Mineração e Classificação de Mensagens em Redes Sociais Utilizando Processamento de Linguagem Natural e Redes Neurais

Gabriel Machado[1], Eugênio Silva[1], Bruno Costa[2], José Andrade[1], Rafael Monteiro[1]

[1]Ciência da Computação – Centro Universitário Serra dos Órgãos (UNIFESO) – Teresópolis, RJ, Brasil

[2]Instituto Federal do Rio de Janeiro (IFRJ) – Rio de Janeiro, RJ, Brasil

gabriel.rmachado10@gmail.com, eugsilva@gmail.com, brunocosta.dsn@gmail.com, jrobert.andrade@unifeso.edu.br, rafaelgomesmonteiro@gmail.com

Mining and Classification of Messages in Social Media Using Natural Language Processing and Artificial Neural Networks

Abstract: *People throughout the world are increasingly using the Online Social Media in order to share, in an extremely agile way, every kind of information, even information about health conditions. This paper proposes a methodology of collecting messages from Twitter Social Media where it's applied a filter, so as to consider only messages related to Aedes Aegypti borne diseases. In the following, these messages are distributed in two classes: one that gathers suspected cases of disease and the other not. The obtained results show the proposed solution can be a very useful complement to the current government surveillance mechanisms.*

Keywords: artificial neural networks, social media, Aedes Aegypti

Resumo: *As redes sociais online são cada vez mais utilizadas por pessoas em todo mundo para compartilhar, de uma forma extremamente ágil, todo tipo de informação, inclusive informações sobre a condição de saúde. Este trabalho propõe uma metodologia de coleta de mensagens da rede social Twitter em que se aplica um filtro para que sejam consideradas apenas as mensagens relacionadas às doenças causadas pelo mosquito Aedes Aegypti. Em seguida, essas mensagens são distribuídas em duas classes: uma que reúne as mensagens em que há a suspeita de ocorrência de uma doença e outra que reúne aquelas em que não há a suspeita. Os resultados obtidos mostram que a solução proposta pode ser um complemento bastante útil para os atuais mecanismos governamentais de vigilância.*

Palavras-chave: Redes Neurais Artificiais, Redes Sociais, *Aedes Aegypti*.

Introdução

As redes sociais, desde o seu surgimento, tornaram a *Web* um meio de comunicação mais dinâmico e aberto, onde milhões de pessoas em todo o mundo, em tempo real, opinam, avaliam, compartilham ou trocam informações a respeito de diversos assuntos. Com a disponibilidade de Terabytes de informação, as redes sociais têm chamado a atenção de pesquisadores. Uma das razões é a velocidade com que informações relacionadas a eventos ocorridos são percebidos e divulgados pelos seus usuários. Tais informações são difundidas em tempo real, proporcionando um cenário instantâneo que pode ser de grande auxílio para previsão e monitoramento de informações, juntamente com os meios de comunicação tradicionais. Como exemplo, em Sakaki *et al.,* (2009) é mencionado que, "quando um terremoto ocorre, várias pessoas publicam mensagens (*tweets*) relacionadas ao incidente natural, que permitem prontamente a detecção do mesmo, simplesmente pela observância dos *tweets*".

Além de desastres naturais, outro grande nicho de atuação de pesquisas em redes sociais tem sido a epidemiologia, como apresentado nos trabalhos de Meira *et al,* (2011), Lampos *et al,* (2010), Chew & Eysenbach (2010) e Ackrekar *et al,* (2011), principalmente porque seus usuários expõem publicamente sua condição de saúde, produzindo um meio auxiliar de rastreamento de epidemias e surtos de doenças. No Brasil, segundo o Ministério da Saúde, no ano de 2015[1] foram registrados mais de 1,5 milhão de casos prováveis de dengue no país, sendo quase 1 milhão na Região Sudeste, representando 1.171,7 casos por 100 mil habitantes. De janeiro ao final de maio de 2016[2] já foram registrados mais de 1,2 milhão de casos de dengue no país com 288 óbitos causados pela doença, o que, juntamente com os dados de 2015, caracteriza o estado de epidemia, que segundo a OMS[3] representa 300 casos por 100 mil habitantes.

Perante tal cenário alarmante, trabalhos como os de Gomide & Meira (2011) e W. Meira, *et al*. (2011) consideraram a possibilidade de adotar um modelo de pesquisa em redes sociais para auxiliar no monitoramento da dengue. Com isso, os meios de

78

[1] Disponível em: <http://portalsaude.saude.gov.br/images/pdf/2016/janeiro/07/2015-svs-be-pncd-se48.pdf>
[2] Disponível em: <http://portalsaude.saude.gov.br/images/pdf/2016/junho/30/2016-021.pdf>
[3] Disponível em: <http://g1.globo.com/bemestar/dengue/noticia/2015/05/brasil-tem-7459-mil-casos-de-dengue-no-pais-ate-18-de-abril.html>

comunicação tradicionais, que dependem dos boletins epidemiológicos do Ministério da Saúde sobre os números e a situação da doença no Brasil, poderiam informar a população de forma mais ágil.

