o modelo. O procedimento é repetido n vezes, utilizando como dados de teste cada um dos n partes apenas uma vez. Os resultados de teste são n finalmente em média, para produzir uma estimativa global (D'ANDREA, DUCANGE, LAZZERINI, 2015, tradução nossa).

6.3.3 matriz de confusão e Indicadores do classificador.

A matriz de confusão apresenta um apanhado das ocorrências das mensagens que foram classificadas, proporcionando uma medida efetiva do modelo de classificação, muito utilizada em algoritmos supervisionados como o caso do SVM, também conhecida como matriz de erro, as linhas representam a classe real e as colunas a classe prevista (WITTEN; FRANK; HALL, 2011 Tradução nossa).

Figura 14 - Matriz de confusão Classe Prevista e Classe Real

		Classe Prevista	
		Positivo	Negativo
Classe Real	Positivo	Verdadeiro positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Adaptado de WITTEN; FRANK; HALL (2011).

Na figura 14 observa-se que na matriz de confusão o valores são simbolizados por:

Na figura 14 observa-se que na matriz de confusão o valores são simbolizados por:

Verdadeiro Positivo (VP) são os resultados previstos como positivo e o algoritmo verdadeiramente classificou como positivo;

Verdadeiro Negativo (VN) são os resultados negativos que o classificador verdadeiramente identificou como negativo;

Falso Positivo (FP) são os resultados em que é esperado o resultado positivo mas foi classificado como negativo;

Falso-Negativo (FN) são os resultados que é esperado o resultado negativo

mas o classificador identificou como positivo;

Conforme Deng et al. (2016) as medidas de desempenho de um classificador pode ser definida com base na matriz de confusão.

Acurácia é a proporção do número total das previsões corretas.

$$Acur = \frac{VP + VN}{VP\ + FP + FN + VN}$$

Precisão é a medida da previsão de uma classe especifica como foi prevista.

$$Prec = \frac{VP}{VP+FP}$$

Sensibilidade (*Recall)* é a capacidade do teste classificar como positivo uma informação que realmente é positiva.

$$Sens = \frac{VP}{VP + FN}$$

Especificidade (*Specificity*) é a capacidade do teste classificar como negativa uma informação que realmente é negativa.

$$Spec = \frac{VN}{VN + FP}$$

F-Score ou *F-measure* é a média ponderada da precisão e recall

Na tabela 4 é apresentado as equações descritas nessas medias.

$$F\text{-}score = \frac{2 \times Precision\ \times Recall}{Precision+Recall}$$

Tabela 4. Indicadores Estatísticos.

Descrição	Equação
Acurácia (Accuracy)	$Acur = \frac{VP + VN}{VP + FP + FN + VN}$
Precisão (Precision)	$Prec = \frac{VP}{VP+FP}$
Sensibilidade (Recall)	$Sens = \frac{TP}{TP + FN}$
Especificidade (Specificity)	$Spec = \frac{TN}{TN + FP}$
F-score	$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision+Recall}$

Fonte: Adaptado de D'Andrea; Ducange; Lazzerini (2015).

6.3.4 O ambiente para análise Weka, as bibliotecas LibSVM e SMO, e o parametro RBFKernel.

Waikato Environment for Knowledge Analysis (Weka), em português Waikato Ambiente para a análise do conhecimento desenvolvido pela Universidade de Waikato na Nova Zelândia[4], é um software que fornece diversas implementações de pré-processamento de dados e algoritmos de aprendizado, O sistema é escrito em java ,é possível analisar e alterar o seu código fonte uma vez que é distribuído sob os termos da *General Public License* (GNU)[5] (WITTEN; FRANK; HALL, 2011, tradução nossa). O ambiente possui uma interface gráfica chamada Explore que dá acesso as funções. É possível ler o conjunto de dados de um arquivo *Attribute Relation File Format (ARFF)* que é a extensão nativa do Weka e consiste na lista de atributos e os valores dos atributos sendo cada instância separada por vírgula, para separar o conjunto de dados é usado a *tag @relation* para as informações do conjunto de dados, *@attribute* para as informações de atributos e *@data* com os dados. Os pacotes LibSVM e Sequential Minimal Optimization(SMO) estão disponibilizados no software, e foram escolhidos para analisar os resultados do classificador SVM.

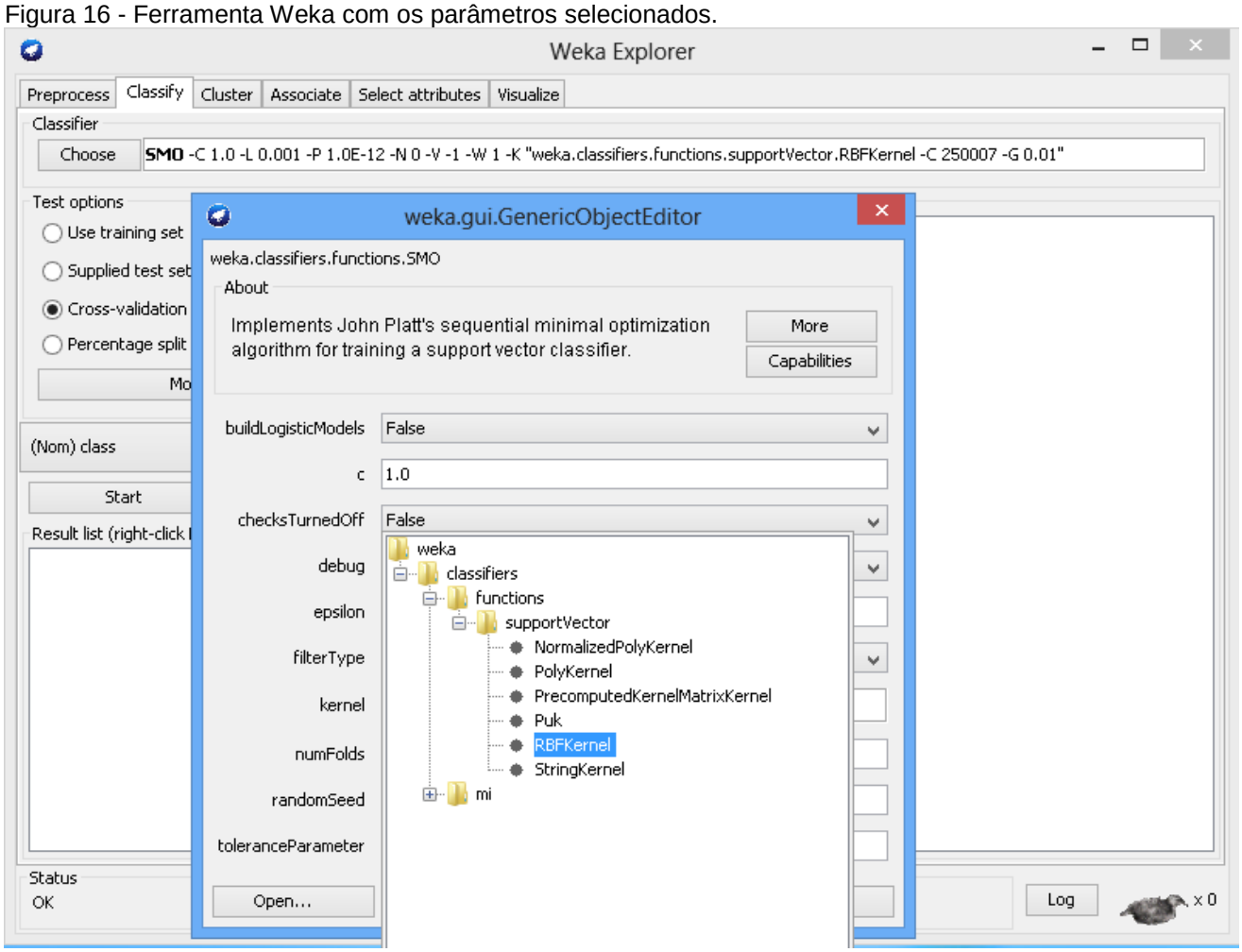
[4] O símbolo do software é a Mecca (uma ave curiosa que não voa nativa da Nova Zelândia).
[5] Disponível para download no seguinte endereço eletrônico: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

A biblioteca LibSVM[6] implementa o SVM que permite a utilização para classificação e também permite selecionar no parâmetro o kernel *radial basis function* (RBF), a LibSVM envolve duas etapas, primeiro a formação de um conjunto de dados para a obtenção do modelo, e segundo um modelo para prever a informação do conjunto de dados. As funções suportadas como estimativas de classificação com funções necessária para utilização do SVM (Chang; Lin, 2013, tradução nossa).

