UNIVERSIDADE NOVE DE JULHO – UNINOVE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO

MARIA SHEILA CARNEIRO

APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA ANÁLISE DE SENTIMENTOS PARA A DESCOBERTA DE CONHECIMENTO SOBRE O CLIENTE

SÃO PAULO
2023

MARIA SHEILA CARNEIRO

APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA ANÁLISE DE SENTIMENTOS PARA A DESCOBERTA DE CONHECIMENTO SOBRE O CLIENTE

Pesquisa de dissertação apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho – UNINOVE, como requisito parcial para a obtenção do grau de Mestre em Informática e Gestão do Conhecimento.

Prof. Orientador: Dr. Marcos Antonio Gaspar
Prof. Coorientador: Dr. Renato José Sassi

São Paulo
2023

Carneiro, Maria Sheila.
Aplicação de redes neurais convolucionais para análise de sentimentos para a descoberta de conhecimento sobre o cliente. / Maria Sheila Carneiro. 2023.
137 f.
Dissertação (Mestrado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2023.
Orientador (a): Prof. Dr. Marcos Antonio Gaspar.
Coorientador (a): Prof. Dr. Renato José Sassi.
1. Inteligência artificial. 2. Análise de sentimentos. 3. Redes neurais convolucionais. 4. Descoberta de conhecimento em bases de dados.
I. Gaspar, Marcos Antonio. II. Sassi, Renato José. III. Título.

CDU 004

UNINOVE
Universidade Nove de Julho

PARECER – EXAME DE DEFESA

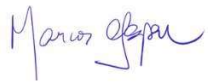
Parecer da Comissão Examinadora designada para o exame de defesa do Programa de Pós-Graduação em Informática e Gestão do Conhecimento a qual se submeteu a aluna Maria Sheila Carneiro.

Tendo examinado o trabalho apresentado para obtenção do título de “Mestre em Informática e Gestão do Conhecimento”, com Dissertação intitulada “APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA ANÁLISE DE SENTIMENTOS PARA A DESCOBERTA DE CONHECIMENTO SOBRE O CLIENTE”, a Comissão Examinadora considerou o trabalho:

(X) Aprovado ( ) Aprovado condicionalmente
( ) Reprovado com direito a novo exame ( ) Reprovado

EXAMINADORES

Prof. Dr. Marcos Antonio Gaspar - Uninove (Orientador)



Prof. Dr. Renato José Sassi - Uninove (Coorientador)

Prof. Dr. Fabio Kazuo Ohashi - Mackenzie (Membro Externo)

Prof. Dr. Cleber Gustavo Dias - Uninove (Membro Interno)

São Paulo, 19 de abril de 2023.

Dedico este trabalho primeiramente a Deus que me deu forças e sabedoria, em especial a minha mãe Maria de Fátima e ao meu irmão Michel, que sempre me apoiaram com palavras de incentivo para que eu pudesse enfrentar essa desafiadora e linda jornada.

AGRADECIMENTOS

A trajetória de uma mestranda é repleta de desafios e muitas vezes parecendo até inatingível, mas com o apoio da família torna-se possível. A primeira pessoa que quero agradecer é minha mãe Maria de Fátima Carneiro Silva, que me apoiou neste desafio desde o começo. Ao meu irmão Michel Silva Carneiro, que sempre me ajuda com conselhos e palavras de incentivo, sendo muito um grande amigo que está ao meu lado em todos os momentos da minha vida.

Além da minha família, quero agradecer primeiramente a Uninove Nove de Julho (UNINOVE) pelo apoio, pela oportunidade de crescimento, aprimoramento acadêmico, pessoal e profissional, assim como pela bolsa de estudos. Nesta instituição passei toda a minha jornada de estudo e aprendizado, tenho enorme gratidão.

Outra pessoa primordial para esta conquista foi o meu orientador Prof. Dr. Marcos Gaspar, pois sem ele também não teria sido possível chegar até aqui. Com muita paciência, compreensão e sabedoria me desafiou e orientou para que eu entregasse o meu melhor. Suas aulas me despertaram total admiração, não só pelo conhecimento que me trouxe, mas especialmente pelo exemplo de ser humano. educação e dedicação, profissional exemplar como orientador sempre me apoiando em todos os momentos ao longo desta jornada.

Agradeço, ao meu coorientador Prof. Dr. Renato José Sassi, que sempre me apoio, me salvou de vários equívocos neste trabalho, trouxe ideias muito valiosas e agradeço por sempre estar presente, e por sempre me indicar a direção correta que o trabalho deveria tomar.

Agradeço, aos professores da minha banca de qualificação, Prof. Dr. Cleber Gustavo Dias e Prof. Dr. Fabio Kazuo Ohashi, que com os seus comentários e indicações pontuais de grande pertinência a esta dissertação, me fizeram reavaliar e melhorar a pesquisa. Agradeço a todos os professores do PPGI, que me deram a oportunidade de assistir aulas e aprender muito com eles.

Agradeço também a todos meus amigos que me ajudaram ao longo desta linda caminhada. Por fim, agradeço à Universidade Nove de Julho por toda a estrutura e investimento na pesquisa, o qual me permitiu concretizar um sonho de titular mestra.

RESUMO

Os clientes trocam informações, opiniões e sentimentos diariamente sobre diversos temas atrelados aos produtos e serviços ofertados pelas empresas. Porém, os dados gerados nesses depoimentos precisam ser analisados visando assim a extração de conhecimento útil à empresa para a melhor compreensão do cliente, de modo a proporcionar um melhor atendimento. A Inteligência Artificial (IA) dispõe de métodos e técnicas aplicáveis à análise de sentimentos expressados por clientes em textos livres com baixa estruturação. Assim, a aplicação da IA para descoberta de conhecimentos em bases de dados pode auxiliar na compreensão dos sentimentos expressados pelo cliente. O objetivo desta pesquisa é aplicar técnicas de redes neurais convolucionais para a análise e classificação de sentimentos em comentários de clientes visando a descoberta de conhecimento do cliente em empresa varejista. Em adição, buscou-se ainda comparar os resultados dos experimentos com os resultados dos indicadores detratores do NPS *(Net Promoter Score*) de empresa varejista. Para tanto, esta pesquisa exploratória e experimental foi viabilizada por meio da execução de experimentos embasados nas etapas do processo de descoberta de conhecimentos em bases de dados (KDD). Técnicas de redes neurais convolucionais foram aplicadas para a descoberta de conhecimento do cliente. Os principais resultados da pesquisa indicam que muitos comentários de clientes estão relacionais a determinados aspectos dos produtos e serviços da empresa, dentre os quais destacam-se: cartão, limite, pagamento, aumento e dificuldade de atendimento nos canais disponibilizados pela empresa. Quanto ao cruzamento dos atributos relativos ao perfil do cliente com os resultados de NPS foi possível identificar os comentários e principais argumentos segregados por gênero, faixa etária, nível de renda e localidade de domicílio de clientes com NPS detratores. Como conclusão da pesquisa é possível afirmar que a solução desenvolvida é capaz de proporcionar descoberta de conhecimento do cliente a partir de textos elaborados pelos mesmos. Em complemento, assevera-se ainda que a solução desenvolvida é capaz de auxiliar na gestão do cliente em empresas varejistas com grande volume de dados de clientes.

