Submission date: 30/09/2021
Resubmission date: 25/02/2022
Camera ready submission date: 10/06/2022

1st round notification: 25/01/2022
2nd round notification: 12/04/2022
Available online: 05/08/2022
Publication date: 05/08/2022

Section: Special Issue
Information Systems in Life and Health Sciences

Natural Language Processing for Clinical Data Classification

Orrana L. V. de Sousa[1], Deborah M. V. Magalhães[1,2], Victor E. S. Campelo[3], Romuere R. V. e Silva[1,2]

[1]Electrical Engineering - PPGEE/UFPI
Picos, Piauí – Brazil

[2]Information Systems - CSHNB/UFPI
Picos, Piauí – Brazil

[3]Specialized Medicine - DME/UFPI
Teresina, Piauí – Brazil

{orranalhaynherv, dr.vcampelo}@gmail.com, {deborah.vm, romuere}@ufpi.edu.br

Abstract. *The widespread adoption of systems for managing and recording medical documents (MD) has generated a large volume of unstructured data. It corresponds to free text containing ambiguous expressions to describe conditions or procedures. It makes the task of manually categorizing MD error-prone. This work aims to label and classify MD in Portuguese using binary labeling (Recipes and Others) and multi-class (Recipes, Exams, Certificates, and Others). The n-gram and term frequency - inverse document frequency (TF–IDF) were used in the text vectorization step. The results achieved are promising: they presented 0.99 and 0.97 for Kappa in the binary and multi-class classification, respectively. Thus, with the classification of MD, it is possible to provide segmentation of information to manage prescription drugs.*

Keywords. *named-entity recognition; clinical dataset; term frequency - inverse Document Frequency.*

Resumo. *A ampla adoção de sistemas para o gerenciamento e registro de documentos médicos (MD) têm gerado um grande volume de dados não estruturados. Tais dados correspondem a texto livre contendo expressões ambíguas para relatar a mesma condição clínica ou procedimentos. Isso torna a tarefa de categorização manual do MD sujeita a erros. Este trabalho visa rotular e classificar MD em português utilizando a rotulação binária (Receita e Outros) e a multiclasse (Receitas, Exames, Atestados e Outros). O n-grama e a frequência do termo - frequência inversa do documento (TF–IDF) foram utilizados na etapa de vetorização do texto. Os resultados alcançados são promissores: apresentaram 0,99 e 0,97 para o Kappa na classificação binária e multiclasse, respectivamente. Assim, com a classificação do MD, é possível fornecer segmentação das informações para gerenciar medicamentos prescritos.*

Cite as: Sousa, O. L. V., Magalhaes, D. M. V., Campelo, V. E. S. & Silva, R. R. V. (2022). An automatic method to medical documents labeling and categorization (Um método automático para rotulagem de documentos médicos e categorização). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 15(1), 13:1-13:17. DOI: 10.5753/isys.2022.2260.

Palavras-Chave. *reconhecimento de entidade mencionada; base de dados clínica; termo - frequência inversa de documento.*

1. Introdução

Com o aumento do uso de soluções tecnológicas para a automação de processos hospitalares, como a utilização de sistemas para o gerenciamento e registro de documentos médicos, um grande número de dados não estruturados vem sendo produzidos [Assale et al. 2019, Wulff et al. 2020, Tayefi et al. 2021]. Eles são criados geralmente na forma de texto livre e são caracterizados por uma multiplicidade intrínseca de expressões pelas quais médicos podem relatar a mesma condição clínica ou procedimentos [Cabitza et al. 2019].

Muitas das informações existentes nesses registros contêm abreviações, termos ambíguos e erros de digitação. Juntos, esses fatores tornam a categorização manual uma tarefa cara, demorada e sujeita a erros [Reys et al. 2020]. Assim, a classificação automática de dados médicos em categorias informativas clínicas pode reduzir enormemente o custo do trabalho humano em serviços médicos. As aplicações de classificação de textos clínicos não só podem ajudar os médicos a melhorar a qualidade do serviço, mas também ajudam a construir a base para aplicações mais avançadas, como sistemas de autopreenchimento de prescrições clínicas [Liu et al. 2020].

A classificação pode ser realizada com o uso de técnicas de processamento de linguagem natural e aprendizado de máquina. Desse modo, esses documentos clínicos podem ser utilizados com maior facilidade no uso secundário (uso orientado para pesquisa e análises) e em uma maior gama de tarefas, como na extração de informações segundo a categoria do texto, utilização no treinamento de sistemas hospitalares de preenchimento automático, gerenciamento de drogas receitadas a pacientes com testes de interação medicamentosa, etc, [Cui et al. 2019].

