UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Setorização automática no atendimento de telecomunicações para aprimorar a experiência do cliente

Radakian Maurity Sousa Lino

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Setorização automática no atendimento de telecomunicações para aprimorar a experiência do cliente

USP - São Carlos
2023

Radakian Maurity Sousa Lino

Setorização automática no atendimento de telecomunicações para aprimorar a experiência do cliente

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Prof. Dr. Tiago A. Almeida

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Lino, Radakian Maurity Sousa
L735s Setorização automática no atendimento de telecomunicações para aprimorar a experiência do cliente / Radakian Maurity Sousa Lino; orientador Tiago A Almeida. -- São Carlos, 2023.
50 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. . I. Almeida, Tiago A, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

Este trabalho de conclusão de curso é dedicado a todos que me apoiaram neste desafio. Aos professores por suas orientações, paciência e compartilhamento de conhecimento. E em especial, a minha amada esposa Daniely, apoiando, incentivando e entendendo os momentos de ausência para estudos e trabalhos.

A minha esposa Daniely pela compreensão, incentivo e otimismo incansável.

AGRADECIMENTOS

Agradeço antes de tudo a Deus por sempre me conceber saúde e disciplina, para seguir em frente nas adversidades da vida, combinando família, trabalho e estudos, sempre me proporcionando oportunidades e desafios que me instigam e motivam para seguir em frente.

A minha família, minha esposa Daniely por seu otimismo e incentivos inabaláveis, e meus filhos Letícia e Heitor, pela compreensão, amor e carinho que sempre me dedicaram.

Ao meu orientador, Prof. Dr. Tiago A. Almeida (UFSCar), que me recebeu como aluno orientando, em tempos de pós-pandemia, trabalhos remotos, se ajustando a agendas difíceis e sempre prestativo a me oferecer conhecimento e orientação.

EPÍGRAFE

"As questões mais importantes da vida de fato, são, na maior parte, apenas problemas de probabilidade."

Pierre-Simon Laplace (1749-1827)

RESUMO

Lino, Radakian Maurity Sousa. **Setorização automática no atendimento de telecomunicações para aprimorar a experiência do cliente**. 2023. 52 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Desde a privatização do setor de telecomunicações no Brasil, em 1998, a demanda por serviços em telefonia fixa, celulares, TV por assinatura e principalmente banda larga, crescem regularmente, gerando uma série de desafios, como expansão de redes, universalização de serviços e atendimento ao cliente. Este trabalho apresenta um estudo de caso, no atendimento ao cliente do setor de Telecom, em operadora de telefonia móvel. Ele trata a classificação automática de textos recebidos por meios de pesquisas juntos aos clientes, onde os mesmos indicam suas necessidades e se o problema foi resolvido ou não. A título de exemplo as necessidades dos clientes são relacionadas a serviços que conhecemos em nosso dia a dia, como solicitações de mudança de endereços, ajustes em faturas e/ou redução de valores, informações diversas, suporte técnico ou mesmo pedidos de cancelamento. Por meio desta classificação, a empresa poderá sistematizar o direcionamento da solução para uma área especifica de tratamento, visando reduzir o tempo de resposta, melhorar processos internos e aprimorar a experiência do cliente. O problema em questão foi abordado como um cenário de classificação multirrótulo, pois uma mesma reclamação do cliente pode ter mais de um motivo associado. Neste contexto, é proposto que a setorização para tratamento da necessidade do cliente ocorra de forma automática após a análise do texto recebido. Neste contexto, foram coletados, tratados e rotulados dados reais e, para geração dos modelos, foi utilizada uma abordagem combinando as transformações *Binary Relevance* (BR) e *Label Powerset* (LP) com os métodos Regressão Logística (RL) e Máquinas de Vetores de Suporte (SVM).

Palavras-chave: Classificação multirrótulo, Aprendizado de máquina, Classificação de Textos; Atendimento ao Cliente, Experiência do Cliente; Telecomunicações.

ABSTRACT

Lino, Radakian Maurity Sousa. **Automatic sectorization in telecommunications service to improve the customer experience**. 2023. 52 f. Completion of course work (MBA in Artificial Intelligence and Big Data) – Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, 2023.

Since the privatization of the telecommunications sector in Brazil, in 1998, the demand for services in fixed lines, mobile phones, pay TV and broadband, grows regularly, generating a series of challenges, such as network expansion, universalization of services and customer service. This work presents a case study, in customer service in the Telecom sector, in a mobile telephone operator. It handles the automatic classification of texts received through customer surveys, where they indicate their needs and whether the problem has been resolved or not. For example, customer needs are related to services that we know in our daily lives, such as requests to change addresses, adjustments to invoices and/or reductions in values, various information, technical support or even cancellation requests. Through this classification, the company will be able to systematize the targeting of the solution to a specific treatment area, aiming to reduce response time, improve internal processes and improve the customer experience. The problem in question was approached as a multi-label classification scenario, as the same customer complaint can have more than one reason associated with it. In this context, it is proposed that the sectorization to address the customer's needs occurs automatically after analyzing the text received. In this context, real data were collected, treated and labeled and, to generate the models, an approach was used combining the Binary Relevance (BR) and Label Powerset (LP) transformations with the Logistic Regression (RL) and Support Vector Machines methods. (SVM).

Keywords: Multilabel classification, Machine learning, Text Classification; Customer Service, Customer Experience; Telecommunications.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BR	*Binary Relevance*
LP	*Label Powerset*
M.NB	*Multilabel Naïve* Bayes
NB	*Naïve* Bayes
PW	*Pairwise*
RF	*Random Forest*
RL	*Regressão Logistica*
SVM	*Support Vector Machines*
TF	*Term Frequency*
TF-IDF	*Term Frequency-Inverse Document Frequency*

SUMÁRIO

1 INTRODUÇÃO

Em meados dos anos 50, a telefonia no Brasil era composta essencialmente de telefonia fixa, operada por intermédio de telefonistas e mais de mil companhias. Entretanto, as concessões eram desordenadas, complicando atividades operacionais e também as evoluções dos padrões técnicos das telecomunicações (TELECO, 2023). Ao longo de décadas, o mercado de Telecom foi se consolidando, mas muitas dificuldades ainda existiam, impedindo que o setor pudesse de fato evoluir na velocidade que a sociedade Brasileira precisava. Christiano (1998, p.23) aponta um conjunto de fatores que justificavam este cenário:

- A diluição da competência entre União, os estados e municípios;
- Interesses partidários em detrimento das questões técnicas;
- Desinteresses das empresas privadas frente o cenário inflacionário pós Segunda Guerra Mundial;
- Carência de mão de obra especializada no país.

As primeiras tentativas de organização do setor tiveram início em 1962, com a instituição do Código Brasileiro de Telecomunicações (CBT). Em seguida, no ano de 1965, surge a Empresa Brasileira de Telecomunicações (Embratel), destinada essencialmente a explorar comercialmente as telecomunicações do país (CHRISTIANO,1998, p.26). A instituição de ambas as organizações marca a estatização do setor no Brasil. Em 1972, ocorreu outra mudança importante, a criação da Telebrás, incorporando as empresas existentes e gerando uma grande empresa de economia mista (TELECO, 2023). O regime administrativo centralizado da Telebrás e suas subsidiárias estaduais, se estendeu por toda década de 70 até os anos 90 (MELLO, 2010, p.10). Tais subsidiárias ficam então conhecidas como “teles”.

A década de 90 foi de grandes transformações e uma tendência neoliberal. As propostas de privatizações ganhavam força e eram bem aceitas nas relações político-sociais. Essa tendência veio de encontro ao mercado de Telecomunicações Brasileiro, onde o governo tinha um setor ultrapassado e que precisava de muitos investimentos, sendo inevitável abrir este mercado ao capital privado (QUEIROZ NETO, 2008, p.15). Faltavam meios legais para iniciar este processo sendo que a pressão interna e externa para abertura do mercado culminou com a Lei Geral de Telecomunicações (LGT). Tal lei foi responsável pela quebra de paradigma do monopólio estatal nas Telecomunicações, liberação do mercado e a criação da Agência Nacional de Telecomunicações (ANATEL). A ANATEL passou a atuar no país

como agência regulatória que norteia como as empresas devem agir no ramo de Telecomunicações (PENNA FILHO, 2009, p.194).

Com um contexto político, econômico e social favorável, o setor de telecomunicações no Brasil passou por uma importante mudança. Até 1998, as telecomunicações eram controladas pelo Estado, com serviços de baixa qualidade. Diante do aumento da demanda e da falta de capacidade para expandir as linhas fixas e móveis, além do acesso limitado e custoso à Internet, houve a abertura para o capital privado em 29 de julho de 1998, por meio da privatização das empresas estatais. A partir dessa data, ocorreram diversas transformações significativas.

No ano de 2003, o setor de telecomunicações no Brasil já havia passado pela fase de privatização, resultando em um ambiente de negócios dinâmico, com a presença de diversos concorrentes no mercado. As empresas concentravam seus esforços na consolidação dos serviços, investindo em qualidade para enfrentar a concorrência e expandindo suas redes fixas e móveis para atender à demanda crescente de clientes. Nesse período, surgiram avanços tecnológicos, como a tecnologia GSM e o lançamento da banda larga conhecida como ADSL. No entanto, o atendimento ao cliente ainda seguia uma abordagem tradicional, com o uso de URA's (Unidades de Resposta Audível), *Call Centers* e pontos de atendimento presenciais, como lojas próprias.

Após cinco anos, em 2008, podia ser percebida uma nova configuração no setor de telecomunicações. As empresas estavam em um momento de consolidação do mercado, resultando em uma série de fusões corporativas. Enquanto as linhas fixas estavam em um cenário de “mercado maduro e estável”, ocorria o crescimento dos serviços móveis e de Internet, incluindo o surgimento dos pacotes de serviços. Nessa época, ocorreu também o lançamento do 3G e a expansão dos serviços de TV (TV a Cabo e Digital).

Em 2011, o mercado estava em alto nível de competição, vivendo a maturidade dos serviços móveis. Havia a tendência de possuir mais de um chip por assinante e os serviços de TV e Banda Larga eram os impulsionadores de crescimento do setor. O 3G se consolidava como uma nova tecnologia, enquanto surgia o Programa Nacional de Acesso à Internet para todas as classes sociais.

Entre 2019 e 2021, os *smartphones* se tornaram muito populares, impulsionados por planos de vendas que combinavam a compra do aparelho com a oferta de linhas telefônicas. Entre 2022 e 2023, o grande destaque foi o início da comercialização das operações 5G, com o mercado de telecomunicações buscando novas receitas e a redução de custos, enquanto enfrenta a pressão de expandir suas redes de dados móveis e banda larga. Esse efeito,

provocado principalmente pelo alto consumo de dados, é o resultado da combinação de vários fatores, como mais pessoas com acesso a telecomunicações, a IoT *(Internet Of Things*), a transmissão de conteúdos audiovisuais por *streaming* e os novos meios de comunicação de voz baseados em dados.

Neste cenário de desafios e competição acirrada, aliado a um mercado que necessita de investimentos massivos, os dados são o principal ativo. Perceber rapidamente a necessidade do cliente, decifrar o comportamento do consumidor, produzir ofertas e campanhas baseadas em hiper personalização e, principalmente, entender a "voz do cliente" traduzida em seus *feedbacks* são o novo caminho para uma indústria que vai se afastando do tradicional *modus operandi* de telecomunicações. Com o propósito de se tornarem empresas de tecnologia, temas como Inteligência Artificial, Aprendizado de Máquina e Estatística Multivariada passaram a ser fundamentais para transformar os dados em informações estratégicas e novos negócios.

Diversas áreas das empresas de telecomunicações estão passando por processo de automatização, em alguns casos também chamados de digitalização. Por exemplo, os departamentos de atendimento ao cliente podem ser amparados pelo uso de técnicas de aprendizado de máquina e processamento de língua natural para darem agilidade na tomada de decisões visando a otimização na resolução de problemas e a maximização da satisfação do cliente. Esses departamentos precisam lidar com grandes volumes de documentos sobre assuntos de problemas variados.

1.1 A estrutura do atendimento ao cliente

Atualmente, estamos vivenciando rápidas mudanças no comportamento do consumidor, impulsionadas pelo fácil acesso às informações, especialmente quando se trata da escolha de produtos, serviços e comparação entre empresas e profissionais de diferentes setores. Diante deste cenário, surge um questionamento importante: como as empresas podem buscar melhores margens? De fato, não existe uma resposta única para este desafio, mas certamente alguns pilares são aceitos, e um deles é oferecer a melhor experiência possível para o cliente.

Experiência do Cliente, do inglês, *Customer Experience* (CX), é o conjunto de sentimentos, percepções e impressões que o cliente forma sobre uma determinada empresa (SALESFORCE, 2023). Após interagir com a empresa, podendo ter consumido ou não seus produtos ou serviços, o cliente constrói uma impressão sobre a empresa e classifica, ainda que intuitivamente, como foi sua experiência. Muitos contatos não necessariamente são convertidos em vendas, mas sempre formam uma impressão.

Uma boa CX para o cliente, de forma agradável, simples, rápida e ótima para a empresa têm ganhado cada vez mais espaço em empresas de diferentes tamanhos e setores de atividade econômica. Ela faz parte do planejamento estratégico, plano de investimentos, pesquisa, melhorias internas, revisões de processos, lançamento ou melhoria de produtos e serviços, e demanda muita análise de dados para entendimento de comportamento do consumidor, segmentação e personalização.

Implementar jornadas de vendas ou pós-vendas que gerem uma CX positiva, demanda uma boa estratégia e estrutura de atendimento ao cliente. Ele pode ser entendido como o conjunto de ações para esclarecimento e suporte que o cliente precisar, seja, antes, depois ou durante uma interação que resulte em compra ou não. Estas interações podem ocorrem por diferentes formas em empresas de telecomunicações, tais como:

1 – Atendimento presencial, representado por lojas próprias e parceiros autorizados, onde o cliente pode buscar apoio em problemas, esclarecer dúvidas e conhecer produtos, bem como fazer novas compras.

