Universidade Federal da Paraíba

Centro de Informática

Programa de Pós-Graduação em Informática

Uma investigação de aspectos da classificação de tópicos para textos curtos

EWERTON LOPES SILVA DE OLIVEIRA

João Pessoa, Paraíba, Brasil

23 de fevereiro de 2015

EWERTON LOPES SILVA DE OLIVEIRA

Uma investigação de aspectos da classificação de tópicos para textos curtos

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Informática da Universidade Federal da Paraíba como parte dos requisitos necessários para obtenção do grau de Mestre em Informática.

Área de Concentração: Ciência da Computação

Linha de Pesquisa: Computação Distribuída

Prof. Dr. Andrei de Araújo Formiga

(Orientador)

Prof. Dr[a]. Natasha Correia Queiroz Lino

(Coorientadora)

João Pessoa, Paraíba, Brasil

O48u Oliveira, Ewerton Lopes Silva de.
Uma investigação de aspectos da classificação de tópicos para textos curtos / Ewerton Lopes Silva de Oliveira.- João Pessoa, 2015.
82f. : il.
Orientador: Andrei de Araújo Formiga
Coorientadora: Natasha Correia Queiroz Lino
Dissertação (Mestrado) - UFPB/CI
1. Informática. 2. Computação distribuída.
3. Aprendizagem de máquina. 4. Classificação de texto.
5. Mensagens sociais.

UFPB/BC CDU: 004(043)

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **Ewerton Lopes Silva de Oliveira**, candidato ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 23 de fevereiro de 2015.

Ao vigésimo terceiro dia do mês de fevereiro do ano de dois mil e quinze, às dez horas, no laboratório 2 da Escola Superior de Redes - Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar o candidato ao grau de Mestre em Informática, na área de "*Sistemas de Computação*", na linha de pesquisa *"Computação Distribuída"*, **Sr. Ewerton Lopes Silva de Oliveira**. A comissão examinadora foi composta pelos professores doutores: **Andrei de Araújo Formiga (PPGI-UFPB)**, orientador e presidente da Banca, **Natasha Correia Queiroz Lino (PPGI-UFPB)**, examinadora interna, **Thais Gaudencio do Rego,** examinadora interna e **Thiago Pereira Falcão,** examinador externo à Instituição. Dando início aos trabalhos, o professor Andrei de Araújo Formiga cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que o mesmo fizesse, oralmente, a exposição do trabalho de dissertação intitulado *"Classificação de Tópicos para Textos Curtos em Redes Sociais"*. Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: "*Aprovado*". Eu, Tadéa Maria da Silva, técnica em secretariado, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 23 de fevereiro de 2015.

Tadéa Maria da Silva

Prof. Dr. Andrei de Araújo Formiga
Orientador (PPGI-UFPB)

Prof[a]. Dr[a]. Natasha Correia Queiroz Lino
Examinadora Interna (PPGI-UFPB)

Prof[a]. Dr[a]. Thais Gaudencio do Rego
Examinadora Interna (PPGI-UFPB)

Prof. Dr.Thiago Pereira Falcão
Examinador Externo à Instituição

À minha esposa, Isabel Vitorino. Aos meus pais, Ednaldo Lopes e Luzimar Maria, e às minhas irmãs, Maria Andrea e Maria de Lourdes.

(AD MAJOREM DEI GLORIAM)

Agradecimentos

Em primeiro lugar, devoto meus agradecimentos à Santíssima Trindade sob o lema *"ad majorem Dei gloriam"*. Em seguida, agradeço à minha esposa, Isabel Vitorino, pelo seu cuidado e por compartilhar cada etapa de minha vida presente.

Agradeço aos meus pais por estarem diariamente incentivando e suportando minhas imperfeições. Aos meus sobrinhos, Pedro Henrique e Ana Júlia, pelas alegrias de fim de semana, bem como às minhas irmãs por todo o afeto familiar.

Em especial, agradeço ao Profº Andrei Formiga, por acreditar no meu trabalho e confiar na minha capacidade. De modo semelhante, agradeço os votos de sucesso dos professores Tiago Nascimento, Clauirton Siebra, Alisson Brito e Thaís Gaudêncio.

Por fim, dirijo meus agradecimentos aos inúmeros amigos e colaboradores.

Resumo

Nos últimos anos um grande número de pesquisas científicas fomentou o uso de informações da web como insumos para a vigilância epidemiológica e descoberta/mineração de conhecimentos relacionados a saúde pública em geral. Ao fazerem uso de conteúdo das mídias sociais, principalmente *tweets*, as abordagens propostas transformam o problema de identificação de conteúdo em um problema de classificação de texto, seguindo o cenário de aprendizagem supervisionada. Neste processo, algumas limitações atribuídas à representação das mensagens, atualização de modelo assim como a extração de atributos discriminativos, surgem. Partido disso, a presente pesquisa propõe investigar o impacto no desempenho de classificação de mensagens sociais curtas através da expansão contínua do conjunto de treinamento tendo como referência a medida de confiança nas predições realizadas. Paralelamente, a pesquisa também teve como objetivo avaliar alternativas para ponderação e extração de termos utilizados para a classificação, de modo a reduzir a dependência em métricas baseadas em frequência de termos. Restringindo-se à classificação binária de *tweets* relacionados a eventos de saúde e escritos em língua inglesa, os resultados obtidos revelaram uma melhoria de F1 de 9%, em relação a linha de base utilizada, evidenciando que a ação de expandir o classificador eleva o desempenho de classificação, também para o caso da classificação de mensagens curtas em domínio de saúde. Sobre a ponderação de termos, tem-se que a principal contribuição obtida, está na capacidade de levantar termos característicos do conjunto de dados e suas classes de interesse automaticamente, sem sofrer com limitações de frequência de termos, o que pode, por exemplo, ser capaz de ajudar a construir processos de classificação mais robustos e dinâmicos ao qual façam uso de listas de termos específicos para indexação em consultas à bancos de dados externos (*background knowledge*). No geral, os resultados apresentados podem beneficiar, pelo aprimoramento das hipóteses levantadas, o surgimento de aplicações mais robustas no campo da vigilância, controle e contrapartida à eventos reais de saúde (epidemiologia, campanhas de saúde, etc.), por meio da tarefa de classificação de mensagens sociais curtas.

Palavras-chave: Aprendizagem de máquina, classificação de texto, mensagens sociais.

Abstract

In recent years a large number of scientific research has stimulated the use of web data as inputs for the epidemiological surveillance and knowledge discovery/mining related to public health in general. In order to make use of social media content, especially tweets, some approaches proposed before transform a content identification problem to a text classification problem, following the supervised learning scenario. However, during this process, some limitations attributed to the representation of messages as well as the extraction of attributes arise. From this, the present research is aimed to investigate the performance impact in the short social messages classification task using a continuous expansion of the training set approach with support of a measure of confidence in the predictions made. At the same time, the survey also aimed to evaluate alternatives for consideration and extraction of terms used for the classification in order to reduce dependencies on term-frequency based metrics. Restricted to the binary classification of tweets related to health events and written in English, the results showed a 9% improvement in F1, compared to the baseline used, showing that the action of expanding the classifier increases the performance, even in the case of short message classification task for health concerns. For the term weighting objective, the main contribution obtained is the ability to automatically indentify high discriminative terms in the dataset, without suffering limitations regarding term-frequency. This may, for example, be able to help build more robust and dynamic classification processes which make use of lists of specific terms for indexing contents on external database (textit background knowledge). Overall, the results can benefit, by the improvement of the discussed hypotheses, the emergence of more robust applications in the field of surveillance, control and decision making to real health events (epidemiology, health campaigns, etc.), through the task of classifying short social messages.

Keywords: Machine learning, text classification, social messages.

Sumário

Lista de acrónimos e siglas

AM : *Aprendizagem de máquina*

AUC : *Área abaixo da curva (do inglês, Area under the curve)*

BOW : *Saco de palavras (do inglês, Bag-Of-Words)*

ROC : *Característica de Operação do Receptor (do inglês, Receiver Operating Characteristic, ROC)*

SVM : *Máquina de vetor de suporte (do inglês, Support Vector Machine)*

TF : *Frequência de Termos, (do inglês, term frequency)*

TF-IDF : *Frequência de termos - Frequência inversa de documentos (do inglês, Term Frequency - Inverse Document Frequency)*

Notação

Φ : *Classificador*

$\mathcal{D}$: *Conjunto de dados* $\mathcal{D} = (x_1, y_1), ..., (x_N, y_N)$*; tecnicamente não é um conjunto, mas sim um vetor de elementos* (x_n, y_n)*.* $\mathcal{D}$ *é tradicionalmente chamado de conjunto de treinamento, sendo frequentemente particionado para a geração de conjuntos de validação/teste*

$\mathcal{D}_{treino}$: *Subconjunto de* $\mathcal{D}$ *utilizado para treinamento quando um conjunto de validação ou teste é utilizado*

$\mathcal{D}_{val}$: *Conjunto de validação, subconjunto de* $\mathcal{D}$

$\overrightarrow{x}^{(i)}$: *entrada* $\overrightarrow{x} \in \mathcal{X}$*. Comumente* $\overrightarrow{x}^{(i)} \in \mathbb{R}^d$ *ou* $\overrightarrow{x} \in \{1\} \times \mathbb{R}^d$*. Apenas* x *é utilizado quando a entrada é um escalar*

$\mathcal{X}$: *Espaço de entrada, cujo elementos são* $\overrightarrow{x} \in \mathcal{X}$

$y^{(i)}$: *Classe ou rótulo* $y \in \mathcal{Y}$

$\widehat{y}$: *Estimativa de y através de* Φ

t : *termo/palavra*

$x^{(i)}$: *documento*

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Capítulo 1

Introdução

1.1 Motivação

Segundo Bermingham e Smeaton (2010), a "*WEB* em tempo real" refere-se à parcela da *WEB* onde a informação está disponível logo após ser criada e onde está ligada de alguma forma aos eventos ocorridos no mundo real, seja no tempo exato ou próximo a ele. O conteúdo produzido por usuários deste ambiente tem sido explorado como uma fonte de informação promissora, tanto do ponto de vista factual quanto do ponto de vista subjetivo, tendo inspirado pesquisas em diversas áreas, como: a análise automática de sentimento (métodos para detecção automática de opiniões e outras avaliações textuais), a estatística, a aprendizagem de máquina, a mineração de dados e a recuperação da informação.

Inserido neste meio, as Redes Sociais *Online* (RSO), tais como Facebook e Twitter, possibilitam aos usuários a criação de mensagens que podem estar relacionadas a uma ampla variedade de tópicos e que usualmente revelam substanciais sobreposições com eventos do cotidiano. Sendo assim, o potencial desse meio de comunicação como base para a coleta de estatísticas e para a exploração de informação sobre comportamento de usuários é atrativo, dado que tanto a diversidade quanto a ampla quantidade de conteúdo disponível proveem uma oportunidade única de investigação de vários fenômenos humanos (SCELLATO et al.,

2010). Por exemplo, a análise de *tweets*[1] em tempo real é considerada como um veículo produtivo para a predição de fenômenos naturais, tais como: terremotos (SAKAKI; OKAZAKI; MATSUO, 2010); (CARAGEA et al., 2011) e emergência de epidemias (GINSBERG et al., 2008); (CULOTTA, 2010); (GOMIDE, 2012).

Especificamente nos últimos anos, um elevado número de pesquisas científicas fomentou o uso de informações da *WEB* como insumos para a vigilância epidemiológica e descoberta de conhecimentos relacionados a saúde pública em geral. Essas pesquisas objetivaram essencialmente correlacionar o volume de informação coletada (mensagens de redes sociais, registros em motores de busca, *blogs*, fórums, etc.) sobre uma determinada doença (por exemplo, gripe/dengue) com as informações oficiais divulgadas por agências do governo, corroborando a utilidade da primeira como recurso complementar válido e de baixo custo para predição de taxas de doenças (CULOTTA, 2010); (LAMPOS; CRISTIANINI, 2010); (CHEN et al., 2010); (ACHREKAR et al., 2011); (LAMPOS; CRISTIANINI, 2012); (GOMIDE, 2012). Outros focos estiveram relacionados à consideração de um maior nível de atributos sociais (as interações interpessoais, influência, mobilidade humana, escolaridade e proximidade à fontes de poluição) pela exploração de informações de geolocalização anexadas ao conteúdo produzido pelos usuários, visando, desta forma, estimar as chances de um indivíduo contrair uma doença em um futuro próximo dado exposição a pessoas contaminadas, incluindo amigos (SADILEK; KAUTZ; SILENZIO, 2012a); (SADILEK; KAUTZ; SILENZIO, 2012b); (SADILEK; KAUTZ, 2013); (BRENNAN; SADILEK; KAUTZ, 2013).

Comumente, sistemas que tentam extrair informações de mensagens sociais curtas, como o caso dos *tweets*, transformam o problema de identificação de conteúdo em um problema de classificação de documentos (texto) e, desse modo, utilizam técnicas de aprendizagem de máquina para classificar as informações em classes de interesse, geralmente, seguindo a abordagem de aprendizagem supervisionada. No entanto, dada a brevidade das mensagens, a esparsividade e os ruídos associados, tem-se a necessidade de buscar alternativas metodológicas que venham atenuar, ou até mesmo superar, limitações decorrentes do uso

[1]Mensagens geradas por usuários do serviço de *microblogging* twitter.com. Possui um limite máximo de 160 caracteres.

de técnicas tradicionais de representação de texto para fins de classificação automática no domínio citado. Por exemplo, de acordo com Sriram et al. (2010) e Tuarob et al. (2014), a utilização de métodos de técnicas de classificação tradicionais as quais representam o documento como um saco de palavras (em inglês, *bag-of-words, BOW*), na qual o mesmo é representado como uma lista de palavras e suas respectivas frequências de ocorrência (ordem das palavras é ignorada), sofrem bastante em cenários onde os documentos de interesse não são ricos em texto bem como não são conformes com a linguagem padrão, como no caso dos *tweets*. Além disso, os sistemas propostos neste cenário devem considerar mecanismos que incorporem novas informações (modelos adaptativos) com o mínimo de intervenção humana[2], e, paralelamente, identificar quais palavras são mais discriminativas nos documentos de interesse, o que pode ser alcançado pela ponderação de termos (TIMONEN, 2013).

Partindo disso, a presente dissertação descreve resultados preliminares na busca de melhorias no processo de classificação automática de mensagens curtas em redes sociais, restringindo o foco à classificação de *tweets* no contexto da descoberta de conteúdos relacionados a eventos de saúde. Seguindo este propósito, investigou-se uma metodologia básica semi-supervisionada de classificação adaptativa, fundamentada no modelo *Naïve Bayes*, visando elevar o desempenho de classificação a partir da confiança na predição de novas mensagens de teste e a expansão do conjunto de treinamento. Para isso, fez-se uso dos estudos preliminares de Silva (2012) e Zimmermann, Ntoutsi e Spiliopoulou (2014), porém, diferentemente dos autores, optou-se pelo foco no modelo de classificação em tópicos (de saúde), ao invés de opiniões e sentimentos. Paralelamente, a presente pesquisa revisitou algumas hipóteses levantadas por Timonen (2013) no que diz respeito a busca por novas formas de ponderação de termos presentes em mensagens curtas, porém, diferentemente do foco dos autores no campo da classificação de opiniões de usuário, o escopo de pesquisa esteve diretamente relacionado com a classificação de documentos curtos (*tweets*), no que diz respeito à descoberta de conhecimento em saúde.

Assim como em Tuarob et al. (2014), no decorrer deste documento de pesquisa, tem-se

[2]já que estes enfrentam o desafio de serem treinados com uma quantidade limitada de informações rotuladas (ZIMMERMANN; NTOUTSI; SPILIOPOULOU, 2014), (SILVA, 2012).

por definição que uma mensagem é dita estar relacionada com eventos de saúde (ou simplesmente relacionada à saúde) se esta segue pelo menos uma das seguintes condições: 1) A mensagem indica que o autor tem problema/sintomas de saúde; 2) a mensagem indica que uma outra pessoa está doente ou expressa preocupações com a saúde em geral.

1.2 Objetivos

O objetivo geral deste trabalho é investigar o impacto no desempenho de classificação de mensagens sociais curtas através da expansão contínua do conjunto de treinamento tendo como referência a confiança da predição realizada nas mensagens de teste. Para o cálculo da confiança, utilizou-se o limiar adaptativo proposto por Silva (2012), porém, o presente trabalho de pesquisa diferencia do autor no contexto de que o modelo de classificação investigado baseia-se no método *Naïve Bayes*, e não em regras de associação. Além disso, o escopo do projeto foca em classificação de (tópicos) documentos, diferente do foco dos autores em análise de sentimentos (um caso mais específico de classificação).

A avaliação experimental restringe-se à classificação binária de *tweets* relacionados a eventos de saúde e escritos em língua inglesa[3] , sendo as classes: “pos”, indicadora de relação com saúde; e “neg” indicadora de que a mensagem não tem relação com saúde.

Paralelamente, a pesquisa também teve como objetivo avaliar alternativas para ponderação e extração de termos utilizados para a classificação, de modo a reduzir a dependência em métricas baseadas em "frequência de termos"do inglês, *term frequency* (TF). Dado que em documentos curtos[4] raramente se encontram palavras que ocorram mais de uma vez, revisitou-se o trabalho de Timonen (2013) na tentativa de ponderar termos eficientemente, tendo como referência a distribuição dos mesmos em nível tanto inter como intra classes,

[3]A opção pelo uso de tweets em língua inglesa é uma decisão de projeto justificada pelo fato da disponibilidade massiva de ferramentas de processamento de linguagem natural voltada para o inglês e as inúmeras aplicações que usam os dados citados em inglês. Porém, os resultados não estão limitados ao inglês sendo passíveis de investigação futura tendo outros idiomas, tais como o português, como recurso principal.