Entretanto, existe uma carência de trabalhos que utilizam dados clínicos em português para este fim. Na literatura, a maioria dos estudos tem como base a língua inglesa. Isso se dá principalmente pelo número maior de iniciativas de disponibilização para desafios de processamento de linguagem natural, como o compartilhamento de conjuntos de dados como o Monitoramento Inteligente Multiparâmetro em Terapia Intensiva (MIMIC) [Lee et al. 2011, Johnson et al. 2016]. Assim, a falta dessas iniciativas e o nível de confidencialidade dos dados envolvidos faz com esses dados sejam de difícil acesso para pesquisadores.

Nesse contexto, o propósito deste trabalho é rotular e classificar um conjunto de dados de textos clínicos em português, utilizando técnicas de aprendizado de máquina. Esse processo tornará possível a utilização dessas informações em tarefas posteriores, como no uso dessas informações para o reconhecimento de entidade mencionada (NER). Na etapa de pré-processamento, alguns métodos são utilizados, como a remoção de alguns caracteres específicos do texto e a remoção da pontuação. Depois, na rotulação, duas categorias diferentes são utilizadas: a rotulação binária e a multiclasse. Após isso, a classificação é realizada com o uso de técnicas de representação vetorial de texto e algoritmos do estado da arte, como a Máquina de Vetores de Suporte, Floresta Aleatória,

XGBoost e Perceptron Multicamadas.

Portanto, as principais contribuições desse trabalho são: (1) a rotulação de um base de dados em português relativos à rotina clínica brasileira envolvendo especialidades como: dermatologia, nutrição e otorrinolaringologia, (2) a detecção automática de diferentes tipos de textos clínicos, como receitas, laudos, solicitação de exames. Desse modo, é possível estruturar uma base de dados que vai auxiliar a segmentação de informações para gerenciamento de drogas receitadas.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta alguns trabalhos relacionados à classificação de textos médicos. A Seção 3 descreve a metodologia empregada neste trabalho. A Seção 4 apresenta os resultados. Por fim, a Seção 5 traz as considerações finais e trabalhos futuros.

2. Trabalhos Relacionados

A classificação automática de texto é um método eficaz para categorizar arquivos em rótulos predefinidos em diferentes níveis como, documento, sentença e caractere [Sebastiani 2002]. Com isso, essa técnica tem sido utilizada para rotular documentos médicos, facilitando assim a organização e extração de informações. Esta seção apresenta e discute estudos da literatura sobre o uso da classificação de texto relativos aos arquivos médicos.

No trabalho de Weng et al. (2017), anotações clínicas foram classificadas conforme o subdomínio médico, como cardiologia ou neurologia, através da construção de um pipeline de processamento de linguagem natural e algoritmos de classificação. Eles compararam o desempenho dos classificadores usando diferentes representações de dados, estratégias de ponderação e algoritmos de aprendizado supervisionado, e utilizaram dois conjuntos de dados adquiridos do repositório de dados *Integrating Data for Analysis, Anonymization and Sharing* (iDASH) [Ohno-Machado et al. 2012] e anotações clínicas do *Massachusetts General Hospital* (MGH) [Murphy and Chueh 2002]. Como resultados, nas duas bases de dados, respectivamente, o classificador Naive Bayes alcançou um F1-score de 0,893 e 0,755, e AUC de 0,935 e 0,867, e a rede neural recorrente convolucional (CRNN) um F1-score de 0,845 e 0,881, e AUC de 0,975 e 0,990.

Baumel et al. (2018) investigaram quatro modelos para atribuição de vários códigos da Classificação Internacional de Doenças (CID) em resumos de altas hospitalares com os conjuntos de dados clínicos MIMIC II [Lee et al. 2011] e III [Johnson et al. 2016]. Dentre estas abordagens de classificação estão o modelo um-vs-todos baseado em SVM, o modelo de saco de palavras contínuo (CBOW), o modelo de rede neural convolucional (CNN) e o modelo proposto *Hierarchical Attention-bidirecional Gated Recurrent Unit* (HA-GRU), uma abordagem hierárquica para rotular um documento. Como resultados, utilizando os dois *datasets*, foram obtidos um Micro-F com a SVM de 0,325 e 0,530, com o CBOW de 0,421 e 0,433, com a CNN de 0,464 e 0,526, e com HA-GRU de 0,539 e 0,559.