2 – Telefone, este é o atendimento mais tradicional e conhecido das empresas de Telecomunicações, combinando URA's e Centrais de atendimento humano, comumente chamados de *call centers*. As URA's, segundo Madruga, R (2009, p. 114) *são as responsáveis pela identificação do cliente por meio de chave primária de acesso e pelo roteamento da ligação para atendimento humano*.

3 – *Canais baseados em textos*, como e-mails e ouvidorias, oferecem comunicação direta para problemas específicos e são de fácil acesso, mas podem perder em velocidade em casos específicos que dependam de ação humana. Estes precisam ter procedimentos muito claros e eficientes em treinamento de equipes, pois perdem a possibilidade de contato humanizado, mas oferecem oportunidade de automatização de respostas.

4 – *Canais baseados em Inteligência Artificial*. As tecnologias voltadas ao atendimento ao cliente evoluem constantemente, mas certamente IA é a que tem recebido maior destaque com os avanços na capacidade de extrair de textos fornecidos pelo cliente, os principais motivos de contato, dificuldades e necessidades de consumo.

As estruturas de atendimento em grandes empresas são compostas por um grande número de células nas centrais de atendimento, tanto em atendimento por voz ou nos

chamados *backoffices*. A representação da estrutura de atendimento é exemplificada na Tabela 1, que descreve brevemente os motivos de contato.

Tabela 1 –Principais motivos de contato no atendimento ao cliente de telefonia móvel

Tipo de serviço	Célula de atendimento	Motivos de Contato
Controle	Serviço	Obter informações sobre recarga Resolver problemas com saldo ou recarga
Pós-pago	Faturas	Contestar cobrança indevida Solicitar 2ª via de fatura Negociar pagamento/acordos Resolver problemas sobre o pagamento da fatura Resolver problemas com ativação ou habilitação do plano Mudança não reconhecida no valor do plano
	Suporte técnico	Resolver problemas de sinal de internet (3G/4G/4.5G) Resolver problemas de sinal das ligações Falar sobre dúvidas ou problemas com o chip Mudança de plano (diminuir, aumentar ou migração)
	Vendas	Contratar ou renovar produtos/serviços Falar sobre portabilidade Aumentar o plano Diminuir o plano Dúvidas sobre compra de smartphones na loja ou site da Operadora
	Retenção/ Cancelamento	Cancelar a linha Cancelar algum produto ou serviço Solicitar bloqueio por perda ou roubo

1.2 O problema

O problema tratado neste trabalho é a classificação automática de textos recebidos pelo departamento de atendimento ao cliente de uma empresa do setor de telecomunicações. Evoluir a atual classificação humana para um método automático poderá permitir o direcionamento mais rápido para as áreas de solução específicas de cada contato.

1.3 Hipótese e objetivos

Este trabalho está fundamentando na hipótese de que é possível realizar a classificação automática de textos, com acurácia suficiente para tomada de decisão, permitindo assim a setorização automática de atendimentos por meio do roteamento baseado em algoritmos de aprendizado de máquina e processamento de linguagem natural (PLN).

A principal contribuição deste trabalho consiste em desenvolver um modelo que seja sustentável em termos de custo computacional e viável para implementação *offline*. Esse modelo tem como objetivo agilizar o tratamento das necessidades dos clientes que enviam mensagens de texto para áreas especializadas de atendimento. Isso resultará em uma melhor experiência para os clientes ao utilizarem os serviços de telefonia móvel da empresa. Além disso, espera-se que essa abordagem possa reduzir os custos operacionais das centrais de atendimento e outros departamentos relacionados ao atendimento ao cliente.

Esta proposta de setorização automática, tem potencial para reduzir o tempo necessário para resolver as solicitações dos clientes. Uma das maneiras de alcançar isso é através da implementação de um modelo de classificação multirrótulo que possua acurácia suficiente que possibilite sua utilização para tomada de decisões. Como parâmetro de comparação para avaliar a acurácia será adotada a probabilidade de categorização de um texto, de forma aleatória em um dos rótulos, essa sugestão de critério comparativo será desenvolvida numericamente nos próximos capítulos.

Esse modelo tem a perspectiva de se tornar um "serviço de TI exposto" interno da empresa no futuro. Ele receberia solicitações de vários canais de atendimento e, com base no texto fornecido, classificaria automaticamente essas solicitações. Isso possibilitaria um direcionamento mais rápido para os setores responsáveis pelo tratamento das demandas ou até mesmo respostas automáticas mais precisas aos clientes. Além disso, essa abordagem também pode melhorar o processo de triagem para equipes especializadas nos *backoffices* e até permitir a detecção mais inteligente de problemas na sua origem. Isso seria alcançado ao identificar alterações nos padrões históricos das razões pelas quais os clientes entram em contato.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são introduzidos os assuntos essenciais para a compreensão do tema abordado, em especial, os tópicos que envolvem pré-processamento de texto. Na Seção 2.1, é apresentada a **representação computacional de textos**, com uma visão introdutória de linguagem verbal e não-verbal, e na Seção 2.2, é abordado o tema **preparação e pré-processamento de textos**.

2.1 Representação computacional de textos

Os seres humanos estão habituados com diversos tipos de linguagem, classificadas como verbais, não-verbais e mistas. A linguagem verbal, que é a comunicação baseada em escrita ou fala, é a linguagem que precisamos converter para formatos específicos, de forma que os computadores possam entender. Linguagens não-verbais, que trataremos aqui como aquelas onde não se aplica a escrita, (e.g. movimento das mãos, expressões faciais, linguagem do corpo, posturas, gestos, placas de trânsito, obras de artes) não serão objeto da nossa análise, pois justamente não possuem ainda, uma forma direta, de tradução para computadores. Para evitar confusões de entendimento, vale o destaque para Libras (Língua Brasileira de Sinais), que não é uma linguagem nem escrita ou falada, e por isso, classifica-se como linguagem não-verbal.

Com o avanço dos computadores e a consequente redução de custos, os textos, que são produtos das linguagens verbais, passaram a ser armazenados digitalmente. Esse processo trouxe inúmeras vantagens, como a facilidade no armazenamento, acesso, recuperação e distribuição das informações. Esta combinação de fatores de acesso aos computadores e a redução de custos, teve um excepcional efeito no aumento vertiginoso do número de obras das mais diferentes naturezas disponibilizadas na internet. Isso inclui jornais, revistas, artigos científicos, livros, bem como textos oriundos de tarefas cotidianas, como envio de e-mails, documentos e mensagens em redes sociais. Essa produção textual contínua ao longo do tempo, proporcionou a criação de um espaço enorme de aplicações de PLN, com as mais diferentes oportunidades de aplicações.

A quantidade crescente de fontes de dados textuais, que consistem em dados não-estruturados, requer o uso de técnicas automatizadas para a recuperação de informações. Uma forma de auxiliar nessa tarefa é através da categorização de textos, também conhecida como

classificação de textos. Essa abordagem envolve a atribuição de categorias ou rótulos aos documentos textuais, com o objetivo de agrupar documentos que compartilham características semelhantes (FACELI et al., 2021).

A conversão dos documentos de texto em formato computacional pode ser realizada por meio de diferentes técnicas, sendo que Turian et al. (2010) agruparam essas técnicas em três categorias: distributiva, por agrupamento e distribuída.

Representação distributiva: baseada em uma matriz de co-ocorrência. Um exemplo conhecido dessa abordagem é a chamada *bag of words* (saco de palavras).

Representação por agrupamento: fundamentada na utilização de um mapa semântico auto organizável (RITTER E KOHONEN et al., 1989), LSA (LANDAUER et al., 1998) e HAL (LUND E BURGESS et al., 1996). Essa abordagem busca induzir grupos sobre palavras, sendo que um dos trabalhos mais relevantes nesse contexto é o agrupamento de Brown et al. (1992). Essas técnicas geralmente derivam um modelo de linguagem baseado em classes.

Representação distribuída: é uma abordagem que analisa palavras ou conceitos em seus contextos de uso. A ideia principal é que palavras com significados semelhantes ocorrem em contextos semelhantes, ou seja, é capaz de capturar similaridades semânticas ao analisar um grande volume de dados, representando unidades de texto em vetores densos de tamanho fixo (Bengio, 2009).

Sobre utilização de n-gramas, cada termo pode representar uma ou mais palavras. Neste caso, unigrama refere-se a representação por palavra. Bigramas refere-se a representação por pares de palavras e, da mesma maneira, um n-grama refere-se a representação de n palavras do texto.

Seja $X = \{X_1, X_2, ... X_n\}$ um conjunto com *n* documentos de textos.

$T = \{t_1, t_2, ... t_d\}$ um conjunto com d termos (atributos) de um vocabulário pré-definido.

No modelo espaço vetorial, cada documento pode ser representado por uma matriz documento-termo, em que cada posição contém um peso $w(t_i , Xn)$ associado a cada termo (FACELI et al., 2021).

A Tabela 2 ilustra uma representação do modelo espaço-vetorial, com os documentos Xn, os termos Td e os pesos $w(t_i, Xn)$.

Tabela 2 – Representação de documentos usando o modelo espaço-vetorial

Documentos	t_1	t_2	t_3	....	t_d
X_1	$w(t_1, X_1)$	$w(t_2, X_1)$	$w(t_3, X_1)$	....	$w(t_d, X_1)$
X_2	$w(t_1, X_2)$	$w(t_2, X_2)$	$w(t_3, X_2)$	....	$w(t_d, X_2)$
X_3	$w(t_1, X_3)$	$w(t_2, X_3)$	$w(t_3, X_3)$	....	$w(t_d, X_3)$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
X_n	$w(t_1, X_n)$	$w(t_2, X_n)$	$w(t_3, X_n)$	....	$w(t_d, X_n)$

Os valores usados para representar os termos de um documento ($w(t_i, X)$) geralmente são baseados na frequência da ocorrência de t_i em X e podem variar dependendo da estratégia de atribuição de pesos aplicada, tais como:

Binária: representação simples e intuitiva. Caso o termo apareça no documento, recebe o valor "1". Se não aparece, recebe o valor "0". Assim, não é representado pela frequência do termo, mas pela sua ocorrência.

$$w(t_i, X) = \begin{cases} 1, \text{ se } t_i \text{ aparece em X;} \\ 0, \text{ se } t_i \text{ não aparece em X;} \end{cases}$$

Frequência do termo (TF – term frequency): mede a frequência (quantidade de vezes) com que um termo específico aparece em um determinado documento, sendo utilizado como um peso. Diferente do peso binário, o peso TF atribui um número que representa a quantidade de vezes que o termo ocorreu. A ideia é que quanto maior for o seu valor, maior será a relevância ou importância desse termo em relação ao conteúdo do documento. A Figura 1 ilustra um exemplo de vetor de frequência de termos.

Figura 1 - Exemplo de um vetor de frequência de termos

d_1 jazz tem um ritmo de swing

d_2 swing é difícil de explicar

d_3 ritmo de swing é um ritmo natural

	um	explicar	difícil	tem	é	jazz	musica	natural	ritmo	swing	para
d_1	1	0	0	1	0	1	1	0	1	1	0
d_2	0	1	1	0	1	0	0	0	0	1	1
d_3	1	0	0	0	1	0	0	1	2	1	0

Frequência do termo-frequência inversa dos documentos (TF-IDF – term frequency-inverse document frequency): Combina a Frequência do Termo (TF) e a Frequência Inversa do Documento (IDF). O TF já explicado anteriormente, mede a frequência com que um termo específico aparece em um documento. Por sua vez, a Frequência Inversa do Termo (IDF) representa quantitativamente o quão comum ou raro é um termo em todo o conjunto de documentos. A ideia é que termos muito comuns em todos os documentos não fornecem capacidade discriminativa, enquanto termos mais raros podem ter maior poder de diferenciação entre os documentos. O IDF é calculado por:

$IDF\ (t_i) = \log\ (N\ /DF(t_i))$

Onde:

t_i é o termo específico considerado

N é o número total de documentos no conjunto de dados

$DF(t_i)$ é o número de documentos que contêm o termo t.

Para construir o peso do termo (t_i), consideramos que este será igual ao produto da sua frequência pela frequência inversa dos documentos (IDF – inverse document frequency) (SALTON et al., 1975). Um alto valor de TF-IDF é obtido quando o termo tem uma alta frequência no documento que está sendo avaliado e uma baixa frequência no conjunto de dados de treinamento (WILBUR E KIM, 2009; RENNIE et al., 2003). Para calcular o peso TF-IDF de um termo t_i qualquer, pode ser aplicada a seguinte equação:

$$w(t_i\ , \mathrm{X}) = \ TF(t_i\ , \mathrm{X}) IDF_{t_i}$$

$$w(t_i\ , \mathrm{X}) = \ log\ (1 + \ TF(t_i\ , \mathrm{X}) \log(\frac{\mathrm{N} + 1}{DF_{t_i} + 1})$$

$TF(t_i\ , \mathrm{X})$ é a frequência do termo t_i no documento *x*, N é a quantidade de documentos no conjunto de treinamento e DF_{t_i} é o número de documentos de treinamento que contém o termo t_i (WILBUR E KIM, 2009). Cada palavra do texto assume uma posição no vetor, e pesos são atribuídos pelo número de vezes que a mesma ocorre. O conjunto de todos os atributos usados na representação, é chamado de dicionário. Abaixo um exemplo ilustrativo (PROVOST F, FAWCETT T, 2016).

2.2 Preparação e pré-processamento de textos

O pré-processamento de textos é uma etapa essencial em PLN. Nesta fase, os textos selecionados passam por adequações que têm como objetivo "limpar" as bases de dados, tornando a extração de informações relevantes e mais eficiente. A seguir, são listadas as etapas mais comuns.