O objetivo deste trabalho é propor um protótipo de software para classificar e avaliar mensagens de teor pessoal publicadas na rede social *Twitter*, especificamente aquelas com conteúdo referente às doenças transmitidas pelo mosquito *Aedes Aegypti*, de forma a auxiliar na estimativa do número de possíveis indivíduos infectados por dengue, *chikungunya* e zika no país. Para atingir esse objetivo é proposto um método de captura das mensagens em tempo real, onde posteriormente são aplicadas técnicas de Processamento de Linguagem Natural (PLN) para pré-processamento dos dados e de Redes Neurais Artificiais (RNA) para a classificação de mensagens. Essa metodologia pode ser um complemento para os atuais mecanismos governamentais de vigilância.

Trabalhos Relacionados

O volume de informação e a diversidade de assuntos veiculados em redes sociais tornam esses meios de comunicação fontes de informação muito ricas para a realização de diversos tipos de estudos. Dentre eles, destacam-se as pesquisas para fins de utilidade pública, em especial aquelas voltadas para o levantamento de estatísticas de ocorrências de algumas doenças tropicais.

Em Lampos; De Bie; Cristianini (2010) as mensagens do *Twitter* (*tweets*) contendo palavras relacionadas com a Influenza foram usadas como um indicador em tempo real de atividade da gripe no Reino Unido. O volume de mensagens foi comparado com os dados oriundos da *Health Protection Agency of United Kingdom*, e foi obtida uma correlação linear maior do que 95%.

Em Achrekar *et al*. (2011) foi desenvolvido o framework SNEFT (do inglês, *Social Network Enabled Flu Trends*), com o objetivo de associar as capacidades de detecção e previsão das redes sociais para o descobrimento de tendências reais da gripe no mundo e, além disso, como correlacionar os *tweets* detectados com os dados oficiais.

Em Chew e Eysenbach (2010), o conteúdo dos tweets relacionados à gripe H1N1 foram estudados e uma análise quantitativa das mensagens foi conduzida. Nessa análise

foi avaliada e validada a viabilidade do *Twitter* como uma ferramenta de tempo real para monitoramento de tendências e sentimentos.

Em Sakaki *et al*. (2010), pesquisas no Twitter foram conduzidas com o objetivo de usar os usuários da rede social como sensores e, através dos *tweets* produzidos por esses usuários, foi possível criar um mecanismo de detecção de terremotos. Para a classificação das mensagens, foi utilizada a técnica de SVM (*Support Vector Machine*) que, juntamente com a produção de um modelo probabilístico espaço-temporal, possibilitou encontrar o epicentro do terremoto e sua trajetória. Com isso, alertas por e-mail e celular foram enviados aos usuários cadastrados em um site denominado *Toretter,* chamando a atenção para a ocorrência dos terremotos de maneira mais rápida que as agências meteorológicas.

Em Gomide & Meira (2011) e Meira *et al.*, (2011), foram realizadas pesquisas no *Twitter* durante cinco anos (2006 a 2011). Os autores efetuaram o monitoramento e análise dos *tweets* capturados na rede social com a finalidade de construir um sistema *online* de vigilância em tempo real da situação epidemiológica da dengue no Brasil. Treinamentos foram realizados utilizando o algoritmo LAC (*Lazy Associative Classification*) (VELOSO *et al*., 2006), que é uma abordagem baseada em Árvores de Decisão, porém adotando uma abordagem paciente (*lazy*), capaz de lidar com um grande volume de dados e um pequeno conjunto de treino, além das classes poderem estar desbalanceadas. Ao final, análises de correlação com os dados emitidos pelo Ministério da Saúde obtiveram um resultado de 95,78%, comprovando a eficácia e a veracidade dos dados colhidos pela metodologia.

Apesar do volume significativo de pesquisas voltadas para captura e avaliação automatizada de mensagens oriundas de redes sociais, não foram encontrados trabalhos acadêmicos relevantes que avaliassem o uso de RNA para classificação de dados coletados desse nicho. Segundo Goldberg (2015), as RNA são poderosos modelos de aprendizagem, principalmente porque podem solucionar problemas não-lineares com várias entradas. As RNA são amplamente aplicadas na solução de diferentes tipos de problemas de classificação e, por isso, este trabalho adota esse modelo em busca de resultados que indiquem a eficácia dessa abordagem também no problema de classificação de mensagens.

Metodologia

A metodologia empregada neste trabalho compreende quatro etapas principais: (i) coleta dos dados, (ii) pré-processamento, (iii) classificação e (iv) avaliação dos resultados obtidos. Essas etapas são abordadas detalhadamente nas seções seguintes.

Coleta dos Dados

A coleta dos dados foi realizada principalmente com o auxílio da *Streaming* API[4], que captura *tweets* publicados em tempo real, filtrados por palavras-chave. Para utilizar essa API, primeiramente realizou-se um cadastro no site do *Twitter*[5], gerando códigos (*tokens*) OAuth de acesso, necessários para liberação e uso da *Streaming* API. OAuth é um padrão aberto de autorização, projetado especialmente para trabalhar em conjunto com o protocolo HTTP, gerando *tokens* (ou códigos) de acesso emitidos por um servidor de autorização para uso de aplicações de terceiros. Essas aplicações então utilizam esses *tokens* para acessar os recursos protegidos.