Palavras-chave: Inteligência artificial. Análise de sentimentos. Redes neurais convolucionais. Descoberta de conhecimento em bases de dados.

ABSTRACT

Customers exchange information, opinions and feelings daily on various topics linked to the products and services offered by companies. However, the data generated in these testimonials need to be analyzed to extract useful knowledge for the company to better understand the customer, in order to provide better service. Artificial Intelligence (AI) has methods and techniques applicable to the analysis of feelings expressed by customers in free texts with low structuring. Thus, the application of AI to discover knowledge in databases can help in understanding the feelings expressed by the client. The objective of this research is to apply techniques of convolutional neural networks for the analysis and classification of sentiments in customer comments aiming at discovering customer knowledge in a retail company. In addition, we also sought to compare the results of the experiments with the results of the detractor indicators of the NPS (Net Promoter Score) of a retailer. Therefore, this exploratory and experimental research was made possible through the execution of experiments based on the stages of the Knowledge Discovery Databases (KDD). Convolutional neural network techniques were applied for customer knowledge discovery. The main results of the research indicate that many customer comments are related to certain aspects of the company's products and services, where the following stand out: card, limit, payment, increase and difficulty in attending to the channels made available by the company. As for the crossing of attributes related to the client's profile with the NPS results, it was possible to identify the comments and main arguments segregated by gender, age group, income level and place of residence of clients with NPS detractors. As a conclusion of the research, it is possible to state that the developed solution can provide knowledge discovery of the client from texts elaborated by them. In addition, it is also asserted that the developed solution can assist in customer management in retail companies with a large volume of customer data.

Keywords: Artificial intelligence. Sentiment analysis. Convolutional neural networks. Discovery of knowledge in databases.

SUMÁRIO

Lista de Siglas

CNNs	Redes Neurais Convolucionais
QWST	Quest Manager
IA	Inteligência Artificial
LGPD	Lei Geral de Proteção de Dados
RNRs	Redes Neurais Recorrentes
SA	Análise de Sentimentos
KDD	Knowledge Discovery in Databases
ML	Machine Learning
NPS	Net Promoter Score

Lista de Tabela

Lista de Figuras

Lista de Quadros

1. Introdução

1.1 Contextualização do tema

Conhecer seus clientes é fundamental para o sucesso de qualquer negócio e empresa, isto porque conhecer características e hábitos do cliente pode proporcionar à empresa a adaptação de produtos e serviços às suas necessidades. Uma forma de viabilizar tal desafio é realizar continuamente a análise das opiniões dos clientes buscando assim extrair informações úteis para a geração de conhecimento sobre o cliente. Entretanto, avaliar a opinião dos clientes sobre os produtos e serviços da empresa não é uma atividade simples, segundo afirmam os autores Yano e Smith (2010), pois deve-se considerar o volume de dados gerados pelos clientes, principalmente em mídias sociais, bem como a complexidade de analisar opiniões expressas em forma de texto livre elaborado pelos clientes.

No entanto, é muito importante considerar a coleta de dados dos clientes para que a empresa possa entender melhor o que o cliente busca em seus produtos e serviços. Ravi e Ravi (2015) argumentam que os dados gerados pelos clientes que estão disponíveis em mídias e redes sociais em geral apresentam baixa estruturação, sendo, portanto, flexíveis e dinâmicos. Tal contexto dificulta a extração de informações para a geração de conhecimento útil para a tomada de decisões de gestores a respeito das adequações necessárias em produtos e serviços. Além disso, existem diversas redes sociais e plataformas na web, nas quais os usuários/clientes podem expressar suas opiniões sobre produtos, serviços, marcas e empresas.

Hoje em dia, diversas empresas aplicam a metodologia NPS (Net Promoter Score) para mensurar a satisfação do cliente, bem como a fidelidade das relações entre o consumidor e a organização. De acordo, com Reichheld (2011), o principal objetivo da metodologia NPS é definir uma pergunta simples que possa ajudar as organizações a construir relacionamentos e satisfação com seus clientes, de modo a proporcionar à empresa o gerenciamento do relacionamento com seus clientes e planejar estrategicamente seu crescimento. A aplicação da ferramenta NPS ilustra a relação diretamente proporcional entre qualidade do relacionamento/serviço e crescimento organizacional. Nesta metodologia, os consumidores respondem a

perguntas, numa escala de 0 a 10, cujas respostas sinalizam elementos detratores, neutros e promotores da satisfação do cliente.

As opiniões expressas nas redes sociais têm sido utilizadas por diferentes empresas para entender o que os consumidores e o público em geral pensam sobre produtos e serviços. Na visão de Liu (2012) e Pozzi et al. (2016), a mineração de sentimentos desenvolve a tarefa de avaliar opiniões, atitudes e emoções dos clientes. Nesse sentido, Hemmatian e Sohrabi (2017) afirmam que a análise de sentimento baseada na análise de documentos pode ter como resultado a classificação de um sentimento expresso pelo cliente como positivo, negativo ou neutro. Tal análise pode ser feita em nível de documento, frase, aspecto ou conceito. Portanto, é necessário aplicar métodos e ferramentas específicas para realizar análises de modo a extrair informações importantes que possam ser transformadas em conhecimento útil para apoiar a tomada de decisão da empresa.

Tal contexto assume ainda mais importância em diversas áreas, tais como economia, indústria, comércio e varejo. Uma das aplicações mais comuns volta-se à mensuração do NPS de uma empresa. Isto porque é especialmente valioso para pesquisa de mercado, monitoramento da presença digital de uma marca e atendimento ao cliente, dentre outras aplicações. Por meio de um processo de coleta e tratamento bem estruturado, as opiniões de clientes tornam-se um insumo inestimável que abre caminho para a fidelização do cliente e, consequentemente, para o fortalecimento da marca e das relações comerciais. (HUB, 2022).

Assim, verifica-se a importância da empresa realizar a mineração textos para viabilizar a análise de opiniões e sentimentos expressos por clientes em mídias sociais, visando assim a estruturação de informações sobre os clientes para a viabilizar a descoberta de conhecimento voltados a tomada de decisão em uma empresa do ramo varejista.

1.2 Problema de Pesquisa

Tendo como base o contexto apresentado até o momento, a seguinte questão de pesquisa é apresentada:

Como aplicar técnicas de Redes Neurais Convolucionais para a análise e classificação de sentimentos em comentários de cliente visando a descoberta de conhecimento de clientes em empresa varejista?

1.3 Objetivos da Pesquisa

Esta pesquisa tem como objetivo aplicar técnicas de Redes Neurais Convolucionais para a análise e classificação de sentimentos em comentários de clientes visando a descoberta de conhecimento do cliente em empresa varejista.