Outra implementação do algoritimo SVM disponibilizada no software Weka é o *Sequential Minimal Optimization* (SMO), é uma implementação simplificada e rápida do algoritmo SVM, os vetores de suporte utilizando as funções do kernel como o *radial basis function* (RBF) escolhido no parâmetro do software Weka, na função implementada do SMO os valores em falta são substituídos globalmente, os atributos nominais são transformados em binários, e os atributos são normalizados pelo valor do coeficiente na saída dos dados, optando pela solução otimizada a cada interação, buscando os melhores valores, utiliza a heurística e o método analítico para resolver os multiplicadores de Lagrange, com o método próprio para encontra o parâmetro da margem de decisão (Platt, 1998, tradução nossa).

[6] Disponível para download no endereço: http://www.csie.ntu.edu.tw/~cjlin/libsvm
Documentação da implementação: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf

Figura 16 - Ferramenta Weka com os parâmetros selecionados.



Fonte: Do autor.

Para utilização do Classificador SVM é necessário indicar no parâmetro o kernel, existe diferentes opções disponível no Weka para esse trabalho será escolhido o kenelRBF que é muito citado na literatura, e tem obtido ótimos resultados em utilizações na área acadêmica (WITTEN; FRANK; HALL, 2011, tradução nossa).

Na tela de parametrização deve ser informado o parâmetro C que determina o ponto máximo e mínimo de margem para o equilíbrio do erro de classificação e melhorar a acurácia (Yadava et al, 2015, tradução nossa).

6.4 RESULTADOS OBTIDOS

Finalizada a etapa de coleta dos dados da rede social Twitter por meio da software desenvolvido Coletor de Dados, realizou-se as consultas via SQL meio da ferramenta HSQL *Database Maneger* para extrair os relatórios; geração do arquivo com o texto das mensagens; pré-processamento do texto; transformação no arquivo

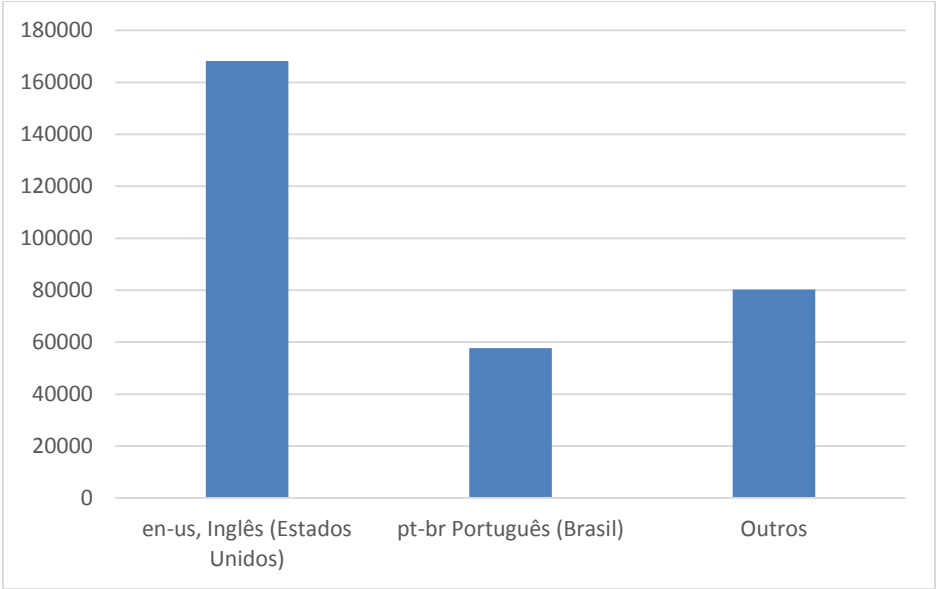
com extensão Attribute Relation File Format (ARRF), extensão utilizada no WEKA; aplicação do método SVM para classificação das mensagens que tem ou não relação com um caso da doença; e a comparação com os registros do Boletim Epidemiológico Secretaria de Vigilância em Saúde do Ministério da Saúde.

Para a realização da etapa de coleta de dados com o software Coletor de Dados foi necessário um microcomputador com processador intel I3,4GB de memória RAM e sistema operacional Windows7, com acesso à internet de 10MB dedicada, e uma vez que a etapa de coleta de dados exige alta disponibilidade, durante o período da coleta de dados utilizou-se um nobreak de 600VA, para evitar interrupções por falta de energia. Na etapa de geração dos relatórios e aplicação do método que identifica se as mensagens fazem parte da classe que tem relação com um caso das doenças foi utilizado um notebook com processador Intel dual core, 2GB de memória RAM e sistema operacional Windows 7.

6.4.1 Consulta das mensagens Coletadas e geração do arquivo para o classificador

No período de 08 de março de 2016 até 22 de abril de 2016 foram coletadas 295985 mensagens de texto do Twitter que contém as palavras chave dengue, zika, Chikungunya no idioma em que a mensagem foi postada, e sendo essas, 47597 das mensagens estão no idioma português brasileiro.

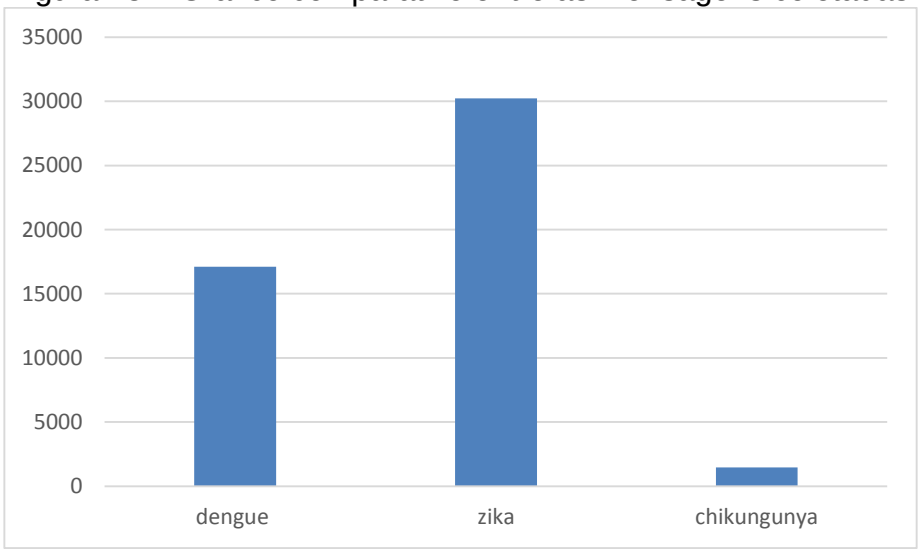
Figura 17 – Gráfico das mensagens coletadas sobre dengue, zika e chikungunya no período de 08/04 até 22/04



Fonte: Do Autor.

Observa-se que a grande maioria das mensagens coletadas estão no idioma inglês nesse trabalho será tradado apenas os casos que ocorreram no brasil, utilizou-se filtros em comandos SQL para retornar apenas mensagens no idioma pt-br e foram exportados em arquivo texto com extensão txt. Obteve 3 arquivo sendo um das mensagens que contem a palavra dengue com 17116 mensagem, zika com 30288 mensagens e chikungunya com 1469 mensagens.

Figura 18 – Gráfico comparativo entre as mensagens coletadas.



Fonte: Do Autor.

Depois da exportação para o arquivo no formato texto é necessário transformar o arquivo para tipo arff antes de aplicar a função.