Após o cadastro e a obtenção dos códigos de acesso pelo Twitter, o framework *Tweetinvi* foi o intermediador entre a API e a aplicação, sendo o principal componente para a implementação do código para captura de mensagens via *Streaming* API. Para realizar a captura de mensagens com conteúdo relacionado ao tema de pesquisa, a API disponibiliza filtragem de mensagens que contenham palavras-chave.As palavras-chave utilizadas para captura de mensagens foram os nomes das doenças “dengue”, “zika” e “chikungunya” e suas principais variações informais como “zica” ou “xicongunha”. As mensagens foram capturadas no final de novembro de 2015 durante um dia inteiro (24 horas) e, em seguida, foram persistidas em um banco de dados para tratamento posterior.

Pré-Processamento dos Dados

Antes que sejam submetidos à etapa de classificação, é fundamental que os dados brutos provenientes da rede social sejam devidamente pré-processados, o que

81

[4] Disponível em <https://dev.twitter.com/streaming/overview>

[5] Disponível em <https://apps.twitter.com/>

compreende a rotulação e a extração de características das mensagens coletadas. A rotulação consiste em associar cada mensagem a uma classe de acordo com o seu conteúdo. Por outro lado, a extração de características consiste em converter as mensagens em alguma representação numérica que promova uma boa diferenciação entre mensagens pertencentes a classes diferentes.

Neste trabalho, cada mensagem coletada foi associada manualmente, com base na interpretação de seu conteúdo, a uma dentre duas classes possíveis: "suspeita de dengue" ou "não dengue", representadas pelos vetores [1 0] ou [0 1], respectivamente. É importante salientar que nessa classificação a palavra "dengue" pode significar dengue, zika ou chikungunya Após a rotulação manual, notou-se que muitas mensagens da classe 0 eram semelhantes em conteúdo, porque foram reenviadas ("*retweetadas*") entre os usuários da rede social. Essas mensagens são apenas cópias de outras mensagens e podem interferir nos resultados da classificação. Por isso, essas repetições foram excluídas, deixando apenas uma amostra de cada mensagem. A eliminação das repetições foi feita de forma automática com o auxílio de um algoritmo que calcula a distância de edição entre duas cadeias de caracteres. A distância de edição, ou distância de *Levenshtein* (LEVENSHTEIN, 1966), é um algoritmo de programação dinâmica que calcula o número de operações elementares necessárias para transformar uma cadeia em outra. As três operações elementares possíveis são: inserção, remoção e substituição. A distância entre duas mensagens quaisquer foi normalizada entre 0 e 100, com o valor 0 indicando mensagens idênticas e o valor 100 indicando mensagens totalmente diferentes. A distância normalizada foi calculada segundo a Equação 1:

Equação 1: Distância de edição normalizada

$$d_{norm} = \frac{100 * d}{d_{max}}$$

Onde:

$\boldsymbol{d_{norm}}$ é a distância normalizada entre duas mensagens;

$\boldsymbol{d}$ é a distância entre duas mensagens retornada pelo algoritmo de *Levenshtein*;

$\boldsymbol{d_{max}}$ é maior distância possível entre duas mensagens, ou seja, é o comprimento da mensagem mais longa.

Para o critério de exclusão foram assumidas como repetições as mensagens cuja distância normalizada entre elas foi menor ou igual a 50. Nessas situações, uma das mensagens foi desconsiderada. É importante destacar que o valor 50 foi definido experimentalmente após a análise das distâncias entre um conjunto de mensagens e seus respectivos "*retweets*".

No caso das mensagens da classe [1 0] ("suspeita de dengue") não foi adotado o algoritmo de *Levenshtein* para a identificação das cópias. Isso se deveu ao fato de as mensagens apresentarem conteúdos geralmente curtos e semelhantes, que não necessariamente representam cópias. Para essa classe foram definidas como repetições as mensagens publicadas pelo mesmo autor num período de 30 dias. Portanto, de todas essas mensagens, apenas uma foi mantida.

Para a extração de características, primeiramente as mensagens foram submetidas a um processo de lematização. A lematização é um processo da linguística, amplamente explorado na área de PLN, que visa agrupar flexões de palavras à sua forma original, para que possam ser analisadas unicamente. Como exemplo, o verbo "correu" após o processo torna-se "correr", uma vez que esta é sua forma original (ou infinitivo). Neste trabalho, o processo de lematização não foi aplicado somente a verbos, mas também a adjetivos, diminutivos, aumentativos, plurais e advérbios de modo com sufixo "-mente". Foi preciso também realizar algumas transformações especiais: palavras que representam as variações de dengue foram reduzidas à palavra "dengue", grandezas numéricas foram reduzidas para a palavra "numeral" e termos chulos reduzidos para "chulo". A lematização foi necessária para que uma análise de frequência de palavras pudesse ser realizada em cada mensagem, identificando termos mais utilizados pelos autores dos *tweets*. Devido ao fato de não ser encontrada nenhuma ferramenta que realizasse o processo descrito automaticamente, toda a lematização precisou ser feita de maneira manual.