Em complemento, os seguintes objetivos específicos de pesquisa são indicados:

- Selecionar bases de dados de atendimento extraídas dos canais de atendimento ao cliente de empresa varejista;
- Executar o pré-processamento e transformação da base de dados de atendimento de empresa varejista;
- Aplicar a técnica de redes neurais convolucionais para a análise e classificação de sentimentos de comentários de clientes de empresa varejista;
- Avaliar e interpretar o conhecimento descoberto com base na técnica de redes neurais convolucionais;
- Comparar os resultados dos experimentos com os resultados dos indicadores detratores do NPS *(Net Promoter Score*) de empresa varejista.

Para atingir os objetivos acima indicados foi considerada uma base de dados com registros provenientes de central de atendimento aos clientes, bem como os resultados do NPS *(Net Promoter Score*) da empresa varejista abordada nesta pesquisa. Tais dados foram utilizados para a realização de experimentos computacionais baseados em métodos e técnicas para a análise e classificação de sentimentos. Os experimentos realizados visam desenvolver uma solução eficiente de descoberta de conhecimento útil para a tomada de decisões a respeito de produtos e serviços oferecidos aos clientes de empresas varejistas.

1.4 Justificativa da pesquisa

A inteligência artificial tem experimentado grande crescimento de aplicações voltadas a análise de grandes massas de dados relacionadas aos negócios das empresas contemporâneas. Mais especificamente, as técnicas de análise de sentimentos têm sido empregadas em questões que envolvem a gestão do negócio na empresa, relativamente à análise de textos escritos por clientes sobre produtos e serviços ofertados pela companhia (KAUFFMANN *et al.*, 2020).

A empresa varejista enfocada nesta dissertação possui milhões de clientes no país e conta com um banco de dados com dezenas de milhões de comentários que precisam ser avaliados para a extração de conhecimento acerca do perfil e comportamento de seus clientes. Tal volume de dados inviabiliza a leitura e interpretação por seres humanos de todos os comentários, buscando assim extrair a correta interpretação de seu teor, ou seja, se são comentários positivos, negativos ou neutros à empresa, seus produtos e serviços.

Além disso, os comentários incluídos no referido banco de dados não estão estruturados de forma a propiciar a extração e análise para fins de criação de novos conhecimentos a serem aplicados ao desenvolvimento e evolução do negócio, uma vez que são formados por textos que contêm erros ortográficos e incompatibilidades de sintaxe. Assim, dada a inviabilidade de avaliar esses comentários, dado seu volume e inconsistências, a empresa não consegue tratar e responder as reclamações dos usuários em tempo hábil de modo a responder prontamente o cliente. A atual abordagem da empresa não fornece uma visão fiel da qualidade dos serviços e produtos prestados pela empresa ao cliente, uma vez que esta não pode abranger todos os clientes e seus comentários.

Tal contexto justifica a aplicação de uma abordagem inteligente de análise de sentimentos em comentários extraídos em mídias sociais para avaliar a opinião de clientes sobre os serviços e produtos oferecidos pela empresa. Assim, técnicas de redes neurais foram aplicadas para analisar e classificar sentimentos em textos de clientes que expressam sua opinião sobre os produtos e serviços da empresa. A empresa varejista enfocada nesta dissertação possui uma grande quantidade de dados gerados pelos clientes em mídias sociais, central de atendimento, estandes em lojas físicas, aplicativos e portais disponibilizados pela empresa.

2. Referencial teórico

Este capítulo apresenta o referencial teórico para os tópicos abordados nesta dissertação. Primeiramente é apresentado o referencial teórico acerca do conhecimento sobre o cliente oriundo de mídias sociais e redes sociais como canais de comunicação com o cliente. Na sequência são apresentados os tópicos sobre descoberta de conhecimento em bancos de dados (KDD), Inteligência Artificial, Análise de Sentimentos, e por fim, a técnica de Redes Neurais Convolucionais (RNC).

2.1 Conhecimento acerca do cliente

O conhecimento do cliente é a informação que é organizada e analisada a fim de compreendê-la e aplicá-la para a resolução de problemas e tomada de decisões nas relações das organizações com os seus clientes, conforme afirma Buchnowska (2011). Segundo Gebert *et al.* (2002), o conhecimento acerca do cliente permitir uma exploração mais rápida das oportunidades apresentadas ao negócio, bem como a consequente resposta mais rápida às ameaças ambientais, tornando a organização mais competitiva.

O conhecimento é um importante gerador de valor para as empresas inovadoras, uma vez o conhecimento do cliente é reconhecido como um elemento importante para a vantagem competitiva das organizações e um fator chave na economia deste século (YIYI; RONGQIU, 2008; LAK; LYU *et al.,* 2009; SEDIGHI *et al.*, 2012; KHOSRAVI *et al.*, 2017; REZAEENOUR, 2017). As empresas contemporâneas buscam agregar valor aos clientes para melhor atender suas necessidades sendo, portanto, importante compreender o comportamento e as demandas do consumidor (TANG *et al.*, 2011; JIEBING *et al.*, 2013). Assim, o conhecimento acerca do cliente ganhou destaque na ciência com o surgimento de pesquisas em diferentes áreas científicas (FIDEL *et al.*, 2015; (KHOSRAVI *et al*., 2017; (XUELIAN *et al.*, 2015).

Conhecer bem seus clientes, segundo Pandey *et al.,* 2014, torna-se elemento vital para a empresa, uma vez que as organizações precisam desenvolver estratégias baseadas no conhecimento acerca do cliente para se manterem competitivas.

Segundo Valacherry e Pakkerappa (2018), o conhecimento acerca do cliente adiciona mais valor à proposta de valor a ser ofertada ao cliente. Khosravi e Hussin (2017), Fidel *et al.* (2015) e Taghizadeh *et al.* (2017) sinalizam que o conhecimento acerca do cliente possibilita que a empresa crie e sustente diferenciais competitivos frente aos concorrentes em seu mercado de atuação. Outra indicação da importância do conhecimento acerca do cliente para a empresa é expressa por Jaziri (2019), ao argumentar que esse conhecimento é um pré-requisito para a inovação.

No entanto, nem todas as empresas são capazes de gerenciar e utilizar esse conhecimento acerca do cliente para oferecer produtos/serviços diferenciados no mercado (XUELIAN *et al.*, 2015). Isto porque, o conhecimento que os gestores têm sobre seus clientes e suas necessidades muitas vezes está longe da realidade e, em razão disso, Daneshgar *et al.* (2012) e Khosravi e Hussin (2016) argumentam que conhecer os clientes é visto com algo muito desafiador para empresas e gestores.

Bo e Ying-Jiao (2009), Chen *et al.* (2009), ZanjaniI *et al.* (2008), Semmelrock-Picej e Kandutsch (2010); e Xu (2011) enfatizam que o conhecimento acerca do cliente tem sido severamente negligenciado pelas organizações, mesmo por aquelas que têm se concentrado cada vez mais na gestão do conhecimento organizacional. No entanto, essa realidade mudou nos últimos anos, pois o conhecimento do cliente passou a desempenhar papel crucial no desenvolvimento de novos produtos e serviços nas empresas modernas (LYU *et al.*, 2009; MA; QI, 2009; XU, 2011; SACHAMANOROM *et al.*, 2016).