Figura 19 – Imagem das declarações no arquivo no formato arrf.

```
@relation twitter

@attribute mensagem string
@attribute class {pos, neg}

@data

'vc ta  dor  cabeca: culpa  celulardor  olhos: culpa  cellresfriado: culpa  celldengue: culpa  cell todamaedizque'        neg
'vespera  aniversrio e to  dengue kkkkk maravilhoso'              pos
'rt mulherdacabello: vc ta  dor  cabeca: culpa  celulardor  olhos: culpa  cellresfriado: culpa  celldengue: culpa  cell'   neg
'anne t  dengue vou ter  ir l amanha urgente'              pos
'itapiuna tem  notificacoes  casos suspeitos  dengue zika virus e chikungunya. clique  imagem e saiba + https:t.commfeblis'   neg
'vo ter q ir  hospital  a  ve  to  dengue :'     pos
'swiftlacrei peguei dengue   e vc'              pos
'bradesco   cidade e amanha( pirapora)  n vou poder ir to  dengue'          pos
'to  medo  ta  dengue :'            pos
'nn sei   cnsg rir   musicas  mosquito  dengue kkkkkkkkkkkkkkkkkkkkk'      neg
```

Fonte. Do autor.

O arquivo contendo a base de teste foi dividido em 10 partes iguais, de forma aleatória. Dessas 9 partes são utilizadas para o teste e uma é usada como base

de treino, o processo é repetido por 10 vezes, sendo cada uma das partes usada como treino.

6.4.2 Classificação das Mensagens de Texto.

No software WEKA utilizou-se duas diferentes bibliotecas, para realizar os testes com SVM, a SMO e a LibSVM, conforme descrito na seção 6.3.3, são bibliotecas implementadas disponíveis no software Weka, realizou-se o teste de classificação das mensagens de texto, separando em duas classes, sendo “pos” indicando positivo, para as mensagens que tem no conteúdo a menção de um caso da doença, e “neg” que indica negativo, referenciando as mensagens que não tem em seu conteúdo o assunto de um caso da doença. Executou-se as funções SMO e LibSVM utilizando os seguintes parametros:

Variável de custo C = 1.0 (Peso para identificar as classificações incorretas)

O parâmetro Épsilon foi ajustado para 10^{-3} (Taxa de erro de arredondamento)

O Kernel utilizado foi o radial basis function (RBF)

O primeiro teste realizado com o arquivo que contém as mensagens de texto com menção a doença dengue, obteve-se a seguinte tabela de confusão do teste realizado usando os parâmetros descritos anteriormente.

Teste realizado com a função SMO resultou na tabela de confusão (figura):

Figura 20 - tabela de confusão sobre o alvo dengue por meio da função SMO

	Positivo	Negativo
Positivo	5858	719
Negativo	46	10493

Fonte: Do Autor.

A partir dos valores da tabela de confusão é possível calcular as métricas:

Calculo da acurácia:

$$Acur = \frac{VP + VN}{VP + FP + FN + VN} = \frac{5858 + 10493}{5858 + 46 + 719 + 10493} = 0{,}955305$$

Calculo da precisão:

$$Prec = \frac{VP}{VP+FP} = \frac{5858}{5858+46} = 0{,}992208$$

Calculo da sensibilidade (Recall):

$$Sens = \frac{VP}{VP\ +\ FN} = \frac{5858}{5858+719} = 0{,}89067$$

Calculo da especificidade:

$$Spec = \frac{TN}{TN+FP} = \frac{10493}{10493+46} = 0{,}995918$$

Calculo do *F-score* ou *F-Measure:*

$$F\text{-}score = \frac{2 \times Precision\ \times Recall}{Precision+Recall} = \frac{2 \times 0{,}99221\ \times 0{,}89067}{0{,}99221+0{,}89067} = 0{,}93870$$

Repetindo o teste com o mesmo arquivo que contém a menção a doença dengue nas mensagem agora realizado com a função LibSVM resultou na tabela de confusão (figura 20):

Figura 21 - tabela de confusão sobre o alvo dengue por meio da função LIBSVM.

	Positivo	Negativo
Positivo	5206	1371
Negativo	2	10537

Fonte: Do Autor.

Calculo da acurácia:

$$Acur = \frac{VP + VN}{VP + FP + FN + VN} = \frac{5206 + 10537}{5206 + 2 + 1371 + 10537} = 0{,}919783$$

Calculo da precisão:

$$Prec = \frac{VP}{VP+FP} = \frac{5206}{5206+2} = 0{,}999616$$

Calculo da sensibilidade (Recall):

$$Sens = \frac{VP}{VP+FN} = \frac{5206}{5206+1371} = 0{,}791546$$

Calculo da especificidade:

$$Spec = \frac{VN}{VN+FP} = \frac{10537}{10537+2} = 0{,}999810$$

Calculo do *F-score* ou *F-Measure:*

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision+Recall} = \frac{2 \times 0{,}999616 \times 0{,}791546}{0{,}999616+0{,}791546} = 0{,}883449$$

Na tabela 5 é demonstrado os resultados obtidos, podemos observar os resultados da função SMO e a LibSVM quando aplicado o mesmo arquivo de teste contendo os textos relacionados com dengue,

Tabela 5. Indicadores do classificador, textos relacionados com dengue

Função	Acurácia	Precisão	Sensibilidade	Especificidade	F-score
SMO	0,955305	0,992208	0,89067	0.995918	0,9387
LIbSVM	0,919783	0,999616	0,791546	0,99981	0,883449

Fonte: Do Autor.

No teste demonstrado na tabela 5 a função SMO obteve melhores resultados para acurácia, sensibilidade e f-score enquanto a LibSVM apresentou resultados superiores para precisão e especificidade.

O teste foi repetido agora com o arquivo de teste contendo menções a palavra zika utilizando as funções SMO e LibSVM demonstrado respectivamente na figura 21 para função SMO e figura 22 para função LibSVM

Figura 22 - tabela de confusão sobre o alvo zika por meio da função SMO.

	Positivo	Negativo
Positivo	4960	1524
Negativo	205	23599

Fonte: Do Autor.

Figura 23 - tabela de confusão sobre o alvo zika por meio da função LIBSVM.

	Positivo	Negativo
Positivo	4824	1421
Negativo	19	24024

Fonte: Do Autor.

A partir dos valores exibidos na tabela de confusão obteve-se os resultados demonstrados na tabela 6 relacionado a palavra zika.

Tabela 6. Indicadores do classificador, textos relacionados com Zika.

Função	Acurácia	Precisão	Sensibilidade	Especificidade	F-score
SMO	0,952456	0,996077	0,772458	0,999209	0,870130
LIbSVM	0,942915	0,960310	0,764960	0,993880	0,851575

Fonte: Do Autor.

Os resultados expostos na tabela 6 observa-se que a função SMO obteve melhores resultados em todos os indicadores quando comparado a função LibSVM.

Um novo teste utilizando os mesmos parâmetros de entrada, foi realizado para as mensagens com conteúdo relacionado a chikungunya, resultando na tabelas representadas na figura 23 e 24.

Figura 24 - tabela de confusão sobre o alvo chikungunya por meio da função SMO.

	Positivo	Negativo
Positivo	270	67
Negativo	6	1120

Fonte: Do Autor.

Figura 25 - tabela de confusão sobre o alvo chikungunya por meio da função LIBSVM.

	Positivo	Negativo
Positivo	253	73
Negativo	1	1139

Fonte: Do Autor.

A partir da tabela confusão foi calculado os indicadores com resultado apresentado na tabela 7.

Tabela 7. Indicadores do classificador, textos relacionados com Chikungunya.

Função	Acurácia	Precisão	Sensibilidade	Especificidade	F-score
SMO	0,950103	0,978261	0,801187	0,994671	0,880914
LIbSVM	0,949523	0,996063	0,776074	0,999123	0,872414

Fonte: Do Autor.

Na tabela 7 observa-se que para chikungunya a função SMO conseguiu melhor desempenho para acurácia, sensibilidade e f-score, enquanto LibSVM apresenta melhor resultado para precisão e especificidade.

Os testes demonstram que os resultados do método SVM, nas implementações SMO e LibSVM, possui alta precisão. Entretanto essa técnica apresenta menor sensibilidade. a função SMO apresentou maior quantidade de indicadores com resultados superior a LibSVM, sendo considerado o modelo que

mostrou-se melhor que a LibSVM. Para as doenças dengue, zika e chikungunya o resultado encontrado da classificação foi que na maior parte das mensagens é verdadeiro negativo. O que indica que a maioria das mensagens coletadas para cada uma das doenças foram classificadas como não relacionadas a um caso da doença.