Finalizada a lematização, todas as mensagens contidas no repositório foram submetidas ao algoritmo de contagem de *n*-gramas. Um *n*-grama é uma sequência contígua de *n* itens que, no caso de um texto, podem ser letras, palavras, sílabas, fonemas etc. Neste trabalho, os itens considerados para a formação dos *n*-gramas foram as palavras e os valores de *n* adotados foram 1 (unigramas), 2 (bigramas) e 3 (trigramas). Esse

algoritmo teve como finalidade não só a contagem, mas também o cálculo da frequência dos uni, bi e trigramas encontrados nas mensagens para a definição da lista *L* de *n*-gramas, levando em consideração a classe de cada mensagem. Dentre os *n*-gramas selecionados, alguns exemplos são: "*estar*", "*dengue*" e "*com*" (unigramas); "*pegar dengue*", "*com dengue*" e "*com suspeita*" (bigramas) e "*achar estar dengue*", "*estar com dengue*" e "*com suspeita dengue*" (trigramas). O critério adotado para a seleção dos *n*-gramas considerou aqueles que eram mais frequentes em uma classe e menos frequentes em outra e vice-versa. Com isso, buscou-se representar as mensagens com características que promovessem uma melhor diferenciação entre mensagens de uma classe e de outra. A boa diferenciação entre mensagens de classes diferentes é determinante para o bom aprendizado da rede neural. Para a melhor diferenciação possível, o ideal seria garantir que *n*-gramas que ocorrem em mensagens de uma classe não ocorram em mensagens de outra classe, mas, devido à complexidade da linguagem natural, isso nem sempre é possível. Assim, o critério de seleção de *n*-gramas baseado na frequência buscou minimizar a ocorrência de *n*-gramas coincidentes em mensagens de classes diferentes.

O último processo de tratamento dos dados consistiu em representar numericamente as mensagens, o que é necessário para a execução dos treinamentos e testes da rede neural. A transformação consistiu basicamente em um processo iterativo que verificou a ocorrência de cada *n*-grama *j* da lista *L* para cada mensagem *i* do repositório. Para a presença foi atribuído o valor "1" e para a ausência o valor "0", formando ao final um vetor numérico contendo $|L|$ posições que representa a mensagem *i*.

Como exemplo, supõe-se que a lista de *n*-gramas *L* tenha os seguintes elementos: *L* = {"*estar dengue*", "*mosquito dengue*", "*com dengue*", "*estar com dengue*", "*caso dengue*", "*numeral*"} e que duas mensagens, (a) "*tô mal com dengue*" classificada como "dengue (1)" e (b) "*100 novos casos de dengue são confirmados em Vitória*" classificada como "não-dengue (0)" sejam capturadas. O processo de tratamento dos dados começa com a lematização que transforma a mensagem (a) em "*estar mal com dengue*" e (b) em "*numeral novo caso dengue ser confirmado Vitória*". Após a lematização, em cada mensagem é verificada a presença do *n*-grama $L[j]$, onde $0 \leq j < |L|$, atribuindo "0" à ausência ou "1" à presença. Ao final do processo, são criados dois vetores contendo as representações numéricas das mensagens (a) = [1, 0, 1, 1, 0, 0] e (b) = [0, 0, 0, 0, 1, 1].

Nesse exemplo, percebe-se que a escolha dos *n*-gramas proporcionou uma boa diferenciação entre as mensagens, pois, apenas o *n*-grama "*mosquito dengue*" teve o mesmo valor em ambas as classes.

Classificação dos Dados

Dentre as várias alternativas de modelos computacionais que podem ser aplicados à solução de problemas que envolvem classificação, neste trabalho optou-se pela utilização das RNA. Essa escolha se baseou, principalmente, na amplitude de áreas em que as RNA podem ser aplicadas e também no bom desempenho que se tem obtido nessas diversas áreas.

Em linhas gerais, as RNA são um modelo matemático computacional inspirado na estrutura neurológica do cérebro humano voltado, principalmente, para a solução de problemas que envolvem classificação de padrões e aproximação de funções. Essa estrutura é formada por um conjunto de neurônios artificiais altamente interconectados e, associado a ela, há um algoritmo de aprendizagem. Esse algoritmo altera iterativamente os pesos associados às conexões entre os neurônios de forma a encontrar a melhor relação entre os dados de entrada e de saída. Espera-se que ao final desse processo de aprendizagem a rede neural adquira a capacidade de generalização e, com isso, seja capaz de associar as saídas corretas a dados de entrada que não participaram do processo de aprendizagem.

O modelo de rede neural adotado foi o do tipo MLP (*multilayer perceptron*), treinada com o algoritmo de retropropagação do erro utilizando taxa de aprendizado adaptativa e termo de *momentum*. A taxa adaptativa e o *momentum* contribuem tanto para acelerar o processo de aprendizagem, quanto para evitar regiões de mínimos locais que possam estar presentes na superfície de erro (HAYKIN, 2009). Para a configuração, treinamento, teste e visualização de resultados da rede foi utilizada a ferramenta gráfica NNTOOL disponível no *Neural Network Toolbox* do ambiente MATLAB (MATLAB, 2016). O NNTOOL apresenta uma interface bastante amigável que permite a manipulação de redes neurais sem a necessidade de programação.