A grande quantidade de dados armazenados pelas empresas vem aumentando significativamente, e com isso os problemas de como melhor tratar esses dados armazenados a partir das transações e interações realizadas cotidianamente com o cliente. Atualmente é imprescindível que os gestores, administradores ou até mesmo colaboradores de uma empresa conheçam os resultados gerados a partir da prestação de serviços e disponibilização de produtos da empresa ao cliente. Para Funchal, Madsen e Adamatti (2015), a mineração de dados é uma técnica de descoberta de conhecimento que se refere a grandes quantidades de dados que podem estar ocultos em bancos de dados corporativos. Segundo os autores, a mineração de dados é cada vez mais uma das soluções utilizadas por empresas de médio e grande porte para extrair conhecimento de grandes quantidades de dados do cliente.

Segundo Hollanda (2019), o conhecimento do cliente é um obstáculo muito comum às organizações contemporâneas, pois a maioria das empresas não consegue desenvolver esta competência antes de desenvolver uma estratégia de experiência do cliente, o que acaba tornando iniciativas nesse sentido inócuas, por se basearem apenas em suposições da empresa. Assim, as empresas têm muitas oportunidades de conhecer seus clientes, mas não conseguem obter *insights* profundos sobre seus clientes pela forma inadequada de extrair informações das bases de dados para a construção de conhecimento acerca do cliente.

Para Aacinol (2019), conhecer seus clientes é a melhor maneira de apresentar a marca da empresa a eles. Afinal, é esse conhecimento do cliente que garante que a empresa saiba exatamente o que comunicar para convencer seu público-alvo, razão pela qual é importante que a empresa saiba como seus clientes preferem ser atendidos. Em outras palavras, a empresa precisa mostrar ao cliente que pode ajudá-lo com o que ela oferece para resolver um problema que ele tem. Mais precisamente, a empresa precisa saber para quem está oferecendo seus produtos e serviços, ou seja, conhecer o perfil e características de seus clientes atuais e potenciais (AACINOL, 2019).

2.2 Mídias sociais e redes sociais como canais de comunicação com o cliente

Nos últimos anos, as mídias sociais tornaram-se uma nova parte das ferramentas de marketing à disposição das empresas, permitindo-lhes construir novas formas de se relacionar com os seus clientes. Segundo Seller e Laurindo (2018), as mídias sociais abrem um canal de comunicação bidirecional entre a organização e seus consumidores, além de estimularem a interação entre eles. Os autores argumentam ainda que as comunidades de marca têm um efeito positivo no cliente/produto, cliente/marca, cliente/empresa e entre os clientes, o que por sua vez tem um efeito positivo na confiança do cliente na empresa. Em função disso, os relacionamentos se tornaram muito mais simplificados e ágeis por meio do uso das mídias sociais para promover conexão entre empresa e clientes.

De acordo com Oliveira e Bermejo (2017), o compartilhamento de texto, imagens, áudio e vídeo, muitas vezes utilizados pelas organizações para fins

comerciais, são intensamente utilizados pela empresa, quanto pelos clientes. As mídias sociais e redes sociais são termos comumente presentes no cotidiano da sociedade, mas há confusão quanto às suas definições. Tanto as mídias sociais quanto as redes sociais são termos que existiam antes do advento da internet, mas com a introdução das plataformas digitais e a interatividade por elas proporcionada, esses termos foram adaptados para a web.

Segundo Drude (2021), um dos meios de comunicação eletrônica mais populares nos últimos tempos são as mídias sociais, que englobam uma ampla gama de diferentes formas de tecnologia de comunicação as mídias sociais são sites e aplicativos de comunicação que conectam pessoas ao redor do mundo. Essas plataformas permitem que os usuários criem, editem e compartilhem conteúdo eletrônico.

De acordo com Hazarika, Mousavizadeh e Tarn (2019), as mídias sociais visam melhorar o relacionamento das empresas com seus consumidores e, principalmente, aumentar a visibilidade da empresa, melhorar a divulgação de seus produtos e serviços, ajudar o estabelecimento a levar conteúdo relevante aos clientes e à população para que a empresa esteja cada vez mais próxima e aproxima mais de seus consumidores e atinge seus objetivos. Portanto, muitas empresas utilizam aplicativos móveis como canais de negócios tradicionais.

As mídias sociais também possibilitaram o surgimento de redes sociais que apresentam textos inseridos em conversas e depoimentos de clientes acerca de produtos e serviços de empresas. Segundo Recuero (2011), tais plataformas deixam rastros de interação entre os usuários, tais como comentários positivos ou negativos, entre outros, e estabelecem interação entre clientes e empresas. Na visão de Telles (2010, p. 19), as mídias sociais são sites na Internet que possibilitam a criação de conteúdo colaborativo (participação), interação social (relacionamento entre pessoas) e troca de informações em múltiplos formatos. Ainda segundo o autor, as redes sociais são “ambientes que focam em aproximar as pessoas, ou seja, construir redes (redes de amizade) entre os participantes”, o que pode incluir clientes de empresas com interações a respeito de seus produtos e serviços.

Segundo Aggarwal e Zhai (2012), as redes sociais são uma fonte muito comum de texto na web, pois possibilitam que atores humanos se expressem e se comuniquem de forma rápida e livre sobre os mais diversos tipos de temas. Os

recursos disponibilizados pelas redes sociais subsidiam a criação e troca de grande quantidade de dados entre clientes e empresas.

De acordo com Statista (2020), com a popularização do acesso à Internet, as redes sociais estão entre as plataformas mais importantes em número de usuários. Pesquisas realizadas em abril/2020 mostram que a principal rede social, o Facebook, tem cerca de 2,5 bilhões de usuários. Já o Twitter tem cerca de 386 milhões de usuários ativos, incluindo quase 14,5 milhões de usuários no Brasil. Segundo o Twitter (2020), desde quando começou a pandemia de Covid-19 as redes sociais estão sendo cada vez mais utilizadas. Em relatório emitido pela empresa verificou-se aumento de 24% ano a ano nos usuários ativos monetizáveis, sendo o maior aumento registrado anualmente e um crescimento de 14% em relação ao trimestre anterior.

Entretanto, a geração de grandes quantidades de dados, apenas, não é suficiente para a empresa compreender melhor seu cliente. Isto porque tais dados em seu formato bruto ou visualizados de forma individualizada não são capazes de gerar descoberta de conhecimento para a empresa. Não obstante, o conjunto desses dados tem enorme potencial para geração de conhecimentos úteis. Para tanto, é necessário realizar análises para extrair informações importantes que possam ser transformadas em conhecimento útil para apoiar o processo decisório da empresa em prol de adequação da proposta de valor a ser ofertada ao cliente. Silva *et al.* (2013) afirmam que essas análises podem ser realizadas por meio de técnicas de mineração de dados, que incluem um conjunto de técnicas baseadas em modelos capazes de encontrar padrões, resumir dados, extrair novas descobertas de conhecimentos a partir de grandes quantidades de dados.