[4]Como o é o caso de *tweets* onde há no máximo 140 caracteres e cerca de 20 palavras (TIMONEN, 2013).

percebendo, desta forma, se é possível obter resultados equivalentes ou superiores a modelos usando TF e TF-IDF, este do inglês, *Term Frequency - Inverse Document Frequency* (SALTON; BUCKLEY, 1988).

1.2.1 Questões de pesquisa

Sabendo que as técnicas de aprendizado semisupervisionado permitem melhorar a eficácia de classificadores utilizando os dados não rotulados (CHAPELLE; SCHLKOPF; ZIEN, 2010), enumerou-se a primeira questão de pesquisa:

- *É possível, usando uma estratégia de expansão de conjunto de treinamento, obter um desempenho de classificação melhor que um modelo tradicional (estático), para o caso da classificação de tweets relacionados à saúde?*.

Disso, a hipótese levantada é a de que: através de uma estimativa de confiança na predição realizada, é possível ampliar o conhecimento do classificador, e, consequentemente, elevar o desempenho de classificação para o caso de mensagens sociais curtas (SILVA, 2012); (ZIMMERMANN; NTOUTSI; SPILIOPOULOU, 2014).

Subsequentemente, no intuito de focar na descoberta de termos mais discriminativos para fins de classificação em mensagens curtas, observou-se formas de enfatizar palavras/termos presentes em *tweets* de modo a atenuar limitações de frequência de termos no contexto da tarefa de classificação citada. Assim, formulou-se a segunda questão de pesquisa:

- *É possível, utilizando as métricas de "contagem inversa de categoria", "probabilidade de categoria"e "probabilidade de documentos", retratadas em Timonen (2013), obter uma discriminação de termos consistente*[5] *no contexto da classificação de mensagens relacionadas à saúde?*

[5]No que diz respeito à relação destes com termos de saúde.

1.3 Justificativa

Conforme discutido na seção 1.1, sistemas que fazem uso de conteúdo de mídias sociais, principalmente *tweets*, para a detecção e monitoramento de eventos de saúde, utilizam técnicas de aprendizagem de máquina para classificar as informações em classes de interesse. Nestas técnicas, é comum que o conteúdo seja representado usando BOW, onde a ordem das palavras/termos é completamente ignorada, restando apenas uma lista de palavras e suas respectivas frequências de ocorrência[6]. Porém, no caso da mineração de conhecimento em saúde, por exemplo, essa técnica possui limitações na qual é possível enumerar duas principais, segundo o trabalho de Tuarob et al. (2014): 1) palavras com múltiplos significados têm tendência a serem tratadas igualmente (ex. *cold*, *sick*, as quais podem ser usadas em múltiplos contextos) e 2) palavras-chave importantes são tratadas como palavras normais devido a baixa representatividade (frequência) no conjunto de treinamento, o que dificulta a possibilidade de a mesma ser identificada como uma palavra discriminativa, ex. *Xeroderma pigmentosum*.

Sendo assim, tem-se a necessidade de investir especificamente em métodos de ponderação/seleção de termos discriminativos, de modo a oferecer um processo de classificação mais robusto e significativo em face do expansivo crescimento de aplicações, tanto científicas quanto comerciais, voltadas à mineração de fenômenos do mundo real (epidemias, desastres naturais e campanhas publicitárias, etc.). Além disso, conforme apresentado em Silva (2012) e Zimmermann, Ntoutsi e Spiliopoulou (2014), um modelo de classificação estático não é adequado para o conteúdo gerado por mídias sociais. Isso porque os dados mudam constantemente e esta técnica não dispõe de recursos para manter-se atualizada e, em adição, existe uma escassez de dados rotulados para treinamento constante do classificador.

[6] Algoritmos de aprendizagem geralmente utilizam vetores numéricos como representação interna dos dados de interesse.

1.4 Estrutura da Dissertação

O restante deste documento de pesquisa possui a seguinte estrutura: Capítulo 2, fornece a fundamentação teórica associada. Capítulo 3, os trabalhos relacionados. A abordagem de classificação expansiva é apresentada no Capítulo 4 assim como a de ponderação de termos no Capítulo 5, respectivamente. Capítulo 6 traz a análise experimental e discussão conclusiva. Por fim, o capítulo 7 fornece o apontamento para os trabalhos futuros.

Capítulo 2

Fundamentação Teórica

O presente capítulo oferece uma fundamentação teórica dos assuntos diretamente ligados ao escopo de pesquisa. Apresenta-se uma visão geral sobre tweets, sobre o problema de classificação de documentos, ponderação de termos e medidas de avaliação de algoritmos.

2.1 Mensagens sociais curtas: *tweets*

A experimentação desenvolvida nesta pesquisa é baseada em dados obtidos do Twitter, um serviço de *microblogging* onde os usuários têm um limite de até 140 caracteres para escrever mensagens livres, expressando opiniões pessoais e/ou qualquer tópico que lhe seja de interesse. Tais mensagens são conhecidas como *tweets* e as relações entre usuários na plataforma não são necessariamente simétricas, significando que um usuário pode seguir[1] outro usuário sem que haja necessidade deste último seguir o primeiro de volta. Quando a relação entre dois usuários é simétrica, diz-se que estes são "amigos" (SADILEK; KAUTZ; SILENZIO, 2012a). Segundo Gruzd, Wellman e Takhteyev (2011), há evidências empíricas de que relações de amizade no Twitter tem uma substancial sobreposição com amizades no mundo real, o que sustenta a utilização de *tweets* em pesquisas em tópicos de saúde, principalmente epi-

[1] "Seguir", neste contexto, é a ação de passar a receber mensagens de atualização de outro usuário, isto é, passar a acompanhar as postagens de outro usuário.

demiologia. O Twitter, lançado em 2006, tem experimentado um crescimento exponencial nos últimos anos. Em março de 2011, aproximadamente 200 milhões de contas registradas no serviço[2]. Em 2014, o Twitter registrou mais de 241 milhões de usuários ativos por mês e uma média de 500 milhões de tweets enviados por dia[3].

2.1.1 Estrutura de um tweet

Apesar da brevidade das mensagens (140 caracteres), estas possuem elementos especiais dos quais tornaram-se, devido a popularidade do serviço, amplamente conhecidos:

Retweet Caracterizado pelas letras "RT" no início das mensagens, indica que a mensagem é cópia de outra já existente. Trata-se também da ação de compartilhar uma mensagem existente.

Usuário (*User*) Uma *string* única associada a cada conta de usuário, usualmente interpretado como nome do usuário (*username*).

Resposta (*Reply*) e menções o símbolo de arroba "@"seguido pelo nome do usuário (ex.: *@ewerlopes*) indica que a mensagem é direcionada a um usuário específico.

Hashtags o sinal de cerquilha '#' seguido por uma palavra (ex.: #ebola, #flu) é utilizado para organizar mensagem em categorias indexáveis por motores de busca. Pode ser visto também como um processo de metaclassificação.

Emoticon Forma de comunicação paralinguística, é uma sequência de caracteres tipográficos, tais como: :), :(, :D , >:(, :3 e :-); ou, também, uma imagem (usualmente, pequena), que traduz ou quer transmitir o estado psicológico, emotivo, do usuário, por meio de ícones ilustrativos de uma expressão facial. Exemplos: :) (isto é, sorrindo, estou alegre); :((estou triste, chorando), etc.

URLs Hiperlinks genéricos compartilhados pelos usuários.

---

[2] <http://goo.gl/mefeht>. Acesso em 02 de agosto de 2014.

[3] <http://goo.gl/94cPjE>. Acesso em 02 de agosto de 2014.

É possível coletar *tweets* por meio de diversas APIs[4] fornecidas pelo serviço. Através destas, pode-se obter a rede de seguidores, os dados de perfil, as últimas mensagens de um usuário em particular, etc. Usando a *API streaming*[5] é possível coletar, em tempo real, *tweets* públicos que contenham termos específicos, configurados no momento da busca. Esse método tem sido amplamente usado para a criação de conjuntos de dados em pesquisas recentes.

2.2 Classificação de documentos

Classificação, no contexto de Aprendizagem de Máquina (AM), trata-se de uma tarefa onde o objetivo é construir um modelo que seja capaz de estimar classes (rótulos ou ainda categorias) y_j pertencentes a um conjunto discreto $\mathcal{Y}$, onde $j = 1, ..., |\mathcal{Y}|$, a partir de vetores de características. Classificação de documentos (texto), por sua vez, refere-se a tarefa de classificação onde os vetores de características são formados a partir de informações textuais.

Supondo a existência de um conjunto de documentos $\mathcal{D}_{train}$ onde cada documento $x \in \mathcal{D}$ está associado a uma classe discreta[6] $y_j \in \mathcal{Y}$, a questão central se baseia na busca de como aprender a prever as classes de novos documentos $x^{(\tau)} \notin \mathcal{D}_{train}$, em função dos termos (palavras) t que este possui.

Visando estabelecer uma notação, usa-se $x^{(i)}$ para identificar as variáveis de entrada (documentos), também chamados **características** de entrada, e $y^{(i)}$ para identificar as **classes**, ou "saída"alvo que se pretende prever (categoria). Um par $(x^{(i)}, y^{(i)})$ é chamado de **exemplo de treinamento**, e o conjunto de dados utilizado para aprendizagem - uma lista de N exemplos de treinamento $\{(x^{(i)}, y^{(i)}); i = 1, ..., N\}$ - é chamado de *conjunto de treinamento* $\mathcal{D}_{train}$[7]. Além disso, usa-se $\mathcal{X}$ como o espaço de valores de entrada e $\mathcal{Y}$ como espaço de

[4] `https://dev.twitter.com/docs/api/1.1`

[5] `https://dev.twitter.com/docs/streaming-apis`

[6] Por exemplo: Um conjunto de documentos jornalísticos, associados a classes como: esportes, novela, ciência, natureza, etc.

[7] Note-se que o sobrescrito "(i)"na notação é simplesmente um índice para o conjunto de treinamento, e

valores de saída.

O processo de aprendizagem de um modelo de classificação Φ possui etapas bem definidas. Inicialmente, cada documento x pertencente a uma coleção de documentos $\mathcal{D}$ forma um vetor de característica $\overrightarrow{x}$ (cf. seção 2.2.1) a partir do mapeamento de cada termo $t_j \in x$ em uma característica $x_j \in \overrightarrow{x}^{(i)}$. Esse mapeamento pode ser feito através de um método de extração de características (extrator), ao qual converte texto em vetores $\overrightarrow{x}^{(i)}; i = 1, ..., N$. Em seguida, tem-se o processo de treinamento ao qual toma um conjunto de exemplos $\{(x^{(i)}, y^{(i)}); i = 1, ..., |\mathcal{D}_{train}|\}$, onde $\mathcal{D}_{train} \subset \mathcal{D}$, como entrada para Φ. Frequentemente, Φ é uma função (modelos descritivos) que estima classes a partir de vetores de características. O tipo de função (complexidade), no entanto, varia com o modelo de selecionado. Por fim, após treino, Φ pode ser utilizado para prever classes $\widehat{y} \in \mathcal{Y}$ para vetores com classes desconhecidas, isto é, novos exemplos $\{(x^{(i)}, ?)\}$. Durante o treinamento, o processo de fornecer na entrada, além dos documentos, as classes a qual cada um pertence, é chamado de *Classificação Supervisionada*. A Figura 2.1 retrata o processo citado.

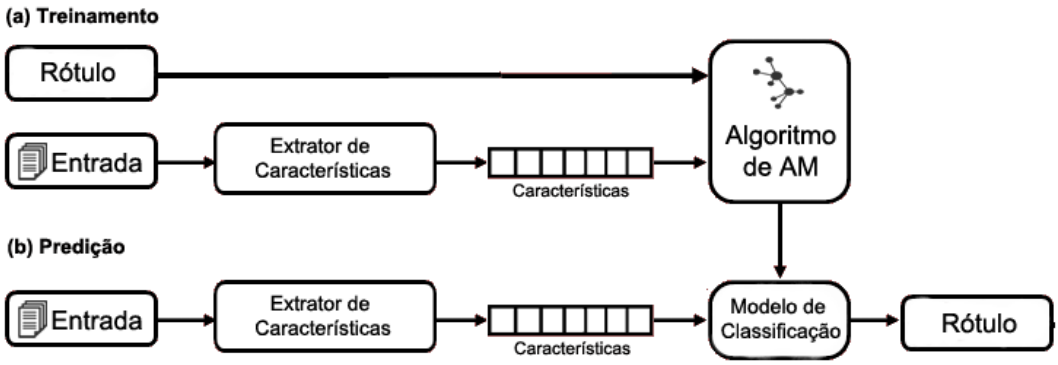


Figura 2.1: Processo de classificação supervisionado. (a) Durante o treinamento, um extrator de características é usado para converter cada valor de entrada (documentos) em seu respectivo vetor de características. Estes, capturam as informações básicas sobre cada entrada as quais devem ser utilizadas para fins de representação (frequência da palavra, por exemplo). Os vetores, assim como os rótulos, são fornecidos ao algoritmo de aprendizagem de máquina para gerar um modelo. (b) Durante a previsão (ou classificação), o mesmo processo de extração de característica acontece. Os vetores, então, são apresentados ao modelo treinado, gerando o rótulo de classificação (BIRD; KLEIN; LOPER, 2009).

não possui relação com a operação de exponenciação.

A performance de Φ, por sua vez é estimada através de um conjunto de teste $\mathcal{D}_{teste}$, geralmente $\mathcal{D}_{teste} = \overline{\mathcal{D}_{train}}$, onde as classes para cada vetor $\overrightarrow{x}^{(i)}, i = 1, ...|\mathcal{D}_{teste}|$ também são conhecidas previamente, embora não dispostas ao modelo. É comum que a medida de desempenho empregada na avaliação de Φ seja realizada através de **taxas de erros**. Isso se dá de maneira simples pela comparação de predições em cada $\overrightarrow{x}^{(i)} \in \mathcal{D}_{teste}$: se correta, conta-se um sucesso; se não, tem-se um erro. A taxa de erro neste cenário apresenta-se, então, como a proporção de erros obtidos em um conjunto de dados independente e mensura a qualidade geral do classificador. Formalmente, como exposto por Faceli et al. (2011), a equação 2.1 define o processo pelo qual a taxa de erro (err) é definida, sendo que $[\![\cdot]\!]$ é um teste[8] pelo qual assume 1, se verdadeiro; e 0, caso contrário. Para um conjunto de dados com n objetos a serem classificados, a taxa descreve a proporção de objetos classificados incorretamente por Φ e é calculada a partir da comparação da classe predita ($\Phi(\overrightarrow{x}^{(i)}) = \widehat{y}^{(i)}$) e a classe verdadeira de $\overrightarrow{x}^{(i)}$, isto é, $y^i \in \mathcal{Y}$.

$$err(\Phi) = \frac{1}{n}\sum_{i=1}^{n}[\![y^i \neq \widehat{y}^{(i)}]\!] \tag{2.1}$$

Sendo que a taxa de erro obtida tem uma variação entre 0 e 1, quanto mais próximos do extremo 0, melhor o desempenho de Φ. Pelo complemento da Equação, tem-se a taxa de acerto ou *acurácia* (ac) do classificador. Simetricamente, quanto mais próximo de 1 melhor o desempenho (Equação 2.2).

$$ac(\Phi) = 1 - err(\Phi) \tag{2.2}$$

Na maioria dos casos, a quantidade de informações de exemplo disponível é limitada. O que por sua vez limita a quantidade de informações presentes nos conjuntos de treinamento e teste. A ação de separar uma certa partição $\mathcal{D}_{teste} \subset \mathcal{D}$ para teste enquanto usa o resto para treinamento é conhecida como método de *holdout*. De acordo com Witten e Frank (2005),

[8]Faceli et al. (2011) chama este teste de "função de custo zero-um (0-1)". embora não se tratar formalmente de uma função, este conceito visa atribuir um valor para o custo de uma previsão incorreta, $(\widehat{y}^{(i)} \neq y^{(i)}) = 1$, e o custo de uma previsão correta, $(\widehat{y}^{(i)} = y^{(i)}) = 0$.

é comum que seja separado $\frac{1}{3}$ dos dados para teste e o restante ($\frac{2}{3}$) para treinamento. Além disso, na intenção de atenuar qualquer problema envolvendo o balanceamento de classes nos conjuntos de dados disponíveis, é aconselhável usar um processo de amostragem que preserve a mesma proporção de classes tanto no conjunto de treinamento quanto no de teste[9].

Uma técnica estatística derivada do método *holdout* e que é bastante utilizada na literatura é a **validação cruzada** (*cross-validation*). Nesta técnica, $\mathcal{D}$ é subdividido em ϵ partições, onde, em ϵ iterações, $(\epsilon - 1)$ partições são utilizadas para treinamento e o restante para teste. Ainda segundo Witten e Frank (2005), o procedimento padrão para a predição de taxas de erro através dessa técnica, é usar $\epsilon = 10$ onde, em 10 iterações, 1 partição é reservada para teste (obtenção de resultados de desempenho usando estimativas de erro e métricas mais específicas, cf. Seção 2.3) e as outras 9 são utilizadas para treinamento. Ao final, calcula-se a média aritmética dos resultados obtidos durante as iterações, resultando em uma estimativa de desempenho geral.