Do mesmo modo que outras pesquisas nessa área. Os resultados indicam que o texto proveniente das redes sociais pode ser trabalhado de forma a fornecer uma base de dados para realização de experimentos com algoritmos de aprendizado de máquina.

7 CONCLUSÃO

O constante avanço da tecnologia e das redes de dados permite maior disponibilidade para acesso à internet, e diariamente o acervo das mais diversas fontes de publicações da rede é alimentado. Tem-se discussões acerca de vários assuntos, e grande parte é expressa nas redes sociais, tornando esse um espaço com potencial para ser explorado com técnicas computacionais, permitindo a extração de conhecimento como no caso das publicações na rede social sobre o assunto relacionado com as doenças dengue, zika, chikungunya, onde grande parte das publicações são cômicas no intuito de fazer o comentário engraçado e também publicidades e campanhas de prevenção das doenças, são textos que não tem relação com um caso da doença, mas a classificação das publicações que tem em seu conteúdo a relação com um caso da doença é importante para a tomada de decisões e planos de ação.

A tarefa de extração do conhecimento das redes sociais é composto por etapas de coleta de dados da rede social, armazenamento das publicações e classificação por meio de algoritmos de aprendizagem de máquinas supervisionadas, sendo que nessa pesquisa acadêmica aprofundou-se no entendimento das redes sociais, no desenvolvimento da aplicação Coleta de Dados que coleta em tempo real as publicações do Twitter. A fase que demandou o maior esforço e tempo foi a etapa de pré-processamento de texto. Após essa fase já é possível por meio do software Weka criar a tabela de valor atributo a qual é utilizada para realizar a classificação das mensagens de texto em duas classes que são “pos” que indica positivo caso a mensagem tenha em seu conteúdo o assunto relacionado com um caso das doenças, ou “neg” caso o conteúdo não tem relação com uma ocorrência da doença.

Durante a realização da pesquisa encontrou-se dificuldades, dentre as quais destaca-se a etapa de pré-processamento do texto ocupando o maior tempo na mineração de dados sendo a mais importante para a obtenção de uma base adequada para a exploração, as funções de separação das palavras em tokens, o algoritmo Porter que permite encontrar o radical das palavras, a remoção de palavras que não são significativas para o algoritmo de aprendizagem, o método *Term frequency Inverse Document Frequency* que determina peso para selecionar as palavras que são relevantes e tem importância para o texto. O método de classificação *support vector machine* é supervisionado requisito a criação da base de treino para teste do

algoritmo, a técnica de validação cruzada do treino n etapas foi utilizada para garantir que cada classe seja representada no conjunto de dados, o procedimento os dados são selecionados aleatoriamente para o treino e o restante é utilizado para o teste a aplicação das 3 bases as quais contém separadamente os textos com menção as doenças dengue, zika, chikungunya foi aplicada no software Weka para análise das duas bibliotecas disponíveis para o *support vector machine* que são a SMO e a LibSVM.

Apesar das dificuldades encontradas no decorrer da pesquisa, principalmente na fase de pré-processamento devido a quantidade de caracteres especiais utilizados na rede social. Ainda conforme demonstram os resultados obtidos por meio dos testes realizados o SVM realizou a classificação das mensagens coletadas e pré-processadas com sucesso, apresentando o correto funcionamento, uma vez que se atingiram os objetivos propostos para a pesquisa.

Por fim, são apresentadas algumas sugestões de trabalhos futuros, tendo por base o conhecimento adquirido por meio da pesquisa:

A) desenvolver o estudo de um sistema híbrido do classificador SVM junto a outro classificador a fim de obter melhores resultados da separação das classes na base de dados proveniente da rede social sobre as mensagens que contém menção aos casos das doenças dengue, zika e chikungunya.

B) implementar na aplicação Coletor de Dados as funções de pré-processamento e exportação da base de dados, com o objetivo de facilitar e reduzir o tempo dispensado na aplicação dos algoritmos de classificação em bases de texto provenientes das redes sociais.

c) Estudar outros classificadores para obter resultado mais preciso sobre as mensagens que contenham no assunto a relação com casos das doenças dengue, zika e chikungunya.

d) Disponibilizar na aplicação Coletor de Dados a opção de coleta de publicações de outras redes sociais.

e) Avaliar a existência de correlação dos dados obtidos da rede social com menção aos casos das doenças, com os dados oficiais provenientes do ministério da saúde sobre a incidência das doenças dengue, zika e chikungunya.

REFERÊNCIAS

ABDEL-ATY, Yu. Utilizing support vector machine in real-time crash risk evaluation, **Accident Analysis and Prevention**, 2013, p. 252– 259.

ACHREKAR, Harshavardhan; GANDHE, Avinash; LAZARUS, Ross; YU, Ssu-Hsin; LIU, Benyuan. **Predicting** Networking Systems, 2011.**Flu Trends using Twitter Data.** The Firts International Workshop on Cyber-Physical

ALONSO J., VILLA A., BAHAMONDE A.Improved estimation of bovine weight trajectories using Support Vector Machine Classification, **Computers and Electronics in Agriculture**, 2015, p. 36–41.

ANTUNES, M. N. et al. Monitoramento de informação em mídias sociais: o e-Monitor Dengue. **TransInformação, Campinas**, 26(1): p.9-18, jan./abr., 2014 Disponível em:<http://www.scielo.br/>. Acesso em: 04 out. 2015.

ARANHA C., PASSOS E. A tecnologia de mineração de textos.**RESI-Revista Eletrônica de Sistemas de Informação**, n. 2, 8f, 2006. Disponível em:<http://revistas.facecla.com.br/ >. Acesso em: 22 set. 2015.

BISOGNIN, Gustavo. **Utilização de Máquinas de Suporte Vetorial Para Predição de Estruturas Terciárias de proteínas**. Ciências exatas e Tecnológicas da UNISINOS, São Leopoldo, 2007.

BRASIL Ministério da Saúde. Secretaria de Vigilância em Saúde. Diretoria Técnica de Gestão. **Dengue**: diagnóstico e manejo clínico: adulto e criança / Ministério da Saúde, Secretaria de Vigilância em Saúde, Diretoria Técnica de Gestão. – 4. ed. – Brasília : Ministério da Saúde, 2013.

Saiba mais sobre dengue, chikungunya e zika. 2016. Disponível em: <http://www.brasil.gov.br/saude/2016/01/saiba-mais-sobre-dengue-chikungunya-e-zika>. Acesso em: 19 maio 2016.

BRITO, Denise et al. **Um sistema de alarme para vigilância epidemiológica de rumores utilizando redes sociais.** Simpósio Brasileiro de Bancos de Dados – SBBD, Universidade Federal de Minas Gerais 2012.

CASTELLS, Manuel; CARDOSO, Gustavo. **A sociedade em rede do conhecimento à ação política**. Centro Cultural de Belém, 2005.

CORDEIRO, João Paulo da Costa. **Extracção de elementos relevantes em texto/páginas da World Wide Web**. 2004. 174 f. Tese (Mestrado em Inteligência Artificial e Computação). Faculdade de Ciências da Universidade do Porto - Departamento de Ciência de Computadores, Porto, Portugal.

CORRÊA, Ulisses. Mineração de dados de help desk usando Rattle – o caso da Petrobras. 2007. 92 f. Dissertação (Mestrado em Administração). Faculdade De

Economia E Finanças Ibmec Programa de Pós-Graduação e Pesquisa em Administração e Economia, Rio de Janeiro.

CORTES, C.; VAPINIK, V.N. **Support-vector networks**, Mach. Learn. 20, 1995, p. 273–297.

D'Andrea E., Ducange P., Lazzerini. Real-Time Detection of Traffic From Twitter Stream Analysis, IEEE Transactions on On Intelligent Transportation Systems, 2015.
DILRUKSHI, I.; ZOYSA K. ; CALDERA, A. Twitter News Classification Using SVM. **The 8th International Conference on Computer Science & Education**, 2013.

FELDMAN, Ronen; SANGER, James. **The text mining handbook:** advanced approaches in analyzing unstructured data. New York, 2007.