2.3 Descoberta de conhecimento em bases de dados

Shapiro *et al.* (1994) afirma que a descoberta de conhecimento em bases de dados seja a extração de dados não triviais que anteriormente eram desconhecidas, criando assim informações potencialmente úteis à empresa. De acordo com Zhong *et al.* (1997), o processo de descoberta de conhecimento em bases de dados, também conhecido por KDD (*Knowledge Discovery in Databases*) se dá diferentes níveis de análise e etapas. Assim, diversas técnicas alternativas podem ser aplicadas no

processo e as iterações podem ser repetidas em diferentes intervalos a partir dos dados à medida que são atualizados a cada fase.

Nos últimos anos, as técnicas de aprendizado de máquina KDD (*Knowledge Discovery in Database*) ou descoberta de conhecimento em bases de dados - têm sido usadas em vários campos. O uso de bancos de dados aumentou com o advento da Internet e, assim, informações que antes eram apenas armazenadas agora são utilizadas para descobertas com aplicação de análise de dados. De acordo com Fayyad *et al.* (1996):

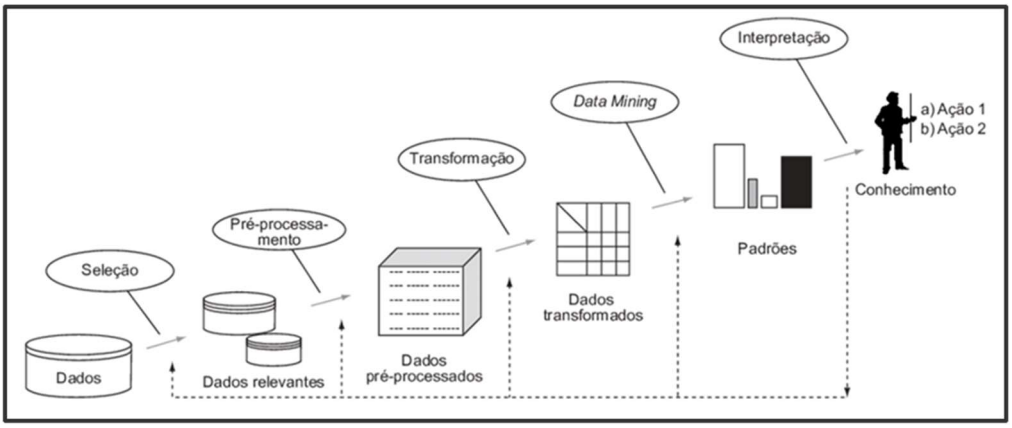
> A descoberta de conhecimento em bancos de dados (KDD) tem como principal objetivo a extração de conhecimento de alto nível de grandes bancos de dados, e de pequenas bases com baixo nível, que consiste em um processo global de descoberta de conhecimento (FAYYAD et al., 1996, p. 39).

Goldschmidt e Passos (2005) afirmam que a aplicação de KDD possui três tipos básicos de componentes: o problema ao qual o processo KDD é aplicado, os recursos disponíveis para resolver o problema e, por fim, os resultados obtidos pela aplicação dos recursos disponíveis na busca de uma solução para o problema. Segundo Freitas (2000), o conhecimento a ser descoberto deve atender a três propriedades: deve ser (tanto quanto possível) correto; deve ser compreensível para usuários humanos e; deve ser interessante, útil e novo. Além disso, ainda segundo o autor, o método de descoberta de conhecimento KDD deve ter as três características a seguir: deve ser eficaz (preciso), genérico (aplicável a diferentes tipos de dados) e flexível (facilmente modificável).

Segundo Fayyad *et al.* (1996) existem vários níveis de análise e diferentes etapas e técnicas de aplicação que podem ser utilizadas no processo de KDD para extração de conhecimento em bancos de dados sem restrições. Ou seja, de todos os domínios possíveis que possuem um conjunto de dados, o KDD se torna uma ferramenta importante para analisar dados de diferentes indústrias e setores.

O processo de KDD, conforme indicado por Fayyad *et al.* (1996), procura encontrar conhecimento em bases de dados, sendo composto pelas seguintes etapas: i) seleção de dados, ii) pré-processamento e limpeza de dados, iii) transformação de dados, iv) mineração de dados e, por fim, v) interpretação e avaliação dos dados. O processo de KDD de Fayyad *et al.* (1996) é ilustrado na Figura 1.

Figura 1 – Processo de KDD



Fonte: Fayyad *et al.* (1996).

As etapas do processo de descoberta de conhecimento (KDD) proposto por Fayyad *et al.* (1996) são:

1) Seleção de dados: visa criar ou selecionar um conjunto de dados de interesse para ser usado no processo de descoberta de conhecimento. É necessário que o conjunto de dados selecionado contenha variáveis previsíveis para o desenvolvimento da análise para que possam levar a um padrão. É importante que essa fase seja realizada com cuidado para garantir um banco de dados de alta qualidade, seja um conjunto de dados existente ou um banco de dados criado. Segundo Ristoski e Paulheim (2016), nesta fase inicial é importante selecionar apenas atributos que sejam relevantes e correspondam aos objetivos da empresa. O subconjunto selecionado é então fornecido ao algoritmo de mineração de dados para que sejam transformados de modo a torná-los mais compreensíveis;
2) Pré-processamento: consiste em usar técnicas para preparar o banco de dados para análise. Esta etapa, que também inclui limpeza de dados, é responsável por filtrar os dados dentro do conjunto de dados alvo para eliminar o ruído encontrado no banco de dados e, assim, selecionar apenas os dados relevantes para a próxima etapa. Segundo (Tenfen, 2003), o processo nesta etapa de limpeza acaba eliminando consultas desnecessárias que o algoritmo faria na fase de mineração de dados, como atribuir valores limite em um banco de dados para que valores desnecessários sejam deixados de lado. Para Phridviraj e Gururao

(2014), o processo de limpeza de dados conhecido como eliminação de ruído ou redução de ruído pode ser realizado utilizando-se diferentes técnicas inteligentes;