Cui et al. (2019) propuseram uma abordagem heurística construtiva para gerar um conjunto de expressões regulares que foram usadas na classificação de textos médicos. Para isso, 13 departamentos clínicos com 776 categorias médicas foram predefinidos por

especialistas e um total de 4.634.742 registros efetivos foram coletados. Duas alternativas de experimentos foram estudadas: o classificador de texto baseado em expressão regular e os modelos de aprendizado de máquina, Naive Bayes, Máquina de Vetores de Suporte, Rede Neural Recorrente (RNN) e Rede Neural Convolucional (CNN). Como melhores resultados, no primeiro experimento obtiveram uma precisão de 0,89 e uma recall de 0,57 com o classificador baseado em expressão regular, e no segundo experimento obtiveram acurácia de 0,95, precisão de 0,94, recall de 0,87, macro F0.5 de 0,92 e micro F0.5 de 0,94 com a CNN.

Teng et al. (2020) utilizaram uma abordagem de aprendizado profundo e um método de mineração de tópicos médicos com a utilização de códigos de metadados da Classificação Internacional de Doenças (CID) para classificar registros textuais clínicos. Os *datasets* utilizados foram o MIMIC III e um conjunto de dados de EHRs ambulatoriais chineses construído pelos autores. Inicialmente, foram selecionados os parágrafos mais úteis de todos os registros médicos. Depois, esses registros pré-processados alimentaram a arquitetura proposta Codificação ICD de Atenção Cruzada (CAIC) para agregar informações usando redes neurais profundas. Como resultados obtidos, tiveram uma AUC de 0,885, precisão de 0,640, *recall* de 0,577 e F1-*score* de 0,607.

Os trabalhos mencionados acima realizaram a classificação de *datasets* em inglês com diferentes intuitos, desde a rotulação utilizando subdomínios médicos até o uso dos códigos da Classificação Internacional de Doenças (CDI). Neste trabalho, a classificação de textos médicos é realizada com o uso de um conjunto privado de dados em português. Dois tipos de rotulação são consideradas: a binária (Receitas e Outros) e a multiclasse (Receitas, Exames, Atestados/Declarações e Outros). Inspirados na literatura, quatro algoritmos de machine learning do estado da arte foram utilizados para classificar esse *dataset* de acordo com essas rotulações.

3. Metodologia

Esta seção apresenta a metodologia de classificação de textos médicos e descreve sua aplicação. Ela é baseada no fluxo de classificação de textos apresentado no Guia de Aprendizado de Máquina desenvolvido pelo Google [1] e possui seis componentes: o banco de dados, a etapa de pré-processamento, a rotulação, a vetorização, a classificação desses vetores (textos) e a validação. A Figura 1 ilustra tais componentes.

3.1. Base de dados

O banco de dados utilizado neste trabalho é de fonte privada, e foi formado por diversos documentos médicos, como receitas, solicitações de exames, notas clínicas, atestados/declarações, encaminhamentos, etc, produzidos por diversos médicos durante consultas presenciais. No total, são 11219 amostras coletadas no período de 01 de janeiro de 2019 até 31 de dezembro de 2020. Essas amostras estão no formato rico de texto (*Rich Text Format* - RTF), que possui muitos elementos de formatação, como fontes, caracteres, características como negrito, itálico e sublinhado, além de imagens ou gráficos. A Tabela 1 apresenta exemplos de amostras para cada uma das categorias da base de dados.

[1] **Step 2.5: Choose a model - Text classification guide.** Disponível em `https://developers.google.com/machine-learning/guides/text-classification/step-2-5`

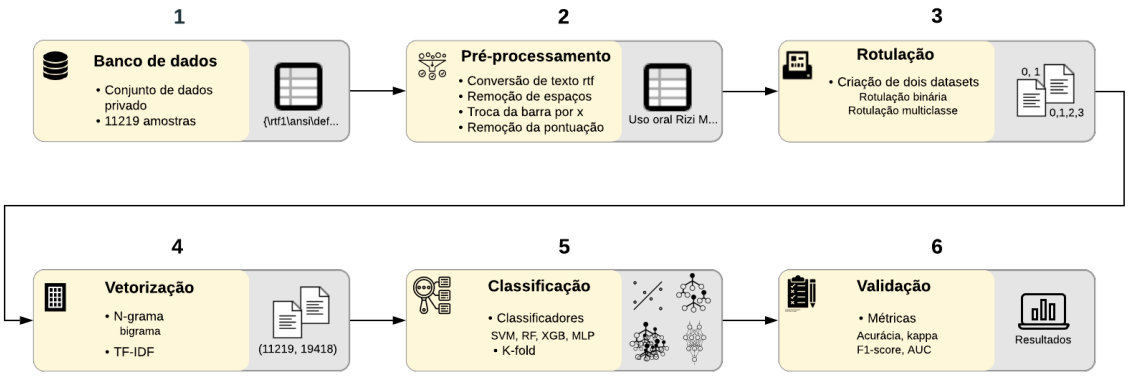

Figura 1. Metodologia de classificação de documentos médicos.