1 – **Limpeza de texto**: retirada de caracteres indesejados, como pontuação, caracteres especiais, números. Ocorre também a uniformização do texto, levando toda a base para um formato único, como letras minúsculas. Isso evita que duas palavras idênticas sejam consideradas diferentes pelo fato de suas letras estarem escritas em maiúsculo ou minúsculo (UYSAL E GUNAL, 2014).

2 – **Remoção de palavras raras**: remove os termos com baixa frequência no conjunto de documentos. A intuição é que termos que aparecem raras vezes no conjunto de treinamento não contribuem na identificação da categoria do documento e, portanto, podem ser descartados (WEISS et al., 2004).

3 – **Tokenização**: é a divisão do texto em trechos de interesse, como palavras, frases ou outros tokens significativos para o problema em análise. Isto ajuda a identificar a estrutura base e prepara a base para análise.

4 – **Retirada de *Stopwords***: são termos muito comuns, tais como preposições, conjunções e artigos, que fornecem pouca ou nenhuma informação (WEISS et al., 2004). Portanto, eles podem ser removidos antes dos processos de treinamento e categorização. Exemplo: "o", "a", "de", "para" ... .

5 – **Stemming (stematização) e lematização**: reduz as palavras a suas raízes (stemming), ou seja, reduz o termo ao seu radical (Uysal e Gunal, 2014), ou forma básica (lematização), para facilitar a análise. Exemplo: "ventania", "ventoinha" e "ventilador", todas podem ser reduzidas a "vento".

6- **Parts-of-speech (POS) Tagging**: classifica as diferentes classes gramaticais de um texto. Segundo Monica, S (2021), "A marcação de POS resulta em várias tuplas, onde cada uma delas contém a palavra e a sua *tag* que classifica gramaticalmente a palavra como verbo, adjetivo, substantivo, etc"

7 – **NER (Named Entity Recognition)**: realiza a identificação de entidades nomeadas do texto, como nome de pessoas, organizações e locais.

8 – **Desambiguação de significado**: trata das palavras com múltiplos significados. Pode ser uma etapa bastante desafiadora dependendo do contexto em análise. Veja os exemplos abaixo:

“João foi atrás do táxi correndo”

“Ana encontrou o gerente da loja com seu irmão”

“Ele sentou na cadeira e quebrou o braço”

Em todos os casos a interpretação é ambígua e muda de acordo com o referencial adotado.

3 CONTEXTUALIZAÇÃO BIBLIOGRÁFICA

Uma diversidade de métodos tem sido aplicada na categorização de textos, cada um se diferenciando pela estratégia utilizada para se obter a função hipótese. As principais estratégias são descritas a seguir (Silva, 2017; Sebastiani, 2002):

1) Métodos baseados em distâncias: consideram a proximidade entre os documentos para realizar as predições. O método dos k-vizinhos mais próximos (Cover e Hart, 1967) é o mais conhecido. Nesse método, a classificação de um documento desconhecido é feita com base nos rótulos dos *k* documentos mais próximos a ele no espaço de características. A proximidade entre documentos é geralmente medida usando alguma métrica de distância, como a distância euclidiana ou a distância de cosseno.

2) Métodos probabilísticos: se baseiam na probabilidade de o documento pertencer a cada uma das classes possíveis do problema. O Naive Bayes (NB), as redes bayesianas (McCallum e Nigam, 1998) e o MDLText (Silva et al., 2017c; Freitas et al., 2019), estes são exemplos de métodos probabilísticos.

3) Métodos baseados em árvores de decisão: constituídos de métodos que dividem um problema complexo em subproblemas mais simples, sob uma estrutura de árvore. As árvores de classificação e regressão (CART – classification and regression trees) (Breiman et al., 1984) e o C4.5 (Quinlan, 1993) são os métodos mais tradicionais baseados em árvore de decisão.

4) Métodos baseados em otimização: a hipótese é encontrada a partir da otimização de alguma função que avalia a capacidade de predição. As máquinas de vetores de suporte (SVM – support vector machines) (Cortes e Vapnik, 1995), a regressão logística e as redes neurais artificias (Haykin, 1999a) são três técnicas muito empregadas com sucesso em diversas aplicações.

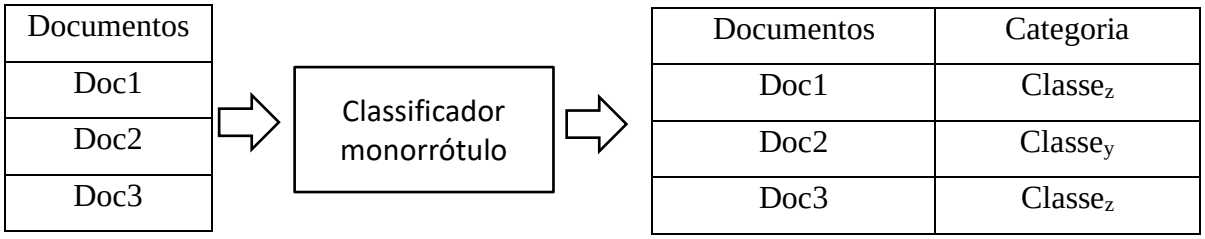
5) Métodos ensemble: treinam diferentes classificadores para a mesma tarefa de classificação e combinam as predições individuais para gerar a predição final (Sebastiani, 2002). *Random Forest* e o AdaBoost (Freund e Schapire, 1996) são exemplos bastante utilizados.

3.1 Categorização monorrótulo

A classificação monorrótulo de textos é caracterizada quando um documento recebe uma única categoria de classificação, ou seja, uma relação de um para um. Isso é conhecido como um problema de classificação de rótulo único ou monorrótulo, onde cada documento é atribuído a apenas uma categoria (CARVALHO; FREITAS, 2009).

Um exemplo típico de problema monorrótulo é o filtro de spam, onde as mensagens de e-mail podem ser rotuladas como spam ou não spam. Quando a classificação envolve mais do que duas categorias (ou seja, não é binária), ela é denominada classificação multiclasse (TSOUMAKAS; KATAKIS, 2007). É importante observar que, nessas situações, as classes são mutuamente exclusivas, o que significa que não podem ocorrer simultaneamente. A Figura 2 apresenta fluxo de classificação monorrótulo.

Figura 2 – Classificação monorrótulo



Cada documento foi classificado para apenas uma classe!

3.2 Categorização multirrótulo

Na classificação multirrótulo, cada documento é associado a uma ou mais classes conforme ilustra a Figura 3. Um desafio da classificação multirrótulo é que as classes não são mutuamente exclusivas.

Figura 3 – Classificação multirrótulo

Q = {A, B, C, D} Base de dados multirrótulo (T)	
D	Y
d_1	{A,C}
d_2	{A,C,D}
d_3	{B,D}
...	...
d_m	{B}

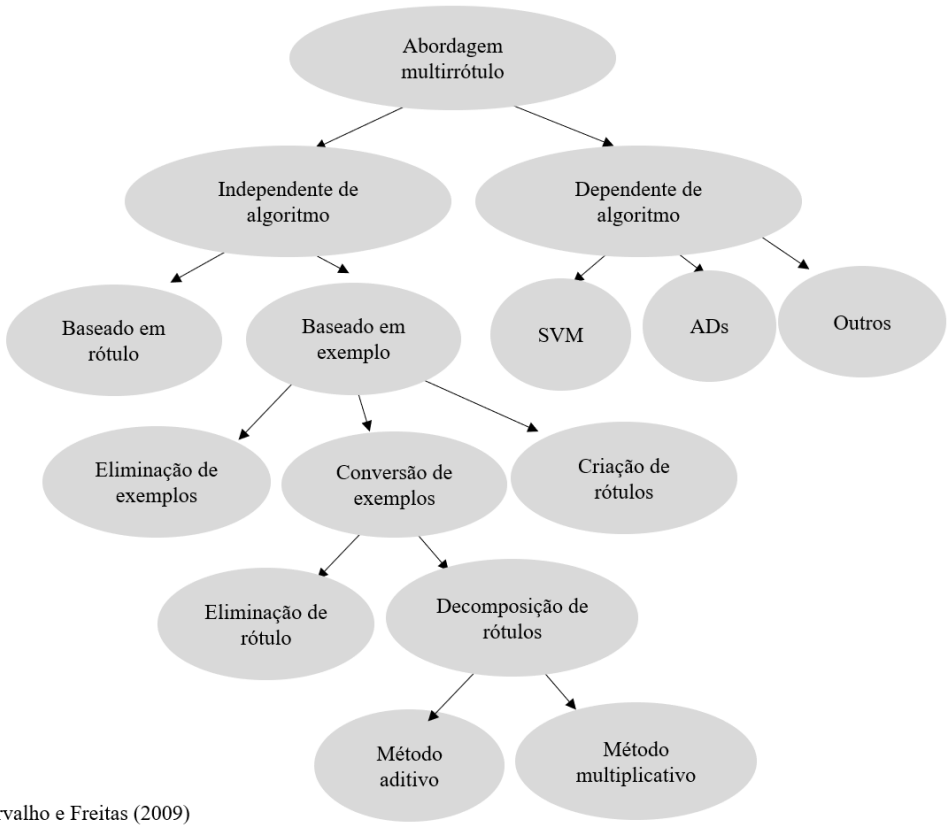
Q = {A, B, C, D} Base de dados multirrótulo (T)				
D	A	B	C	D
d_1	1	0	1	0
d_2	1	0	1	1
d_3	0	1	0	1
...	...	....	...	...
d_m	0	1	0	0

De acordo com Tsoumakas, Katakis e Vlahavas (2010), os métodos de classificação multirrótulo podem ser divididos em dois grupos principais: **transformação de problemas** e **adaptação de algoritmos**.

O primeiro grupo, conhecido como **transformação de problemas**, visa converter problemas multirrótulo em um ou mais problemas monorrótulo. Para isso, existem três abordagens distintas: transformação BR, transformação LP e transformação emparelhada. Em seguida, métodos tradicionais de classificação podem ser aplicados aos problemas transformados. Isso torna os métodos desse grupo independentes de um único algoritmo como base, permitindo que, tanto abordagens binárias quanto multiclasse, sejam testadas e utilizadas.

O segundo grupo, conhecido como **adaptação de algoritmos**, consiste em métodos de classificação monorrótulo ajustados para lidar diretamente com dados multirrótulo, sem a necessidade de aplicar transformações nos dados originais. Essas abordagens têm a vantagem de trabalhar com as características intrínsecas do problema multirrótulo (GIBAJA; VENTURA, 2015; TSOUMAKAS; KATAKIS; VLAHAVAS, 2010; CARVALHO; FREITAS, 2009).A Figura 4 ilustra os dois grupos:

Figura 4 – Abordagens multirrótulo



Fonte: Carvalho e Freitas (2009)

A Figura 4 apresenta as diferentes abordagens que a literatura define para problemas de classificação multirrótulo (Carvalho e Freitas, 2009): **abordagem independente do algoritmo** e **abordagem dependente de algoritmo**. A seguir, são apresentados mais detalhes de cada uma destas abordagens.

Abordagem independente do algoritmo tem como objetivo converter problemas multirrótulo em um conjunto de problemas monorrótulo. Nessa abordagem, qualquer algoritmo de classificação monorrótulo tradicional pode ser considerado como uma alternativa para lidar com problemas multirrótulo.

Uma das formas de fazer isto, é a *transformação baseada nos rótulos***,** onde são utilizados *K* classificadores, sendo *K* o número de classes do problema. Cada classificador é então associado a uma classe e treinado para resolver um problema de classificação binária, na qual é considerada a classe que ele está associado contra todas as outras classes envolvidas. Esse método é também chamado de método binário ou um-contra-todos (Tsoumakas e Vlahavas, 2007). Um ponto fraco desse método é que ele assume que as classes atribuídas a um exemplo são independentes entre si. Isso nem sempre é verdade.

Outra transformação possível, é a chamada, *transformação baseada nos exemplos*, . Nesta, o conjunto de classes associado a cada exemplo é redefinido, de maneira a converter o problema multirrótulo original em um ou mais problemas monorrótulo. Ao contrário do método anterior, esse método não produz apenas problemas de classificação binária, podendo também produzir problemas multiclasse (Carvalho e Freitas, 2009).

Abordagem dependente de algoritmo – objetivo é criar algoritmos específicos para problemas multirrótulo, mesmo que estes algoritmos sejam baseados em métodos de classificação tradicionais, como SVM ou árvores de decisão. Em resumo, novos algoritmos são construídos para tratar a classificação multirrótulo como um problema único, um todo indivisível, ou seja, em uma etapa única.

Um método de classificação baseado em árvores de decisão, chamado de "Árvore de Decisão Alternada" (ADT), foi proposto por Freund e Mason (1999). Trata-se de uma generalização das árvores de decisão, sendo que seu princípio indutivo é baseado no método *boosting* (Freund e Schapire, 1999).