3) Transformação de dados: também chamada de redução e projeção de dados. De acordo com Fayyad *et al.* (1996), nesta fase são os recursos úteis para representar os dados, bem como as representações invariantes considerando o objetivo final do processo. Na visão de Panwar e Raiwani (2014), as vantagens desse processo é que os resultados são apresentados de forma compacta e de fácil compreensão, observando-se padrões gerais. Já a desvantagem dessa transformação de dados é a perda dos dados originais, que é um processo irreversível. Ainda conforme os autores, a redução de dados pode reduzir custos e aumentar a eficiência do armazenamento. Existem três tipos de estratégias que podem ser usadas para redução de dados: redução de dimensão, *clustering* (agrupamento) e amostragem;
4) Mineração de dados: Castro e Ferrari (2016) afirmam que a fase de mineração de dados inclui a etapa de exploração da base de dados pré-processada. A mineração de dados é responsável por extrair informações da base utilizando algoritmos apropriados e técnicas específicas ao objetivo da análise. Segundo Jothi, Rashid e Husain (2015), a mineração de dados é considerada a fase mais importante do processo, uma vez que será capaz de extrair efetivamente o conhecimento implícito e útil da base de dados considerada. Assim, nesta fase é realizada uma busca concreta por conhecimento útil no contexto da aplicação KDD. Nela são definidas as técnicas e algoritmos a serem aplicados ao problema em questão. A escolha da técnica depende, muitas vezes, do tipo de tarefa do KDD que será realizada, conforme afirmam Piatetsky-Shapiro, Matheus e Chan (1993); Linoff e Berry (2011 e Peres e Boscarioli (2016);
5) Interpretação e avaliação de dados: visa selecionar modelos que sejam válidos e úteis para a tomada de futuras decisões de negócios. Existem diferentes métricas de avaliação de desempenho, para isso dependendo do tipo de tarefa de mineração de dados que você está aplicando. Segundo Ristoski e Paulheim (2016), a interpretação e avaliação do conhecimento visa garantir um bom entendimento do conhecimento descoberto pelo algoritmo de mineração aplicado, além de validá-lo por meio de medidas da qualidade da solução e da percepção do analista de dados sobre a medição do conhecimento recém-descoberto. Esta etapa também pode incluir a visualização dos padrões e

modelos extraídos, bem como a visualização dos dados aplicando-se tais padrões e modelos.

Fayyad, Piatetsky-Shapiro e Smith (1996) afirmam que existem várias maneiras de interpretar os dados do KDD, a partir da execução de diferentes tarefas. As mais comuns são a seguir indicadas:

a) Regras de associação: são consideradas importantes não só para associações triviais, mas também para associações não óbvias, as quais jamais se imaginaria a existência de uma relação, que se tornam uma importante fonte de informação na tomada de decisão (WITTEN *et al.,* 2016);
b) Classificação: o banco de dados aprende uma função para mapear e classificar classes, a partir da qual um diagrama pode ser elaborado para indicar a posição dos dados que define sua classe, por exemplo, traçando uma linha linear simples (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007);
c) Clusterização/agrupamento: é uma tarefa descritiva que tenta identificar um conjunto finito de categorias ou grupos comuns, para descrever os dados, com instâncias naturalmente agrupadas com base nos valores de seus atributos. A clusterização transforma conjuntos de dados com muitos atributos em conjuntos relativamente menores (BRAGA; CARVALHO; LUDEMIR, 2011);
d) Visualização de dados: facilita o entendimento e é muito utilizada na mineração de dados. Existem diferentes tipos de gráficos para a visualização de dados (WITTEN *et. al.,* 2016). A representação gráfica auxilia na visualização para a interpretação dos dados, contribuindo assim para estudos primários, além de auxiliar na tarefa de mineração de dados e ajudar na interpretação destes (LI *et al.,* 2016).

2.4 Inteligência Artificial

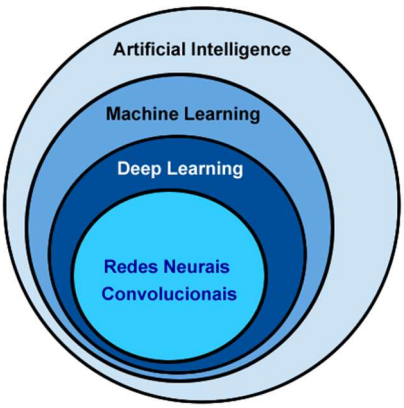
Rich e Knight (1991) afirmam que o objetivo da Inteligência Artificial (IA) é desenvolver sistemas para executar tarefas que são melhor executadas por humanos do que por máquinas ou tarefas para as quais nenhuma solução algorítmica por

computadores tradicionais seja possível. De acordo com Fernandes (2003), o termo inteligência artificial vem do latim *inter* (entre) e *legere* (escolher). O autor indica que é a inteligência que permite ao homem poder escolher entre uma coisa e outra, sendo a IA a maneira de resolver problemas em tarefas complexas. Assim, a IA é vista como um tipo de inteligência que é produzida por humanos para promover às máquinas um tipo de habilidade que simula a inteligência natural dos humanos.

A partir da década de 1980, sistemas de inteligência artificial foram desenvolvidos com base em sistemas especialistas, que abriram as portas a diminuição de custos nas empresas. Nos últimos anos, a inteligência artificial amadureceu a ponto de ser elevada a ciência em 1987, voltada à melhor compreensão dos problemas e sua complexidade. É importante diferenciar a inteligência artificial de outros termos largamente associados a ela em diferentes contextos, como aprendizado de máquina e aprendizado profundo. Simplificando, o aprendizado de máquina (*Machine Learning* - ML) é uma subárea da IA, assim como o *Deep Learning* é uma subárea do aprendizado de máquina.

O Deep Learning em si não é uma técnica, mas uma rede neural de várias camadas com algumas especializações. Algumas técnicas tiveram resultados muito promissores, como as Redes Neurais Convolucionais, que perfazem um algoritmo de aprendizado de máquina. Isso fica mais claro na Figura 2, que ilustra a relação das diferentes disciplinas de IA e mostra que o *Deep Learning* é um tipo de aprendizado de máquina usado por muitas abordagens (mas não todas) de Inteligência Artificial, conforme indicado por Kenji (2019).

Figura 2 – Relação ente a inteligência artificial, aprendizado de máquina e aprendizado profundo



Fonte: adaptado de Kenji (2019).

De acordo com o autor Kenji (2019), a Inteligência Artificial explora maneiras de simular a inteligência humana. A Inteligência Artificial possui diferentes subcampos, tais como a Visão Computacional, Processamento de Linguagem Natural (PNL) e Aprendizado de Máquina.

Segundo Willian *et al.* (2017) e Hariri (2019), a inteligência artificial é utilizada para apoiar e melhorar a qualidade da tomada de decisão e resolução de problemas de empresas. Para tanto, são utilizadas várias técnicas para fornecer informações baseadas em grandes quantidades de dados, incluindo Mineração de Textos, Processamento de Linguagem Natural e Aprendizado de Máquina, sendo que todas consideram técnicas de classificação, dentre as quais destaca-se as redes neurais artificiais.

Segundo Ludermir (2021), as técnicas de IA mais largamente bem-sucedidas exigem maior poder computacional e elevado volume de dados, possibilitando assim que a IA consiga resolver problemas cada vez mais complexos no ambiente de negócios, como é o caso da análise de sentimentos de opiniões de clientes, foco desta pesquisa. Bianchi (2020) argumenta que o aprendizado de máquina é um ramo da Inteligência Artificial que se baseia na ideia de que os sistemas podem aprender com dados, reconhecer padrões e tomar decisões com pouca intervenção humana. Segundo o autor, esse tipo de construção de conhecimento conta com a elaboração de algoritmos que aprendem com seus erros e fazem previsões sobre dados de diferentes origens e abordagens.