Tabela 1. Exemplos do conteúdo das amostras por categoria da base de dados. Os asteriscos (*) indicam informações ocultadas para garantir a confidencialidade dos dados do médico e paciente. Já os \n e \t indicam as formatações do formato rico de texto.

Categoria	Texto
Receita	Uso oral \n Betina 24 mg ——————— Tomar um comp via oral de 12/12 horas por 2 meses
Solicitação de exame	No **** Convênio ***** Matrícula ********* \n Solicito RM da coluna lombar
Atestado	ATESTADO MÉDICO \n\t Atesto para os devidos fins que o paciente supracitado necessita de 03 três dias afastado das atividades profissionais \n\n CID 10 j040xr490
Encaminhamento	Encaminho ao Dermatologista \n Paciente com xantelasma e olheira \n Solicito avaliação
Laudo médico	LAUDO MÉDICO \n Paciente atópica portadora de alergia respiratória e dermatite alérgicas a disposição

3.2. Pré-processamento

Antes que o texto seja vetorizado, os dados devem ser tratados para aumentar a qualidade de representação deles. Então, diferentes técnicas de pré-processamento foram utilizadas. Inicialmente, quatro amostras do *dataset* foram removidas por conter dados nulos e a conversão dos arquivos no formato *rich text* (rtf) em strings Python foi feita. Muitos documentos médicos são escritos em formato rtf, o que não é ideal para análise e processamento posterior. Em seguida, foram removidos todos os caracteres representantes de espaços em branco identificados nas amostras, como o \t, \n e o \r, e as barras encontradas na posologia de receitas foram substituídas pelo caractere x, sendo posteriormente removidas. Finalmente, a última etapa consiste na remoção da pontuação. Na Tabela 2, exemplos de documentos clínicos que constituem o conjunto de dados são mostrados antes e depois do pré-processamento.

Tabela 2. Exemplos dos documentos clínicos antes e depois do pré-processamento. Os asteriscos (*) indicam informações ocultadas para garantir a confidencialidade dos dados do médico e paciente. Já os \n e \t indicam as formatações do formato rico de texto.

Categoria	Antes do pré-processamento	Depois do pré-processamento
Receita	Uso oral \n Betina 24 mg ——— ——————— Tomar um comp via oral de 12/12 horas por 2 meses	Uso oral Betina 24 mg Tomar um comp via oral de 12x12 horas por 2 meses
Exames	Solicitação de Exame: No. **** Convênio: ***** Matrícula: ******** \n Solicito: RM DA COLUNA LOMBAR	Solicitação de Exame No **** Convênio ***** Matrícula ******** Solicito RM DA COLUNA LOMBAR
Atestados	ATESTADO MÉDICO \n\t Atesto, para os devidos fins, que o paciente supracitado necessita de 01 (um) dias afastado das atividades profissionais \n\n CID 10 j00	ATESTADO MÉDICO Atesto para os devidos fins que o paciente supracitado necessita de 01 um dias afastado das atividades profissionais CID 10 j00
Outros	Ao Dentista \n\n\n\t Encaminho a paciente com roncos e apnéia leve para colocação de aparelho intra-oral. \n	Ao Dentista Encaminho a paciente com roncos e apnéia leve para colocação de aparelho intra-oral

3.3. Rotulação

Com as amostras já pré-processadas, a rotulação da base de dados foi feita baseada na criação de dois *datasets*: um com rotulação binária e outro com rotulação multiclasse. Na rotulação binária, o *dataset* foi dividido em amostras de receitas e outros. Já na rotulação multiclasse, as amostras foram divididas nas seguintes classes: receitas, exames, atestados/declarações e outros. Esse processo foi feito utilizando expressões específicas identificadas como representantes de cada uma das classes definidas. Em receitas, por exemplo, palavras como 'uso', 'fórmula' e 'tratamento' são frequentes. Após isso, a rotulação foi verificada manualmente e os rótulos finais foram confirmados. Como resultado dessa rotulação, tivemos a primeira base formada por 7617 e 3602 amostras, respectivamente, e a segunda formada por 7617, 2002, 258 e 1342 amostras, respectivamente.