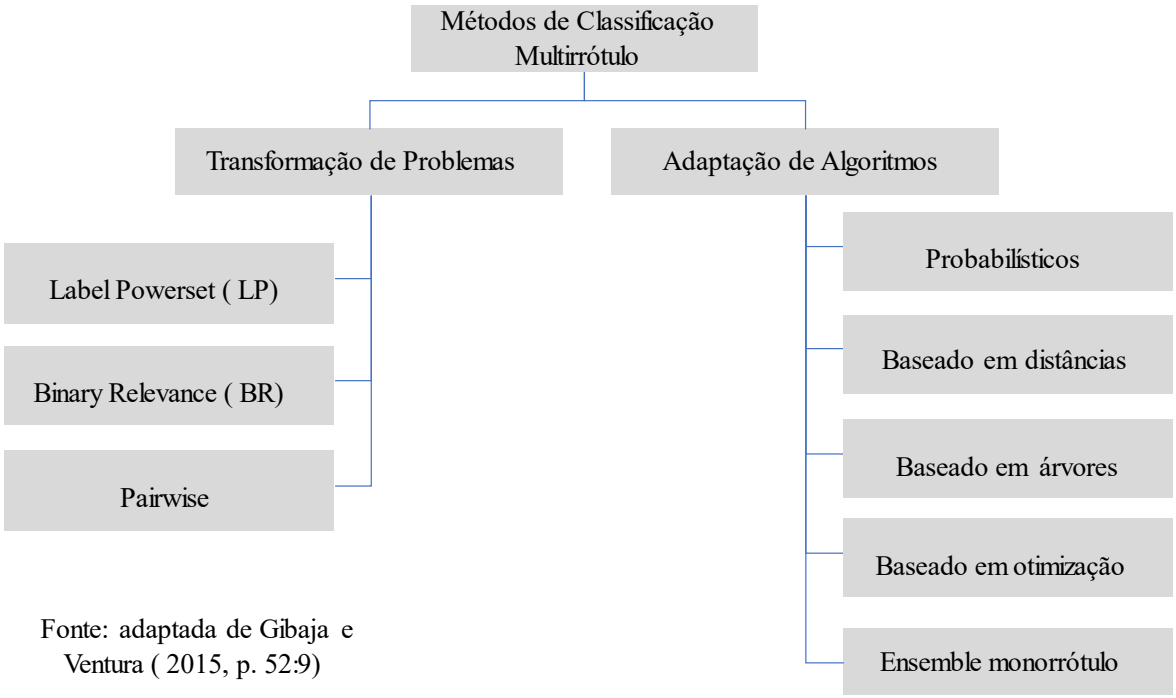
Uma extensão do método ADT foi proposta por de Comite et al. (2003) e é baseada nos métodos *AdaBoost* (Freund e Schapire, 1995) e *ADTBoost* (Freund e Mason, 1999). Esse algoritmo estende o ADT pela decomposição de problemas multiclasse usando a abordagem

um-contra-todos (Carvalho e Freitas, 2009). Em Zhang e Zhou (2005), é proposto um método para classificação multirrótulo baseado no algoritmo *K*-NN, chamado ML-*K*-NN. Nesse método, para cada exemplo, as classes associadas aos *K* exemplos vizinhos mais próximos são recuperados, e é feita uma contagem dos vizinhos associados a cada classe.

O princípio *maximum a posteriori* (Saridis, 1983) é utilizado para definir o conjunto de classes de um novo exemplo. Em Bittencourt et al. (2019; 2020), é apresentado um método denominado ML-MDLText. Trata-se de uma adaptação do método de categorização de texto multiclasse MDLText (Silva et al., 2017b) para lidar com o problema multirrótulo.

A vantagem desse método com relação às alternativas existentes, é que ele não requer transformação do problema multirrótulo, e suporta naturalmente o aprendizado incremental. Essas características permitem que o método possa ser aplicado em problemas dinâmicos e de grande porte. A Figura 5 ilustra os dois grupos de classificação multirrótulo; **transformação de problemas** e **adaptação de algoritmos**.

Figura 5 – Métodos de classificação multirrótulo



Fonte: adaptada de Gibaja e Ventura (2015, p. 52:9)

Transformação de problemas – A seguir, são apresentadas as principais características dos métodos Label Powerset, Binary Relevance e Pairwise.

No método de transformação *Label Powerset (LP)*, cada subconjunto de rótulos das amostras de treinamento é tratado como se fosse uma classe independente. Esta técnica LP,

transforma os diferentes *labelsets* das amostras de treinamento em classes únicas. Depois dessa transformação, um classificador multiclasse é treinado com este conjunto e é utilizado para predizer o *labelset* mais provável. Em situações nas quais existem muitos *labelsets* distintos e um alto desbalanceamento entre eles, o desempenho do método LP é prejudicado (TSOUMAKAS; KATAKIS; VLAHAVAS, 2010; GIBAJA; VENTURA, 2015). A Figura 6 apresenta a aplicação da Transformação LP, para o conjunto de dados D, com classificação multirrótulo.

Figura 6 –Transformação LP aplicada no conjunto de dados da Figura 3.

Q = {AC,ACD,BD,...,B} Base de dados multirrótulo (T)	
D	Y
d_1	{AC}
d_2	{ACD}
d_3	{BD}
...	...
d_m	{B}

O método *Binary Relevance (BR)*, entende o problema de classificação multirrótulo como um conjunto de problemas de classificação binária (BOUTELL at al., 2004). Cada classificador define se o seu rótulo é relevante ou não para o problema analisado e, dessa forma, a predição final é obtida através da combinação das predições individuais desses classificadores (TSOUMAKAS; KATAKIS; VLAHAVAS, 2010; GIBAJA; VENTURA, 2015). A Figura 7 apresenta a aplicação da Transformação BR, para o conjunto de dados D, com classificação multirrótulo.

Figura 7 – Transformação BR aplicada no conjunto de dados da Figura 3

Base de dados A	
D	*Y*
d_1	*A*
d_2	*A*
d_3	*~A*
...	...
d_m	*~A*

Base de dados B	
D	*Y*
d_1	*~B*
d_2	*~B*
d_3	*B*
...	...
d_m	*B*

Base de dados C	
D	Y
d_1	*C*
d_2	*C*
d_3	*~C*
...	...
d_m	*~C*

Base de dados D	
D	*Y*
d_1	*~D*
d_2	*D*
d_3	*D*
...	...
d_m	*~D*

O método *Pairwise (PW),* na decomposição do tipo todos-contra-todos, também denominada um-contra-um (OAO, do inglês *one-against-one*) e em pares (*pairwise*), dadas *k* classes, (K (K-1)) /2 classificadores binários são gerados. Da mesma forma que nos métodos de transformação BR, um classificador binário é empregado para resolver cada um desses subproblemas binários. Na classificação, os rótulos que mais apareceram nas saídas dos classificadores binários (selecionados através da aplicação de um limiar na frequência de ocorrências) são escolhidos para a predição.

Adaptação de algoritmos

Esse método consiste em ajustar algoritmos tradicionais, desenvolvidos para classificação monorrótulo, para que possam tratar problemas de classificação multirrótulo. A essência é a transformação do problema multirrótulo para vários problemas de classificação binária, cada um buscando classificação de um rótulo específico. Dessa forma os algoritmos tradicionais podem tratar cada parte do problema, como subconjuntos específicos de classificação. Isto facilita a adaptação para tarefas mais complexas com muitos rótulos.

3.3 Medidas de desempenho

No contexto de problemas multirrótulo, é natural testar diferentes algoritmos, e a seleção do mais adequado requer um processo de comparação de resultados. Nessa comparação, são utilizadas diferentes métricas de avaliação, e é importante ressaltar que problemas multirrótulo demandam medidas distintas das utilizadas em problemas monorrótulo.

A seguir, apresentaremos as principais medidas de avaliação utilizadas para problemas multirrótulo. Essas métricas são essenciais para compreender a eficácia dos algoritmos de classificação e sua capacidade de lidar com as complexidades inerentes aos dados multirrótulo.

Como múltiplos rótulos podem ser atribuídos à um único documento, então, uma predição pode estar parcialmente correta, totalmente correta ou totalmente errada (Gibaja; Ventura, 2015), o que torna a avaliação dos classificadores multirrótulo um desafio mais complexo. Outra questão relevante, é que os conjuntos de dados não são todos igualmente multirrotulados.

Em alguns casos, o número de classes de cada exemplo pode ser pequeno se comparado ao número total de exemplos, enquanto em outros, pode ser grande. Esse úmero pode ser um parâmetro que influencia o desempenho dos diferentes métodos de classificação multirrótulo

(Tsoumakas e Katakis, 2007). Para representar essa variação, duas métricas são essenciais, cardinalidade e densidade. Sendo **X** um conjunto de dados multirrótulo com *n* exemplos ($\mathbf{x}_i$, $\mathbf{y}_i$), com i =1,2,...,*n*, a cardinalidade e densidade de **X** são definidas nas equações 3.1 e 3.2 respectivamente:

Cardinalidade é dada pelo número médio de rótulos dos exemplos de **X**:

$$CR(X) = \frac{1}{n}\sum_{i=1}^{n}(y_i) \tag{3.1}$$

Densidade é dada pelo número médio de rótulos dos exemplos de **X** dividido por *k*, o número total de classes:

$$DR(X) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i)}{k} \tag{3.2}$$

Nas equações, *sum*($\mathbf{y}_i$) é o número de rótulos do objeto $\mathbf{x}_i$. A cardinalidade de rótulo é independente do número de possíveis classes (*k*) e é utilizada para quantificar o número de rótulos alternativos que caracterizam os exemplos de um conjunto de dados multirrótulo. A densidade de rótulo, por sua vez, leva em conta o número de possíveis rótulos.

Dois conjuntos de dados com a mesma cardinalidade de rótulo, mas com uma grande diferença no número de rótulos (diferentes densidades), podem apresentar propriedades diferentes, que podem afetar o desempenho dos algoritmos de classificação multirrótulo. Suponha, por exemplo, dois conjuntos de dados com a mesma cardinalidade de dois rótulos por objeto, mas com diferentes densidades: (1) dois rótulos por objeto dentre quatro possíveis classes e (2) dois rótulos por objeto dentre 40 possíveis classes. O número de possíveis combinações no segundo caso é bem maior que no primeiro. Assim, essas duas métricas são relacionadas umas com a outra: $CR(\mathbf{X}) = k \times DR(\mathbf{X})$ (Tsoumakas e Katakis, 2007). Contudo, cardinalidade e densidade não são suficientes para avaliação de um método sozinhas e, por isso são empregadas medidas específicas para avaliação de classificadores destinados a problemas multirrótulo.

Diferentemente da classificação monorrótulo, em que um exemplo é classificado de maneira errada ou correta, na classificação multirrótulo, um exemplo pode ser classificado de maneira parcialmente errada ou parcialmente correta. Esses casos acontecem quando um

classificador atribui corretamente a um exemplo pelo menos uma das classes a que ele pertence, mas também não atribui ao exemplo uma ou mais classes às quais ele pertence. Pode acontecer também de o classificador atribuir a um exemplo uma ou mais classes às quais ele não pertence.

Segundo Gibaja e Ventura (2015) e Tsoumakas, Katakis e Vlahavas (2010), as medidas que avaliam os classificadores multirrótulo podem ser divididas em dois grupos: *medidas baseadas em rótulos* e *medidas baseadas em exemplos*.

Medidas baseadas em rótulo: são aquelas que avaliam a capacidade de predizer cada rótulo individualmente, através de alguma medição binária, e calculam a média sobre todos os rótulos, para uma estimativa geral. As métricas binárias que podem ser utilizadas são aquelas baseadas nas tabelas de contingência, que levam em conta a presença e a ausência de um rótulo no conjunto verdadeiro e no conjunto predito, como a precisão, revocação, F-score e acurácia (Sokolova e Lapalme, 2009). Já a média pode ser obtida através de duas técnicas populares: a *média macro* e a *média micro*, descritas pelas equações 3.3 e 3.4 respectivamente (TSOUMAKAS; KATAKIS; VLAHAVAS, 2010; GIBAJA; VENTURA, 2015):

$$F_{Macro}(\mathrm{H}) = \frac{1}{|Q|}\sum_{J=1}^{|Q|} Medida\left(TP_{cj}, FP_{cj}, TN_{cj}, FN_{cj}\right) \tag{3.3}$$

$$F_{Micro}(\mathrm{H}) = Medida\left(\sum_{J=1}^{|Q|} TP_{cj}, \sum_{J=1}^{|Q|} fP_{cj}, \sum_{J=1}^{|Q|} TN_{cj}, \sum_{J=1}^{|Q|} FN_{cj}\right) \tag{3.4}$$

Nas equações acima, *TPcj* , *FPcj* , *TNcj* e *FNcj* é o número de verdadeiros e falsos positivos e de verdadeiros e falsos negativos relacionadas as predições do rótulo *cj* da tabela de contingência. *F-score* refere-se a métrica binária empregada. A média *macro* considera pesos iguais para cada rótulo e é influenciada pelo desempenho em classes raras. Já a média *micro* tende a ser dominada pelo desempenho em classes mais frequentes, pois ela considera pesos iguais para cada amostra (GIBAJA; VENTURA, 2015).

Não existe um consenso sobre a melhor medida a ser utilizada na avaliação, porém, o comportamento da média macro pode ser desejado para avaliar a classificação em problemas com rótulos desbalanceados (ALVARES-CHERMAN, 2014). Em ambos os casos, os melhores desempenhos apontam valores próximos de 1.

Medidas baseadas em exemplos: indicam a capacidade de predição em cada amostra e geram uma média sobre todas as amostras. Definindo uma função ranking, para cada exemplo (amostra), o classificador produz um *ranking* de rótulos.

Seja **X** um conjunto de dados multirrótulo com *n* exemplos ($\mathbf{x}_i$, $\mathbf{y}_i$), com $i = 1, 2, ..., n$ e $sum(\mathbf{y}_i) < k$, em que *k* é o conjunto de possíveis classes. Sejam ainda $\hat{f}$ um classificador multirrótulo e $\mathbf{z}_i = \hat{f}(\mathbf{x}_i)$ um vetor binário com *k* elementos representando o conjunto de classes preditas por $\hat{f}$ para um dado exemplo $\mathbf{x}_i$. Uma medida comumente utilizada para realizar uma avaliação baseada na classificação é o *Hamming Loss* (Schapire e Singer, 2000), definida na equação 3.5.

$$Hamming\ Loss(\hat{f}, X) = \frac{1}{n}\sum_{i=1}^{n}\frac{a(y_i, z_i)}{k} \tag{3.5}$$

Esta medida considera predições parcialmente corretas em seu cálculo. Ela analisa a quantidade de vezes que, em média, um par de rótulos é classificado incorretamente. Esta medida leva em consideração tanto os erros de predição quanto as omissões de predição, sendo que valores próximos de zero caracterizam um melhor poder de predição (SCHAPIRE; SINGER, 2000; GIBAJA; VENTURA, 2015).