2.2.1 Representação textual e ponderação de termos

Ao analisar informações textuais, pode-se observar que as palavras formam a unidade básica de construção de texto. O conceito de *n-gramas*, nas áreas de recuperação da informação e de classificação automática de documentos, é usado para determinar uma sequência de n palavras. Porém, a literatura pode também associar o conceito à uma cadeia de caracteres de comprimento n, deve-se então ficar claro a partir de qual contexto se pretende usar. Nesta dissertação, o conceito está atrelado à sua relação com palavras, muitas vezes chamado de *termos*, sendo, então, a unidade atômica de cadeia de texto, denotada por t. Partindo disso, tem-se que *1-grama* (unigrama), representa termos isolados, isto é, usado para determinar que uma única palavra possui seu próprio poder discriminativo e, assim, é utilizado como uma característica básica de representação, ex.: *"car", "flu", "baby", etc*. Por sua vez, 2-gramas (bigrama), denota uma sequência de 2 termos, por exemplo, *"my house", "my car",*

[9]Esse procedimento é chamado como *estratificação*. Com ele é possível falar em método de *holdout* estratificado. Porém, embora bem aceito, a estratificação provê apenas uma proteção básica contra representações desbalanceadas nos conjuntos de treinamento e teste (WITTEN; FRANK, 2005).

"the words" e, neste caso, os termos não são considerados separadamente, mas em sequência de dois termos. A generalização segue naturalmente com 3-gramas (trigrama), ex.: *"The white house", "The blue shoes"*.

Supondo o conhecimento de todos os termos que podem aparecer em uma determinada linguagem e denominando tal conjunto de vocabulário $\mathcal{V}$, transforma-se qualquer documento $x^{(i)}$ em uma representação de vetor $\overrightarrow{x}^{(i)}$, onde $\overrightarrow{x}^{(i)} = \langle x_1, ..., x_{|\mathcal{V}|} \rangle$[10], através da ponderação de cada $t_j \in x^{(i)}$, obedecendo uma predefinida ordem lexicográfica predefinida em $\mathcal{V}$. Vale salientar que $\mathcal{V}$ pode ser formado só por unigramas, por bigramas, por trigramas, etc; e/ou apresentar características híbridas de n-grams.

Intuitivamente, a ponderação de termos objetiva calcular um grau de relação entre termos $t_j \in x^{(i)}$ e documentos $x^{(i)} \in \mathcal{D}_{train}; i = 1, ..., |\mathcal{D}_{train}|$ e, desta forma, levantar estimativas discriminantes que estime uma classe $\widehat{y} \in \mathcal{Y}$ para todo novo documento $x^{(\tau)} \in D_{teste}$, tendo como referencia $t_j \in x^{(\tau)}$. Nesse contexto, a abordagem mais simples é atribuir o peso de t como sendo igual ao número de ocorrências deste em documentos $x^{(i)} \in \mathcal{D}_{train}$ (documentos de treinamento). Este sistema de ponderação é referido como *frequência de termos* e é denotado por $TF_{t,d}$ (MANNING; RAGHAVAN; SCHÜTZE, 2008). Partindo disso, o espaço de documentos $\mathcal{X}$, o domínio da função de classificação, assume $\mathbb{R}^{|\mathcal{V}|}$. Outra alternativa, no entanto, é considerar $\overrightarrow{x}^{(i)} \in \{0, 1\}^{|\mathcal{V}|}$ (o que é conhecido como *vetor de documento binário*, ou simplesmente *vetor binário*), considerando apenas a existência do termo t em d.

De fato, para um documento $(x^{(i)})$, o conjunto de pesos determinados por *TF* (ou na verdade, qualquer função de ponderação, que mapeie o número de ocorrências de t em $(x^{(i)})$ a um valor real positivo) pode ser visto como um resumo quantitativo do documento. Neste ponto de vista, conhecido na literatura como técnica de representação BOW, a ordem exata dos termos em um documento é ignorada, mas o número de ocorrências de cada termo é mantido (em contraste com recuperação binária). Retem-se apenas informações sobre o número de ocorrências de cada termo. Assim, o documento *"Mary is quicker than John"* é, neste ponto de vista, idêntico ao documento *"John is quicker than Mary"*. É conhecido que

[10] também chamado de vetor de características ou ainda vetor de documento.

Tabela 2.1: Dois vetores de características de um mesmo documento: um vetor booleano (binário) e outro baseado em frequência. O vocabulário ($\mathcal{V}$) foi severamente limitada para fins de simplicidade. Em um ambiente prático, $|\mathcal{V}| \gg 100.000$ e, neste caso, os vetores resultantes são bastante esparsos. Capitalização é ignorada visando redução de termos em $\mathcal{V}$.

Vocabulário ($\mathcal{V}$)	'a', 'brown', 'dog', 'fox', 'i', 'is', 'jumped', lazy', 'over', 'quick', 'the', 'this'
Documento ($x^{(i)}$)	"The quick brown fox jumped over the lazy dog"
Termos (ordem)	'brown', 'dog', 'fox', 'jumped', 'lazy', 'over', 'quick', 'the', 'the'
Vetor binário ($\overrightarrow{x}_{bin}$)	(0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0)
Vetor de frequência ($\overrightarrow{x}_{freq}$)	(0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 2, 0)

dois documentos com representações BOW semelhantes são também similares em conteúdo (MANNING; RAGHAVAN; SCHÜTZE, 2008).

A Tabela 2.1 mostra um exemplo do processo. Na prática, não se tem conhecimento de todos os possível termos que aparecerão em um texto, mas, como solução, pode-se ignorar palavras que não foram vistas anteriormente ou adicioná-las ao vocabulário elevando a dimensão dos vetores.

Além dos métodos de ponderação citados, é comum a utilização de um esquema chamado *frequência de termo - frequência inversa de documentos* (do inglês, *Term frequency – inverse document frequency*, TF-IDF). Trata-se de estatística numérica que visa refletir a importância de uma palavra para um documento em uma coleção ou *corpus*. O valor dessa estatística aumenta proporcionalmente ao número de vezes em que uma palavra aparece no documento, mas é compensado pela frequência da palavra na coleção, o que ajuda a ajustar para o fato de que algumas palavras aparecem com mais frequência na maioria dos documentos idependentemente das classes a qual estes pertençam. Estas, são chamadas de *stopwords*. Apesar de ser amplamente utilizada como um fator de ponderação na recuperação de informação (motores de busca) e mineração de texto, TF-IDF pode ser usado com sucesso para a filtragem de *stopwords* em vários campos, incluindo a sumarização de texto e classificação.

Em Tuarob et al. (2014), este esquema foi utilizado na formulação de vetores de características durante o processo de classificação de *tweets*.

A fórmula do esquema TF-IDF (Equação 2.3) baseia-se nas definições de *frequência de termo* e *frequência inversa de documento* para produzir uma combinação que pondere cada termo em cada documento.

$$\text{TF-IDF}_{t,d} = TF_{t,d} \times IDF_t \tag{2.3}$$

onde, IDF (do inglês, *Inverse Document Frequency*) é a frequência inversa de documento, definida como:

$$IDF_t = log\frac{N}{df_t} \tag{2.4}$$

e, por sua vez, N denota o número total de documentos na coleção e df_t, o número de documentos na coleção que contêm um termo t (MANNING; RAGHAVAN; SCHÜTZE, 2008).

Ainda segundo Manning, Raghavan e SchÜtze (2008), tem-se que TF-IDF:

1. maior, quando t ocorre muitas vezes dentro de um pequeno número de documentos (aumentando o poder de discriminação desses documentos);
2. menor, quando o termo ocorre poucas vezes em um documento, ou ocorre em muitos documentos (oferecendo assim um sinal relevante menos acentuado);
3. menor, quando o termo ocorre em praticamente todos os documentos.

2.3 Medidas de avaliação de algoritmos em AM

Conforme apontado anteriormente, o modelo de AM pertencente ao escopo desta pesquisa se refere ao problema de classificação binária. Neste, uma classe $y \in \mathcal{Y}$ é denotada como

"pos" (positiva) e a outra como "neg" (negativa)[11], isto é, $\{pos, neg\} \subseteq \mathcal{Y}$. Como estrutura de visualização de desempenho, tem-se a Matriz de Confusão (MC). Esta oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas *versus* as classificações preditas por Φ para cada classe, sobre o conjunto de exemplos $\mathcal{D}_{teste}$. Uma matriz de confusão possui tamanho $L \times L$, onde L é o número de diferentes classes $y \in \mathcal{Y}$ (FACELI et al., 2011). A Tabela 2.2 ilustra uma matriz de confusão para o caso da classificação binária, em que:

- **Vendadeiro Positivo (VP)** corresponde ao número de verdadeiros positivos, isto é, o número de exemplos positivos preditos como corretamente como positivos;
- **Verdadeiro Negativo (VN)** corresponde ao número de verdadeiros negativos. Exemplos da classe negativa classificados corretamente como negativos;
- **Falso Positivo (FP)** corresponde ao número de falsos positivos. Exemplos da classe negativa preditos incorretamente como pertencentes a classe positiva;
- **Falso Negativo (FN)** corresponde ao número de falsos negativos, isto é, o número de exemplos pertencentes originalmente à classe positiva e que foram incorretamente classificados como negativos.

A partir da matriz de confusão é possível determinar a quantidade de exemplos disponíveis N, onde $N = VP + VN + FP + FN$.

Tabela 2.2: Matriz de confusão para um problema da classificação binária

	$\widehat{y} = neg$	$\widehat{y} = pos$
$y = neg$	VN	FP
$y = pos$	FN	VP

A partir da matriz de confusão, uma série de outras medidas de desempenho são derivadas (FACELI et al., 2011):

[11] Também é comum a utilização dos escalares: 0, para negativa, e 1, para positiva.

- **Taxa de erro na classe positiva** também conhecido como *Taxa de Falso Negativo* (TFN), corresponde a proporção de exemplos da classe positiva incorretamente classificados por Φ como pertencentes a classe negativa:

$$err_{pos}(\Phi) = \frac{FN}{VP + FN} \tag{2.5}$$

- **Taxa de erro na classe negativa** também conhecida como *Taxa de Falso Positivos* (TFP), corresponde a proporção de exemplos da classe negativa incorretamente classificados como pertencentes a classe positiva:

$$err_{neg}(\Phi) = \frac{FP}{FP + VN} \tag{2.6}$$

- **Taxa de erro total** apresentada como a soma dos valores das diagonais secundárias da matriz de confusão, dividida pela soma de todos os valores da matriz:

$$err(\Phi) = \frac{FP + FN}{N} \tag{2.7}$$

- **Taxa de acerto ou acurácia total** calculada pela soma dos valores da diagonal principal, divido pelo número total de elementos da matriz:

$$ac(\Phi) = \frac{VP + VN}{N} \tag{2.8}$$

- **Precisão (*precision*)** proporção de exemplos originalmente positivos classificados corretamente em relação a todos os elementos apontados por Φ, como sendo pertencentes a classe:

$$prec(\Phi) = pr(\Phi) = \frac{VP}{VP + FP} \tag{2.9}$$

- **Sensibilidade ou Revocação (*Recall*)** equivale à taxa de acerto na classe positiva, isto é, a *Taxa de Verdadeiros Positivos* (TVP):

$$sens(\Phi) = rev(\Phi) = re(\Phi) = \frac{VP}{VP + FN} \tag{2.10}$$

- **Especificidade** corresponde a taxa de acerto na classe negativa cujo complemento corresponde à *TFP*:

$$esp(\Phi) = \frac{VN}{VN + FP} = 1 - TFP(\Phi) \tag{2.11}$$

A precisão pode ser vista como uma medida de exatidão do modelo, enquanto que a revocação pode ser vista como uma medida de completude. De modo geral, essas duas medidas não são discutidas separadamente, isso porque ao obter um desempenho de Φ igual a (1,0) (*precision/recall*) para uma classe qualquer $y \in \mathcal{Y}$, tem-se que cada item rotulado pertence à classe y, mas não há informação a respeito do número de exemplos originalmente em y que não foram classificados corretamente. O oposto ocorre quando para a classe y o desempenho de Φ é igual a (0,1) (*precision/recall*). Assim, as duas métricas são combinadas em uma única medida de performance, conhecida como **medida-F**, que é a média harmônica ponderada, como ilustra a equação 2.12 (FACELI et al., 2011).

$$F_m(\Phi) = \frac{(\beta + 1) \times rev(\Phi) \times prec(\Phi)}{rev(\Phi) + \beta \times prec(\Phi)} \tag{2.12}$$

Quando há o interesse em atribuir um mesmo grau de importância tanto para a precisão quanto para a revocação, usa-se $\beta = 1$, o que é conhecido como **medida F1** (Equação 2.13).

$$F_1(\Phi) = \frac{2 \times prec(\Phi) \times rev(\Phi)}{prec(\Phi) + rev(\Phi)} \tag{2.13}$$

2.4 Considerações Finais

Este capítulo forneceu elementos conceituais fundamentais para o presente documento de pesquisa, no sentido de presentear o leitor com uma descrição autocontida das terminologias e conceitos empregados durante a experimentação de pesquisa, obtenção e discussão de resultados. O capítulo seguinte, por sua vez, aponta os trabalhos mais populares, divulgados na literatura, e relacionados aos objetivos de pesquisa, no intuito de fornecer, ao mesmo leitor, uma ideia da importância do domínio sob investigação.

Capítulo 3

Trabalhos Relacionados

Ao realizar a análise de literatura, divide-se o conhecimento adquirido em três subseções distintas: Subseção 3.1 apresenta discussão sobre trabalhos relacionados à descoberta de informações de saúde tanto em redes sociais quanto em serviços da WEB em geral dando ênfase ao uso de abordagens de classificação e/ou processo de filtragem de mensagens relevantes; Na subseção 3.2, discorre-se brevemente sobre trabalhos relacionados a classificação automática de mensagens curtas evidenciando o interesse da comunidade científica no assunto.

3.1 Descoberta e identificação de conhecimento relacionado à saúde

Duas estratégias principais foram amplamente exploradas no contexto da descoberta e identificação de conhecimentos de saúde em serviço da *WEB* através de pesquisas recentes: métodos baseados em palavras-chaves e métodos baseados em aprendizagem de máquina. Na primeira, as mensagens filtradas são ditas relevantes se contém uma ou mais palavras-chave pertencentes a um dicionário específico. Nesta ideia, Ginsberg et al. (2008) demonstraram a utilização de modelos de regressão para a predição de taxas de doenças semelhantes a

gripe[1] através da proporção de consultas relacionadas no Google em um mesmo período de tempo. A metodologia utilizada classifica registros de consulta através da presença de palavras-chave ligadas aos sintomas da doença e é implementada pelo Google Flu Trends para prover estimativas de taxas de prevalência da gripe e outras doenças em diversos países. Usando vários modelos de regressão, Culotta (2010) filtrou mensagens de usuários no Twitter via palavras chaves específicas, tais como $\{flu, cough, sorethroat, headache\}$, como meio de detecção de surtos de gripe H1N1. Segundo o autor, perfis do Twitter são passíveis de conterem metadados semiestruturados (cidade, estado, sexo, idade), possibilitando uma análise estatística mais detalhada. Em Corley et al. (2009), foi realizado um estudo similar desta vez usando as palavras-chaves *influenza* e *flu* para a detecção de posts em *WEB blogs* relacionados à gripe.

Em Achrekar et al. (2011), mensagens do Twitter contendo termos relacionados a sintomas de gripe foram monitoradas como indicadores para o rastreamento e previsão de uma possível epidemia na população. Durante o processo de análise dos dados, características semânticas mais complexas presentes nas mensagens foram totalmente ignoradas, sendo consideradas como relevantes todas as mensagens contendo termos relacionados com a gripe. Apesar disso, os resultados obtidos indicaram que o número de *tweets* relacionados à gripe está altamente correlacionado com dados oficiais divulgados por agências de proteção a doença.

Apesar de ser simples e não consumir demasiados recursos computacionais, a estratégia baseada em palavras-chaves, conforme Tuarob et al. (2014), tende a sofrer no tratamento de palavras polissêmicas e/ou palavras de domínio específico (domínio de saúde) raramente utilizadas. Em outras palavras, um modelo baseado nesta estratégia é altamente propenso a sofrer de *overfitting*. Neste sentido, Culotta (2010) foi um dos idealizadores do uso de técnicas de classificação automática de dados no cenário em questão, ao afirmar que o uso de um processo de classificação mais robusto, ao invés apenas de busca por palavras-chaves específicas no conteúdo das mensagens, é um caminho promissor na redução da vulnera-

[1]Neste contexto entende-se por doenças semelhantes a gripe, em inglês *influenza-like illness* ou *flu-like illness*, doenças que apresentam sintomas parecidos à gripe: *fever, cough, headache,*etc.

bilidade em relação ao processo de análise correlacional, comumente realizado após coleta de informações, e que pode sofrer com as flutuações de frequência em termos que tenham alguma relação com os eventos de saúde, tal como o *recall* de medicamentos[2].

Segundo Tuarob et al. (2014), o uso de abordagens baseadas em aprendizagem de máquina tende a resolver o problema de desambiguação de termos (comuns em métodos baseados em filtragem de palavras-chaves) e são capazes de aprenderem um nível de semântica mais elevado para determinadas palavras a partir do contexto no qual elas aparecem. A partir desta ideia, Collier, Son e Nguyen (2011) e Doan, Ohno-Machado e Collier (2012) estiveram interessados na classificação de *tweets* em classes relacionadas a vários tipos de doenças: respiratórias, gastrointestinais, hemorrágicas, dermatológicas, esqueléticas ou neurológicas. Após filtrarem as mensagens que contenham termos sintomáticos definidos pela ontologia de saúde BioCaster[3], os autores classificam as mensagens usando unigrams, juntamente com o modelo de ponderação binária) como características para algoritmos de aprendizagem.

Em um estudo relacionado, Collier, Son e Nguyen (2011), por sua vez, empregou aprendizado supervisionado utilizando unigramas, bigramas e expressões regulares como vetor de características para dois classificadores supervisionados (SVM e *Naïve Bayes*), no intuito de classificar *tweets* em quatro categorias de comportamento de prevenção auto relatados, além de um diagnóstico auto referido. Resultados indicaram um relativo nível de correlação ao comparar a saída do classificador e os dados laboratoriais da Agência Mundial de Saúde (do inglês, *World Health Organization*, WHO) para a H1N1 nos EUA durante a temporada de gripe de 2009-2010.