3.4. Vetorização

Seguindo o fluxo apresentado no Guia de Aprendizado de Máquina, a definição do modo de tokenização ocorre nessa etapa. Essa tokenização é feita por palavra. Depois, é utilizado uma técnica conhecida como n-grama, em que uma amostra de texto é representada por uma sequência contígua de n itens, podendo ser definidos 1,2,3,...,n itens para essa representação [Brown et al. 1992]. Neste trabalho, foi utilizado o unigrama e bigrama, ou seja, cada texto foi dividido em conjuntos de um e dois elementos adjacentes

de cada sequência de *tokens*. Na Tabela 3, exemplos de amostras com o processamento do n-grama são apresentados.

Tabela 3. Exemplos das classes do *dataset* na criação de unigramas e bigramas.

Categoria	Texto	Bigrama
Receita	Uso oral Betina 24 mg Tomar um comp via oral de 12x12 horas por 2 meses	['12x12', '12x12 horas', '24', '24 mg', 'betina', 'betina 24', 'comp', 'comp via', 'de', 'de 12x12', 'horas', 'horas por', 'meses', 'mg', 'mg tomar', 'oral', 'oral betina', 'oral de', 'por', 'por meses', 'tomar', 'tomar um', 'um', 'um comp', 'uso', 'uso oral', 'via','via oral']
Exames	Solicitação de Exame No **** Convênio ****** Matrícula ******** Solicito RM DA COLUNA LOMBAR	['********', '******** solicito', '****', '**** convênio', 'coluna', 'coluna lombar', 'convênio', 'convênio ******', 'da', 'da coluna', 'de', 'de exame', 'exame', 'exame no', 'lombar', 'matrícula', 'matrícula ********', '******', '****** matrícula', 'no', 'no ****', 'rm', 'rm da', 'solicitação', 'solicitação de', 'solicito', 'solicito rm']
Atestados	ATESTADO MÉDICO Atesto para os devidos fins que o paciente supracitado necessita de 01 um dias afastado das atividades profissionais CID 10 j00	['01', '01 um', '10', '10 j00', 'afastado', 'afastado das', 'atestado', 'atestado médico', 'atesto', 'atesto para', 'atividades', 'atividades profissionais', 'cid', 'cid 10', 'das', 'das atividades', 'de', 'de 01', 'devidos', 'devidos fins', 'dias', 'dias afastado', 'fins', 'fins que', 'j00', 'médico', 'médico atesto', 'necessita', 'necessita de', 'os', 'os devidos', 'paciente', 'paciente supracitado', 'para', 'para os', 'profissionais', 'profissionais cid', 'que', 'que paciente', 'supracitado', 'supracitado necessita', 'um', 'um dias']
Outros	Ao Dentista Encaminho a paciente com roncos e apnéia leve para colocação de aparelho intraoral	['ao', 'ao dentista', 'aparelho', 'aparelho intraoral', 'apnéia', 'apnéia leve', 'colocação', 'colocação de', 'com', 'com roncos', 'de', 'de aparelho', 'dentista', 'dentista encaminho', 'encaminho', 'encaminho paciente', 'intraoral', 'leve', 'leve para', 'paciente', 'paciente com', 'para', 'para colocação', 'roncos', 'roncos apnéia']

Em seguida, a medida estatística Frequência do Termo - Frequência Inversa do Documento (TF–IDF) é usada. Essa medida é uma abordagem comumente usada para indicar a importância de uma palavra em um documento em relação a uma coleção de documentos ou um corpus linguístico [Yun-tao et al. 2005]. Essa abordagem captura a relevância entre palavras, documentos de texto e categorias particulares. Assim, essa medida foi empregada para indicar a importância de cada unigrama e bigrama numa amostra

em relação a coleção de amostras de cada base, resultando em uma matriz de valores numéricos de 11219x19418 correspondente ao texto vetorizado.

3.5. Classificação

Na classificação, quatro algoritmos foram utilizados: Máquina de Vetores de Suporte, Floresta Aleatória, XGBoost e Perceptron Multicamadas. A Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) é um algoritmo do estado da arte que é uma técnica de aprendizado estatístico não paramétrico supervisionado [Mountrakis et al. 2011]. Ele se baseia no princípio de minimização do risco estrutural (SRM), que visa selecionar uma função de hipótese com baixa capacidade de uma sequência aninhada de funções que minimiza simultaneamente a taxa de erro verdadeira e a taxa de erro empírica [Burges 1998].