De acordo com Waltrick (2020), os algoritmos de aprendizado de máquina podem ser divididos em três grupos principais: aprendizado supervisionado, aprendizado não supervisionado e por reforço. Em seguida são apresentados cada um desses algoritmos de IA, conforme Waltrick (2020):

- **Aprendizado supervisionado:** algoritmos que fornecem às máquinas os dados e rótulos esperados. O modelo deve ser capaz de classificar corretamente os dados. Um conjunto de dados rotulado é fornecido ao modelo para aprendizado, com a indicação de cada classe/categoria, sendo estes dados divididos em partes para treinamento e teste;
- **Aprendizado não supervisionado:** algoritmos nos quais se espera que a solução detecte tendências nos dados sem identificação prévia. Esse modelo aprende a realizar uma tarefa a partir de dados não rotulados (sem resultado conhecido) com base apenas em suas propriedades e padrões semelhantes.

Ou seja, o modelo não supervisionado deriva estruturas de uma amostra do problema, adequado em situações com muitas observações e muitas funções, tais como áudios, imagens e vídeos;

- **Aprendizagem por reforço:** algoritmos que ensinam o modelo, dando reforços positivos para comportamentos esperados e/ou reforços negativos para comportamentos indesejáveis. Nesse tipo de aprendizado, o paradigma muda em relação aos outros dois anteriores. Geralmente é aplicado quando as regras são conhecidas, mas a melhor ordem de ações a serem tomadas não está identificada.

Ludermir (2021) argumenta que aplicar aprendizado de máquina para resolver problemas nem sempre é fácil, pois é necessário um bom conjunto de exemplos (dados). A base de padrões geralmente precisa ser constantemente construída e atualizada. Como os dados nem sempre são adequados, se faz necessário empregar técnicas de Inteligência Artificial para melhorar a qualidade dos dados sendo, portanto, muito importante escolher a melhor técnica a ser utilizada em cada situação específica.

2.5 Análise de Sentimentos

A Análise de Sentimentos é a tarefa de reconhecer, extrair e classificar opiniões, sentimentos e atitudes sobre diversos temas expressos em textos. Segundo Gonçalvez *et al.* (2012), a análise de sentimentos tem como objetivo extrair informações úteis sobre publicações com a finalidade de analisar textos e identificar a opinião do usuário. Conforme indicado por Tebaldi (2019), se a empresa deseja entender o que os clientes estão falando sobre ela e qual a reputação de seus produtos e serviços, umas das formas de se fazer isso é por meio do aprendizado de máquina. Isto porque o objetivo desta técnica é classificar sentenças (ou um conjunto de sentenças) dos clientes como positiva, negativa ou neutra.

Análise de Sentimentos (SA) visa identificar as opiniões postadas pelos usuários na web, pois a quantidade de comentários no ambiente digital é muito grande, sendo necessário e muito útil resumir os diferentes comentários disponíveis

em um simples resultado. Tal expediente economiza tempo e ajuda os usuários a tomar decisões com mais confiança, conforme indicado por Santos (2014). Além disso, a análise de humor avalia e classifica as opiniões como positivas, negativas ou neutras, disponibilizando assim o resultado ao usuário da empresa que utiliza análise de sentimentos (LIU, 2012).

Análise de Sentimentos é um problema cada vez mais desafiador, pois além de apresentar múltiplos subproblemas, requer técnicas cada vez melhores que ensinem programas de computador a identificarem efetivamente opiniões e sentimentos sobre as entidades envolvidas num texto, de acordo com Ravi e Ravi (2015). Assim, muitas empresas têm valorizado este tipo de análise de opinião, não só porque é mais rentável, mas também porque deposita grandes exigências nos resultados como, por exemplo, aceitação de um produto, serviço ou imagem própria. Desta forma, os resultados são recolhidos sem que seja necessária a realização de entrevistas junto aos autores dos comentários, o que facilita a coleta e interpretação das opiniões postadas sobre um tema específico.

Antes de tudo é necessário coletar os dados dos textos produzidos pelos clientes da empresa, visando assim iniciar um processo de mineração desses dados. Os dados coletados podem vir de diferentes fontes, devendo ser armazenados e analisados posteriormente. Na etapa de pré-processamento, os dados são verificados, utilizando-se algoritmos para padronizar as palavras e termos presentes nos textos produzidos pelos clientes. De acordo Santos (2014), a etapa de reestruturação de palavras pode ser facilmente corrigida usando um dicionário. A remoção dessas palavras reduz a dimensionalidade do problema, diminui o tempo de resposta dos algoritmos de classificação e aumenta seu desempenho.

Liu (2012) indica três níveis que abrangem a análise de sentimentos: nível de documento, de sentença e de entidade e aspectos. Segue abaixo a descrição de cada nível:

- Nível de documento: tem como objetivo classifica se o um documento de opinião expressa um sentimento positivo ou negativo. Este nível de análise pressupõe que o documento como um todo se refere a apenas uma entidade.
- Nível de sentença: tem como objetivo analisa as sentenças de um documento separadamente e as classifica individualmente como positivos, negativos ou neutros.

- Nível de entidade e aspecto: tem como objetivo realizar uma análise mais detalhada. Este nível difere dos demais por não se basear em documentos, parágrafos ou frases, mas se concentrar na opinião real, ou seja, baseada na medida em que a opinião é formada por sentimento positivo ou negativo.

Ainda segundo (Liu, 2012), existem diferentes tipos de opiniões, sendo possível classificá-las de acordo com a forma como são apresentadas no texto. As opiniões também podem ser classificadas como opiniões regulares, comparativas, explícitas ou implícitas, conforme explicado a seguir:

- Opinião regular: é muitas vezes referida como opinião literária e se divide em dois subtipos: opinião direta e indireta;
- Opinião comparativa: apresenta uma lista de diferenças e semelhanças entre duas ou mais entidades ou com base em algum aspecto da entidade;
- Opinião explícita: é fácil de entender, ou seja, fica claro qual é a intenção do texto. É uma declaração subjetiva que dá uma opinião regular ou comparativa;
- Opinião implícita: é uma declaração objetiva que implica uma opinião justa ou comparativa.

De acordo com Ohashi (2019), a análise de sentimento é a ferramenta de classificação de texto comumente empregada para analisar uma mensagem recebida e relatar seu sentimento implícito. Com o reconhecimento de emoções do cliente é possível obter informações mais detalhadas que mostrem o que o usuário realmente está sentindo naquele exato momento ao postar ou comentar sobre um produto e serviço.

Tebaldi (2019) afirma que a análise de sentimentos auxilia os especialistas a analisar dados de grandes empresas para avaliar a opinião pública, permitindo assim que as empresas realizem pesquisas de mercado, monitorem a reputação de marcas, produtos, serviços para melhor compreender seus clientes.

A análise de sentimentos é uma tecnologia valiosa, especialmente para as empresas contemporâneas inseridas num mundo com grande geração de dados a todo o momento. Isto porque a empresa recebe *feedbacks* de seus clientes de forma imparcial ou relativamente menos tendenciosa em função das técnicas de análise de sentimentos aplicadas. Quando feito corretamente, a análise de sentimentos pode

proporcionar criação de valor para uma organização, além de ser capaz de fornecer fatos e dados mensuráveis para tomada de decisões futuras em prol do negócio (TIBCO, 2022). Para tanto, há de se considerar técnicas de redes neurais convolucionais para a análise de sentimentos de clientes, conforme abordado nos próximos tópicos.