O trabalho de Doan, Ohno-Machado e Collier (2012) apresentou um método de filtragem de mensagens relacionadas com sintomas da gripe, usando 587 milhões de mensagens do Twitter. Os autores primeiramente filtram os dados com base em palavras-chave rela-

[2]Por exemplo, a palavra-chave "Tylenol" pode ser um termo válido para a descoberta de informação relativas à gripe, mas um possível *recall* do produto pode levar a um pico de frequência deste termo, sem corresponder a picos em taxas de gripe.

[3]um modelo de terminologia multilíngue de saúde pública projetada para a vigilância de eventos relacionados em meios de comunicação.

cionadas a sintomas a partir da ontologia BioCaster. Em seguida, as mensagens passaram por uma análise de traços semânticos e estruturais, como critério de validação dos dados, identificando sua relação em classes de interesse: negação, *hashtags, emoticons*, humor e geografia. Se uma relação gramatical direta ou indireta entre palavras indicando negação e a palavra "flu" for percebida, o *tweet* é descartado[4]. Outros critérios de descarte são: presença de *hashtags* não relacionadas a sintomas da gripe; *emoticons* contendo expressões felizes; características de humor (tais como risos) bem como características geográficas fora dos Estados Unidos. Segundo os autores, os resultados indicaram que simples melhorias baseadas em processamento de linguagem natural podem culminar em avanços das abordagens existentes para extração de dados do Twitter no domínio do monitoramento de surtos epidemiológicos.

Por sua vez, Paul e Dredze (2011a) e Paul e Dredze (2011b) investiram no uso de algoritmos de aprendizagem de máquina para a identificação de *tweets* relacionados a saúde. Como características utilizadas para o treinamento de um classificador SVM de kernel linear, os autores utilizaram *unigrams, bigrams e trigrams*. Em adição, os autores mineraram informações de saúde pública usando técnicas baseadas em LDA (*Latent Dirichlet Allocation*) (PAUL; DREDZE, 2011b).

Semelhantemente, Signorini, Segre e Polgreen (2011) usaram o Twitter para rastrear níveis de atividades de doenças e preocupações públicas nos EUA durante a pandemia de gripe H1N1. De modo a estimar taxas da doença, primeiramente coletaram mensagens contendo palavras-chave como *"swine", "flu", "influenza" ou "h1n1"* e então construíram um modelo de estimação usando SVM. Os resultados mostraram uma alta taxa relativa da doença, com um erro médio de $0,28\%$ para níveis nacionais durante as semanas alvo e $0,37\%$ para níveis regionais. Focando no rastreamento da preocupação pública, os pesquisadores adicionaram palavras chaves tais como: *"travel", "trip", "flight"* (para a transmissão de doenças) ou *"wash", "hand", "hygiene"* (para contramedidas) ou *"guillain", "infection"* (para efeitos de vacina). Calculando a porcentagem dos *tweets* observados, os resultados reforçaram

[4]Durante esse processo foi utilizado o analisador gramatical RASP descrito em Briscoe, Carroll e Watson (2006).

a ideia de que o Twitter pode ser usado como uma medida do interesse/preocupação pública sobre eventos relacionados à saúde.

Em Sadilek, Kautz e Silenzio (2012a), os pesquisadores estudaram o papel que os laços sociais e interações entre indivíduos específicos desempenham no progresso do contágio de doenças relacionadas com a gripe. Em um trabalho seguinte, Sadilek, Kautz e Silenzio (2012b) usaram um modelo baseado em campos aleatórios condicionais (do inglês, *Conditional Random Field* - CRF) para a predição do status de saúde individual do usuário, através de características derivadas de *tweets* e localizações de outras pessoas. Assim, o modelo gerado pelos autores é capaz de capturar o papel da localização no espalhamento de doenças infecciosas, o impacto da duração de co-localização na transmissão de doenças, bem como o atraso entre o evento de contágio e o início dos sintomas. Os autores afirmam que são capazes de prever os dias em que uma pessoa está doente com uma precisão de $0,94$ e um *recall* de $0,18$.

Por sua vez, a dissertação de Gomide (2012) apresenta uma metodologia para predição de eventos do mundo real a partir de dados minerados no Twitter tendo como estudo de caso a predição das taxas da dengue no Brasil. Durante o estudo, a autora utiliza o algoritmo Classificação Associativa Preguiçosa (em inglês, *Lazy Associative Classification*) (VELOSO; MEIRA; ZAKI, 2006) em uma abordagem de treinamento supervisionado com o objetivo de classificar as mensagens coletadas em cinco categorias diferentes: informação, experiência pessoal, opinião, piada ou ironia e campanha ou propaganda. Segundo a autora, as categorias definidas devem fornecer a informação necessária para eliminar as mensagens que, apesar de conterem pelo menos alguma palavra-chave sobre o evento, não estão relacionadas com a sua ocorrência. Dessa forma, segundo resultados obtidos e conforme expectativas intuitivas, os *tweets* classificados como "experiência pessoal" são mais representativos no contexto da predição de taxas da dengue, sendo que estes descrevem a vivência da própria pessoa que publicou a mensagem indicando possíveis casos reais.

Atacando o problema geral da identificação de mensagens relacionadas com a saúde em um conjunto heterogêneo de dados advindos da mídia social, Tuarob et al. (2014), abordam as limitações impostas por métodos de classificação tradicionais, quando estes representam documentos através da técnica BOW, alegando que tais métodos funcionam bem quando do-

cumentos são ricos em texto e adequados ao estilo padrão de escrita (tanto gramaticalmente, como sintaticamente), porém não são alternativas ótimas para o cenário das redes sociais onde esparsividade e ruídos são comuns. Os autores propõem a combinação de múltiplos classificadores baseados em técnicas de aprendizagem de máquina (*emsemble methods*) com os quais são treinados a aprender diferentes aspectos sobre o conteúdo das mensagens coletadas, tais como características semânticas, emocionais e físicas.

3.2 Classificação de mensagens curtas em mídias sociais

Mensagens em ambiente de redes sociais diferem de um texto convencional geralmente no fator tamanho e informalidade da língua. Neste trabalho, considera-se como mensagem curta qualquer mensagem cuja restrição de tamanho seja menor ou igual a 200 caracteres. Sendo assim, *tweets* fazem parte do domínio citado.

Conforme discutido na Seção 1.1, algoritmos de classificação empregados no domínio de textos convencionais, isto é, cuja a restrição de tamanho é muito superior a 200 caracteres e o estilo formal de escrita é recorrente, podem não apresentar um desempenho equivalente no domínio de mensagens curtas devido à baixa dimensão e ruídos característicos (expressões idiomáticas, abreviações, etc). Sriram et al. (2010) propuseram a característica "8F" para a classificação de *tweets*, visando a redução de limitações de BOW. Na abordagem dos autores, informações sobre o autor da mensagem, bem como interação entre usuários são consideradas. Caragea et al. (2011), por sua vez, propuseram um sistema chamado EMERSE para a agregação de *tweets* relacionados ao terremoto no Haiti. Para tanto, os autores implementaram o treinamento de um classificador SVM a partir da combinação de 4 conjuntos diferentes de características: *unigramas*, *unigramas* com seleção de características RELIEF[5], abstrações, e tópicos gerados a partir de LDA. Em uma abordagem semelhante, no entanto, complexa, Tuarob et al. (2014) focando na combinação de 5 outros conjuntos de características: *N-gramas*; características de sentimento; características de distribuição de tó-

[5]RELIEF é um algoritmo de seleção de recursos utilizados na classificação binária (generalizável a classificação polinomial por decomposição em uma série de problemas binários) proposto por Kira e Rendell (1992).

picos; características compostas baseadas em dicionário específico; e a combinação de todas as outras quatro características anteriores. Embora tenham obtido resultados interessantes usando conjunto (*ensemble methods*) de classificadores, nenhum estudo sobre extração de características discriminativas ou adaptação de modelo foi realizada.

Estudos em adaptação de modelo para classificação de mensagens curtas, foram recentemente, no entanto, investigadas nos trabalhos de Silva (2012) e Zimmermann, Ntoutsi e Spiliopoulou (2014). No primeiro, o autor faz uso de regras de associação em modelo de classificação preditiva, através do qual categoriza *tweets* em função de seu sentimento associado. Ainda neste trabalho, os autores estudaram hipóteses baseadas em mudança de conceitos (em inglês, *Concept Drift*), por meio do qual a tarefa principal é tornar o modelo robusto (não obsoleto) frente a dinâmica do espaço de mensagem. Esse trabalho está intimamente relacionado a pesquisa descrita neste documento, uma vez que se revisita à medida de confiança proposta pelos autores no Capítulo 4.

O segundo trabalho (ZIMMERMANN; NTOUTSI; SPILIOPOULOU, 2014) foca na mesma linha que Silva (2012), isto é, na classificação de sentimentos em *tweets*. No entanto, este propõe o uso do método *Naïve Bayes*, através do qual afere a confiança da predição realizada via diferença de entropia entre o conjunto de treinamento e a informação sendo avaliada. O trabalho do autor fornece elementos teóricos básicos para o capítulo 4.

No que diz respeito a ponderação de termos para a classificação de *tweets*, a pesquisa de Timonen (2013), fundamenta a abordagem descrita no Capítulo 5 uma vez que ela fornece modelos de equações para determinar a importância de termos (contidos nos *tweets*) em aspecto tanto inter como intra classes, sem considerar uma forte dependência na frequência de termos. A pesquisa do autor também está voltada à análise de sentimentos (em um ponto de vista mais comercial) das mensagens.

Apesar dos trabalhos de Silva (2012), Zimmermann, Ntoutsi e Spiliopoulou (2014), assim como Timonen (2013), estarem intimamente relacionados à pesquisa descrita neste documento, estes divergem quanto à tarefa de classificação, contexto e/ou esquema de classificação. Por exemplo, apesar de se reutilizar modelos matemáticos propostos em Timonen (2013), a abordagem proposta (Capítulo 5) diverge no ponto de que, além da modificação do uso combinado destes modelos, o esquema de classificação é fundamentado em *Naive*

Bayes.

3.3 Considerações Finais

Durante este capítulo foram descritos os trabalhos relacionados ao domínio de pesquisa em questão, evidenciando o interesse da comunidade científica em relação a temática discutida.

Capítulo 4

Um modelo de classificação expansivo

Neste capítulo será apresentado uma proposta de aprendizado de um modelo de classificação expansivo para a classificação de tweets relacionados com saúde.

4.1 Domínio de aplicação

A tarefa de classificação de tweets, no contexto deste trabalho, é definida como a seguir: Tem-se como entrada para um modelo de classificação Φ, um conjunto inicial de treinamento, $\mathcal{D}_{train} \subset \mathcal{D}$, o qual consiste em uma lista de exemplos na forma $\{(x^{(i)}, y^{(i)}); i = 1, 2, ..., |\mathcal{D}_{train}|\}$, onde $x^{(i)}$ é um *tweet* (representado como uma lista de termos, $< t_1, t_2, ..., t_n >$) e $y^{(i)}$ é a categoria atribuída a $x^{(i)}$. Mensagens em $\mathcal{D}_{train}$ possuem no máximo 140 caracteres, em torno de 20 palavras (TIMONEN, 2013), e as classes assumem seus valores a partir de um conjunto pré-definido e discreto $\mathcal{Y}$ de possibilidades $\{pos, neg\}$.

Caracterizando-se como um problema de classificação binária, conforme mencionado na Seção 1.1, tem-se que $x^{(i)}$ está relacionada com *eventos de saúde*[1] $(y =$ 'pos'$)$ se este segue pelo menos uma das seguintes condições (TUAROB et al., 2014):

[1] ou simplesmente relacionada à saúde.

1. a mensagem indica que o autor tem problema/sintomas de saúde;

2. a mensagem indica que uma outra pessoa está doente ou expressa preocupações com a saúde em geral.

As tabelas 4.1 e 4.2 fornecem exemplos de mensagens atribuídas às classes citadas. Nota-se a brevidade recorrente, bem como variações ortográficas que visam dar ênfase a certos termos. É possível perceber, ainda, que algumas mensagens, apesar de apresentarem termos discriminantes - *"hysteria", "sick"*, não possuem um envolvimento real com conceitos de saúde em face da definição adotada.

Tabela 4.1: Exemplos de *tweets* associados a $y =$ 'pos'.

ID	Mensagem
3697505306	*i think im getting sick, why i oughta!!! i never ever ever EVER get sick so when i do its for a looong time. Maybe its just a cold :)*
9241287301	*Taking my fever out for a walk.*
16622006002	*sick sick*
2034084803	*Headache and cramps..... no fun!*
11869412900	*Ian Dury died of cancer WAY before Malcolm Maclaren did.*

Semelhante ao apresentado na Seção 2.2, durante o treinamento de Φ, o conjunto $\mathcal{D}_{train}$ é utilizado para construir uma função relacionando padrões textuais nas mensagens às suas respectivas classes. Na fase de teste, uma sequência de mensagens futuras, $\mathcal{D}_{teste}$, consistindo em uma lista de exemplos $\{(x^{(\tau)}, ?); \tau = 1, 2, ..., |\mathcal{D}_{teste}|\}$ para o qual somente elementos $x^{(\tau)}$ são conhecidos, é fornecida à Φ, no intuito de que este possa estimar, a partir dos relacionamentos formados na fase de treinamento, as classes $\widehat{y}^{(\tau)} \forall x^{(\tau)}$, até então ausentes.

Apesar de modelos de classificação automática proverem uma alternativa para a mineração de informações em recursos não estuturados em grande escala, é possível perceber, através de uma análise mais específica, que os *tweets* possuem características próprias que dificultam o processo. Isto porque há, intrinsecamente, um alto nível de ruído associado e,

Tabela 4.2: Exemplos de *tweets* associados a $y =$ 'neg'.

ID	Mensagem
9377975912	*muse - hysteria*
23993897009	*I smell AIDS.*
10552475900	*I am so beyond sick of being infected with Malware. What ever happened to good, old fashion trojans? Spyware?*
7633914207	*Ok... I'm sick of this now... when is summer coming again?*
10126183108	*sick of thisss*

ainda, os usuários expressam um alto grau de liberdade de escrita, criando seus próprios estilos, abreviações, gírias e expressões idiomáticas (sem falar que as mensagens muitas vezes são curtas demais: 1 ou 2 palavras). Adicionalmente, em função da quantidade de usuários ativos, um elevado número de mensagens é produzido diariamente, fazendo com que o serviço forneça um fluxo contínuo de conteúdo, pelo qual o classificador deve operar.

Um importante aspecto observado no fluxo de informações disponível é o de um fator aqui definido como *mutabilidade* de termos ou tópico, isto é, a tendência de que assuntos discutidos no Twitter, bem como no ambiente de mídias sociais em geral, variem constantemente, quer seja sazonalmente ou esporadicamente - através do surgimento de um novo tópico de discussão. Por exemplo, no que diz respeito ao contexto de saúde, a pandemia de gripe A (H1N1) de 2009 foi bastante comentada naquele ano, no entanto, perdeu força no ano seguinte. Por outro lado, a gripe comum apresenta um volume de discussão em aspectos sazonais. Semelhantemente, o termo "Ebola"não foi tão enfatizado nas discussões do Twitter no ano de 2013 tanto quanto a partir da epidemia em 2014 até o presente.

Observando-se variações como estas, tem-se que o fluxo dinâmico de conteúdo disponível pode tornar a distribuição de termos no conjunto de teste muito diferente daquela presente no conjunto que fora utilizado para treinamento, isso, por sua vez, tende a diminuir o desempenho do classificador com o passar do tempo. Sendo assim, salienta-se a necessidade de implementação de um método que leve em consideração a manutenção/atualização da dis-

tribuição de termos entre as classes y de interesse, o que não foi prática comum entre os métodos propostos recentemente (cf. Capítulo 3).

Embora o foco adaptativo tenha sido discutido (obtendo resultados promissores) em ambiente semelhante nos trabalhos de Silva (2012) e Zimmermann, Ntoutsi e Spiliopoulou (2014), tem-se, aqui, uma direção única: **a classificação de *tweets* mediante a relação destes com saúde**, tendo, como elemento norteador, a possibilidade de a incorporação de novos termos discriminantes (atuais) oferecer ganhos superiores ou equivalentes no domínio citado e, assim, justificar a sua implementação. Essa direção vai de encontro à obtenção de resposta para primeira questão de pesquisa, levantada na Seção 1.2.1. A Seção a seguir, por sua vez, expõe detalhes teóricos do processo investigado.

4.2 A estratégia de expansão do modelo de classificação

Considerando a dinâmica do domínio estudado, é fácil perceber que um classificador Φ treinado a partir de $\mathcal{D}_{train}$ pode ter seu desempenho afetado quando a distribuição de dados no conjunto de testes, $\mathcal{D}_{teste}$, torna-se diferente da que fora treinado (SILVA, 2012). Esse problema tende a aumentar ao longo do tempo e, em algum ponto, Φ pode se tornar obsoleto (BIFET, 2010 apud SILVA, 2012). As subseções a seguir mostram o direcionamento teórico desenvolvido com a finalidade de atenuar esse problema, visando a tarefa de classificação antes mencionada.

4.2.1 O modelo básico

O ponto central da estratégia, assim como discutido em Silva (2012), consiste em coletar informações adequadas que visem expandir o conhecimento do modelo de classificação de modo que este se mantenha robusto, frente ao fluxo de mensagens do qual ele opera. Para fins de implementação da estratégia, percebeu-se a tendência em fundamentar a criação do modelo Φ, via hipóteses probabilísticas. Isto porque a distribuição de termos nas mensagens pode ser modelada através da relação probabilística, em um certo período de classificação, com as classes de interesse. Em outras palavras, é comum que haja mudança na probabili-

dade condicional dos termos em relação as classes consideradas.