O algoritmo Floresta Aleatória (*Random Forest* - RF) é um método de aprendizagem de conjunto para classificação, regressão e outras tarefas. Esse algoritmo é uma combinação de preditores de árvore de decisão - caracterizados pelo uso de uma ou várias funções de decisão de forma sucessiva (diagrama de árvore) [Swain and Hauska 1977] - de modo que cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta [Breiman 2001]. Cada árvore individual na floresta aleatória exibe uma previsão de classe e a classe com mais votos torna-se a previsão do modelo.

Já o XGBoost (*eXtreme Gradient Boosting* - XGB) é uma aplicação extensível e de ponta de máquinas de aumento de gradiente baseadas em árvores de decisão que geralmente produz alta precisão e tempo de processamento rápido [Chen and Guestrin 2016]. Ele é um algoritmo em que novos modelos são criados para prever os resíduos de modelos anteriores e, em seguida, somados para fazer a previsão final [Ogunleye and Wang 2019]. Para minimizar a perda ao adicionar novos modelos, ele usa um algoritmo de descida de gradiente.

O Perceptron Multicamadas (*Multilayer Perceptron* - MLP) é uma classe de rede neural artificial feedforward (ANN) [Hornik et al. 1989]. Ele consiste em um sistema de neurônios simples interconectados, ou nós, sendo um modelo que representa um mapeamento não linear entre um vetor de entrada e um vetor de saída [Gardner and Dorling 1998]. Ao contrário de outras técnicas estatísticas, o perceptron multicamadas não faz suposições anteriores sobre a distribuição de dados. Ele pode modelar funções não lineares e pode ser treinado para generalizar com precisão quando apresentado a novos dados [Lins and Ludermir 2005].

Com o uso desses quatro classificadores, dois experimentos foram realizados: um com a base binária e outro com a base multiclasse. Na SVM e RF, os hiperparâmetros padrões foram utilizados. Enquanto no XGB, a métrica de avaliação foi definida como a mlogloss, e na MLP o número de iterações foi determinado como 100. A validação cruzada com o método k-folds foi utilizada para dividir o conjunto total de dados em 10 subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, os modelos foram treinados e avaliados.

3.6. Validação

Para avaliar os classificadores, selecionamos a acurácia (Equação 1), o kappa (Equação 2), o F1-score (Equação 3) e a área sob a curva ROC (AUC) como as métricas. O cálculo da acurácia (Acc) [Baratloo et al. 2015] é apresentado na equação 1:

$$Acc = \frac{VP + VN}{VP + VN + FP + FN}, \tag{1}$$

onde VP, VN, FP, e FN correspondem a Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo, respectivamente. Quanto mais próximo de 1, melhor o resultado da acurácia.

A métrica kappa de Cohen [Cohen 1960] indica como os classificadores selecionados superaram o classificador que simplesmente adivinha aleatoriamente conforme a frequência de cada classe. Os valores do índice são categorizados como: ruim ($\kappa \leq 0.2$), razoável ($0.21 \leq \kappa \leq 0.4$), bom ($0.41 \leq \kappa \leq 0.6$), muito bom ($0.61 \leq \kappa \leq 0.8$), e excelente ($\kappa \geq 0.81$). Ele é definido como:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{2}$$

onde p_o é a concordância observada, e p_e é a concordância esperada.

O F1-score corresponde a média harmônica entre precisão e sensibilidade. Essa métrica é comumente utilizada em problemas com classes desequilibradas, como na base adotada neste trabalho. O valor mais alto possível do F1-score é 1, indicando precisão e sensibilidade perfeitos, e o valor mais baixo possível é 0, se a precisão ou sensibilidade for zero.

$$F1 - Score = \frac{VP}{VP + \frac{1}{2}(FP + FN)}. \tag{3}$$

Por fim, usamos a métrica área sob a curva ROC (AUC) [Hanley and McNeil 1982]. O melhor desempenho é alcançado quando o valor AUC se aproxima de 1 (equivalente a 100 %).

4. Resultados

Nesta seção, são apresentados e discutidos os resultados da classificação dos textos médicos. Essa classificação se divide em duas categorias de acordo com as rotulações feitas na base de dados: binária e multiclasse.

Na Figura 2, a incorporação de vizinho estocástico distribuído t (*t-Distributed Stochastic Neighbor Embedding* - t-SNE) foi utilizado para a visualização dos *datasets* do trabalho, desde que esta técnica é empregada na visualização de dados de alta dimensionalidade, conseguindo capturar a estrutura local dos dados, enquanto revela a estrutura global, como a presença de clusters em várias escalas [Van der Maaten and Hinton 2008].