Existem ainda, métricas já tradicionais para problemas monorrótulo que foram adaptadas para os problemas de classificação multirrótulo. A precisão é a porção de rótulos corretamente classificados dentre os rótulos positivos preditos, representada na equação 3.6. Já a acurácia é a porção de rótulos classificados corretamente dentre todos os rótulos verdadeiros e os preditos da amostra, representada na equação 3.7, enquanto que a *F-medida* é a média harmônica entre a precisão e a revocação, representada na equação 3.8 (GODBOLE; SARAWAGI, 2004; GIBAJA; VENTURA, 2015).

$$Precisão(\hat{f}, X) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i\ AND\ z_i}{sum(z_i)} \tag{3.6}$$

$$Acurácia(\hat{f}, X) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i\ AND\ z_i}{y_i\ OR\ z_i} \tag{3.7}$$

$$Revocação(\hat{f}, X) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \; AND \; z_i}{sum(y_i)} \tag{3.8}$$

Nas equações acima, AND e OR representam as operações lógicas booleanas AND e OR aplicadas a dois vetores binários.

4 MATERIAIS E MÉTODOS

Neste projeto, o estudo de caso se concentra nos clientes de serviços móveis, abrangendo tanto o segmento pré-pago quanto o pós-pago. Periodicamente, são conduzidas pesquisas para avaliar a satisfação desses clientes após interagirem com o *call center*, relatando seus problemas e recebendo uma resposta da empresa.

Para esta análise, foi considerado um conjunto de 2.104 respostas de usuários, coletadas no período de novembro de 2022 a fevereiro de 2023. Durante o processo de pesquisa, os clientes foram questionados sobre sua satisfação com o serviço, se recomendariam a empresa e se suas solicitações foram resolvidas. Neste último aspecto, os clientes tiveram a oportunidade de fornecer respostas abertas, descrevendo a situação e indicando se o problema foi efetivamente resolvido, bem como destacando qualquer área que necessite de aprimoramento.

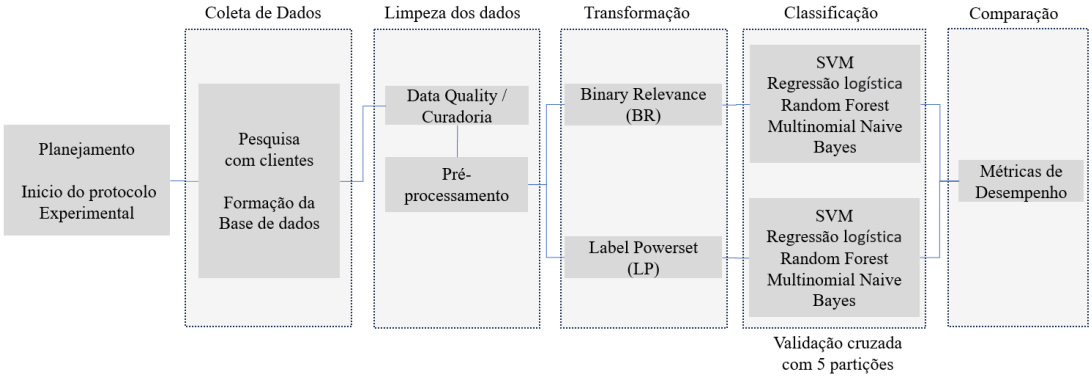
Nas respostas destas pesquisas de satisfação, o cliente também atribui uma nota de 0 a 10 que é traduzida para um *score* chamado de NPS (*Net Promoter Score*). Todas as respostas são reais, com falhas de escrita, abreviações, jargões e nomes específicos de produtos, tornando desafiador a manipulação desses dados.

É importante destacar, que a base de dados utilizada no estudo segue rigorosamente as normas da Lei Geral de Proteção aos Dados (LGPD). A identidade dos clientes foi anonimizada, e quaisquer outros dados considerados sensíveis foram removidos.

4.1 Metodologia

Para melhor entendimento do protocolo experimental, a Figura 8 sumariza o fluxo com as etapas realizadas neste estudo.

Figura 8 – Protocolo experimental



4.2 Bases de dados

Este experimento utilizou uma base de dados, que contem respostas sobre a experiência do cliente, termo conhecido do inglês como *Customer Experience – CX*. As respostas foram obtidas de situações de atendimento reais, onde o cliente utilizou serviços de telecomunicações, fornecendo seu *feedback* por meio de textos livres, oriundos de entrevistas eletrônicas (formulários).

Os textos possuem de uma grande quantidade de abreviações, linguagens comuns em chats de internet, termos de baixo calão, gírias, nomes de produtos, promoções e palavras específicas do mercado de telecomunicações. Essas percepções sobre o atendimento e os problemas enfrentados, formaram a base de dados que será chamada de *Telecom Customer Experience Dataset* (TCXD).

4.2.1 TCXD-ML-Full

A base de dados *Telecom Customer Experience Dataset,* ou apenas TCXD, contém 2.104 amostras. Ao todo, foram 4.101 classificações, ou seja, em média 1,94 rótulos por resposta obtida, justificando a abordagem multirrótulo. Como todos os registros foram utilizados, chamaremos esta base de TCXD-ML-Full. O desbalanceamento entre os rótulos é uma questão que merece ser destacada. O motivo mais citado "*Resolver problemas de sinal de internet*", aparece em 36,64% das amostras. Por outro lado, "*Dúvidas sobre compra na loja ou site da Operadora*", aparece em apenas 0,14%.

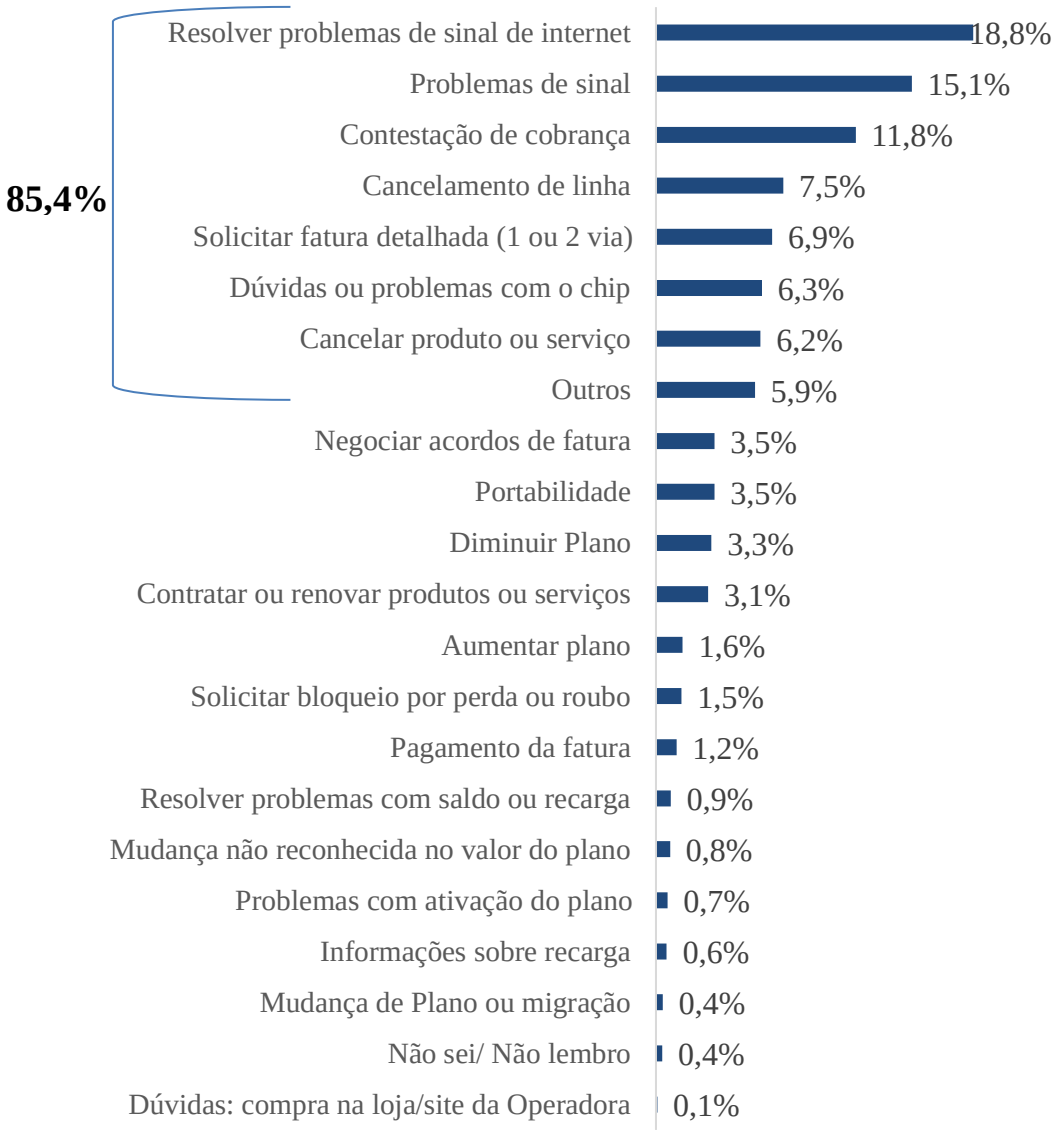
É possível observar, com base na distribuição de frequência dos rótulos na Tabela 3, que o rótulo mais presente tem frequência 200 vezes maior que o rótulo menos observado. Na base de dados, os rótulos não aparecem igualmente distribuídos, esse desbalanceamento dificulta a acurácia dos modelos de classificação empregados.

Tabela 3 – Sumarização da base de TCXD-ML-Full

Categorias para classificação (multirrótulo)	Qtd Respostas	% No Corpus
Resolver problemas de sinal de internet	771	36,64%
Problemas de sinal	621	29,52%
Contestação de cobrança	485	23,05%
Cancelamento de linha	309	14,69%
Solicitar fatura detalhada (1 ou 2ª Via)	282	13,40%
Dúvidas ou problemas com o chip	257	12,21%
Cancelar produto ou serviço	254	12,07%
Outros	240	11,41%
Negociar acordos de fatura	142	6,75%
Portabilidade	142	6,75%
Diminuir Plano	135	6,42%
Contratar ou renovar produtos ou serviços	127	6,04%
Aumentar plano	65	3,09%
Solicitar bloqueio por perda ou roubo	62	2,95%
Pagamento da fatura	50	2,38%
Resolver problemas com saldo ou recarga	36	1,71%
Mudança não reconhecida no valor do plano	34	1,62%
Problemas com ativação do plano	28	1,33%
Informações sobre recarga	26	1,24%
Mudança de Plano ou migração	17	0,81%
Não sei/ Não lembro	15	0,71%
Dúvidas sobre compra na loja/site da Claro	3	0,14%
Qtd de Labels total	4.101	
Qtd de Respostas	2.104	
Número médio de labels por resposta	1,9491	

A Figura 9 apresenta a distribuição de frequência simples dos motivos de contato (rótulos) para classificação das respostas fornecidas, com relação ao total de rótulos (4.101) e não do total de entrevistas (2.104), onde sua soma totaliza 100%. Por este motivo os percentuais da Figura 9 e da Tabela 3 são diferentes.

Figura 9 - Distribuição de frequência das respostas obtidas (rótulos)



Nota-se que os primeiros 10 motivos somam 85,4% dos problemas reportados pelos clientes, e estes são concentrados em temas de sinal (internet ou celular), cobrança fatura, planos ou contestações e cancelamentos.

Exemplo de algumas respostas dos clientes e seus motivos de classificação. Em alguns casos mais de um motivo é citado na mesma resposta, justificando novamente a abordagem Multirrótulo.

Tabela 4 – Exemplo de respostas dos clientes

Comentário	Contestação de cobrança	Solicitar fatura detalhada (1ª ou 2ª via)	Diminuir Plano	Cancelamento de linha
Tentei contestar cobranças na minha fatura e não consigo fazer por telefone. Cobrança de serviços que não solicitei. Quero cancelar minha assinatura.	✓	-	✓	✓
A operadora não envia a fatura ai se não correrrmos atras pra pagar a fatura em dias eles ja cobram juros, sendo que o erro de nao enviar a fatura é deles, pq se a gente coloca que quer receber por email, eles tem que enviar o email com a fatura	✓	✓	-	-
Tenho dois planos e estou tentando cancelar um plano tem três meses e não consigo	-	-	-	✓
Operadora é boa! Só que tem que melhorar os planos pras pessoas de baixa renda. Neste preço vou cancelar.	-	-	✓	✓
Cobrança de seguro sem ter feito seguro..péssimo funcionários ñ sabem resolver nada.. tenho anos com esta operadora tô saindo fora	✓	-	-	✓
Atendimento péssimo. está bom de melhorar o sistema de atendimento virtual e presencial. Vou cancelar	-	-	-	✓
Fiz o pagamento porém, cortaram minha internet, não posso ficar sem internet. Vou cancelar.	✓	-	-	✓
Já era pra existir no aplicativo, no site, uma opção do cliente cancelar sua conta/plano. Os atendentes deveriam ter mais empatia com seus clientes, Quero cancelar ou reduzir meu plano	✓	-	✓	✓

A Figura 10 apresenta duas nuvens de palavras, representando os termos mais frequentes utilizados pelos clientes. Destaque para os termos “resolver”, “cobrança”,

“operadora”, “fatura”, “problema”, “indevido”. A alta frequência de termos relacionados a problemas na nuvem de palavras era esperada dado que o *call center* recebe maior número de ligações com pedidos de tratamento de ocorrências em discordância com o cliente.

Figura 10 – Nuvem de Palavras para as repostas dos clientes

Removendo *stopwords*, pontos, números + *stemming*.