GOLDSCHMIDT Ronaldo, PASSOS Emmanuel**. Data Mining**: um guia prático. Rio de Janeiro: Campus, 2005.

GOMIDE, J. et al. **Dengue surveillance based on a computational model of spatio-temporal locality of twitter.**Em*Proceedings of the 3rd International Web Science Conference*, WebSci '11, pp. 3:1--3:8, New York, NY, USA. ACM. 2011.

GUEDES, Gilleanes T. A.. UML: uma abordagem prática. 3. ed. São Paulo: Novatec, 2008.

KONCHADY, Manu. **Text mining application programming.** Massachusetts: Charles River Media, 2006.

LEITE, E.M. Análise da correlação entre dengue e indicadores sociais a partir do sig. **HYGEIA**, v.6, n.11, p 44-59, 2010. Disponível em: <www.hygeia.ig.ufu.br/>. Acesso em: 27 set. 2015.

LEITE, George Salomão; LEMOS, Ronaldo. **O marco civil da Internet**. São Paulo: atlas, 2014.

LO, Siaw Ling; CHIONG, Raymond; CORNFORTH, David. **Using Support Vector Machine Ensembles for Target Audience Classification on Twitter**. Plos One, Vol. 10(4), April 2015.

LUTSA, Jan et al. A tutorial on support vector machine-based methods for classification problems in chemometrics, **AnalyticaChimicaActa** 665, 2010, p. 129–145.

MACHADO, Aydano P. Mineração de texto em redes sociais aplicada à educação a distância. Colabora – Revista digital da CVA – Ricesu, vol. 6, n. 23, jul, 2010.

MARTINO, Luís Mauro Sá. **Teoria das mídias digitais:** linguagens, ambientes, redes. Petrópolis: Vozes, 2014.

MILLER, Thomas W. **Data and text mining**: a business applications approach. New Jersey: Pearson Prentice Hall, 2005.

MIRAGEM, Bruno. Aspectos característicos da disciplina do comércio eletrônico de consumo: comentários ao Dec. 7.962, de 15.03.2013. **Revista de Direito do Consumidor,** São Paulo, v. 22, n. 86 , p.287-300,, abr. 2013.

MIRANDA FILHO, Renato. **Um arcabouço para pesquisas de opinião em redes sociais.** Belo Horizonte, 2012.

PAESANI, Liliana Minardi. **Direito e Internet**: liberdade de informação, privacidade e responsabilidade civil. 6. ed. São Paulo: Atlas, 2013.

PARK J. LEE, G., A robust algorithm for text region detection in natural scene images. **Can. J. Elect. Comput. Eng**., v.. 33, n. 3/4, Summer/Fall, 2008.

PEREIRA, Sara; PEREIRA, Luís; PINTO, Manuel. **Internet e redes sociais**. São Paulo: Centro de Estudos de Comunicação e Sociedade, 2011.

PESSANHA, José Eduardo Marques. . Avaliação do Plano Nacional de Controle da Denguen. **Cadernos de Saúde Pública,** Rio de Janeiro , v.25, n.7 , p.1637-1641, jul. 2009.

PINHEIRO, Carlos André Reis**. Inteligência analítica**: mineração de dados e descoberta do conhecimento. Rio de Janeiro: Ciência Moderna, 2008.

PORTER, M. F. .Analgorithm for suffixstripping. **Program**: electronic library and information systems 40 (3), 211–218, 2006.

RECUERO, Raquel. **Redes Sociais na internet**. 2 ed. Porto Alegre: Sulina, 2011.

SAKAKI, T.; OKAZAKI, M.; and MATSUO, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In WWW.

SANTOS, B. N. et al. Monitoramento para prever epidemias utilizando redes sociais. **ColloquiumExactarum**, v. 6, n.1, p.65 – 79, Jan-Jun. 2013. Disponível em: <http://revistas.unoeste.br/>. Acesso em 22/09/2015.

SILVA, Edson; SILVA, Elissa Gonçalves de Oliveira. Consequências da ação do vírus da dengue no organismo humano. **Infarma:** informativo profissional do Conselho Federal de Farmácia, Brasília, v.21, n.3/4 , p.24-29, abr. 2009.

SILVEIRA, Sergio Amadeu da. **Cidadania e redes digitais**, 2 ed. São Paulo: Comitê Gestor da Internet no Brasil: Maracá – Educação e Tecnologias, 2010.

STITSON, M. O et al. Theory of support Vector Machines. **Technical Report CSD-TR-96-17**, Computational Intelligence Group, Royal Holloway, University of London, 1996.

TAN, Pan-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining:** mineração de dados**.** Rio de Janeiro: Ciência Moderna, 2009.

TEIXEIRA, Tarcisio. **Comércio eletrônico**: conforme o Marco Civil da Internet e a regulamentação do e-commerce no Brasil. São Paulo: Saraiva, 2015.

TOMAÉL, Maria Inês; ALCARA, Adriana Rosecler; CHIARA, Ivone Guerreiro Di. **Das Redes Sociais à Inovação,** Brasilia, 2006.

VAPNIK, V. N. **Statistical Learning Theory**. New York, USA: John Wiley & Sons, Inc., 1998.

VIEIRA, Valter Afonso; MATOS, Celso Augusto de; SLONGO, Luiz Antônio. Avaliação das relações entre qualidade de serviço do site, satisfação, valor percebido, lealdade e boca a boca por meio de um modelo teórico. **Revista de Administração,** São Paulo , v.44, n.2 , p.131-146, jun. 2009.

WANGA X., WANGA T., BU, E. Color imagesegmentationusing pixel wisesupport vector machine (SVM) classification, **PatternRecognition**, 2011, p. 777–787.

WITTEN, I. H; FRANK, Eibe. **Data mining:** practical machine learning tools and tecniques. 3nd ed San Francisco: Morgan Kaufmann, 2011.

APÊNDICE A - ARTIGO

Mineração de Dados na Rede Social Twitter a Respeito de Casos das Doenças Dengue, Zika e Chikungunya

Inócio Felipe da Costa[1], Paulo João Martins[2]

[1]Acadêmico do Curso de Ciência da Computação – Unidade Acadêmica de Ciências, Engenharias e Tecnologias – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma– SC

[2]Professor do Curso de Ciência da Computação – Unidade Acadêmica de Ciências, Engenharias e Tecnologias – Universidade do Extremo Sul Catarinense (UNESC) – Criciúma– SC

inocio.felipe@gmail.com, pjm@unesc.net

Resumo. *A cada dia com o avanço das tecnologias o acesso as redes sociais aumenta. Todo esse avanço desperta o interesse em revelar o conhecimento sobre o teor das mensagens publicadas nesse meio. Como ocorre na área da saúde que por meio das redes sociais busca-se identificar mensagens que contenham menção com casos das doenças. De tal modo tem-se mensagens publicadas na rede social sobre a dengue, zika, chikungunya. Essas são doenças distintas que possuem o mesmo vetor transmissor, o mosquito aedes aegypti. Este artigo demonstra por meio do método de máquina de suporte vetorial a identificação das mensagens coletada, da rede social Twitter, que contenha relação com casos dessas doenças.*

Abstract. *Every day with the advancement of technology access social networks increases. All this advancement arouses interest in revealing the knowledge of the content of the messages published in this medium. As in health that through social networks seek to identify messages that contain words with cases of disease. So we have posted messages on the social network about dengue, zika, chikungunya. These are different diseases that have the same transmission vector, the mosquito Aedes aegypti. This article demonstrates through support vector machine method of identifying the collected messages, social network Twitter, containing relation to cases of these diseases.*

1. Introdução

A Internet revolucionou a área da computação e comunicações. Ela tornou-se um importante meio de comunicação e de disseminação da informação, onde as pessoas interagem e colaboram independente da sua localização geográfica [Teixeira 2015].

Entre as inovações trazidas pela internet, incluem-se as redes sociais. A rede social online é um espaço que admite a exposição de informação por pessoas ou organizações [Recuero 2011].

Nesse ambiente social é comum o compartilhamento de histórias e experiências [Miragem 2013]. As redes sociais no atual contexto têm a importância de disseminar informações relevantes para a sociedade. A exemplo, pode-se citar o caso da dengue, doença que vem assustando muitas regiões do mundo.