Nas Figuras 2 (a) e 2 (b) são observadas as projeções dos dados na extensão dos componentes principais. Nota-se a variância na densidade dos clusters de acordo com a variância existente nas classes. Os pontos dentro dos clusters individuais são muito

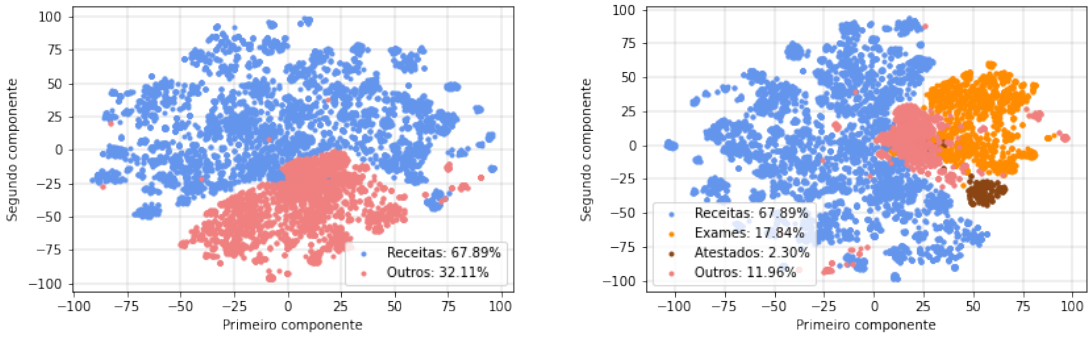

Figura 2. (a) t-SNE do *dataset* com duas classes. (b) t-SNE do *dataset* multi-classe.

semelhantes entre si e distantes dos pontos de outros clusters, o que faz com que seja mais evidente essa separação nas figuras. Além disso, nota-se a existência de alguns valores discrepantes (*outliers*) na classe *Outros* da Figura 2 (b).

Na Tabela 4, todos os resultados mostrados foram obtidos usando k-fold igual a 10, ou seja, esses valores são suas respectivas médias e desvios-padrão. Observa-se grande similaridade nos resultados da classificação binária. Os classificadores SVM, RF e MLP apresentam valores idênticos na acurácia, no kappa e na AUC, enquanto o XBoost apresenta os menores resultados na tabela. Porém, a diferença dos valores das métricas é baixa, assim como a variação do desvio-padrão, o que implica que os classificadores conseguem obter excelente performance na identificação dos padrões encontrados na base de dados.

Tabela 4. Resultados obtidos nas classificações binária e multiclasse.

Classificador	Acurácia	Kappa	F1-Score	AUC
		Binária		
Support Vector Machine	0,999±0,001	0,998±0,002	0,999±0,001	0,999±0,001
Random Forest	0,999±0,001	0,998±0,002	0,998±0,002	0,999±0,001
XGBoost	0,998±0,001	0,996±0,003	0,998±0,002	0,998±0,002
Multilayer Perceptron	0,999±0,001	0,998±0,002	0,999±0,001	0,999±0,001
		Multiclasse		
Support Vector Machine	0,993±0,003	0,985±0,006	0,977±0,008	0,985±0,006
Random Forest	0,992±0,003	0,983±0,006	0,977±0,007	0,986±0,005
XGBoost	0,992±0,002	0,983±0,005	0,982±0,006	0,990±0,004
Multilayer Perceptron	0,994±0,002	0,988±0,005	0,982±0,007	0,988±0,004

Na classificação multiclasse existe uma variação maior nos valores dos resultados. Na acurácia e no kappa, o classificador MLP obteve os melhores resultados, enquanto a XGB alcançou maior performance na F1 e AUC. Neste *dataset*, nota-se também uma alteração maior nos valores do desvio-padrão, que chega a 0.008 no F1 da SVM. Porém, os resultados confirmam que os classificadores ainda apresentam alta efetividade

na predição das classes, mesmo com o aumento do número de classes e o desbalanceamento encontrados nelas.

Esses resultados são explicados pela homogeneidade encontrada nos formatos de amostra de cada classe. Devido à estruturação particular já amplamente utilizada na criação de cada categoria de texto clínico, as amostras têm diferenças distintas entre si. Além disso, com as associações das palavras obtidas através do uso do bigrama, que mantém a sequência dos tokens, a representação vetorial das amostras por classe é bem característica, o que faz com que os classificadores não encontrem dificuldades na predição das classes.

5. Conclusão

Neste artigo, a classificação binária e multiclasse de textos médicos foi realizada utilizando as técnicas de representação de texto n-grama e TF-IDF associadas à algoritmos de classificação amplamente utilizados na literatura, incluindo a rede neural MLP. No primeiro experimento, a base de dados adotada foi rotulada em duas classes: Receitas e Outros. Esse experimento alcançou resultados expressivos, com os modelos SVM e MLP apresentando os melhores desempenhos. No segundo experimento, a base de dados foi rotulada nas classes Receitas, Exames, Atestados/Declarações e Outros e também alcançou resultados significativos, com o XGB e a MLP superando a performance dos outros classificadores.