Removendo *stopwords*, pontos, números + *Lematização*.

Os problemas de escrita dos textos e a baixa precisão na marcação dos rótulos, quando comparados com a informação passada pelo cliente, possuem efeito direto na qualidade de treinamento dos modelos. Por esse motivo, um subconjunto dos dados, formando uma amostra que passou por curadoria detalhada, formou uma segunda base de dados para testes. Esta base é chamada de *Telecom Customer Experience Dataset* (TCXD) – 549, ou apenas, TCXD-ML-549.

4.2.2 TCXD-ML-549

A TCXD-ML-549 é formada por um subconjunto de 549 respostas validadas em curadoria, que verificou correções de escrita e precisão da marcação do rótulo. Ao todo, foram 948 classificações, ou seja, em média 1,72 rótulos por resposta obtida. Nessa base, o desbalanceamento também foi evidente. O motivo mais citado "*Contestação de cobrança*", aparece em 31,69% das amostras. Por outro lado, "*Dúvidas sobre compra na loja ou site da Operadora*", aparece em apenas 0,55%.

Tabela 5 – Sumarização da base TCXD-ML-549

Categorias para classificação (multirrótulo)	Qtd Respostas	% No Corpus
Contestação de cobrança	174	31,69%
Resolver problemas de sinal de internet	141	25,68%
Outros	88	16,03%
Problemas de sinal nas ligações	77	14,03%
Cancelar produto ou serviço	62	11,29%
Portabilidade	51	9,29%
Solicitar fatura detalhada (1 ou 2ª via)	50	9,11%
Cancelar produto ou serviço	44	8,01%
Dúvidas ou problemas com o chip	40	7,29%
Solicitar bloqueio por perda ou roubo	34	6,19%
Negociar acordos de fatura	32	5,83%
Diminuir Plano	32	5,83%
Pagamento da fatura	30	5,46%
Mudança não reconhecida no valor do plano	29	5,28%
Contratar ou renovar produtos ou serviços	15	2,73%
Não sei/ Não lembro	15	2,73%
Aumentar plano	9	1,64%
Problemas com ativação do plano	7	1,28%
Informações sobre recarga	6	1,09%
Resolver problemas com saldo ou recarga	5	0,91%
Mudança de Plano ou migração	4	0,73%
Dúvidas sobre compra na loja/site da Claro	3	0,55%
Qtd de Labels total	948	
Qtd de Respostas	549	
Número médio de labels por resposta	1,7268	

Nessa base, os 10 motivos mais frequentes também concentram mais de 80% dos rótulos de classificação dos problemas.

Um terceiro subconjunto dos dados foi testado, de forma a permitir a comparação das medidas de performance na condição multiclasse, ou seja, cada reposta tendo apenas um rótulo, que seja mutuamente exclusivo em relação ao demais, esta base foi chamada de TCXD-ML-443.

4.2.3 TCXD-MC-443

A terceira e última base de dados possui 443 respostas que passaram por curadoria. Trata-se de um subconjunto da segunda base, com uma restrição, todos os registros possuem um único rótulo. Os três rótulos mais frequentes foram mantidos, apresentados na Tabela 6, sendo eles: contestação de cobrança, resolução de problemas de internet e cancelamento de produtos e outros, que completa as marcações.

Tabela 6 – Sumarização da base TCXD-MC-443

Categorias para classificação (multirrótulo)	Qtd Respostas	% No Corpus
Outros	221	50,00%
Contestação de cobrança	93	21,04%
Resolver problemas de sinal de internet	84	19,00%
Cancelar produto ou serviço	44	9,95%
Qtd de labels total	442	
Qtd de Respostas	442	100,00%
Número médio de labels por resposta	1,0000	

4.2.4 Resumo analítico sobre as três bases de dados

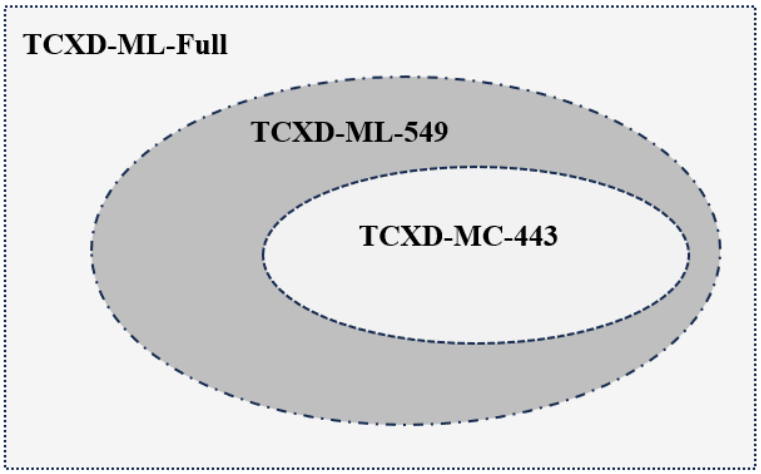
Este experimento considerou três bases de dados, representadas na Figura 11. As bases foram utilizadas para modelagem do problema, mantendo os critérios de transformação e classificação inalterados nos três conjuntos de dados, de forma que permitisse a comparação direta dos resultados, isolando o impacto de curadoria dos dados.

A primeira base, chamada "**TCXD-ML-Full**", possui 2.104 registros e não passou por uma curadoria profunda, onde as repostas dos clientes foram utilizadas sem tratamento prévio, mantendo o formato original recebido.

A segunda base, chamada "**TCXD-ML-549**", consiste em um subconjunto da primeira base, com curadoria completa em 549 registros. Estas amostras foram selecionadas por meio de amostragem aleatória simples, com 95% de confiança e 4% de margem de erro.

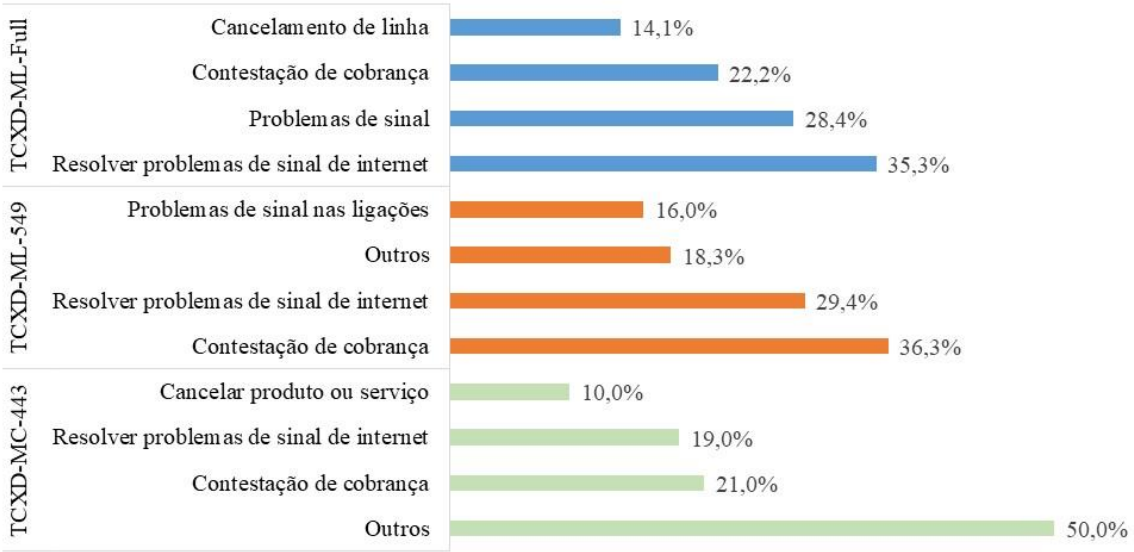
A terceira e última base de dados, chamada TCXD-MC-443 é um subconjunto da segunda base, ou seja, com curadoria completa. Nessa base foram mantidos quatro rótulos únicos (mutuamente exclusivos) para cada registro: "Contestar cobrança indevida", "Cancelamento de linha, produto ou serviço", "Resolver problemas de sinal de Internet (3G/4G/4.5G)" e "Outros". As bases são apresentadas na Figura 11.

Figura 11 – Bases de dados utilizadas no experimento



As três bases de dados mantiveram razoável proporcionalidade entre os principais motivos de contato. A Figura 12 limita os 4 maiores de cada uma para comparação.

Figura 12 – As quatro categorias mais frequentes x base de dados



4.3 Treinamento e teste

Este trabalho utilizou validação cruzada k-fold, que é uma técnica que divide o conjunto de dados em subconjuntos de treinamento e teste, permitindo avaliar o modelo em diferentes partições. Isto facilita verificar a capacidade do modelo de generalizar para outros conjuntos de dados.

No experimento foi utilizado k=5, de forma que o classificador é treinado em quatro dos subconjuntos (treinamento) e avaliado no subconjunto de teste. Esse processo é repetido cinco vezes, com cada subconjunto sendo usado como conjunto de teste uma vez. As partições são escolhidas aleatoriamente e possuem tamanhos aproximadamente iguais.

4.4 Preparação dos dados

Foram aplicadas as técnicas definidas no Capítulo 2. Primeiro, todos os *tokens* foram convertidos para minúsculas. Em seguida, pontos, vírgulas, exclamações, interrogações e pontos finais foram substituídos por espaços. Na sequência, foram removidas *stopwords* e *tokens* com menos de 3 caracteres e aplicado *stemming*, todo este processo foi repetido ignorando stemming e aplicando lematização. Por último, todos os textos dessas bases de dados foram convertidos para a representação espaço-vetorial.

4.5 Métodos de classificação

Foram empregados quatro métodos de classificação: a regressão logística (RL) (COX, D. R.,1958, p.93), o SVM (FAN, R.,2008, p.93 e 128), Multinomial *Naive Bayes* (McCALLUM E NIGAM,1998) e *Random Forest* (BREIMAN, 2001). Eles foram escolhidos por serem tradicionais e possuírem interpretação direta, reconhecida eficiência quando bem utilizados, capacidade em lidar com dados desbalanceados e generalização.

Os métodos RL, SVM, RF e M.NB não podem ser aplicados diretamente em problemas multirrótulo, pois foram propostos para problemas monorrótulo. Para contornar essa limitação, foram realizadas as transformações de problemas BR e LP, explicadas na Seção 3.2, que tem como objetivo converter problemas multirrótulo para monorrótulo. A implementação utilizada para as transformações é proveniente da biblioteca scikit-multilearn, que disponibiliza diversos métodos de classificação multirrótulo.

5 PROPOSTA DE SOLUÇÃO

Nas seções subsequentes, será apresenta a proposta de solução para o problema analisado, incluindo a abordagem adotada para o problema multirrótulo, a organização das simulações que foram implementadas e como será a comparação dos resultados, estabelecendo assim as bases para as conclusões e os passos futuros.

5.1 Abordagem

O problema tratado, possui características de classificação de textos multirrótulo, conforme já apresentado, sendo assim, entre as opções adaptação de algoritmos e transformação de problemas, foi adotada a segunda por sua simplicidade, viabilidade computacional e disposição de vários métodos de classificação possíveis de serem aplicados. Para fins didáticos, serão consideradas duas etapas, a primeira com a transformação do problema e a segunda a classificação dos rótulos.

A primeira etapa, transformação do problema multirrótulo em um conjunto de problemas binários, adotando duas transformações, a *Binary Relevance* (BR) e a *Label Powerset* (LP) para viabilizar uso de classificadores, que foram desenvolvidos originalmente para situações monorrótulo ou multiclasse. Essas transformações possuem características complementares, não cobrem todas as situações possíveis, mas atendem a condições essenciais deste problema, como a independência ou não entre os rótulos e a possibilidade de agrupamento dos mesmos.

Binary Relevance:

- Cada classificador define se o seu rótulo é relevante ou não para o problema analisado.
- O modelo final é a combinação das predições individuais desses classificadores.
- Assumi que os rótulos são independentes.

Label Powerset:

- Cada subconjunto de rótulos das amostras de treinamento como se fosse uma classe independente.
- Transforma os diferentes rótulos das amostras de treinamento em classes únicas.
- O modelo final é a combinação das predições individuais desses classificadores.
- Como os rótulos são analisados em conjunto com as ocorrências dos outros, é possível considerar a presença de correlação entre eles.

A segunda etapa, consiste na combinação de cada transformação adotada, BR e LP, com os classificadores escolhidos, Regressão Logística (RL), SVM, Multinomial *Naive Bayes* e *Random Forest*. Seguindo o mesmo princípio, esses classificadores foram adotados por possuírem mecanismos de funcionamento complementares e cobrirem importantes situações, estão classificados entre os mais conhecidos em métodos baseados em probabilidade, ensemble e otimização.

5.2 Classificação e performance

Cada base de dados, recebeu diferentes agrupamentos de rótulos, e foi classificada com cada um dos classificadores definidos para o experimento, resultando em 72 simulações de classificação de textos, lembrando que foi utilizado particionamento K-fold=5, onde as medidas de performance foram as médias desses agrupamentos. Essa combinação de bases, agrupamento de rótulos, transformação e classificadores está representada na Tabela 7.

Tabela 7 – Combinação de bases x rótulos x transformação x classificador

Bases de dados	Rótulos	Transformações (BR , LP)	Classificadores (M.NB, SVM, RL , RF)	Total de Simulações
TCXD-ML-Full	22	2	4	8
	18	2	4	8
	10	2	4	8
	7	2	4	8
TCXD-ML-549	22	2	4	8
	18	2	4	8
	10	2	4	8
	7	2	4	8
TCXD-MC-443	4	2	4	8
Total				72

Como parâmetro de comparação, foi calculada a probabilidade (*P*) de um documento (*d*) ser classificado aleatoriamente nos seus rótulos corretos, considerando que para a base TCXD-ML-Full, existem 4.101 rotulações, *P*(*d*) = 0.0000244. Esta *P(d)*, pode ser uma referência para a acurácia do processo de classificação automática. No Capítulo 6, os resultados são apresentados para as simulações realizadas, os comparativos são feitos utilizando as medidas de performance definidas para este experimento, Perda Hamming, F1-Macro e Acurácia.