Abordagens e Resultados

O passo seguinte à preparação dos dados consistiu em definir os conjuntos de treinamento e de teste da rede neural. Para isso, analisou-se a quantidade de mensagens disponíveis em cada classe e obteve-se 1137 *tweets* para a classe [0 1] (“não dengue”) e 305 *tweets* para a classe [1 0] (“suspeita de dengue”), totalizando uma base de dados com 1442 mensagens. Para evitar um viés no processo de aprendizagem da rede neural, é preciso haver um equilíbrio entre as quantidades de dados de treinamento para ambas as classes. Por isso, para o conjunto de treinamento foram reservados 80% das mensagens da classe com a menor quantidade de mensagens (“suspeita de dengue”), o que correspondeu a um total de 251 mensagens. Em seguida, a mesma quantidade de mensagens foi reservada da classe “não dengue”. Com isso, formou-se um conjunto de treinamento com 502 mensagens (251 mensagens de cada classe). As 940 mensagens restantes formaram o conjunto de teste, sendo 886 pertencentes à classe “não dengue” e 54 pertencentes à classe “suspeita de dengue”.

Devido à forma como as classes foram representadas, [1 0] para “suspeita de dengue” e [0 1] para “não dengue”, as redes neurais usadas em todos os experimentos têm duas saídas. Além disso, em todas as redes os neurônios, tanto da camada escondida quanto da camada de saída, adotam funções de ativação sigmoidais. Portanto, dada uma mensagem de entrada, o resultado em cada saída da rede pode assumir qualquer valor entre 0 e 1. Por isso, para identificar a qual classe a rede está associando uma mensagem de entrada, é necesário aplicar algum critério de discretização às saídas. Aqui foi adotado o critério “*winner takes all*”, em que a saída de maior valor é convertida para 1 e a de menor valor é convertida para 0.

Os experimentos foram conduzidos segundo três abordagens que se diferenciaram pela composição da lista L de n-gramas e, consequentemente, pela configuração da rede neural. Os detalhes de cada abordagem, bem como os resultados dos melhores experimentos de cada uma, são apresentados a seguir.

Abordagem 1

Nesta abordagem a lista L foi composta por 15 unigramas e 23 bigramas, totalizando 38 n-gramas. A partir disso, foi definida uma rede neural com uma camada escondida com a seguinte configuração: 38 entradas, 10 neurônios na camada escondida

e 2 nerônios na camada de saída. A quantidade de neurônios da camada escondida, bem como os parâmetros de treinamento, foram definidos empiricamente com base em vários experimentos em que esses valores foram modificados. Os parâmetros utilizados para o treinamento da rede neural, bem como o resultado do melhor experimento realizado, são exibidos na Figura 1.

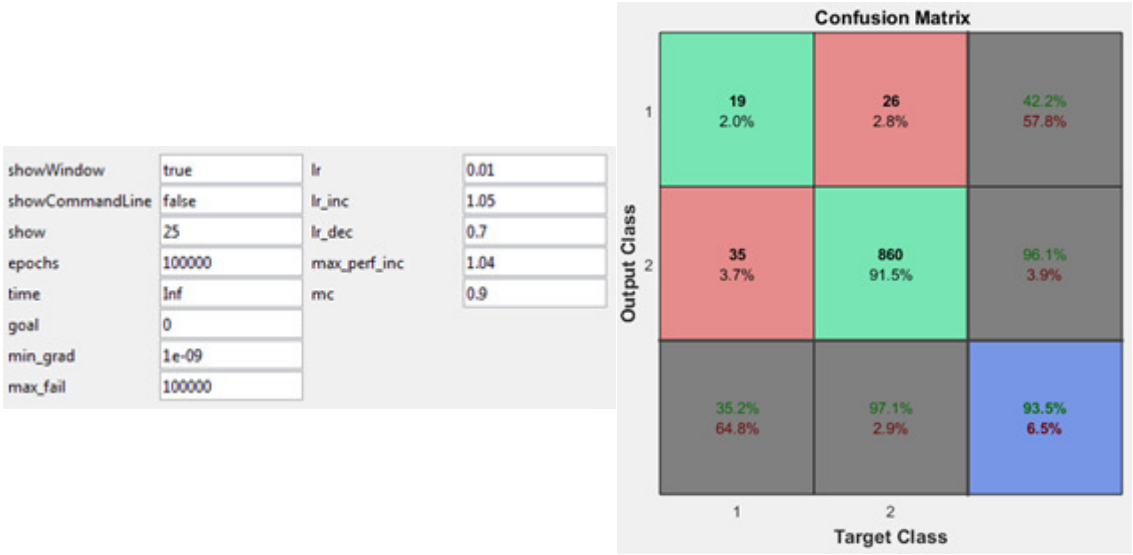

Figura 1: Parâmetros de treinamento e resultado da rede neural (abordagem 1)

Os resultados obtidos pela rede neural no conjunto de teste estão dispostos na forma de uma matriz de confusão em que o eixo *x* representa as respostas desejadas para as mensagens de entrada, enquanto que o eixo *y* representa as respostas emitidas pela rede neural. Portanto, os acertos, ou seja, as mensagens para as quais as respostas emitidas pela rede coincidem com as respostas desejadas, estão dispostas na diagonal principal da matriz. A posição (1, 1) exibe os acertos da classe "suspeita de dengue") e a posição (2, 2) exibe os acertos da classe ("não dengue"). Na posição (1, 2) estão as mensagens da classe "não dengue", mas a rede classificou como sendo da classe "suspeita de dengue". Na posição (2, 1) acontece o inverso do que está representado na posição (1, 2).