Relacionando a estratégia proposta em cenário de classificação probabilística, tem-se que $P(y_j|\overrightarrow{x}^{(i)})$ denota a probabilidade de um exemplo (*tweet*) $\overrightarrow{x}^{(i)}$ pertencer a classe y_j. A partir disso, uma função de custo, que representa o custo de associar $\overrightarrow{x}^{(i)}$ à classe incorreta, é minimizada se, e somente si, $\overrightarrow{x}^{(i)}$ é associado à uma classe $y_k; k = 1, ...|\mathcal{Y}|$, para qual $P(y_k|\overrightarrow{x}^{(i)})$ é máxima (DUDA; HART; STORK, 2000 apud FACELI et al., 2011). Esse método é designado na literatura como MAP (do inglês, *Maximum A Posteriori*) e é formalmente descrito como na expressão a seguir:

$$\widehat{y}_{MAP} = \arg\max_{j} P(y_j|\overrightarrow{x}^{(i)}) \tag{4.1}$$

Naturalmente, o teorema de Bayes (4.2), como uma função discriminante que calcula a probabilidade condicional (*a posteriori*) associada à classe a partir de uma nova mensagem, pode ser usado no cálculo de $P(y_j|\overrightarrow{x}^{(i)})$, como:

$$P(y_j|\overrightarrow{x}^{(i)}) = \frac{P(y_j)P(\overrightarrow{x}^{(i)}|y_j)}{P(\overrightarrow{x}^{(i)})} \tag{4.2}$$

Assumindo que um conjunto de mensagens $\mathcal{D}$ pode ser representado como um conjunto de termos $\mathcal{T}$, através do qual cada $t \in \mathcal{T}$ possui uma relação quantificável com cada classe de interesse, julgou-se conveniente fundamentar a implementação através do modelo de classificação *Naïve Bayes*. Este, assume que os valores dos atributos do vetor de representação $\overrightarrow{x}^{(i)}$ são independentes entre si dado a classe e, por essa hipótese, decompõe $P(\overrightarrow{x}^{(i)}|y_j)$ no produto $P(\overrightarrow{x}^{(i)}_{(1)}|y_j) \times ... \times P(\overrightarrow{x}^{(i)}_{(d)}|y_j)$, em que $\overrightarrow{x}^{(i)}_{k}$ é o *k*-ésimo atributo do exemplo $\overrightarrow{x}^{(i)}$ (FACELI et al., 2011). *Naïve Bayes* usa a Equação 4.2 e a regra de decisão 4.1 como função discriminante, porém, visto que $P(\overrightarrow{x}^{(i)})$ é o mesmo para todas as classes, sua remoção da Equação 4.2 não afeta os valores relativos de suas probabilidades, e, com isso, a chance de $\overrightarrow{x}^{(i)}$ pertencer a uma classe y_j é frequentemente representada pela seguinte proporcionalidade:

$$P(y_j|\overrightarrow{x}^{(i)}) \propto P(y_j) \prod_{k=1}^{d} P(\overrightarrow{x}^{(i)}_{k}|y_j) \tag{4.3}$$

O fato de considerar os valores de atributos independentes é uma característica importante ao qual tomou-se em consideração para a estratégia de adaptação do modelo, visto que cada termo t pode estar associada a um certo grau de importância em relação às classes de interesse. Por exemplo, o unigrama *"lung"* assim como *"ache"* e *"flu"* têm um peso maior em relação à classe *'pos'* do que quando comparada à classe *'neg'*.

A consideração conjunta de bigramas como valores de atributos também foi analisada como uma alternativa para o aumento de informações e descoberta de termos discriminantes. Nesta, a ideia é permitir que, ainda sobre a hipótese de independência de termos, possa-se elucidar termos característicos no contexto de saúde. Por exemplo, o unigrama *"swine"* tem diferente importância para a classes $y =$ 'neg' quando comparada com $y =$ 'pos', no entanto, pode ter seu impacto discriminativo elevado quando comparado na forma do bigrama *"swine flu"*, já que se percebe uma relação adjetiva deste, com o substantivo *"flu"*.

A partir das direções descritas, a estratégia, descrita no pseudocódigo 1, considera: (1) um conjunto de mensagens $\mathcal{D}$, onde cada *tweet* $\overrightarrow{x}^{(i)} \in \mathcal{D}$ assume uma representação de vetor; (2) um conjunto inicial de treinamento $\mathcal{D}_{train} \subset \mathcal{D}$, onde, para cada $\overrightarrow{x}^{(i)} \in \mathcal{D}_{train}$, uma classe $y^{(i)} \in \mathcal{Y}|\{\text{'pos','neg'}\} \subseteq \mathcal{Y}$, é conhecida; (3) um conjunto de mensagens de teste $\mathcal{D}_{teste} \in \mathcal{D}$; e (4) um parâmetro δ para a regulação do nível de confiança da predição $\widehat{y}$ realizada. Outros parâmetro presentes no algoritmo são descritos na tabela seguir:

Tabela 4.3: Descrição de variáveis para a o algoritmo 1

Parâmetro	**Definição**
Φ	Classificador (*Naïve Bayes*)
$\widehat{\mathcal{Y}}$	Lista de predições realizadas
$\Delta_{\text{'pos'}}$	Histórico de predições $\widehat{y} =$ 'pos'
$\Delta_{\text{'neg'}}$	Histórico de predições $\widehat{y} =$ 'neg'
$\overrightarrow{y}$	Vetor de valores de predição para cada classe. $\overrightarrow{y} \in \mathbb{R}^2$ onde cada elemento $\overrightarrow{y}_i \in \overrightarrow{y}$ está definido em $0 \leq \widehat{y} \leq 1$
$\widehat{y}$	Predição de classe para um dado $\overrightarrow{x}^{(\tau)}$

Algoritmo 1: Estratégia de expansão do conjunto de treinamento

Entrada: $\mathcal{D}_{train}, \mathcal{D}_{teste}, \delta$

Saída: $\{\widehat{y}^{(\tau)}; \tau = 1, ..., |\mathcal{D}_{teste}|\}$

1 **início**
2 $\quad \Phi \leftarrow$ TREINAR O CLASSIFICADOR USANDO $\mathcal{D}_{train}$;
3 $\quad \widehat{\mathcal{Y}} \leftarrow [\,]$;
4 $\quad \Delta_{\text{'pos'}} \leftarrow [\,]$;
5 $\quad \Delta_{\text{'neg'}} \leftarrow [\,]$;
6 $\quad$ **para** $\overrightarrow{x}^{(\tau)} \in \mathcal{D}_{teste}; \tau = 1, ..., |\mathcal{D}_{teste}|$ **faça**
7 $\quad\quad \overrightarrow{y} \leftarrow$ CALCULAR ESTIMATIVAS DE CLASSE USANDO Φ EM $\overrightarrow{x}^{(\tau)}$;
8 $\quad\quad \widehat{y} \leftarrow retornaClasse(max(\overrightarrow{y}))$;
9 $\quad\quad \Delta_{\text{'pos'}} \leftarrow adiciona(\Delta_{\text{'pos'}}, \overrightarrow{y}_{\text{'pos'}})$;
10 $\quad\quad \Delta_{\text{'neg'}} \leftarrow adiciona(\Delta_{\text{'pos'}}, \overrightarrow{y}_{\text{'neg'}})$;
11 $\quad\quad$ **se** $\overrightarrow{y}_{\widehat{y}} \geq$ CONF$(\Delta_{\widehat{y}}, \delta)$ **então**
12 $\quad\quad\quad \mathcal{D}_{train} \leftarrow \mathcal{D}_{train} \cup \{(\overrightarrow{x}^{(\tau)}, \widehat{y})\}$;
13 $\quad\quad\quad \widehat{\mathcal{Y}} \leftarrow adiciona(\widehat{\mathcal{Y}}, \widehat{y})$;
14 $\quad\quad\quad \Phi \leftarrow$ RETREINAR Φ USANDO $\mathcal{D}_{train}$;
15 $\quad\quad$ **senão**
16 $\quad\quad\quad \widehat{\mathcal{Y}} \leftarrow adiciona(\widehat{\mathcal{Y}}, \widehat{y})$;
17 $\quad\quad$ **fim**
18 $\quad$ **fim**
19 $\quad$ **retorna** $\widehat{\mathcal{Y}}$
20 **fim**

O conjunto de treinamento é utilizado inicialmente para treinar o classificador Φ (linha 2). Apesar de Φ ser fundamentado no modelo *Naïve Bayes*, percebeu-se a necessidade de normalizar as estimativas de predição $(0 \leq \widehat{y} \leq 1)$ no intuito de facilitar o reuso da métrica de confiança proposta em Silva (2012)(cf. Subseção 4.2.2). Para cada mensagem, a qual deriva-se uma classe $\widehat{y}$ (Algoritmo 1, linha 8), realiza-se o teste de confiança de predição (Algoritmo 1, linha 11) usando como base o histórico de predições para cada classe realizadas até o momento (linha 4 e 5). Se a mensagem é julgada como útil, atualiza-se o conjunto de treinamento pela inserção desta, juntamente com a classes predita. Em seguida,

retreina-se Φ a partir do novo $\mathcal{D}_{train}$ (linha 14).

Nota-se que o método de retreino de Φ pode ser implementado eficientemente, dado que, como os parâmetros do modelo são atualizados com base em $\overrightarrow{x}^{(\tau)}$ e sua classe predita $\widehat{y}^{(\tau)}$, é necessário apenas atualizar as contagens de termos (ou outra métrica associada) $N_{i\widehat{y}}$, para todos os termos $t_i \in \overrightarrow{x}^{(\tau)}$ e a classe $\widehat{y}^{(\tau)}$ (ZIMMERMANN; NTOUTSI; SPILIOPOULOU, 2014). A subseção a seguir detalha a computação da confiança de predição.

4.2.2 Estimativa de confiança de predição e a inclusão de novos dados

A estimativa de confiança, no cenário investigado, tem o propósito de elucidar se a previsão mais recente pode ser utilizada para fins de expansão de $\mathcal{D}_{train}$. Para esta finalidade, observou-se utilizar a função de confiança proposta por Silva (2012), ao qual leva em consideração o histórico de predições realizadas em $\mathcal{D}_{teste}$ e é descrita na Equação 4.4.

$$\textsc{Conf}(\Delta_{\widehat{y}}, \delta) = \delta \times \frac{0,5 + \sum_{u=1}^{|\Delta_{\widehat{y}}|-1} \Delta_{\widehat{y}}^{(u)}}{|\Delta_{\widehat{y}}|} \tag{4.4}$$

Utilizando-se da Equação 4.4 no algoritmo 1, tem-se que uma mensagem arbitrária $\overrightarrow{x}^{(\tau)} \in \mathcal{D}_{teste}$ é confiavelmente predita como sendo da classe $\widehat{y}$ se $\overrightarrow{y}_{\widehat{y}} \geq \textsc{conf}(\Delta_{\widehat{y}}, \delta)$. Segundo os autores, a constante $0,5$ é utilizada para que a primeira mensagem inserida no conjunto de treinamento tenha a confiança maior que a soma das predições realizada anteriormente. δ, por sua vez, é um fator especificado $k \geq \delta \geq 1,0$, onde k é o número de classes. δ ainda pode ser visto como um fator de quanto a estimativa deve desviar da média para ser considerada como importante. Esse fator visa também evitar o alto viés para uma determinada classes, dado um possível aumento do volume de mensagens atribuídas à uma única classe (SILVA, 2012).

Ainda segundo Silva (2012), o esperado é que a confiança seja próxima da média e, uma vez que esta alcança uma valor acima desta, existem indícios que a predição está correta e que a mensagem pode ser usada para expandir o conjunto de treinamento inicial. Intuitivamente, se realmente confiável, os dados de treinamento serão expandidos, aumentando a diversidade dos dados como resposta às mensagens mais atuais.

4.3 Considerações finais

O presente capítulo ofereceu um detalhamento teórico sobre a ideia de um modelo de classificação, fundamento em regra de decisão probabilística e mais precisamente no modelo *Naïve Bayes*, ao qual trabalha sobre o regime de expansão do conjunto de treinamento a partir de predições consideradas confiáveis. Apesar da abordagem trabalhar utilizando o limiar adaptativo, como métrica de confiança, proposto inicialmente por Silva (2012), este trabalho, como já citado anteriormente, diferencia dos autores pelo fato de o modelo de classificação ser diferente e, além disso, a tarefa de classificação como um todo também possui foco diferente.

Acredita-se que a abordagem proposta neste capítulo se apresenta como uma alternativa, ainda não explorada em pesquisas anteriores, no que diz respeito à mineração de mensagens sociais curtas (como é o caso de *tweets*) para o monitoramento e prevenção de eventos de saúde, assim como pesquisas fundamentadas em grandes volumes de dados para fins relacionados. Desse modo, tem-se que a presente pesquisa poderá servir de base para o desenvolvimento de modelos adaptativos de alta performace, mais robustos e que considerem conteúdos externos (*thesaurus*, dicionários de saúde) para o refinamento do processo.

Ainda como citado no texto, este capítulo apresentou uma tentativa de resposta à primeira questão (cf. Capítulo 1) levando em consideração à hipótese de que, através de uma estimativa de confiança na predição realizada, é possível ampliar o conhecimento do classificador, e, consequentemente, elevar o desempenho de classificação para o caso de mensagens sociais curtas no domínio citado. Os resultados numéricos obtidos são descritos no Capítulo 5, juntamente com toda o detalhamento de configuração experimental, isto é, ferramentas e recursos.

Capítulo 5

Descoberta de termos importantes: um foco na classificação de documentos curtos relacionados à saúde

Neste capítulo discorre-se sobre o segundo objetivo deste documento: trata-se da investigação de uma alternativa para o levantamento de termos importantes que auxiliem a classificação de documentos curtos sob a perspectiva da relação destes com eventos de saúde.

5.1 Discussão preliminar

Conforme discutido anteriormente, no campo de documentos curtos, *tweets* tendem a ter uma baixíssima frequência de palavras[1] em cada mensagem - em torno de 20 termos. Dado que o escopo desse documento de pesquisa está relacionado à classificação automática dessas mensagens em sua relação com saúde, é mandatório buscar alternativas que consigam descriminar termos entre as classes e, dessa forma, elevar a performance de predição. Isso porque, como já exposto, devido a brevidade das mensagens, é comum que palavras-chave

[1] Aqui, como em outras partes do texto, tem-se que "termos"também é usado para se referir a uma "palavra". Isto é, "termo"e "palavra"são equivalentes.

importantes sejam tratadas como palavras normais em função da baixa representatividade (frequência) no conjunto de treinamento. Estas palavras-chaves, por exemplo, podem estar ligadas diretamente à um contexto de saúde, mas, em razão desse problema, pode passar despercebida e/ou ocasionar a geração de falsos negativos. Sobre isso, Timonen (2013), ao focar na análise de sentimentos em *tweets*, afirma que por causa da brevidade das mensagens, algumas abordagens existentes, baseadas principalmente em TF-IDF ou outras métricas de TF, não funcionam bem.

A afirmação do autor, neste campo, baseia-se no fato de que quando o documento contém apenas alguns poucas palavras, raramente se tem algumas destas acontecendo mais de uma vez. Essa característica também vai de encontro com a observação realizada no domínio de mensagens relacionadas à saúde, onde algumas palavras discriminantes (como nomes técnicos de doenças, por exemplo) têm uma frequência muito baixa e/ou aparecem apenas uma vez em todo o conjunto de treinamento.

Como muitas abordagens tradicionais baseiam-se em *TF*, torna-se necessário investir em alternativas para ponderar os termos de uma forma mais eficiente e que contribua eficazmente para o desempenho de classificação, atenuando a limitação citada (Questão de pesquisa 2, Capítulo 1). Nesse intuito, baseando-se no trabalho inicial de Timonen (2013), buscou-se aferir o nível de informação de cada palavra através da estimação de sua relevância em nível tanto intra como inter classe. Revisita-se a ideia de comparação de distribuição desenvolvida no trabalho do autor, adaptando-as ao escopo desse documento de pesquisa.

5.2 Detalhamento da abordagem

Em Timonen (2013), conforme citado no Capítulo 3, o autor propõe o uso de uma métrica de ponderação de termos baseada na combinação de quatro medidas distintas ao qual considera a distribuição de palavras em sua relação em nível amplo (em todas as classes) e em nível local (apenas em uma classe). Destas medidas básicas, três são apresentadas a seguir sendo que uma delas, *"Média inversa de comprimento de fragmento"*, foi descartada visto que a mesma diverge ao escopo de pesquisa.

Contagem inversa de categoria: na busca por aferir os termos presentes nas mensagens em nível de conjunto, essa medida foi desenvolvida com a ideia de enfatizar palavras que ocorrem em poucas categorias, isto é, quanto menor for o número de categorias que fazem uso do termo, mais informativo ele é. Formalmente, tem-se que, para um termo t qualquer:

$$icc(t) = \frac{1}{\upsilon_t} \tag{5.1}$$

onde υ_t é o número de categorias onde t está presente. Apesar de ser uma medida independente da tarefa de classificação, Timonen (2013) ressalta que esta possui um poder discriminativo regular em tarefa de classificação binária. No entanto, intuitivamente, em combinação com outras, esta métrica pode ajudar a elucidar termos específicos de saúde.