Segundo Brasil (2016), uma das forma de prevenir as doenças dengue, zika e chikungunya é o compartilhamento de informação e a mobilização social para ampliar as medidas de prevenção.

Em meio a todo conteúdo publicado nas redes sociais com texto mencionando as doenças, pode-se extrair conhecimento. O que desperta a atenção de pesquisas a fim de encontrar informações por meio das técnicas de mineração de texto [Pinheiro 2008].

Para realizar a mineração de texto das redes sociais é necessário coletar as publicações, em seguida preparar a base para realizar a classificação desse conteúdo. Entre diversas técnica de aprendizado disponíveis para realizar a classificação de texto tem-se a máquina de suporte vetorial, uma técnica com vasta área de aplicação. Sendo empregado o modelo em dados não estruturados [Tan, Steinbach e Kumar 2009].

Diante disso, apresentam-se como a mineração de dados e as redes sociais podem atuar em eventos da saúde pública, especificamente nas doenças causadas pelo Aedes aegypti.

2. Mineração de Texto nas Redes Sociais

A mineração de texto é a aplicação das técnicas de *Knowledge Discovery in Database* (KDD) em dados não estruturados como as mensagens provenientes das redes sociais. Os dados armazenados em forma de texto demanda tratamento que transforma a base coletada em uma forma estruturada como por exemplo a tabela de valor atributo. Sendo que essa pode ser utilizadas por métodos de classificação como a MSV [Feldman e Sanger 2007].

O processo inicia pela coleta e armazenamento dos dados, em seguida realiza-se o pré-processamento do texto, e a etapa seguinte pode-se aplicar métodos para obter resultados [Pinheiro 2008]. O pré-processamento é a fase que ocupa maior parte do tempo na mineração de texto. Sendo relevante para obtenção de informações seguras [Miller 2005].

Conforme Golschmidt e Passos (2005), por meio de algoritmos computadorizados pode-se classificar os dados transformando em informações. Esse é um processo em que o usuário interage com a coleção de textos visando extrair informações úteis por meio da identificação de padrões.

Diante do amplo alcance e aplicação nas mais diversas áreas a técnica de máquina de suporte vetorial tem despertado a atenção, principalmente na tarefa de classificação de dados não estruturados [Tan, Steinbach e Kumar 2009].

3. Coleta, Armazenamento e Pré-processamento

A rede social utilizada na presente pesquisa necessita conter interações atualizadas constantemente. Essa por sua vez deve ter um grande número de usuários. Também é necessário que a rede social permita a coleta das publicações em tempo real. entre as redes sociais mais utilizadas no Brasil está o Twitter que é conhecida por seus usuários publicarem mensagens de textos curtos. Sendo constantemente utilizada para expressar situações do dia a dia [Vieira, Matos e Slongo 2009].

O Twitter possui a Application Programming Interface (API) Twitter Streaming, que permite a integração com aplicações para desenvolvedores. Para realizar a conexão a API exige os parâmetros AuthConsumerKey, AuthConsumerSecret, AuthAccessToken, AuthAccessTokenSecret. Esses parâmetros são fornecidos no site da rede social mediante o cadastro no qual é provido o ingresso na rede. De modo que encontra-se na área destinada a desenvolvedores. Sendo que no local destinado a desenvolvedores obtêm-se a documentação da API e os códigos utilizados para autenticação.

O SGBD Hyper SQL Data Base (HSQLDB) é onde as mensagens são armazenadas. Esse é um banco de dados relacional e tem como características possuir um pequeno servidor, é transacional e possuir uma ferramenta para operação por meio de linhas de comando em Structured Query Language (SQL). É um software livre, com base na licença Berkeley Software Distribution (BSD), compatível com as licenças de código aberto.

Conforme Konchady (2006), na etapa de pré-processamento tem-se funções diferentes para tratar e reduzir o volume de dados:

word tokenization: é o processo que limita o texto. Sendo que dependendo da implementação esse, é organizado em uma ou mais palavras. Isso torna-se necessário uma vez que a palavra depende de outra para formar o sentido correto;

stemming: consistem em encontrar o radical das palavras. Assim para realizar a normalização linguística as variações da palavra são reduzidas a forma comum;

Stop words: é uma relação de palavras que podem ser removidas do texto. Isso porque, isoladamente, não são significativas para o algoritmo de aprendizado.

Feldman e Sanger (2007) definem no método Term frequency – Inverse Document Frequency (TF-IDF) a importância do tremo de acordo com a ocorrência no conjunto de texto. Constituindo em definir o peso de cada termo segundo a formula:

$$\boldsymbol{TF - IDF\_Peso}\ (\boldsymbol{w}, \boldsymbol{d}) = \boldsymbol{TermFreq}(\boldsymbol{w}, \boldsymbol{d}).\log(\frac{\boldsymbol{N}}{\boldsymbol{DocFreq}(\boldsymbol{w})}))$$

Onde, TermFreq(w, d) é a frequência das ocorrências do temo nas mensagens. N é o número total de mensagens. DocFreq(w) é o número de mensagens que contém a palavra w. para exemplificar supondo que no conjunto de 10 publicações coletada da rede social dessas o termo "peguei" ocorre 1 vez na primeira mensagem e 4 vezes no total em nossa base de exemplo. Assim podemos definir que

Outra função utilizada na fase de pré-processamento do texto é a distribuição dos termos: também conhecido como lei da Zipf (Figura 1). É utilizado para determinar a frequência de palavras no texto. Sendo que palavras que tem pouca frequência não tem representação para os métodos classificadores. Assim como as palavras que ocorrem com muita frequência, também não são importantes para as técnicas de mineração de texto (Konchady 2006).

No método de distribuição de termos é levantado a frequência de ocorrência de cada termo. Então é realizado o ajuste tratando as pontas ou sejas as palavras que ocorrem com muita frequência e as, que raramente aparecem no texto.

Aplicando as funções na base de texto pode-se reduzir o esforço computacional na aplicação das técnicas de mineração de texto.

3. Máquinas de Suporte Vetorial

Support Vector Machines (SVM) é uma técnica de classificação fundamentada na teoria de aprendizagem com recursos da estatística [Tan, Steinbach e Kumar 2009]. É vasta a área de aplicação dessa técnica com êxito, como exemplo a classificação de imagens, o reconhecimento de áudio e o diagnóstico de doenças com análise e previsões na área de bioinformática [Lutsa et al 2010], [Vapnik 1998).

O Support Vector Machine (SVM) tem recebido atenção por parte de pesquisadores na área acadêmica principalmente na tarefa de mineração de dados. A técnica possui funções de classificação em categorias, regressão e funções de ranking (RankSVM).

Existem duas propriedades especiais do SVM. A primeira é a alta generalização, porque maximiza a margem; segunda é a propriedade de suporte para aprendizado das funções não lineares incorporado ao kernel trick. O SVM utiliza as características da geometria para extrair e calcular o melhor hiperplano e separar os dados do treinamento O melhor hiperplano entre os grupos de dados obtém uma margem maximalizada conforme a A margem é a distância do hiperplano até o ponto de dados mais próximo. O procedimento não generaliza o problema, ele apenas minimiza o erro, tentando encontrar a melhor margem, que é a que tem a maior separação entre a linha que classifica e os pontos dos dados [Dilrukshi, Zoysa e Caldera 2013].

4. O Conhecimento das Opiniões Publicas nas Redes Sociais

Para a viabilidade desta pesquisa foram realizados os procedimentos que compreendem: a coleta das publicações na rede social, processamento e preparação do texto, para, em seguida, proceder-se a classificação desse conteúdo por meio do método de aprendizagem de máquina SVM. E por meio desse, confirmar o funcionamento do método para classificar o texto das opiniões contidas nas redes sociais em duas categorias. Essas, consistem em: publicações de texto que apresenta relação com uma ocorrência da doença; e as que não contêm relação com uma ocorrência real da doença.