6 RESULTADOS

Nas próximas seções, são apresentados os resultados em cada uma das três bases de dados, com seus diferentes agrupamentos de rótulos e a comparação das medidas de desempenho.

6.1 TCXD-ML-Full

A Tabela 8 apresenta os resultados obtidos pelos métodos de classificação no experimento realizado com a base de dados TCXD-ML-Full.

De uma forma geral, as medidas de performance foram “muitas baixas” indicando fraco poder de classificação, apesar de maiores que a probabilidade de uma seleção aleatória correta, não são robustos em todos os cenários avaliados, ou ainda, esses resultados baixos podem ser indicação da complexidade do problema multirrótulo e a necessidade de outros tratamentos em trabalhos futuros. Os resultados são apresentados em escala de tons de cinza, sendo que, quanto mais escuro, melhor é o resultado.

Tabela 8 – TCXD-ML-Full – Médias das medidas de performance

Classificador	Transformação BR			Transformação LP		
	Perda Hamming	F1-Macro	Acurácia	Perda Hamming	F1-Macro	Acurácia
Regressão Logística	0,0880	0,0535	0,0385	0,0985	0,0595	0,1127
SVM	0,0876	0,0765	0,0390	0,0977	0,0837	0,1122
Multinominal Naïve Bayes	0,0880	0,0329	0,0242	0,0994	0,0429	0,1136
Random Forest	0,0906	0,0863	0,0457	0,1060	0,1103	0,1042

Obs: k-folds (k=5) / 22 rótulos

O melhor desempenho foi obtido pela Random Forest combinado com a transformação LP, sendo que as medidas de performance foram: Perda de Hamming = 0,1060, F1-Macro = 0,1103 e Acurácia = 0,1042. Diante do fraco resultado, outros cenários foram experimentados, como redução do número de rótulos, uma vez que a dispersão original é muito alta. A Tabela 9 apresenta os resultados resumidos de vários outros cenários, como com 18 rótulos, onde alguns foram agrupados, com 10 rótulos, pois resumiram mais de 80% das classificações e 7 rótulos mais frequentes.

Na Tabela 9, está ilustrado o resumo dos resultados, obtidos durante o experimento para cada combinação entre transformação e classificador, representados pelas médias das métricas de avaliação F1-Macro, Perda de Hamming. O conjunto de dados foi dividido em cinco partes (5-*folds*), mantendo a proporção de exemplos de cada classe, respeitando o princípio de amostragem aleatória estratificada.

Tabela 9 – TCXD-ML-Full – Médias agrupando rótulos

Classificador	Transformação BR			Transformação LP		
	Perda Hamming	F1-Macro	Acurácia	Perda Hamming	F1-Macro	Acurácia
22 Rótulos						
RL	0,0880	0,0535	0,0385	0,0985	0,0595	0,1127
SVM	0,0876	0,0765	0,0390	0,0977	0,0837	0,1122
M.NB	0,0880	0,0329	0,0242	0,0994	0,0429	0,1136
RF	0,0906	0,0863	0,0457	0,1060	0,1103	0,1042
18 rótulos						
RL	0,1052	0,0642	0,0404	0,1171	0,0826	0,1192
SVM	0,1046	0,0934	0,0447	0,1167	0,1058	0,1187
M.NB	0,1050	0,0448	0,0318	0,1190	0,0522	0,1169
RF	0,1093	0,1006	0,0470	0,1274	0,1349	0,1012
10 rótulos						
RL	0,1670	0,0861	0,0913	0,1853	0,1358	0,1312
SVM	0,1680	0,0927	0,0956	0,1839	0,1372	0,1308
M.NB	0,1665	0,0466	0,0709	0,1895	0,0969	0,1289
RF	0,1727	0,1064	0,0975	0,1980	0,1673	0,1250
7 rótulos						
RL	0,1773	0,0943	0,0956	0,1971	0,1585	0,1374
SVM	0,1764	0,1039	0,0980	0,1968	0,1612	0,1393
M.NB	0,1776	0,0526	0,0751	0,2043	0,1056	0,1322
RF	0,1850	0,1224	0,1022	0,2094	0,1867	0,1303

A redução do número de rótulos, concentrado os mais frequentes, reduz a imprecisão de marcação e ameniza a falta de curadoria da base, mas ainda assim as métricas de desempenho são modestas, com o melhor resultado obtido pelo método Random Forest com Transformação LP, com sete rótulos mais frequentes, com Perda de Hamming = 0,2094, F1-Macro = 0,1867 e Acurácia = 0,1303.

6.2 TCXD-ML-549

A Tabela 10 apresenta os resultados obtidos durante no experimento realizado com cada combinação entre as transformações BR e LP e os classificadores RL, SVM, M.NB e RF, representados pelas médias das métricas de avaliação F-Macro, Perda de Hamming e Acurácia de cada combinação de rótulos, obtido nas execuções de cada K-*fold* (k=5).

De uma forma geral as medidas de performance da base TCXD-ML-549 foram melhores que na TCXD-ML-Full, demonstrando o quanto o processo de curadoria é relevante, mas ainda assim os resultados foram modestos.

Tabela 10 – TCXD-ML-549– médias das medidas de performance

Classificador	Transformação BR			Transformação LP		
	Perda Hamming	F1-Macro	Acurácia	Perda Hamming	F1-Macro	Acurácia
22 Rótulos						
RL	0,0664	0,0907	0,1110	0,0791	0,1167	0,1951
SVM	0,0646	0,1178	0,1256	0,0781	0,1319	0,1986
M.NB	0,0660	0,0955	0,1058	0,0793	0,1093	0,1952
RF	0,0658	0,1218	0,1164	0,0791	0,1400	0,1985
18 rótulos						
RL	0,0791	0,1145	0,1188	0,0940	0,1442	0,2081
SVM	0,0766	0,1486	0,1391	0,0921	0,1600	0,2210
M.NB	0,0782	0,1347	0,1245	0,0944	0,1355	0,2081
RF	0,0780	0,1554	0,1320	0,0930	0,1835	0,2160
10 rótulos						
RL	0,1133	0,1686	0,1229	0,1399	0,2092	0,2162
SVM	0,1101	0,2111	0,1388	0,1365	0,2399	0,2248
M.NB	0,1117	0,1946	0,1301	0,1417	0,1927	0,2145
RF	0,1116	0,2147	0,1310	0,1399	0,2462	0,2205
7 rótulos						
RL	0,1348	0,2390	0,1430	0,1605	0,3090	0,3071
SVM	0,1310	0,2914	0,1668	0,1554	0,3424	0,3181
M.NB	0,1336	0,2856	0,1557	0,1638	0,2641	0,2927
RF	0,1332	0,2951	0,1649	0,1578	0,3391	0,3107

O melhor desempenho foi novamente obtido pela Random Forest combinado com a transformação LP, sendo que as medidas de performance foram: Perda de Hamming = 0,1578, F1-Macro = 0,3391 e Acurácia = 0,3107.

6.3 TCXD-MC-443

Mesmo com a curadoria da base, reduzindo de 2.104 registros para uma amostra validada de 549 registros, não foram obtidas medidas de performance robustas, que pudessem sinalizar uma aplicação prática e segura dos modelos em um cenário de produção.

Um último cenário foi avaliado, isolando um subconjunto multiclasse na amostra validada multirrótulo. O objetivo deste experimento foi isolar o efeito multirrótulo. Os resultados são apresentados na Tabela 11. O classificador RF com a transformação LP novamente obteve o melhor resultado e as medidas de performance tiveram aumento significativo. A medida F1-Macro atingiu o valor de 0,5205 e a Acurácia foi de 0,6697, valores muito superiores as mesmas medidas com a base TCXD-ML-549.

Tabela 11 – TCXD-MC-443

Classificador	Transformação BR		Transformação LP	
	F1-Macro	Acurácia	F1-Macro	Acurácia
RL	0,4917	0,6449	0,5008	0,6584
SVM	0,5089	0,6652	0,5089	0,6652
M.NB	0,3985	0,3298	0,4901	0,6471
RF	0,5120	0,6629	0,5205	0,6697

Como explicado anteriormente, os resultados tratando a base multiclasse foram melhores do que na modelagem multirrótulo, demonstrando que os ganhos da curadoria são imprescindíveis e que a complexidade multirrótulo é fator determinante nesta modelagem.

A comparação de resultados entre os três cenários avaliados consta nas Figuras 13, Figura 14 e Figura 15, que comparam os resultados das bases TCXD-ML-Full, TCXD-ML-549 e TCXD-MC-443 respectivamente. Os eixos (Y) de cada gráfico foram fixamos em tamanho 0,8 para que possam ser comparados visualmente.

Figura 13 – Comparação entre os resultados de classificação - TCXD-ML-Full

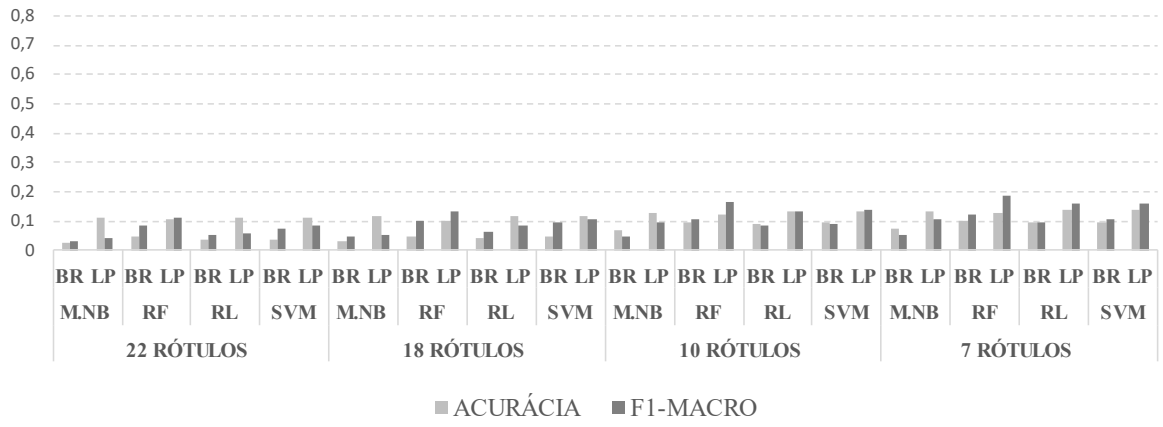


Figura 14 – Comparação entre os resultados de classificação - TCXD-ML-549

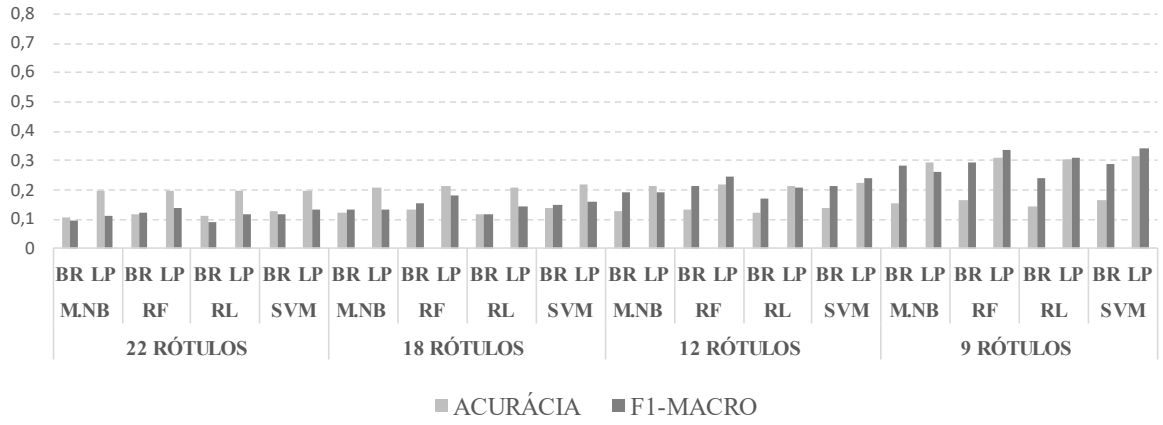
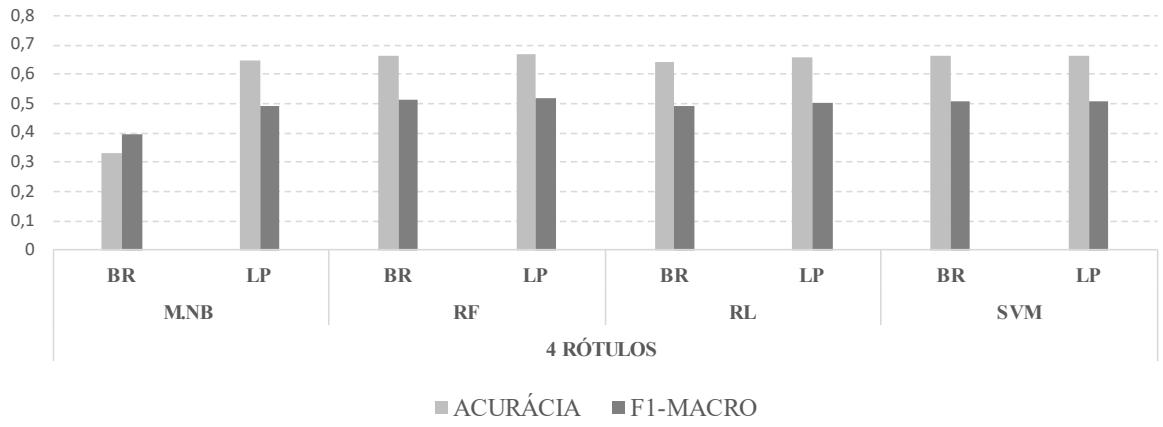


Figura 15 – Comparação entre os resultados de classificação - TCXD-MC-443



7 CONCLUSÃO

O setor de telecomunicações no Brasil passou por uma significativa transformação ao longo das décadas. Inicialmente marcado pela complexidade de concessões desordenadas e pela estatização do setor com a criação da Telebrás na década de 70. A grande mudança ocorreu nos anos 90 com a tendência neoliberal, culminando na privatização e na criação da Agência Nacional de Telecomunicações (ANATEL) pela Lei Geral de Telecomunicações (LGT). A partir de 1998, a abertura para o capital privado desencadeou uma série de mudanças, incluindo fusões corporativas, avanços tecnológicos como o 3G, e, mais recentemente, o advento do 5G.