Apesar de ter alcançado um desempenho total de 93,5% no conjunto de teste, o resultado não foi considerado satisfatório. Isso se deveu, principalmente, ao fato de a rede ter obtido um índice de erro (64,8%) bem maior que o índice de erro (35,2%) para a classe

"suspeita de dengue". Esse resultado sugere que a forma como as mensagens foram representadas não favoreceu a separabilidade espacial entre as classes, prejudicando assim a capacidade de aprendizado da rede neural.

Abordagem 2

Nesta abordagem a lista de *n*-gramas foi ampliada com a intenção de obter uma forma de representação das mensagens que permitisse uma melhor diferenciação entre as classes. Para isso, foram acrescentados 34 trigramas à lista da abordagem anterior, resultanto em uma lista *L* com 72 *n*-gramas. Além disso, os dados de entrada da rede foram submetidos ao método de PCA (*Principal Component Analisys*) (SHLENS, 2003). O PCA é uma transformação linear ortogonal que converte um conjunto de valores de variáveis, que em geral apresentam alguma correlação, em um conjunto de valores de variáveis linearmente descorrelacionadas denominadas componentes principais. Os componentes obedecem uma ordem onde o primeiro apresenta a maior variância possível entre seus valores, o segundo tem a maior variância e é ortogonal (não correlacionado) ao primeiro, e assim sucessivamente até o último componente. Na prática, isso significa que os componentes são extraídos na ordem do mais explicativo para o menos explicativo, ou seja, que os primeiros componentes retêm o máximo possível da informação contida nas variáveis originais e que não há redundância na informação. Por isso, o PCA é comumente utilizado como um método de redução de dimensionalidade, uma vez que os componentes mais explicativos são preservados e os menos explicativos podem ser descartados.

Com a aplicação do PCA, as 72 entradas da rede neural foram reduzidas para 64. Em seguida a rede neural foi configurada com 64 entradas, 25 neurônios na camada escondida e 2 neurônios na camada de saída. Os parâmetros da rede neural e o melhor resultado estão ilustrados na Figura 2.

Nota-se que o desempenho total de 78,4 % obtido na abordagem 2 ficou abaixo do desempenho total de 93,5% alcançado na abordagem 1. No entanto, os índices de acerto em cada classe (64,8% para a classe "suspeita de dengue" e 79,2% para a classe "não dengue") ficaram bem acima dos índices de erro (35,2% e 20,8%). Além disso, a maior proximidade entre os percentuais de erro das duas classes mostra que houve uma

distribuição mais homogênea dos erros entre as classes, o que sugere que o aprendizado da rede neural foi mais equilibrado.

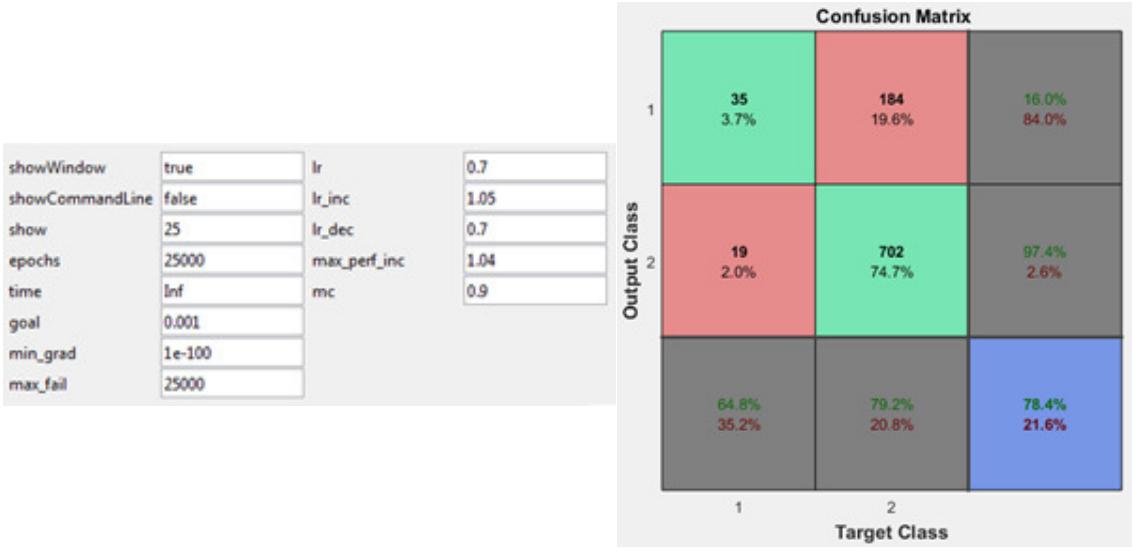

Figura 2: Parâmetros de treinamento e resultado da rede neural (abordagem 2)