Probabilidade de Categoria: trata-se da probabilidade de se encontrar uma palavra dentre uma classe. Aqui a ideia base é a de que uma palavra que está presente muitas vezes em uma classe específica e raramente em outras são as mais importes. Essa medida usa a distribuição de palavras entre as classes e determina que, se uma palavra ocorre apenas em uma delas, a probabilidade assume valor 1, sendo que a probabilidade das classes restantes assume 0. Essa medida é calculada levando em consideração o número de documentos d em uma coleção de documentos $\mathcal{D}$ associados à uma classe y e que contêm o termo t, dividido pelo número total de documentos em $\mathcal{D}$ contendo t:

$$P(y,t) = \frac{|d \in \mathcal{D} : t \in d, d \in y|}{|d \in \mathcal{D} : t \in d} \tag{5.2}$$

Probabilidade de documento: diz respeito a chance de um documento d na classe y conter a palavra t. A intuição derivada é a de que um termo é importante se este ocorre frequentemente em uma classe e menos importante se ocorre muito raramente. Formalmente está definida como:

$$P(t,y) = \frac{|d \in \mathcal{D} : t \in d, d \in y|}{|d \in \mathcal{D} : d \in y|} \tag{5.3}$$

Nota-se que o uso isolado dessa estatística pode enfatizar termos que ocorrem muito frequentemente e não possuem poder discriminativo, como o caso de verbos ("*is*", "*go*"), preposições ("*trough*", "*by*") e artigos ("*the*","*a*") - *stopwords*, porém, o uso

combinado com a medida de *probabilidade de categoria* enfatiza palavras que ocorrem raramente entre as classes, diminuindo a influência daquelas.

A partir das medidas apresentadas, tem-se a combinação destas de modo a formar uma medida compacta de ponderação dos termos (w), tal como na versão original do autor. Semelhantemente, o peso é calculado para cada par $< y, t >$, onde y é uma classe pertencente a um conjunto de classes Y e t é uma termo contido em $\mathcal{V}$, o conjunto de termos derivado do conjunto de treinamento. Partindo disso, se t aparece em duas classes diferentes, seu peso possivelmente será diferente para cada classe. Pela adaptação da função descrita em Timonen (2013), tem-se que:

$$w(t, y) = icc(t) \times (P(y, t) + P(t, y)) \tag{5.4}$$

Na Equação 5.4, as duas últimas medidas descritas acima são combinadas via operação de adição pelo fato de que torna-se mais conveniente oferecer igual ênfase entre elas, isto porque valores pequenos poderiam ter um grande impacto através da operação de multiplicação (TIMONEN, 2013).

Diferente do autor, esta pesquisa não considera nenhum efeito de normalização para w, dado que os pesos obtidos serão avaliados usando o método de classificação *Naïve Bayes*, no Capítulo 6.

5.3 Considerações Finais

Este capítulo apresentou como se pretende revisitar a ideia de ponderação de termos descrita em Timonen (2013) para o caso da descoberta de termos discriminantes visando à classificação de *tweets* em sua relação com saúde. Apesar da ligeira adaptação de cenário, acredita-se que esta abordagem também poderá ser usada para o caso da classificação de documentos em contrapartida à tarefa de análise de sentimentos investigada pelos autores. Resultados experimentais são discutidos no capítulo a seguir.

Capítulo 6

Avaliação Experimental

Este capítulo apresenta detalhes sobre a análise experimental realizada para fins de obtenção de resultados numéricos concernentes as abordagens apontadas nos Capítulos 4 e 5, respectivamente.

6.1 Configuração experimental

A fim de possibilitar respostas às questões de pesquisa, optou-se pela utilização de uma configuração experimentação que possibilite alguma comparação com resultados divulgados em trabalhos relacionados. A presente seção define as decisões de projeto acerca da origem dos dados utilizados durante o estudo, as bibliotecas de programação, assim como os métodos de avaliação do modelo de classificação.

6.1.1 Conjunto de Dados

Conforme mencionado frequentemente, os dados de interesse são *tweets* publicamente disponíveis e coletados através de chamadas à API do Twitter. Observa-se que o limite de caracteres disponíveis (140 caracteres por mensagem) induz o usuário a ser coerente e direto ao expressar sua opinião, mas também induz ao uso livre da escrita padrão, elevando o número de palavras no vocabulário geral e tornando tais mensagens altamente ruidosas.

No entanto, a brevidade induzida facilita o processamento e classificação dos dados no fator tempo de execução, visto que apenas um conjunto tratável τ de termos é possível dentro do limite de caracteres. Sendo assim, um tratamento de ordem polinomial para cada mensagem $t|\tau \subseteq t$ é passível de ser realizado.

Para a realização de experimentação durante a pesquisa, dois conjuntos de dados foram utilizados, onde, em cada um deles, tem-se o inglês como idioma. Estes conjuntos de dados estiveram sendo usados em pesquisas anteriores (TUAROB et al., 2014) (PAUL; DREDZE, 2011a) e são descritos a seguir:

- **DatasetA:** consiste de um conjunto de exemplos contendo 5138 *tweets* manualmente rotulados. Este conjunto foi utilizado em Tuarob et al. (2014) para experimentos de classificação envolvendo o protocolo de validação cruzada e para o treinamento de múltiplos classificadores no cenário de classificação coletiva (*ensemble methods*). O acesso aos dados foi obtido mediante solicitação enviada por *e-mail* aos autores. Cada *tweet* pertencente ao conjunto é uma tupla contendo o ID do *tweet* (fornecido pela API) e o conteúdo da mensagem. Cada instância desse conjunto é rotulada como positiva se está relacionada à saúde, e negativa, caso contrário. Os conceitos de saúde existentes nesse conjunto são diversos. Em termos de suporte[1], o conjunto contém 1832 $(35,73\%)$ instâncias positivas e 3296 $(64,27\%)$ instâncias negativas.

- **DatasetB:** conjunto de *tweets* rotulados e que foram usados no treinamento de classificadores de texto relacionado com a gripe no trabalho de Lamb, Paul e Dredze (2013). Apesar de relacionado diretamente a um tópico de saúde específico (*flu*), esse conjunto de dados apresenta uma diversidade léxica abrangente e, além disso, não foge das características principais consideradas para que um documento seja dito relacionado à saúde (cf. Seção 4.1). O acesso aos dados foi obtido via download disponível no site pessoal de um dos autores. Vale ressaltar que o download disponibiliza um pacote de diferentes conjuntos de dados para a exploração de modelos supervionados relacionados à classificação de mensagens em sua relação com à gripe. Para a definição

[1] Proporção de instância de classes.

do *DatasetB*, porém, observando a tarefa de classificação desta pesquisa, seguiu-se apenas com o conjunto intitulado *RelatedVsNotRelated.txt*. Em se tratando de suporte, este conjunto de dados contém originalmente 2764 ($\approx 57\%$) instâncias positivas e 2086 ($\approx 43\%$) instâncias negativas, porém, devido a existência de mensagens em idioma diferente do inglês, após processo de filtragem, o suporte assume a seguinte proporção: 2784 ($\approx 61,64\%$) instâncias positivas e 1720 ($\approx 38,35\%$) instâncias negativas.

É necessário salientar que, conforme argumentado pelos desenvolvedores, o processo de coleta de mensagens para a geração dos conjuntos leva em consideração a existência de uma lista de palavras-chave relacionadas à saúde, a qual é utilizada para fins de indexação e, desta forma, oferece um processo de meta-seleção de dados frente à diversidade de mensagens disponível na plataforma do serviço. Sobre isso, (SADILEK; KAUTZ; SILENZIO, 2012a) afirmam que, para cada mensagens realmente relacionada à um conceito de saúde, existem milhares de outras não relacionadas e que podem, ao contrário, estar ligadas à qualquer outro tipo de assunto[2].

Sendo que cada mensagem, pertencente tanto à classe negativa quanto positiva, possui pelo menos um termo relacionado a saúde, tem-se que o processo de classificação tem de ser, de certa forma, robusto o suficiente para lidar com a ocorrência de termos em ambas as classes, ressaltando o ganho de informação com o passar do tempo e a melhoria/exploração de técnicas de ponderação de termos. Para ilustrar essa situação, é fácil perceber que a mensagem *Im sick! #Flu #Disease* está claramente associada a classe positiva[3], enquanto que a mensagem *Im sick of this country! #violence #death*, apesar de possuir um termo discriminante (*sick*), não tem uma relação real com eventos de saúde em seu sentido mais restrito.

[2]Isso, segundo o autor, suporta o uso de uma lista de termos de saúde (da qual depende inteiramente da finalidade de classificação) para coleta de dados.

[3]Isto é, no contexto deste trabalho, está relacionada à saúde

6.1.2 Ferramentas e Tecnologia

Para a implementação do modelo de classificação, bem como para o tratamento dos conjuntos de dados disponíveis, utilizou-se bibliotecas de aprendizagem de máquina e processamento de linguagem natural disponíveis em linguagem Python 2.5.7. Especificamente, dispôs-se do módulo *Scikit-learn* (0.15.1)[4], ao qual oferece várias implementações de algoritmos voltados à classificação de dados bem como, ferramentas úteis de pré-processamento e extração de características. O módulo NLTK[5] (do inglês, *Python's Natural Language Toolkit*) também foi utilizado para operações de pré-processamento, tais como: tokenização[6] e stemização (do inglês, *stemming*)[7].

6.2 Experimento 1: abordagem de expansão de conjunto de treinamento

Esta seção descreve a análise experimental realizada no intuito de obter evidências numéricas que suportem o desenvolvimento e uso da abordagem proposta no Capítulo 4. Inicialmente, para o preprocessamento das mensagens do conjunto de dados, optou-se por remover caracteres de pontuação e substituir cada letra maiúscula pela correspondente minúscula (por exemplo, *cOld* é tratado da mesma forma que *cold*). Além disso, aplicou-se as seguintes etapas adicionais:

- **Remoção de HTML:** Todas as *tags* HTML são removidas das mensagens. Algumas mensagens possuem códigos de formatação HTML que podem elevar, desnecessariamente, a quantidade de termos no vocabulário utilizado para classificação.

[4]http://scikit-learn.org/stable/

[5]http://nltk.org/

[6]Processo de quebrar um fluxo de caracteres ou uma sequência de palavras (sentença) em unidades menores. Neste documento, trata-se da ação de reduzir uma sentença à uma lista de palavras.

[7]Trata-se de um processo de reduzir palavras flexionadas (ou às vezes derivadas) eu sua raiz (*stem)*. Por exemplo: *car, cars, car's, cars'* $\Rightarrow$ *car* (MANNING; RAGHAVAN; SCHÜTZE, 2008).

- **Normalização de URLs:** Todos os *URLs* encontrados são substituídos pelo texto url.
- **Normalização de números:** Todos os números encontrados nas mensagens são substituídos pelo termo number. Dessa forma, *3 weeks later* e *2 pills of medicine* são substituídas por *number weeks later* e *number pills of medicine*, respectivamente.
- **Stemização:** Cada palavra nas mensagens são reduzidas à sua forma básica. Por exemplo, "discount", "discounts", "discounted" e "discounting" são consideradas apenas como "discount". As vezes, o processo de stemização retira caracteres adicionais no final da palavra, sendo que "include", "includes", "included", e "including" se tornam includ.
- **Remoção de espaçamento:** além da remoção de pontuação, todos os espaços em branco (tabulações, novas linhas, espaços) foram reduzidos a um único caractere de espaço.

As etapas de preprocessamento adotadas visam diminuir a complexidade do vocabulário de termos considerados para treinamento e estabelecer um meio de extrair o potencial completo do conjunto de dados, isto é, reduzir o ruído associado nas mensagens. Por exemplo, nomes específicos de usuários não possuem relação com características de saúde e, neste caso, dado que a ocorrência discriminativa de um termo representando um nome de usuário é muito baixa, faz sentido normalizar tal entidade e preservar uma distribuição mais uniforme da mesma entre as instâncias do conjunto.

Partindo disso, após o preprocessamento, buscou-se a replição do modelo de classificação baseado em características de *N-gramas* citado em Tuarob et al. (2014), ao qual objetivou estabelecer resultado de base (*baseline*) para que se possa desenvolver alguma comparação empírica com o modelo proposto.

A replicação (baseada em modelo estático) foi inicialmente realizada usando o DatasetA. Esta, tendo como referência a Tabela 6.1, possuiu as seguintes características: <Φ = SVM; Ω = ValidaçãoCruzada(10 Partições); clean = True; stem = True; N = 2; W = TF-IDF>. A Tabela 6.2 apresenta os resultados obtidos.

Tabela 6.1: Tabela de variáveis para avaliação experimental.

Parâmetro	**Descrição**
Φ	algoritmo de classificação
Ω	método de avaliação
clean	remoção de pontuação e substituição de letras maiúsculas por suas respectivas minúsculas
stem	aplicação da algoritmo de stemização de Porter (PORTER, 1997)
N	numéro máximo de termos consecutivos para formar o conceito de *N-grama*
W	esquema de ponderação de termos
δ	fator de desvio de média para cálculo da confiança de predição

Tabela 6.2: Medidas de Precisão (Pr%), Revocação (Rev%) e F1 (F1%) obtidos durante a replicação do modelo de classificação baseado em *N-gramas*, segundo Tuarob et al. (2014).

Pr%	**Rev%**	**F1%**
$67,00$	$52,00$	$58,00$

Porém, tendo que, conforme Bifet e Frank (2010), o método de avaliação *hold-out* é mais indicado para abordagens adaptativas, reavaliou-se a replicação citada, considerando Ω = *hold-out*. A justificativa para essa decisão está ligada ao fato de Tuarob et al. (2014) retratar a configuração citada como a que obteve melhor resultado considerando características de *N-gramas*. Desse modo, realizar a replicação usando *hold-out* torna o resultado passível de comparação. No contexto do número de instâncias observadas para as tarefas de treinamento e teste, a avaliação por *hold-out* foi realizada de modo que a proporção de treinamento assuma $\frac{2}{3}$ do conjunto de dados utilizados e que $\frac{1}{3}$ deste seja direcionada para teste do modelo. Assim, devido as dimensões do conjunto DatasetA, tem-se um suporte de teste com 862 instâncias negativas e 525 positivas.

Neste cenário experimental, a medida *F1* foi pensada para ser o principal meio de comparação direta entre os modelos uma vez que a precisão e a revocação são tratadas com igual importância. É possível observar que a diferença de *F1* em relação as tabelas 6.2 e 6.3, pode ser explicada pela diferença na proporção de dados utilizados para avaliação em cada método. Sob esta configuração, o modelo é definido como Φ = *estático-TUAROB* visando facilitar referência futura. Seguindo o exposto, a Tabela 6.3 aponta os resultados para este caso, assim como a figura 6.1 retrata a matriz de confusão associada.

Tabela 6.3: Medidas de Precisão (Pr%), Revocação (Rev%) e F1 (F1%) do modelo Φ = *estático-TUAROB* , usando Ω = *hold-out*.

Pr%	**Rev%**	**F1%**
$65,00$	$49,00$	$56,00$

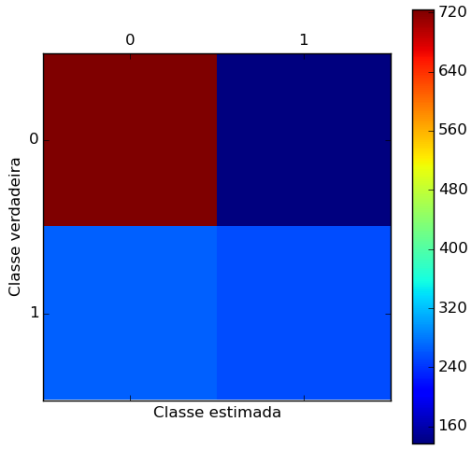


Figura 6.1: Matriz de confusão do resultado de replicação de Φ = *estático-TUAROB* , usando Ω = *hold-out*.

Por sua vez, a experimentação com a abordagem de expansão de conjunto de treinamento, discutida no (Capítulo 4), procurou confrontar o resultado apontado na tabela 6.3, através da seguinte configuração: <Φ = expansivo; Ω = *hold-out*; clean = True; stem = False; N = 1; W = *freq*; $\delta = 2$>. A Tabela 6.4 apresenta os resultados obtidos para a configuração citada em comparação com àqueles apresentados na Tabela 6.3.

A abordagem de expansão, conforme configurações citadas no parágrafo anterior, obtém um ganho de 9% em F1 se comparado ao modelo apontado em Tuarob et al. (2014). Durante o processo experimentação, observou-se que considerar o modelo com $stem = True$ gerou

Tabela 6.4: Comparação entre os resultados de classificação baseados em abordagem estática (*estático-TUAROB*) e expansiva. Pr%, Rev% e F1% denotam a porcentagem de precisão, de revocação e de medida de F, respectivamente. $\Delta F1\%$, por sua vez, se refere a diferença (em porcentagem) em ganho de performance em F1.

Φ	**Nº estimativas confiáveis**	**Pr%**	**Rev%**	**F1%**	$\Delta F1\%$
estático-TUAROB	$--$	$65,00$	$49,00$	$56,00$	$0,00$
expansivo	378	$66,00$	$65,00$	$65,00$	$9,00 \uparrow$

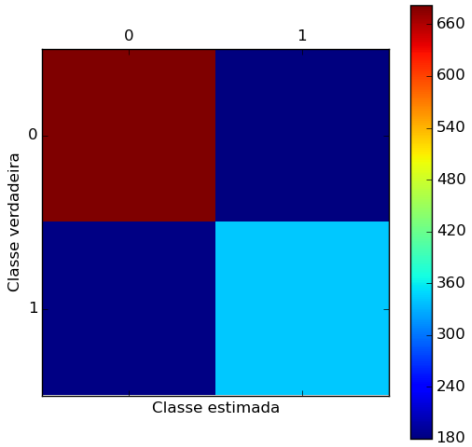


Figura 6.2: Matriz de confusão do modelo de classificação <Φ = expansivo; Ω = *hold-out*; clean = True; stem = False; N = 1; W = *freq*, $\delta = 2$>.

resultados muito similares ao apontado na Tabela 6.4 e por isso não foram considerados em configuração experimentais seguintes. Igualmente, para experimentações envolvendo a proposta de classificação expansiva usando o DatasetA, o valor de $\delta = 2$ foi mantido constante visto que esse valor proporcionou um melhor resultado - identificado através de procedimento de seleção de hiperparâmetros, onde se buscou variar o valor desse atributo uniformemente seguindo escala logarítmica (base = 2) no intervalo $1 \leq \delta \leq 2$ (cf. Subseção 4.2.2). Pela análise da matriz de confusão (Figura 6.2) associada ao modelo, é possível ainda perceber que a mesma apresenta melhor consistência em relação às classes de interesse. A partir desse resultado, buscou-se variar o valor de N e W tendo em vista conhecer o impacto de performance provocado pelos parâmetros. Os resultados desse processo são apresentados na Figura 6.3.