No contexto mais recente, entre 2019 e 2023, o mercado enfrenta desafios relacionados à novas receitas e redução de custos. A popularização dos *smartphones*, o aumento do consumo de dados e a pressão para expandir as redes de dados móveis e banda larga destacam-se como fatores influentes. Diante dessa competição acirrada, as empresas de telecomunicações enfrentam a necessidade de investir massivamente e focar em tecnologias como Inteligência Artificial (IA) e Aprendizado de Máquina para atender às demandas crescentes e se transformarem em empresas de tecnologia.

Além disso, a importância crescente da Experiência do Cliente (CX) no setor tornou-se um pilar estratégico, exigindo uma estrutura de atendimento ao cliente eficiente e personalizada. A implementação de jornadas de vendas ou pós-vendas bem-sucedidas demanda análise de dados detalhada para compreender o comportamento do consumidor. Diferentes canais de atendimento, incluindo atendimento presencial, telefone, canais baseados em texto e IA, são fundamentais para proporcionar uma CX positiva, evidenciando a necessidade de adaptação constante às mudanças no comportamento do consumidor.

Neste cenário, este trabalho de conclusão de curso aborda a implementação de um sistema de classificação automática de textos, recebidos pelo departamento de atendimento ao cliente de uma empresa de telecomunicações. A transição da classificação humana para um método automático, visa agilizar o direcionamento rápido para áreas específicas de solução, aprimorando a eficiência do atendimento. A hipótese central trata da possibilidade de realizar classificação automática dos motivos de contato com acurácia suficiente, utilizando algoritmos de Aprendizado de Máquina e PLN, para problemas multirrótulo. O modelo proposto busca ser sustentável em termos de custo computacional e viável para implementação offline, visando melhorar a experiência do cliente e reduzir os custos operacionais. A perspectiva é transformar esse modelo em um "serviço de TI exposto" interno,

capaz de receber solicitações de vários canais de atendimento e classificar automaticamente, proporcionando respostas mais rápidas e eficazes, além de aprimorar a detecção inteligente de problemas na origem das demandas.

A preparação dos dados, conforme definido no Capítulo 2, incluiu a conversão de *tokens* para minúsculas, substituição de pontuações por espaços, remoção de *stopwords*, *tokens* com menos de 3 caracteres e aplicação de *stemming*, com uma repetição do processo ignorando *stemming* e aplicando lematização.

Todos os textos foram convertidos para representação espaço-vetorial. A representação computacional utilizou o método TF-IDF. Foram considerados quatro métodos de classificação, sendo eles: regressão logística, SVM, Multinomial Naive Bayes e Random Forest – esses foram adaptados para problemas multirrótulo usando transformações BR e LP. Esses métodos foram selecionados devido à sua tradição, interpretação direta, eficiência reconhecida, capacidade de lidar com dados desbalanceados e capacidade de generalização.

Inicialmente foi utilizada a base de dados mais completa disponível, construída a partir de entrevistas com os clientes da empresa de telecomunicações, a base foi chamada de TCXD-ML-Full, com 2.104 registros, mas que não passaram por um processo pleno de curadoria, as baixas medidas de performance nos classificadores, levaram ao próximo passo, reavaliar a qualidade do dado que estava sendo utilizado, e os textos (respostas) fornecidos pelos clientes entrevistados, possuíam diversas inconsistências a serem tratadas, como problemas de escrita, marcação de rótulo inconsistente com a descrição do cliente, reduzindo a capacidade dos classificadores em ‘acertarem’ os rótulos fornecidos para treinamento.

Dessa forma, uma amostra foi isolada para processo curadoria refinada, TCXD-ML-549, e apesar de não podermos comparar com o experimento anterior de forma justa, devido a uma mudança na quantidade de exemplos, foi observado que nesse caso as métricas foram melhores, elevando os resultados das métricas de performance dos modelos de classificação, mas ainda assim foram modestos, então foi realizada comparação com um subconjunto da amostra validada, mas com redução de registros e que possuíam apenas um rótulo, formando a base TCXD-MC-443, ou seja, os classificadores foram testados como multiclasse e não apenas multirrótulo. Nessa simulação multiclasse, foram alcançados resultados satisfatórios, nas métricas de performance dos classificadores, com taxas de “acerto” dos rótulos superiores a 66%.

Assim concluímos que este estudo de caso, com problemas reais de clientes de uma operadora de telecomunicações no Brasil, era extremamente sensível a qualidade de dados, mas que a proposição inicial, que é a construção de um modelo para setorização automática

de atendimento é viável, cabendo algumas recomendações. Para implementação do modelo em produção, deve-se utilizar bases maiores para treinamento e verificação da robustez da qualidade dos dados, e comparação com problemas ou subconjuntos multiclasse, uma vez que a complexidade do tema multirrótulo é evidente e deve continuar sendo aprofundado e aprimorado para uso corporativo. Importante destacar, como evolução deste trabalho, a necessidade de buscar métodos que possam tratar os textos de forma original, pois neste experimento controlado, obtivemos melhora dos resultados apenas após a curadoria dos textos, mas este tratamento em produção seria inviável.

8 REFERÊNCIAS

ALMEIDA, T. A.; YAMAKAMI, A.; ALMEIDA, J. **Filtering spams using the minimum description length principle**. In: Proceedings of the 2010 ACM Symposium on Applied Computing. New York, NY, USA: ACM, 2010. (SAC '10), p. 1854–1858. Citado 2 vezes nas páginas 30 e 60.

ALMEIDA, T. A., GOMEZ HIDALGO, J. M., & SILVA, T. P. (2012). **Towards SMS Spam Filtering** : Results under a New Dataset. *International Journal of Information Security Science T.*, *2*(1).

ALVIM, L., Vilela, P., Motta, E., & Milidiú, R. (2010). **Sentiment of Financial News: A Natural Language Processing Approach**. *1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology*.

BITTENCOURT, M. de M. (2020). *ML-MDLText:* ***um método de classificação de textos multirrótulo de aprendizado incremental***. Usp – Sorocaba.

CÂMARA JÚNIOR, A. (2013). **Processamento de linguagem natural para indexação automática semântico-ontológica**. *Revista Ibero-Americana de Ciência Da Informação*, *9*(2).

CARVALHO, N. R., & Simões, A. (2018). PLN.pt: **Processamento de Linguagem Natural para Português como um Serviço**. *Linguamática*, *10*(1). https://doi.org/10.21814/lm.10.1.267

COX, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 20, n. 2, p. 215–232, 1958. Citado na página 93.

CHRISTIANO, LÉO – **EMBRATEL; Interligando o Brasil ao Infinito: memória histórica da EMBRATEL**, 1965/1997. Editorial, 1998.

DEMSZKY, D., MOVSHOVITZ-ATTIAS, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). *GoEmotions:* ***A Dataset of Fine-Grained Emotions.*** https://doi.org/10.18653/v1/2020.acl-main.372

FACELI, K., LORENA, A. C., GAMA, J., & CARVALHO, A. C. P. L. F. (2011). **Inteligência artificial : uma abordagem de aprendizado de máquina. In *Livros Técnicos e Científicos*.**

FAN, R.-E. et al. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, JMLR.org, v. 9, p. 1871–1874, ago. 2008. ISSN 1532-4435. Citado 2 vezes nas páginas 93 e 128.

FINGER, M. (2021). **Inteligência Artificial e os rumos do processamento do português brasileiro**. *Estudos Avancados*, *35*(101). https://doi.org/10.1590/s0103-4014.2021.35101.005

FLORENCIO, Roberto. **Os Tipos de Linguagens Expressadas pelo Ser Humano no Meio em que Vive**. Revista Científica Multidisciplinar Núcleo do Conhecimento. Ano 03, Ed. 08, Vol. 16, pp. 184-192, Agosto de 2018. ISSN:2448-0959

GOMES, L. M., Sá, J. M. C. de, & Yaohao, P. (2020). TD 2612 **- Línguas Naturais E Máquinas Artificiais**: aplicação de técnicas de mineração de texto para a classificação de sentenças judiciais brasileiras. *Texto Para Discussão*. https://doi.org/10.38116/td2612

GUIMARÃES, L. M. S., MEIRELES, M. R. G., & Almeida, P. E. M. de. (2019). **Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação**. *Perspectivas Em Ciência Da Informação*, *24*(1). https://doi.org/10.1590/1981-5344/3505

GONÇALVES, M., Coheur, L., Baptista, J., & Mineiro, A. (2021). **Avaliação de recursos computacionais para o português**. *Linguamática*, *12*(2). https://doi.org/10.21814/lm.12.2.331

JURAFSKY, D., & Martin, J. H. (2009). **Book Review Speech and Language Processing** (second edition). *Computational Linguistics*, 0–3.

KIM, S. B., RIM, H. C., YOOK, D. S., & Lim, H. S. (2002). Effective methods for improving naive Bayes text classifiers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2417*, 414–423. https://doi.org/10.1007/3-540-45683-X_45

MADRUGA, R. (2009). ***Call Centers de Alta Performance****: Manual indispensável para todos que buscam excelência no atendimento.* Atlas.

MARTIN, J. H. (2008). ***Speech and Language Processing****: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. https://www.researchgate.net/publication/200111340

MELLO, Jorge Luiz Marques de.**A terceirização no setor de telecomunicações.**2010. 67 f. Monografia (Especialização) - Curso de Direito e Processo do Trabalho, Universidade Candido Mendes, Rio de Janeiro, 2010. Disponível em:

http://www.avm.edu.br/docpdf/monografias_publicadas/k214544.pdf

Acesso em: 11 jun. 2014.

PENNA FILHO, Pedro Baptista de Araújo – **Telecomunicações: O Desafio da Integração Nacional, Embratel 1967-2004**, Rio de Janeiro: Editora Ciência Moderna, 2009.

PEREIRA CERQUEIRA, S. (2021). **Estudos de técnicas para processamento de linguagem natural**. *Anais Dos Seminários de Iniciação Científica*, *23*. https://doi.org/10.13102/semic.v0i23.6597

PREDICTION, W., & Models, L. (2007**). But it must be recognized that the notion** "probability of a sen- tence" is an entirely useless one, under any known interpretation of this term. *Language*, 934. https://doi.org/10.1007/s00134-010-1760-5

PROVOST, F., & Fawcett, T. (2016). **Data science para negócios:** *O que você precisa saber sobre mineração de dados e pensamento analítico de dados*. ISBN 978-85-7608-972-8 páginas 254-256

RIBEIRO PEREIRA, F., & Jose Rigo, S. (2013). **Utilização de processamento de linguagem natural e ontologias na análise qualitativa de frases curtas**. *RENOTE*, *11*(3). https://doi.org/10.22456/1679-1916.44431

SARDINHA, T. B. (2000). **Lingüística de Corpus: histórico e problemát**ica. *DELTA: Documentação de Estudos Em Lingüística Teórica e Aplicada*, *16*(2). https://doi.org/10.1590/s0102-44502000000200005

SILVA, B. C. D. da. (2006). **O estudo lingüístico-computacional da linguagem**. *Letras de Hoje*, *41*(2).

SILVA, E. M. da, & Souza, R. R. (2014). **Fundamentos em processamento de linguagem natural**: uma proposta para extração de bigramas. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência Da Informação*, *19*(40). https://doi.org/10.5007/1518-2924.2014v19n40p1

SILVA, R. M., SANTOS, R. L. S., ALMEIDA, T. A., & PARDO, T. A. S. (2020). **Towards automatically filtering fake news in Portuguese**. *Expert Systems with Applications*, *146*. https://doi.org/10.1016/j.eswa.2020.113199

TELEBRAS - TELECOMUNICAÇÕES BRASILEIRAS - S.A (Brasil).**A Telebras e a Evolução das Telecomunicações:** HISTÓRICO. Disponível em:

http://www.telebras.com.br/inst/?page_id=41

Acesso em: 21 jul. 2014.

TELEBRASIL – ASSOCIAÇÃO BRASILEIRA DE TELECOMUNICAÇÕES, **O Desempenho do Setor de Telecomunicações no Brasil - Séries Temporais – 2012,** disponível em:

http://www.telebrasil.org.br/panorama-do-setor/o-setor-de-telecomunicacoes

Acesso em 19/07/14.

TELECO (Brasil).**O Desempenho do Setor de Telecomunicações no Brasil – Séries Temporais, preparado pelo Teleco para a Telebrasil.** 2014. Disponível em:

http://www.teleco.com.br/estatis.asp

Acesso em: 18 jul. 2014.

TELECO (Brasil). **O Histórico do Setor de Telecom e a Privatização**. Disponível em: https://www.teleco.com.br/tutoriais/tutorialgpter1/pagina_2.asp Acesso em: 20 ago. 2023

WAGNER FILHO, J. A., Wilkens, R., Idiart, M., & Villavicencio, A. (2019). The BRWAC corpus: **A new open resource for Brazilian Portuguese**. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.

WEISS, S. M.; INDURKHYA, N.; ZHANG, **T. Fundamentals of predictive text mining**. 2.
ed. [S.l.]: Springer Verlag London, 2015. ISBN 978-1-4471-6749-5. Citado 3 vezes nas páginas 35, 36 e 92.