Abordagem 3

Com o intuito de obter uma base de dados ainda mais representativa que aquela utilizada nas abordagens anteriores, uma nova coleta de mensagens foi realizada em maio de 2016 durante 24 horas. No pré-processamento, após a eliminação das mensagens repetidas, foram apuradas 910 novas mensagens que foram acrescentadas às anteriores, totalizando 2352 mensagens. Isso representou um aumento de 63% no total de mensagens. Como consequência desse aumento, os conjuntos de treinamento e de teste foram reorganizados e passaram a ter, respectivamente, 766 mensagens (383 em cada classe) e 1586 (1490 na classe "não dengue" e 96 na classe "suspeita de dengue"). Em seguida, as frequências dos uni, bi e trigramas foram recalculadas para que uma nova lista *L* de *n*-gramas fosse definida conforme essas novas frequências. A nova lista de *n*-gramas passou a ter 10 unigramas, 15 bigramas e 12 trigramas, totalizando 37 elementos.

A nova base de dados também foi submetida ao método de PCA e as 37 entradas originais foram reduzidas para 32. Com isso, a rede neural foi configurada com 32

entradas, 19 neurônios na camada escondida e 2 neurônios na camada de saída. Os parâmetros da rede assim como o melhor resultado obtido estão ilustrados na Figura 3.

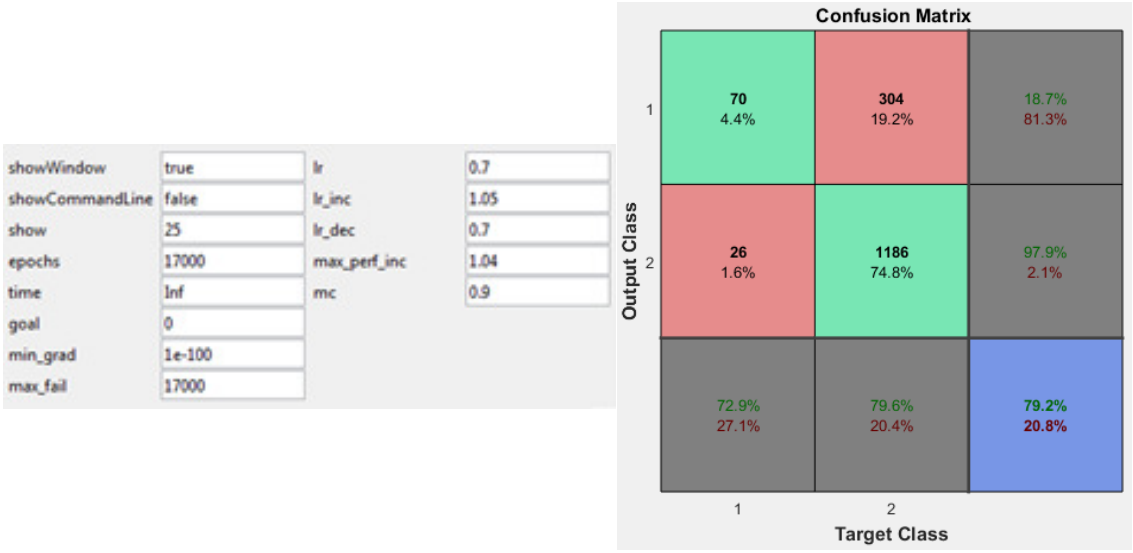

Figura 3: Parâmetros de treinamento e resultado da rede neural (abordagem 3)

Nota-se que o desempenho total de 79,2% alcançado na abordagem 3 ficou muito próximo daquele obtido na abordagem 2 (78,4%). O mesmo aconteceu com o desempenho individual obtido na classe "não dengue", sendo 79,6% obtido na abordagem 3 contra 79,2% obtido na abordagem 2. Por outro lado, cabe destacar o aumento de 8,1% (64,8% para 72,9%) no desempenho individual obtido na classe "suspeita de dengue". Com isso, houve uma proximidade ainda maior entre os desempenhos individuais em relação à abordagem 2, o que sugere um equilíbrio ainda melhor no aprendizado da rede neural. Os resultados obtidos indicam também que a redefinição da lista de *n*-gramas após o acréscimo de novas mensagens favoreceu a separabilidade entre as classes.

Conclusões

O trabalho propôs um protótipo de software para capturar, filtrar e classificar mensagens publicadas na rede social *Twitter*, a fim de identificar, dentre as mensagens com conteúdo relacionado às doenças causadas pelo *Aedes Aegypti*, as que indicam e as que não indicam casos suspeitos de alguma das doenças. A metodologia adotada explorou

técnicas de PLN e RNA para, respectivamente, extração de características e classificação das mensagens.