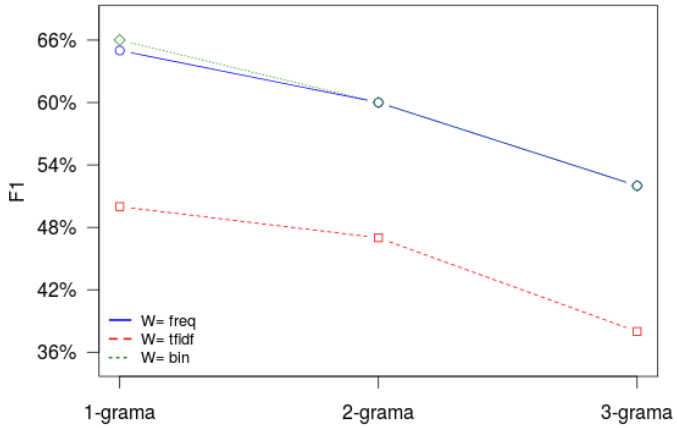


Figura 6.3: Comparação de resultados em função do número de N-gramas e o método de ponderação de termos.

Através da Figura 6.3, tem-se que o modelo expansivo baseado em ponderação de termos binária e unigramas atinge o melhor resultado. Porém, este não é expressivamente diferente para o caso onde se faz uso de ponderação baseada em frequência. No entanto, percebe-se uma diferença considerável em relação ao uso de TF-IDF, o que pode ser explicado através da arquitetura do método de classificação, no caso *Naive Bayes*, e sua função de decisão probabilística[8].

Apesar dos resultados de F1 apresentados na Tabela 6.3 demonstrarem uma melhoria de performance com o modelo proposto, entende-se que fornecer uma análise de curva de Característica de Operação do Receptor (do inglês, *Receiver Operating Characteristic*, ROC) pode dar uma representação visual mais consistente sobre a eficácia do modelo em questão. Esta análise é definida como uma representação gráfica que ilustra a performance de um sistema de classificação binária a partir da variação de seu limiar de discriminação (FACELI et al., 2011).

Curvas ROC apresentam taxa de verdadeiro positivo sobre o eixo Y e a taxa de falsos positivos no eixo X. A partir disso, o canto superior esquerdo é o ponto ideal - uma taxa de falsos positivos igual a 0, e uma taxa verdadeiro positivo igual a 1. Na prática, alcançar o ponto ideal é surreal, porém, geralmente quanto maior área sob a curva (do Inglês, *Area*

[8]SVM tem tendência a funcionar melhor com o uso de TF-IDF dado que sua fronteira de decisão tem poder suficiente para descrever espaços de características altamente complexos.

Under the Curve, AUC) melhor a performance do modelo (WITTEN; FRANK, 2005). A principal vantagem de se obter a AUC é que esta se apresenta como uma medida mais robusta do que a *Precisão* em situações onde há desbalanceamento de classes (JAPKOWICZ; SHAH, 2011), como no caso dos conjuntos de dados utilizados. A Figura 6.4 retrata o resultado da análise de ROC, corroborando a efetividade do modelo de classificação expansivo proposto.

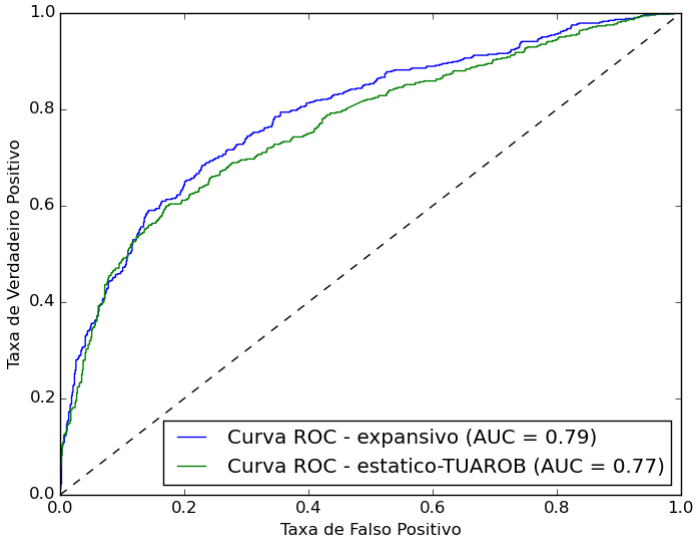


Figura 6.4: Curva ROC para os modelos investigados usando o DatasetA.

Os mesmos procedimentos de análise descritos acima também foram aplicados ao DatasetB. Para este caso, utilizou-se também, como resultado de base para fins de comparação, a configuração Φ = *estático-TUAROB* aplicada para o conjuto DatasetB, o que gerou um resultado 92%,66% e 77%, para precisão, revocação e medida F1, respectivamente. A matriz de confusão para este modelo é apresentada na Figura 6.5. Quanto ao suporte de teste gerado pela avaliação *holdout* no conjunto citado, esta possui 286 instâncias negativas e um total de 1239 instâncias positivas.

A Figura 6.6 apresenta os resultados de performance do modelo expansivo para o conjunto DatasetB no contexto da variação dos parâmetros W e N, Análide de ROC e matriz de confusão. Para este conjunto de dados, ao variar o valor do parâmetro δ, este proporcionou uma variação de desempenho de performance insignicativa[9] e, desta forma, foi mantido o mesmo para todos os casos experimentais envolvendo o conjunto citado.

[9]Usando o mesmo procedimento de seleção de hiperparâmetro descrito para o DatasetA

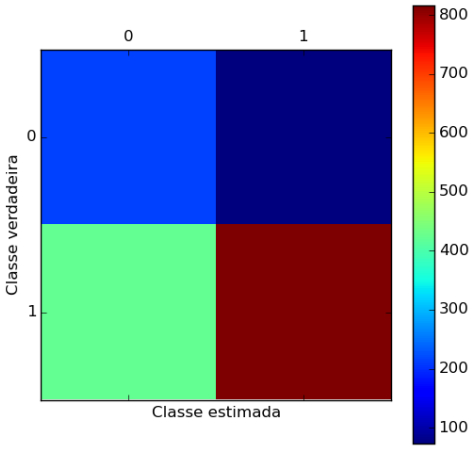


Figura 6.5: Matriz de confusão para o modelo Φ = *estático-TUAROB* no DatasetB.

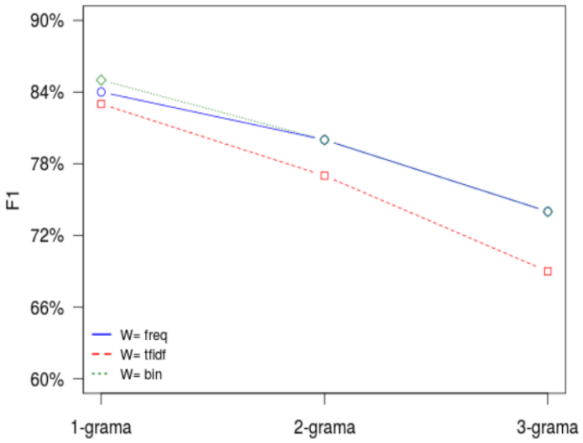
6.2.1 Discussão

Os resultados discutidos acima fornecem evidências de que uma estratégia baseada na confiança de predições realizadas pode elevar estimativas de performance e ajudar na descoberta de mensagens relacionadas à saúde, no domínio de mensagens sociais curtas, através do uso de métodos considerados simples, tal como *Naive Bayes*.

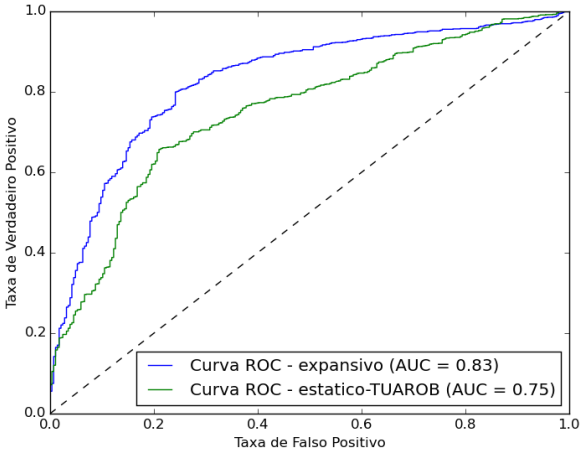
A principal contribuição desse processo, tomando como referências as evidências colhidas, está no fato de o mesmo abrir caminho para estudos mais aprofundados na área, tendo como referência o esquema proposto. Como a abordagem se baseia diretamente na medida de confiança de predição, o desenvolvimento/aprimoramento dessa medida deve conduzir a resultados ainda melhores. Além do mais, devido a dinâmica do espaço de mensagens no domínio investigado, não se faz apropriado o uso exclusivo de técnicas estáticas bem como não é viável a obtenção de uma quantidade massiva de dados rotulados manualmente, o que abre caminho para a consideração de abordagem expansivas como aqui tratada.

Outro aspecto positivo está ligado ao fato de que o uso de classificadores probabilísticos tende a facilitar a adaptação do modelo, uma vez que a atualização das probabilidades condicionais podem ser armazenadas em estruturas de dados de rápido acesso, diferente por exemplo, de classificadores baseados em ajustes de parâmetros multidimensionais e redução de função de custo.

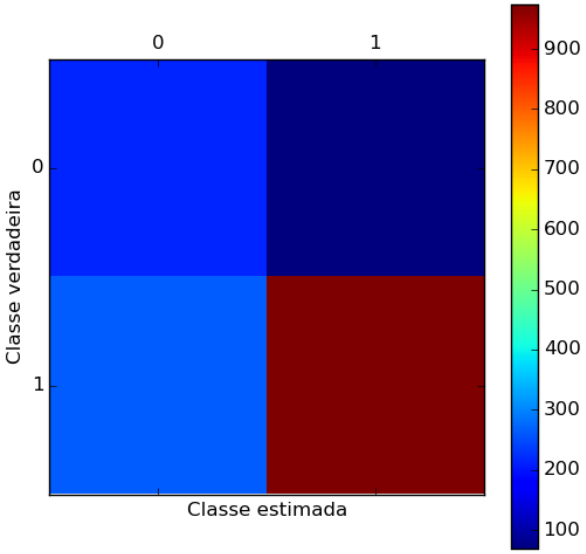
De fato, a análise experimental foi capaz de fornecer uma resposta regular à primeira questão de pesquisa, sendo que se conseguiu demonstrar o ganho de desempenho através da



(a) F1 em função de W e N.



(b) Curva ROC.



(c) Matriz de Confusão para $W = bin$ e $N = 1$.

Figura 6.6: Resultado de experimentação usando modelo expansivo e DatasetB. Em todos os casos experimentais, tem-se que $< \Omega = \text{hold-out}, \text{clean} = \text{True}, \text{stem} = \text{False}, \delta = 2 >$ são constantes.

expansão do conjunto de treinamento.

6.3 Experimento 2: Ponderação de termos

Em se tratando da análise experimental relacionada ao Capítulo 5, buscou-se observar o impacto da abordagem de ponderação de termos nos dois conjuntos de dados disponíveis. Nesses conjuntos, foi possível perceber que a abordagem consegue efetivamente discriminar os termos a partir de sua relação perante as diferentes classes, levantando termos que, à primeira vista, satisfazem a perspectiva intuitiva para o domínio investigado, isto é, a de que termos relacionados à saúde (*cold*, *flu*, *cramps*) possuem um maior *ranking* em relação à classe positiva quando comparado às suas ocorrências na classe negativa. Além disso, é possível obter um balanceamento de poder discriminativo entre os termos de modo que palavras raras ganham significância e palavras tais como artigos, verbos e preposições (*stopwords*), são atenuadas em importância.

As Tabelas 6.5 6.6 apresentam o comportamento da abordagem no que diz respeito ao ranqueamento de termos nas classes de interesse, exibindo as palavras mais importante e menos importante, respectivamente. Partindo das evidências observadas nas tabelas, investigou-se o impacto no desempenho de classificação a partir do uso da abordagem em combinação com o método de classificação *Naive Bayes*[10], foi $0,1$. A melhor configuração de modelo obtida foi: <clean=True; stem=False, N= (1,2)[11];W=freq>, em cenário de avaliação por validação cruzada em 10 partições, cujo resultado rendeu Pr%=54, Rev%=92 e F1%=68, para o DatasetA e Pr%=75, Rev%=89 e F1%=81, para o DatasetB. Figura 6.7, ilustra as matrizes de confusão associadas aos resultados.

[10]O melhor valor de parâmetro de correção de Laplace, utilizado para evitar problemas de contagem nula (quando termos na mensagem são totalmente desconhecido pelo modelo

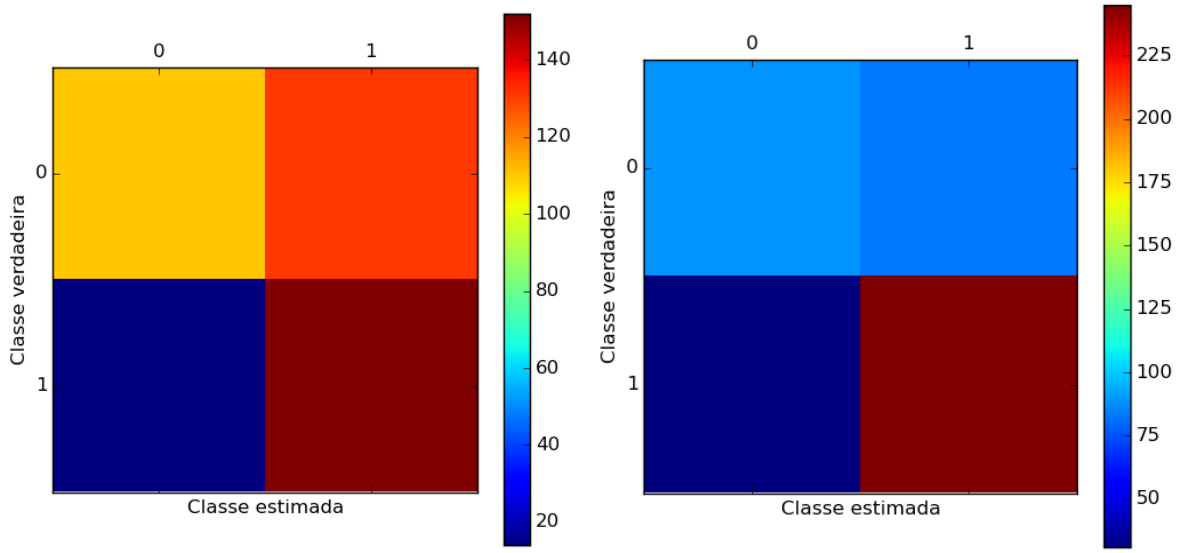
[11]Onde unigramas e bigramas são utilizados conjuntamente.

Tabela 6.5: As 20 palavras *mais* discriminativas segundo a abordagem do Capítulo 5. As palavras são listadas de acordo com as classes e os conjuntos de dados a que pertencem.

DATASETA		**DATASETB**	
Termos $(y = \text{pos})$	*Termos* $(y = \text{neg})$	*Termos* $(y = \text{pos})$	*Termos* $(y = \text{neg})$
cramps	*song*	*flu*	*URL*
(killing, me)	*expresions*	*hopefully*	*(big, worry)*
(sore, throat)	*wind*	*hoping*	*(york, times)*
asthma	*clubbing*	*bring*	*begin*
nose	*alone*	*pray*	*floridas*
ear	*rock*	*sanitizer*	*(floridas, meyer)*
lay	*(under, pressure)*	*(hand, sanitizer)*	*telephone*
shoulder	*twitches*	*dude*	*(from, telephone)*
fell	*tic*	*dying*	*hotlines*
arthritis	*joy*	*(seem, pretty)*	*reuters*
chest	*facial*	*remotely*	*(telephone, hotlines)*
(body, aches)	*guilt*	*(pretty, scared)*	*(will, begin)*
close	*mb*	*stds*	*(health, workers)*
acupuncture	*tia*	*(catching, stds)*	*mail*
bronchitis	*despair*	*(remotely, scared)*	*churches*
miserable	*sing*	*(someone, whos)*	*doses*
aspirin	*ball*	*(well, soon)*	*military*
severe	*turns*	*(not, remotely)*	*herald*
(major, headache)	*mph*	*sickness*	*haj*
ibuprofen	*(winds, are)*	*alergies*	*millions*

Tabela 6.6: As 20 palavras *menos* discriminativas segundo a abordagem do Capítulo 5. As palavras são listadas de acordo com as classes e os conjuntos de dados a que pertencem.

DATASETA		**DATASETB**	
Termos $(y = \text{pos})$	*Termos* $(y = \text{neg})$	*Termos* $(y = \text{pos})$	*Termos* $(y = \text{neg})$
anger	*alergies*	*guide*	*throat*
confusion	*hurts*	*ap*	*coughing*
weakness	*sore*	*h5n1*	*feeling*
bald	*(hate, being)*	*meyer*	*(sore, throat)*
nightmares	*meds*	*ways*	*stupid*
cocaine	*throat*	*(meyer, concerned)*	*went*
mania	*sinus*	*(your, guide)*	*daughter*
snoring	*slept*	*unscathed*	*nose*
heard	*(better, soon)*	*(season, unscathed)*	*(i've, been)*
dream	*dentist*	*washington*	*normal*
become	*antibiotics*	*(pregnant, women)*	*(going, around)*
winds	*gym*	*(over, pregnant)*	*nasty*
sex	*headache*	*york*	*guys*
understand	*hospital*	*(new, york)*	*sitting*
music	*infection*	*states*	*ha*
fake	*stomach*	*(health, oficials)*	*considering*
turned	*ulcer*	*(spreading, widely)*	*(fels, like)*
somebody	*hoping*	*(widely, worry)*	*past*
true	*muscle*	*widely*	*bout*
hearing	*ache*	*women*	*(woke, up)*



(a) Em relação ao DatasetA. **(b)** Em relação ao DatasetB.