Tabela 1. Mensagens a respeito da dengue

Classe: Não tem relação com um caso da doença.	**Classe: Tem relação com um caso da doença.**
"sorte q meu whatsapp não entra água, se não ia tá tendo foco de dengue já, de tão parado"	"Dps q eu sarar dessa dengue, vou passar repelente até na unha do dedin do pé"
"Até a dengue tem casos e eu não"	"Os particulares estão na mesma situação por causa da epidemia de Dengue,não aguento mais sentir dor ☹"
"Piauí mantém redução de casos de dengue, diz Sesapi"	"essa dengue podia sarar logo"

Seguindo esse processo, optou-se por utilizar uma base de dados colhida na rede social Twitter das mensagens publicadas em tempo real. Para realizar a coleta foi desenvolvido o software Coletor de Dados, que permitirá a recolher as mensagens e armazenar em banco de dados relacional. O software Coletor de Dados permite que o usuário informe a palavra-chave, assim, a aplicação somente coleta as mensagens de texto que contenha esse termo.

A próxima etapa é filtrar e preparar as mensagens coletadas. A realização de tal procedimento será por meio do software Weka, utilizaou-se as implementações SMO e LibSVM que estão disponíveis software Weka. Então realiza-se a demonstração dos indicadores obtidos com o método SVM para separação das mensagens em duas categorias que constitui conforme a tabela 2 mensagens que têm a relação com um caso da doença e as que não têm relação com um caso de dengue.

4.1 Implementação do Coletor de Dados e o pré-processamento

Desenvolveu-se a aplicação coletor de dados. Nessa, o usuário deve informar um termo e clicar no botão confirmar. Então o software realiza por meio da API Streaming uma conexão com a rede social Twitter e obtém em tempo real as mensagens que contém o termo indicado pelo usuário. Para realizar a conexão com a API de Streamming do Twitter é necessário os parâmetros AuthConsumerKey, AuthConsumerSecret, AuthAccessToken, AuthAccessTokenSecret, que são fornecidos no site da rede social mediante o cadastro no qual é provido o ingresso na rede, com o devido acesso, encontra-se a área destinada a desenvolvedores, e nesta se obtém as informações das aplicações e os códigos utilizados na autenticação.

O modelo implementado armazenar as mensagens de texto da RSO, utilizou-se o banco de dados e recursos de linguagem padrão SQL para criação das tabelas, inserção dos registros e a realização consultas.

O pré-processamento é uma das fases da mineração de textos. O texto contido nas redes sociais precisa ser estruturado na forma de tabela de atributo-valor (Tabela 3).

Tabela 3. Exemplo da tabela de Atributo Valor.

Mensagem	Pacient	Morad	Mosquito	Grip	...
t1	0,33	0,5	0,4	0,5	...
t2	0,2	0,3	0,33	0,3	...
t3	0	0,4	0	0,1	...
t4	0,4	0,6	1	0,2	...
t5	1	0,1	0,22	0	...
t6	0,6	0	0,33	1	...
tn	...	...	...	...	...

A tabela de atributo valor possibilitar a utilização de textos em algoritmos de aprendizado de máquina. A partir do número de palavras, independente de regras, são contados e esse valor determina o peso de cada atributo.

4.2 O ambiente para análise Weka, as bibliotecas LibSVM e SMO

Waikato Environment for Knowledge Analysis (Weka), é um software que fornece diversas implementações de pré-processamento de dados e algoritmos de aprendizado, O sistema é escrito em java ,é possível analisar e alterar o seu código fonte uma vez que é distribuído sob os termos da General Public License (GNU) [Witten, Frank e Hall 2011].

Os pacotes LibSVM e Sequential Minimal Optimization(SMO) estão disponibilizados no software, e foram escolhidos para analisar os resultados do classificador SVM.

A biblioteca LibSVM implementa o SVM que permite a utilização para classificação e também permite selecionar no parâmetro o kernel radial basis function (RBF), a LibSVM envolve duas etapas, primeiro a formação de um conjunto de dados para a obtenção do modelo, e segundo um modelo para prever a informação do conjunto de dados. As funções suportadas como a estimativas de classificação com funções necessária para utilização do SVM [Chang e Lin 2013].

Outra implementação do algoritmo SVM disponibilizada no software Weka é o Sequential Minimal Optimization (SMO), é uma implementação simplificada e rápida do

algoritmo SVM, os vetores de suporte utilizando as funções do kernel como o radial basis function (RBF) [Platt, 1998].

4.2.1 Base de Treino

Devido à grande quantidade de mensagens coletadas não é viável classificar manualmente todas as mensagens, para determinar a quantidade de mensagens de treino utilizou-se a técnica de validação cruzada. Nessa técnica a amostragem é feita de forma a garantir que cada classe seja devidamente representada em ambos os conjuntos de treino. Com o procedimento os dados são selecionados aleatoriamente para treino, e o restante é usado para teste. As taxas de erro das diferentes interações são calculadas para se obter uma taxa de erro global [Witten, Frank e Hall 2011].

Para realizar o treino da base definiu-se o valor de n igual a 10. Sendo assim base de dados foi dividida aleatoriamente em n partes e as classes em cada parte são representados com a mesma proporção dos dados originais. Em seguida, o modelo de classificação é treinado em n-1 partes, e da parte restante é usado para testar o modelo. O procedimento é repetido n vezes, utilizando como dados de teste cada um dos n partes apenas uma vez. Os resultados de teste são n finalmente em média, para produzir uma estimativa global [D'Andrea, Ducange, Lazzerini 2015].

4.2.2 Matriz de Confusão e Indicadores do Classificador

A matriz de confusão apresenta um apanhado das ocorrências das mensagens que foram classificadas, proporcionando uma medida efetiva do modelo de classificação, muito utilizada em algoritmos supervisionados como o caso do SVM, também conhecida como matriz de erro, as linhas representam a classe real e as colunas a classe prevista [Witten, Frank e Hall 2011].

		Classe Prevista	
		Positivo	Negativo
Classe Real	Positivo	Verdadeiro positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 14 - Matriz de confusão Classe Prevista e Classe Real.

Conforme Deng et al. (2016) as medidas de desempenho de um classificador pode ser definida com base na matriz de confusão.

Acurácia é a proporção do número total das previsões corretas.

$$Acur = \frac{TP + TN}{TP + FP + FN + TN}$$

Precisão é a medida da previsão de uma classe especifica como foi prevista.

$$Prec = \frac{TP}{TP+FP}$$

Sensibilidade é a capacidade do teste classificar como positivo uma informação que realmente é positiva.

$$Sens = \frac{TP}{TP + FN}$$

Especificidade é a capacidade do teste classificar como negativa uma informação que realmente é negativa.

$$Spec = \frac{TN}{TN + FP}$$

F-Score ou *F-measure* é a média ponderada da precisão e recall

Na tabela 4 é apresentado as equações descritas nessas medias.

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

5. Resultados Obtidos

No período de 08 de abril de 2016 até 22 de abril de 2016 foram coletadas 295985 mensagens de texto do Twitter que contém as palavras chave dengue, zika, Chikungunya no idioma em que a mensagem foi postada, e sendo essas, 47597 das mensagens estão no idioma português brasileiro.

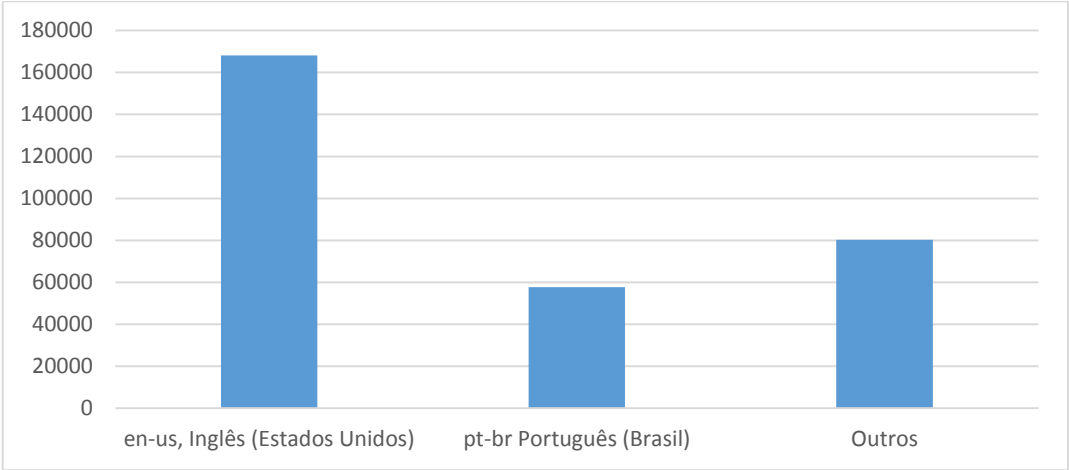


Figura 17 – Gráfico das mensagens coletadas por idioma no período de 08/04 até 22/04

Observa-se que a grande maioria das mensagens coletadas estão no idioma inglês nesse trabalho será tradado apenas os casos que ocorreram no brasil, utilizou-se filtros em comandos SQL para retornar apenas mensagens no idioma pt-br e foram exportados em arquivo texto com extensão txt. Obteve 3 arquivo sendo um das mensagens que contem a palavra dengue com 17116 mensagem, zika com 30288 mensagens e chikungunya com 1469 mensagens.