Na comparação entre os resultados obtidos nas abordagens 1 e 2 observou-se que, apesar do resultado obtido na abordagem 1 ter sido superior ao obtido na abordagem 2, o segundo foi considerado mais consistente, pois, a distribuição dos percentuais de erro entre as duas classes foi mais equilibrada. O desempenho total na abordagem 3 também foi inferior ao da abordagem 1, mas, novamente houve uma melhor distribuição dos erros entre as classes. Além disso, a abordagem 3 alcançou um desempenho total ligeiramente superior ao da abordagem 2. Isso indica que o aumento na quantidade de dados de treinamento, a redefinção da lista de *n*-gramas, bem como a redução de dimensionalidade, contribuíram para a obtenção de um resultado mais confiável.

A pequena elevação do desempenho entre as abordagens 2 e 3 indica que a base de dados formada por mensagens capturadas em duas ocasiões diferentes pode melhorar a representativadade dos padrões a serem aprendidos pela rede neural. Assim, a captura de mensagens em várias ocasiões, especialmente em épocas de maior incidência das doenças de interesse, pode contribuir para melhorar ainda mais os resultados.

Trabalhos Futuros

O trabalho apresentado aqui é a etapa inicial de uma pesquisa cujo objetivo é criar um sistema Web para o monitoramento em tempo real de mensagens compartilhadas na rede social *Twitter* a fim de identificar casos supeitos de doenças causadas pelo mosquito *Aedes Aegypti*. Isso pode auxiliar os meios de comunicação e a população em geral a obter informações instantâneas sobre o cenário dessas doenças no Brasil.

Para viabilizar o monitoramento em tempo real, a próxima etapa do trabalho está direcionada para o desenvolvimento de um lematizador automático para as mensagens capturadas. Assim, não haverá mais a necessidade de intervenção humana no processo de extração de características das mensagens. Com a lematização automática, a captura de mensagens poderá ocorrer continuamente e será possível observar eventuais sazonalidades nas ocorrências das doenças que não são detetáveis em coletas pontuais. Isso possibilitará a apresentação de resultados de classificação acompanhados de estatísticas que agreguem valor (informação) a esses resultados.

Estão previstas também melhorias na forma de extrair as características das mensagens e, consequentemente, novas parametrizações e treinamentos de redes neurais, a fim de alcançar desempenhos de classificação ainda melhores. Também devem ser estudadas e avaliadas estratégias de rejeição para que um classificador não seja "obrigado" a associar cada mensagem de entrada a uma dentre as duas classes possíveis, como acontece com o critério "*winner takes all*". A rejeição permite que a resposta da rede neural seja considerada somente quando a diferença entre os valores de saída para cada classe esteja acima de um determinado limiar. Essa medida também pode aumentar significativamente o desempenho na classificação.

Ao final, pretende-se ainda avaliar o modelo de RNA em comparação a outras técnicas de classificação e dados epidemiológicos oficiais emitidos pelo Ministério da Saúde. Nesse caso, para que possa ser feita uma análise comparativa com os dados oficiais da área de Saúde, a sazonalidade das doenças deverá ser considerada na metodologia de coleta e análise dos dados.

Referências

ACHREKAR , et al. Predicting Flu Trends using Twitter Data. **FIU Computer and Information Sciences**, 2011. Disponivel em: <http://users.cs.fiu.edu/~jli003/papers/trend-event-detection/Predicting%20flu%20trends%20using%20twitter%20data.pdf>. Acesso em: mar. 2016.

GOLDBERG, Y. CS BIU. **A Primer on Neural Network Models for Natural Language Processing**, 2015.

GOMIDE, J.; MEIRA, W. J. DCC-UFMG. **Mineração de redes sociais para detecção e previsão de eventos reais**, 2011.

HAYKIN, S. **Neural networks and learning machines**. 3ª. ed. [S.l.]: Pearson, 2009.

LAMPOS, V.; DE BIE, T.; CRISTIANINI, N. Lampos.net. **Flu Detector - Tracking Epidemics on Twitter**, 2010.

LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics Doklady**, Russia, v. 10, n. 8, p. 845-848, fev. 1966.

MATLAB. Open Network Data Manager, 2016. Disponivel em: <http://www.mathworks.com/help/nnet/ref/nntool.html>. Acesso em: 2016.

MEIRA, W. et al. WebSci. **Dengue surveillance based on a computational model of spatio-temporal locallity of Twitter**, 2011.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Yutaka Matsuo's Homepage. **Earthquake Shakes Twitter Users:** Real-time Event Detection by Social Sensors, 2010.

SAÚDE, M. D. **Portal da Saúde**, set. 2016. Disponivel em: <http://portalsaude.saude.gov.br/images/pdf/2016/maio/17/2016-016---Dengue-SE16-publica----o.pdf>. Acesso em: Maio 2016.

SHLENS, J. A Tutorial on Principal Component Analysis - Derivation, Discussion and Singular Decomposition. **Departament of Computer Science - Princeton University**, v. 1, p. 16, mar. 2003.

VELOSO, A.; JR., W. M.; ZAKI, M. Ressenlaer Computer Science. **Lazy Associative Classification**, 2006.