Figura 6.7: Matriz de Confusão associadas ao teste da abordagem de ponderação de termos.

6.3.1 Discussão

Apesar dos resultados de classificação apontados apresentarem uma performance regular, é possível perceber, através das matrizes de confusão, que o modelo tende a gerar predições ligadas à classe positiva. Essa característica pode ser estuda no futuro, tendo como ponto norteador a avaliação da técnica através de outras famílias de classificadores, tais como as baseados em otimização (SVM, por exemplo) e em distância (KNN). Além disso, vale ressaltar a necessidade de buscar alternativas, isto é, aprimoramentos da Equação 5.4, de modo a lidar com o desbalanceamento no número de palavras em cada classe, ao qual limita a discriminação dos termos, e atualmente, pode ser visto como um ponto fraco da abordagem.

Adicionalmente, tem-se que a principal contribuição desse método de ponderação de termos, é a capacidade de levantar termos característicos do conjunto de dados e suas classes de interesse. Por exemplo é possível apontar itens discriminativos sem estar baseado diretamente no fator *frequência de termos*. Esta, conforme discutido, tem limitações bastante conhecidas para a tarefa de classificação de mensagens curtas relacionadas à saúde. Além disso, uma segunda contribuição está associada ao fato de que a abordagem de ponderação pode ajudar a construir, sem supervisão humana específica, um vocabulário de palavras importantes para uso em processos de classificação semisupervisionadas, ao qual este pode ser

usado para indexação em consultas à bancos de dados externos (*background knowledge*) - atualmente a geração de palavras-chaves é feito por consultas a especialistas de domínio.

Acredita-se, então, que a proposta de ponderação de termos investigada responde a segunda questão de pesquisa no sentido de que as métricas propostas por Timonen (2013) podem ser adaptadas ao processo de descoberta de termos relevantes frente à tarefa de classificação de documentos, principalmente no contexto de saúde. Pelo estudo desse aspecto, a pesquisa abre caminho para estudos posteriores que visem contribuir para o aperfeiçoamento de técnicas relacionadas ao domínio em discussão.

6.4 Considerações Finais

Este capítulo discutiu os resultados das análises experimentais relacionadas às propostas dos Capítulos 4 e 5. Apesar das análises fazerem uso de um conjunto de dados relativamente pequeno, tem-se que conjuntos com características similares, e até mesmo iguais, foram utilizados em pesquisas semelhantes (TUAROB et al., 2014); (PAUL; DREDZE, 2011b); (LAMB; PAUL; DREDZE, 2013). Além dos resultados, o capítulo discutiu as contribuições atreladas e evidenciou abertura para pesquisas futuras tendo como base as evidências colhidas e respostas às questões levantadas.

Capítulo 7

Conclusão e Trabalhos Futuros

O presente trabalho concentrou-se no problema de classificação binária de mensagens sociais curtas, precisamente *tweets*, sob a perspectiva geral de implementação de técnicas de aprendizagem de máquina no que toca a classificação de documentos em categorias (tópicos). O efeito de classificação investigado teve o propósito de categorizar mensagens em sua relação com saúde, determinando que um dado tweet $x^{(i)}$ pertence à classes positiva ($y =$ 'pos') se: (1) este indica que seu autor tem problema de saúde ou sintomas de doença (2) faz referência ao fato de outra pessoa (diferente do autor) estar doente ou mesmo expressa preocupações com a saúde em geral.

A motivação para o desenvolvimento da pesquisa esteve relacionada ao fato de que muitas abordagens e sistemas relacionados usam técnicas tradicionais de classificação bem sucedidas em ambientes onde os textos utilizados são longos, possuem uma alta distribuição de frequência de termos e, além disso, são conforme com a linguagem padrão. Porém, discutiu-se que o uso destas não está adequado ao cenário de classificação de mensagens sociais curtas, onde esparsividade de termos e ruídos associados à utilização livre da linguagem são comuns. Em adição, a motivação também leva em consideração as hipóteses, discutidas inicialmente em Silva (2012) e Zimmermann, Ntoutsi e Spiliopoulou (2014), de que um modelo de classificação estático, isto é, que não sofre retreino com o passar do tempo, não está adequado para o conteúdo gerado por mídias sociais. Isso porque os dados mudam constantemente e uma característica de modelo estático não dispõe de recursos para acompanhar a

evolução de características presentes nas mensagens.

Sendo assim, visando diretamente atacar o problema de utilização de um modelo estático, o primeiro objetivo desse trabalho investigou o impacto no desempenho de classificação de mensagens sociais curtas através da expansão contínua do conjunto de treinamento tendo como referência a confiança da predição realizada nas mensagens de teste. Neste contexto, a abordagem proposta (Capítulo 4) faz uso de aprendizado semisupervisionado, onde mensagens são incorporadas eficientemente ao modelo *Naive Bayes* e o cálculo da predição é feito reutilizando o limiar adaptativo proposto em Silva (2012). Diferente do foco de pesquisa abordado em Silva (2012), a questão de pesquisa atrelada ao objetivo descrito foi: "É possível, usando uma estratégia de expansão de conjunto de treinamento, obter um desempenho de classificação melhor que um modelo tradicional (estático), para o caso da classificação de tweets relacionados à saúde?".

Paralelamente, a pesquisa também objetivou avaliar alternativas para ponderação e extração de termos utilizados na classificação de modo a reduzir a dependência em métricas baseadas em frequência de termos (Capítulo 5), já que estes são relativamente uniformes (isto é, a maioria dos termos acontece apenas uma vez) em mensagens curtas. Para tanto, revisitou-se o trabalho de Timonen (2013) tendo como referência as medidas desenvolvidas pelo autor no que diz respeito a determinação da importância de cada termo em função de sua distribuição tanto em nível inter como intra classes. A questão de pesquisa atrelada à esse objetivo visou responder: É possível, utilizando as métricas de "contagem inversa de categoria", "probabilidade de categoria"e "probabilidade de documentos", retratadas em Timonen (2013), obter uma discriminação de termos consistente no contexto da classificação de mensagens relacionadas à saúde?

Experimentalmente (Capítulo 6), os resultados atrelados ao ataque do primeiro objetivo, considerando um linha de base (*baseline*) disposto em Tuarob et al. (2014), mostraram que a ação de expandir o classificador, por meio uma estimativa de confiança na predição realizada, amplia o conhecimento do classificador e, consequentemente, eleva o desempenho de classificação para o caso da classificação de mensagens curtas em domínio de saúde. Isso porque a abordagem atingiu um desempenho de classificação 9% superior em relação ao resultado de base, sob as mesmas configuração básica de representação, conjunto de dados e

avaliação. Para justificar o resultado, considerando a característica desbalanceada dos dados (SADILEK; KAUTZ; SILENZIO, 2012a), a análise ROC apontou uma área abaixo da curva (AUC) de $0,79$ e $0,83$ para a abordagem proposta, em comparação à $0,77$ e $0,75$ da linha de base, nos dois conjuntos de dados utilizados, respectivamente. Conforme citado na literatura, uma melhor área de AUC geralmente está associada a um melhor modelo de classificação (FACELI et al., 2011) (JAPKOWICZ; SHAH, 2011).

No contexto da experimentação voltada à segunda questão de pesquisa, tem-se que a principal contribuição obtida está na capacidade de levantar termos característicos do conjunto de dados e suas classes de interesse automaticamente, sem sofrer com limitações de frequência de termos. Isto, por exemplo, pode ser capaz de ajudar a construir, sem supervisão humana específica, processos de classificação semisupervisionados mais robustos e dinâmicos ao qual façam uso de listas de termos específicos para indexação em consultas à bancos de dados externos (*background knowledge*) $-$ o que geralmente é feito por consultas a especialistas de domínio. Com base nos resultados, viu-se que a abordagem de ponderação de termos, adaptadas do trabalho de Timonen (2013), é consistente no contexto da discriminação de termos importantes ligados à saúde (nome de doenças, sintomas), salientado diferente nível de importância para um mesmo termo perante as classes consideradas. Por exemplo, as Tabelas 6.5 e 6.6 na Seção 6.3, mostram, respectivamente, que termos têm maior e menor poder discriminativo nas duas classes consideradas. Pela análise desenvolvida, acredita-se que a proposta de ponderação de termos responde a segunda questão de pesquisa no sentido de que as métricas propostas por Timonen (2013) podem ser adaptadas positivamente ao processo de descoberta de termos relevantes frente à tarefa de classificação de documentos, principalmente no contexto de saúde.

Apesar do desempenho de classificação ter apresentado resultados relevantes, acredita-se que ainda há muitos caminhos de aperfeiçoamento e teste das abordagens propostas, principalmente no que diz respeito à análise experimental em conjuntos de dados mais extensos e através de outras famílias de algoritmos. Um aspecto importante, atrelado à experimentação em conjuntos massivos de dados, no entanto, é a questão de acompanhamento da estabilidade do modelo frente as constantes inserções/modificações do conjunto de treinamento. Devido às dificuldades concernetes a geração de uma largo conjunto de dados rotulados, essa

investigação fica determinada como alvo para trabalho futuro.

Outros fatores importantes e que devem ser explorados em trabalhos futuros, são:

- Investigação de novas estimativas de confiança que levem em consideração não só o histórico de predições realizadas, mas também fatores externos relevantes à aplicação de domínio e flutuações acentuadas na distribuição de termos;
- A redução do espaço de memória; uma vez que, pela abordagem discutida neste documento, o método expande o conjunto de treinamento indefinidamente, o que pode não ser prático em longos períodos de tempo e grandes volumes de dados. Para tanto, pode se pensar em modificações no mecanismo de esquecimento discutido em Zimmermann, Ntoutsi e Spiliopoulou (2014), onde mensagens (e seus termos associados) mais antigas, inseridas expansivamente no conjunto de treinamento, possuem influência progressivamente menor no processo de classificação;
- Exploração de modificações na Equação 5.4 visando melhor representação dos termos dentro das hipóteses discutidas;
- Teste da abordagem de ponderação em outros esquemas de classificação: Modelos baseados em Otimização (*SVM*, *Regressão Logística*) e Distância (*KNN*, K-médias).
- Acompanhamento do desempenho da abordagem de ponderação para fins de levantamento de conjunto de palavras-chaves para uso em abordagens fundamentadas em conhecimento externo (*background knowledge*).

Ao final da pesquisa descrita neste documento, acredita-se que esta foi capaz de levantar, testar e discutir hipóteses interessantes para a tarefa da classificação de mensagens sociais curtas, principalmente no domínio de saúde. O desenvolvimento das hipóteses aqui propostas podem beneficiar o surgimento de aplicações mais robustas no campo da vigilância, controle e contrapartida de eventos reais de saúde (epidemiologia, campanhas de saúde, etc). Além disso, muito embora faça menção recorrente à temática de saúde, as abordagens consideradas não estão limitadas à esse domínio, sendo facilmente adaptadas a um outro ambiente de interesse.

Bibliografia

ACHREKAR, H. et al. Predicting flu trends using twitter data. In: *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. [S.l.: s.n.], 2011. p. 702–707.

BERMINGHAM, A.; SMEATON, A. F. Crowdsourced real-world sensing: sentiment analysis and the real-time web. In: *Sentiment Analysis Workshop at Artificial Intelligence and Cognitive Science (AICS 2010)*. Galway, Ireland: [s.n.], 2010.

BIFET, A. *Adaptive stream mining : pattern learning and mining from evolving data streams*. Amsterdam, Netherlands: IOS Press, 2010. (Frontiers in artificial intelligence and applications).

BIFET, A.; FRANK, E. Sentiment knowledge discovery in twitter streaming data. In: *Proceedings of the 13th International Conference on Discovery Science*. Berlin, Heidelberg: Springer-Verlag, 2010. (DS'10), p. 1–15.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2009.

BRENNAN, S.; SADILEK, A.; KAUTZ, H. Towards understanding global spread of disease from everyday interpersonal interactions. In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 2013. (IJCAI '13), p. 2783–2789.

BRISCOE, T.; CARROLL, J. A.; WATSON, R. The second release of the rasp system. In: *ACL*. [S.l.]: The Association for Computer Linguistics, 2006.

CARAGEA, C. et al. Classifying text messages for the haiti earthquake. In: *8th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Lisbon, Portugal: [s.n.], 2011.

CHAPELLE, O.; SCHLKOPF, B.; ZIEN, A. *Semi-Supervised Learning*. 1st. ed. [S.l.]: The MIT Press, 2010.

CHEN, L. et al. Vision: Towards real time epidemic vigilance through online social networks: Introducing sneft – social network enabled flu trends. In: *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*. New York, NY, USA: ACM, 2010. (MCS '10), p. 4:1–4:5.

COLLIER, N.; SON, N. T.; NGUYEN, N. M. Omg u got flu? analysis of shared health messages for bio-surveillance. *CoRR*, abs/1110.3089, 2011.

CORLEY, C. D. et al. Monitoring influenza trends through mining social media. In: *International Conference on Bioinformatics and Computational Biology (BIOCOMP09)*. Las Vegas, NV: [s.n.], 2009.

CULOTTA, A. Towards detecting influenza epidemics by analyzing twitter messages. In: *Proceedings of the First Workshop on Social Media Analytics*. New York, NY, USA: ACM, 2010. (SOMA '10), p. 115–122.

DOAN, S.; OHNO-MACHADO, L.; COLLIER, N. Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In: *Proceedings of the 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*. Washington, DC, USA: IEEE Computer Society, 2012. (HISB '12), p. 62–71.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification (2Nd Edition)*. [S.l.]: Wiley-Interscience, 2000.

FACELI, K. et al. *Inteligência Artificial - Uma abordagem de aprendizado de máquina*. [S.l.]: LTC, 2011.

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

GOMIDE, J. S. *Mineração de redes sociais para detecção e previsão de eventos reais*. Dissertação (Mestrado) — Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 2012.

GRUZD, A.; WELLMAN, B.; TAKHTEYEV, Y. *Imagining Twitter as an Imagined Community*. 2011.

JAPKOWICZ, N.; SHAH, M. *Evaluating Learning Algorithms: a classification perspective*. Cambridge, New York: Cambridge University Press, 2011.

KIRA, K.; RENDELL, L. A. The feature selection problem: Traditional methods and a new algorithm. In: *Proceedings of the Tenth National Conference on Artificial Intelligence*. [S.l.]: AAAI Press, 1992. (AAAI'92), p. 129–134.

LAMB, A.; PAUL, M. J.; DREDZE, M. Separating fact from fear: Tracking flu infections on twitter. In: *In NAACL*. [S.l.: s.n.], 2013.

LAMPOS, V.; CRISTIANINI, N. Tracking the flu pandemic by monitoring the social web. In: IEEE. *CIP '10*. [S.l.]: IEEE, 2010. p. 411–416.

LAMPOS, V.; CRISTIANINI, N. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.*, ACM, New York, NY, USA, v. 3, n. 4, p. 72:1–72:22, set. 2012.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

PAUL, M.; DREDZE, M. You are what you tweet: Analyzing twitter for public health. In: . [S.l.: s.n.], 2011b.

PAUL, M. J.; DREDZE, M. *A Model for Mining Public Health Topics from Twitter*. [S.l.], 2011a.

PORTER, M. F. Readings in information retrieval. In: JONES, K. S.; WILLETT, P. (Ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. cap. An Algorithm for Suffix Stripping, p. 313–316.

SADILEK, A.; KAUTZ, H. Modeling the impact of lifestyle on health at scale. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2013. (WSDM '13), p. 637–646.

SADILEK, A.; KAUTZ, H.; SILENZIO, V. Modeling spread of disease from social interactions. In: *In Sixth AAAI International Conference on Weblogs and Social Media (ICWSM*. [S.l.: s.n.], 2012a.

SADILEK, A.; KAUTZ, H.; SILENZIO, V. Predicting disease transmission from geo-tagged micro-blog data. In: *In Twenty-Sixth AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2012b.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. (WWW '10), p. 851–860.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 24, n. 5, p. 513–523, ago. 1988.

SCELLATO, S. et al. Distance matters: Geo-social metrics for online social networks. In: *Proceedings of the 3rd Wonference on Online Social Networks*. Berkeley, CA, USA: USENIX Association, 2010. (WOSN'10), p. 8–8.

SIGNORINI, A.; SEGRE, A. M.; POLGREEN, P. M. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, Public Library of Science, v. 6, n. 5, 05 2011.

SILVA, I. S. *Análise adaptativa de fluxos de sentimento*. Dissertação (Mestrado) — Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 2012.

SRIRAM, B. et al. Short text classification in twitter to improve information filtering. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2010. (SIGIR '10), p. 841–842.

TIMONEN, M. *Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion*. Tese (article-based) — Department of Computer Science, Faculty of Science, University of Helsinki, VTT Technical Research Centre of Finland, Finland, 2013.

TUAROB, S. et al. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of Biomedical Informatics*, v. 49, n. 0, p. 255 – 268, 2014.

VELOSO, A.; MEIRA, W.; ZAKI, M. J. Lazy associative classification. In: *Proceedings of the Sixth International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2006. (ICDM '06), p. 645–654.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

ZIMMERMANN, M.; NTOUTSI, E.; SPILIOPOULOU, M. Adaptive semi supervised opinion classifier with forgetting mechanism. In: *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2014. (SAC '14), p. 805–812.