Executou-se as funções SMO e LibSVM utilizando os seguintes parametros: Variável de custo C = 1.0 (Peso para identificar as classificações incorretas) O parâmetro Épsilon foi ajustado para 10-3 (Taxa de erro de arredondamento) O Kernel utilizado foi o radial basis function (RBF)

O primeiro teste realizado com o arquivo que contém as mensagens de texto com menção a doença dengue, obteve-se a seguinte tabela de confusão do teste realizado usando os parâmetros descritos anteriormente.

Teste realizado com a função SMO resultou nos resultados: VP = 5858; VN = 10493; FP = 46; FN = 719. E para a função LibSVM: VP = 5206; VN = 10537; FP = 2; FN = 2.

Na tabela 5 é demonstrado os resultados obtidos, podemos observar os resultados da função SMO e a LibSVM quando aplicado o mesmo arquivo de teste contendo os textos relacionados com dengue,

Tabela 5. Indicadores do classificador, textos relacionados com dengue

Função	Acurácia	Precisão	Sensibilidade	Especificidade	F-score
SMO	0,955305	0,992208	0,89067	0.995918	0,9387
LIbSVM	0,919783	0,999616	0,791546	0,99981	0,883449

Fonte: Do Autor.

No teste demonstrado na tabela 5 a função SMO obteve melhores resultados para acurácia, sensibilidade e f-score enquanto a LibSVM apresentou resultados superiores para precisão e especificidade.

O teste foi repetido agora com o arquivo de teste contendo menções a palavra zika utilizando as funções SMO resultou-se em: VP = 4960; VN = 23599; FP = 205; FN = 1524. E para a função LibSVM obteve-se os resultados: VP = 4824; VN = 24024; FP = 19; FN = 1421. A partir dos valores exibidos na tabela de confusão obteve-se os resultados demonstrados na tabela 6 relacionado a palavra zika.

Tabela 6. Indicadores do classificador, textos relacionados com Zika

Função	Acurácia	Precisão	Sensibilidade	Especificidade	F-score
SMO	0,952456	0,996077	0,772458	0,999209	0,870130
LIbSVM	0,942915	0,960310	0,764960	0,993880	0,851575

Os resultados expostos na tabela 6 observa-se que a função SMO obteve melhores resultados em todos os indicadores quando comparado a função LibSVM.

Um novo teste utilizando os mesmos parâmetros de entrada, foi realizado para as mensagens com conteúdo relacionado a chikungunya, resultados obtidos para a função SMO: VP = 270; VN = 1120; FP = 6 e FN = 67. E para a função LibSVM: VP = 253; VN = 67; FP = 6; FP = 73.

Tabela 7. Indicadores do classificador, textos relacionados com Chikungunya.

Função	Acurácia	Precisão	Sensibilidade	Especificidade	F-score
SMO	0,950103	0,978261	0,801187	0,994671	0,880914
LIbSVM	0,949523	0,996063	0,776074	0,999123	0,872414

Na tabela 7 observa-se que para chikungunya a função SMO conseguiu melhor desempenho para acurácia, sensibilidade e f-score, enquanto LibSVM apresenta melhor resultado para precisão e especificidade.

Os testes demonstram que os resultados do método SVM, nas implementações SMO e LibSVM, possui alta precisão. Entretanto essa técnica apresenta menor sensibilidade. a função SMO apresentou maior quantidade de indicadores com resultados superior a LibSVM, sendo considerado o modelo que mostrou-se melhor que a LibSVM. Para as doenças dengue, zika e

chikungunya o resultado encontrado da classificação foi que na maior parte das mensagens é verdadeiro negativo. O que indica que a maioria das mensagens coletadas para cada uma das doenças foram classificadas como não relacionadas a um caso da doença.

Do mesmo modo que outras pesquisas nessa área. Os resultados indicam que o texto proveniente das redes sociais pode ser trabalhado de forma a fornecer uma base de dados para realização de experimentos com algoritmos de aprendizado de máquina.

6. Considerações Finais

Este artigo apresentou os dados minerados da rede social Twiter sobre casos das doenças dengue, zika e chikungunya. As publicações da rede social é coletado e armazenado. Com a base formada, o texto passa por uma etapa tratamento. Sendo essa importante para que se alcance resultados confiáveis. Então foi aplicado o método para classificar o conteúdo em duas classes. O método SVM realizou a classificação das mensagens, observou-se que nas implementações SMO e LibSVM a técnica possui alta precisão, mas um valor menor na sensibilidade. Da mesma maneira que outras pesquisas nessa área constata-se que as publicações das redes sociais tem potencial para fornecer uma base de dados para ser utilizados em algoritmos de aprendizado de máquina. Nos testes realizados revelou-se que a maior quantidade de mensagens publicadas na rede social não tem relação com um caso da doença.

7. Referências

BRASIL Ministério da Saúde. Secretaria de Vigilância em Saúde. Diretoria Técnica de Gestão. **Dengue**: diagnóstico e manejo clínico: adulto e criança / Ministério da Saúde, Secretaria de Vigilância em Saúde, Diretoria Técnica de Gestão. – 4. ed. – Brasília: Ministério da Saúde, 2013.

FELDMAN, Ronen; SANGER, James. **The text mining handbook:** advanced approaches in analyzing unstructured data. New York, 2007.

KONCHADY, Manu. **Text mining application programming.** Massachusetts: Charles River Media, 2006.

LUTSA, Jan et al. A tutorial on support vector machine-based methods for classification problems in chemometrics, **AnalyticaChimicaActa** 665, 2010, p. 129–145.

MILLER, Thomas W. **Data and text mining**: a business applications approach. New Jersey: Pearson Prentice Hall, 2005.

MIRAGEM, Bruno. Aspectos característicos da disciplina do comércio eletrônico de consumo: comentários ao Dec. 7.962, de 15.03.2013. **Revista de Direito do Consumidor,** São Paulo, v. 22, n. 86 , p.287-300,, abr. 2013.

PINHEIRO, Carlos André Reis**. Inteligência analítica**: mineração de dados e descoberta do conhecimento. Rio de Janeiro: Ciência Moderna, 2008.

RECUERO, Raquel. **Redes Sociais na internet**. 2 ed. Porto Alegre: Sulina, 2011.

TAN, Pan-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining:** mineração de dados**.** Rio de Janeiro: Ciência Moderna, 2009.

TEIXEIRA, Tarcisio. **Comércio eletrônico**: conforme o Marco Civil da Internet e a regulamentação do e-commerce no Brasil. São Paulo: Saraiva, 2015.

VAPNIK, V. N. **Statistical Learning Theory**. New York, USA: John Wiley & Sons, Inc., 1998.

VIEIRA, Valter Afonso; MATOS, Celso Augusto de; SLONGO, Luiz Antônio. Avaliação das relações entre qualidade de serviço do site, satisfação, valor percebido, lealdade e boca a boca por meio de um modelo teórico. **Revista de Administração,** São Paulo , v.44, n.2 , p.131-146, jun. 2009.

STITSON, M. O et al. Theory of support Vector Machines. **Technical Report CSD-TR-96-17**, Computational Intelligence Group, Royal Holloway, University of London, 1996.

WITTEN, I. H; FRANK, Eibe. **Data mining:** practical machine learning tools and tecniques. 3nd ed San Francisco: Morgan Kaufmann, 2011.