Assim, as características especifícas de cada classe de documento médico promovem *clusters* bem definidos e, mesmo no classificação multiclasse, onde há uma maior superposição dos *clusters*, modelos como a MLP conseguem obter resultados efetivos na detecção das diferentes classes de textos médicos. Como trabalhos futuros, temos a classificação semisupervisionada de um novo *dataset* não rotulado e a extração de informações, a partir de algumas classes, como posologia, nome do medicamento, nome do exame, etc.

Referências

[Assale et al. 2019] Assale, M., Dui, L. G., Cina, A., Seveso, A., and Cabitza, F. (2019). The revival of the notes field: Leveraging the unstructured content in electronic health records. *Frontiers in Medicine*, 0:66.

[Baratloo et al. 2015] Baratloo, A., Hosseini, M., Negida, A., and El Ashal, G. (2015). Evidence based emergency medicine; part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emergency*, 3:48–49.

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine Learning 2001 45:1*, 45:5–32.

[Brown et al. 1992] Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

[Burges 1998] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

[Cabitza et al. 2019] Cabitza, F., Locoro, A., Alderighi, C., Rasoini, R., Compagnone, D., and Berjano, P. (2019). The elephant in the record: On the multiplicity of data recording work:. *Health Informatics Journal*, 25:475–490.

[Chen and Guestrin 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[Cohen 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[Cui et al. 2019] Cui, M., Bai, R., Lu, Z., Li, X., Aickelin, U., and Ge, P. (2019). Regular expression based medical text classification using constructive heuristic approach. *IEEE Access*, 7:147892–147904.

[Gardner and Dorling 1998] Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

[Hanley and McNeil 1982] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

[Hornik et al. 1989] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366.

[Johnson et al. 2016] Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

[Lee et al. 2011] Lee, J., Scott, D. J., Villarroel, M., Clifford, G. D., Saeed, M., and Mark, R. G. (2011). Open-access mimic-ii database for intensive care research. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 8315–8318. IEEE.

[Lins and Ludermir 2005] Lins, A. and Ludermir, T. B. (2005). Hybrid optimization algorithm for the definition of mlp neural network architectures and weights. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, pages 6–pp. IEEE.

[Liu et al. 2020] Liu, J., Bai, R., Lu, Z., Ge, P., Aickelin, U., and Liu, D. (2020). Data-driven regular expressions evolution for medical text classification using genetic programming. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

[Mountrakis et al. 2011] Mountrakis, G., Im, J., and Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259.

[Murphy and Chueh 2002] Murphy, S. N. and Chueh, H. C. (2002). A security architecture for query tools used to access large biomedical databases. In *Proceedings of the AMIA Symposium*, page 552. American Medical Informatics Association.

[Ogunleye and Wang 2019] Ogunleye, A. and Wang, Q.-G. (2019). Xgboost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6):2131–2140.

[Ohno-Machado et al. 2012] Ohno-Machado, L., Bafna, V., Boxwala, A. A., Chapman, B. E., Chapman, W. W., Chaudhuri, K., Day, M. E., Farcas, C., Heintzman, N. D., Jiang, X., et al. (2012). idash: integrating data for analysis, anonymization, and sharing. *Journal of the American Medical Informatics Association*, 19(2):196–201.

[Reys et al. 2020] Reys, A. D., Silva, D., Severo, D., Pedro, S., e Sá, M. M. d. S., and Salgado, G. A. (2020). Predicting multiple icd-10 codes from brazilian-portuguese clinical notes. In *Brazilian Conference on Intelligent Systems*, pages 566–580. Springer.

[Sebastiani 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

[Swain and Hauska 1977] Swain, P. H. and Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147.

[Tayefi et al. 2021] Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., and Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1549.

[Van der Maaten and Hinton 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[Wulff et al. 2020] Wulff, A., Mast, M., Hassler, M., Montag, S., Marschollek, M., and Jack, T. (2020). Designing an openehr-based pipeline for extracting and standardizing unstructured clinical data using natural language processing. *Methods of Information in Medicine*, 59:e64–e78.

[Yun-tao et al. 2005] Yun-tao, Z., Ling, G., and Yong-cheng, W. (2005). An improved tf-idf approach for text classification. *Journal of Zhejiang University-SCIENCE A 2005 6:1*, 6:49–55.