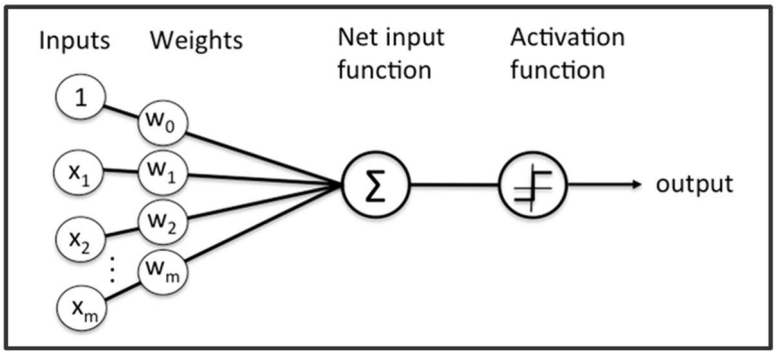
2.6 Redes Neurais

As Redes Neurais são um dos algoritmos de aprendizado de máquina mais populares, porque são modelos computacionais inspirados no sistema nervoso do ser humano. Eles são projetados para ajudar as pessoas a resolver muitos problemas complexos em uma variedade de situações do mundo real (BARONI; ITANIMÁ, 2021).

O aprendizado de rede neural permite aprender e modelar relacionamentos entre entradas e saídas de dados não lineares e complexos, estruturados ou não, bem como imagens, além de séries temporais (MJV Team, 2021).

As Redes Neurais são criadas a partir de algoritmos projetados para um propósito específico, conforme indicado por Baroni e Itanimá (2021). As camadas são compostas de nós. Um nó é apenas um lugar onde os cálculos, vagamente padronizados em um neurônio no cérebro humano que dispara quando recebe estímulos suficientes. Um nó combina os dados de entrada com uma série de coeficientes (ou pesos) que reforçam ou atenuam essa entrada, atribuindo assim significado às entradas em relação à tarefa que o algoritmo está tentando aprender, de acordo com Pathmind (2019). A Figura 3 ilustra um exemplo de estrutura de nós.

Figura 3 - Diagrama exemplo de um nó



Fonte: Pathmind (2019).

De acordo com Pathmind (2019), as Redes Neurais devem reconhecer padrões e interpretar dados sensoriais por meio de algum tipo de percepção de máquina, rotulagem ou agrupamento de dados brutos. Os padrões que eles reconhecem são numéricos, contidos em vetores nos quais todos os dados da vida real, sejam imagens, som, texto ou séries temporais, devem ser traduzidos.

Existem diferentes tipos de redes neurais profundas e cada uma delas tem vantagens e desvantagens. (Herhold, 2022) indica os seguintes tipos de redes neurais:

- **Redes Neurais Convolucionais (CNNs):** possuem cinco tipos de camadas: de entrada, convolucional, cluster, camadas completamente conectadas e de saída. As Redes Neurais Convolucionais podem ser usadas na classificação de imagens e detecção de objetos, bem como podem ser aplicadas em outras áreas, tais como previsão e processamento de linguagem natural;
- **Redes Neurais Recorrentes (RNRs):** são utilizadas informações sequenciais, como dados de carimbo de data/hora de um sensor ou uma frase falada. O RNRs é utilizado em aplicações e previsões de séries temporais, análise de sentimentos e aplicações de texto;
- **Redes Neurais Feedforward:** cada perceptron em uma camada está conectado a todos perceptron da próxima camada. As informações são entregues antecipadamente de uma camada para a próxima e avançam. Não há ciclos de feedback;

- **Redes Neurais de Autoencoder:** usadas para criar abstrações chamadas de codificadores que são criadas a partir de um conjunto fixo de entradas. Os autoencoders tentam modelar as próprias entradas e, portanto, o método é considerado não supervisionado. Essas abstrações podem então ser usadas por classificadores lineares ou não lineares.

Conforme ilustrado na Figura 3, a rede neural aprende os padrões para cada tipo de classe e, quando um padrão desconhecido é fornecido à rede, ela pode determinar a qual classe ela pertence. Para tanto, há um algoritmo de treinamento chamado *Backpropagation*. Este algoritmo se baseia no cálculo do erro ocorrido na camada de saída e em uma rede neural, recalculando assim o valor dos pesos do vetor 'W' da última camada e dos neurônicos. Desta forma, o algoritmo atualizar de trás para frente todos os pesos 'W' das camadas da última camada até atingir a camada de entrada da rede, realizando então a propagação reversa do erro recebido da rede (LEITE, 2018). Na próxima seção são expostas as redes neurais convolucionais consideradas para conduzir os experimentos desta pesquisa.

2.6.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais são arquiteturas que podem ser treinadas e são imutáveis em relação ao dimensionamento, translação e transformações associadas. As redes convolucionais consistem nas seguintes camadas, segundo Mousser (2019):

a) **Camada convolucional:** usa filtros sobre imagem para obter uma série de mapas de características, sendo uma para cada filtro. Tendo a função de aplicar máscaras às imagens de entrada com base em pixels. É possível obter filtros de convolução (matrizes), armazenando assim os pesos das conexões entre os neurônios (SANTOS, 2017);
b) **Camada de agrupamento:** responsáveis por reduzir a amostra para esses mapas de características, reduzindo assim a largura e altura. O *clustering* na maioria das arquiteturas de convolução usa as funções de *Max Pooling* e

Avarage Pooling, que são capazes de determinar o valor máximo e médio de agrupamento;

c) **Camada totalmente conectada:** responsável pela classificação dos dados. Uma função determina a identificação de saídas em classes. As funções mais comumente usadas são a *Softmax* (problemas múltiplos classes) e a *Sigmoid* (problemas binários).

Segundo Alves (2018), a ideia de uma rede neural convolucional é filtrar linhas, curvas e bordas para transformar essa filtragem em uma imagem mais complexa a cada camada adicionada. A arquitetura e o funcionamento das redes convolucionais são semelhantes à linguagem natural aplicada ao tratamento de imagens podendo ser aplicada também em tarefas de classificação de texto como a análise de sentimento, categorização e reconhecimento, dentre outras possibilidades (CARNEIRO, 2020). A seguir são apresentados os conceitos mais comuns e como funcionam as redes neurais aplicadas aos textos de opiniões de clientes enfocados nesta pesquisa.

Primeiramente, os textos devem ser convertidos em matrizes numéricas, pois a operação de convolução trabalha com múltiplas matrizes com sinal de entrada (imagem ou texto) e um *kernel* (filtro). Portanto, além de converter as palavras em números como é usual em tarefas de aprendizado de máquina que envolvam variáveis categóricas, também é necessário criar um *array* de matrizes para elas (CARNEIRO, 2020). As técnicas mais usadas são *Enconding* e *Word Embedding*, que são explicadas a seguir.

Encoding é amplamente utilizado no processo de pré-processamento para transformar palavras em números. Encoding também é conhecido como variável de Dummy, que identifica a palavra pela posição do número 1 e 0, conforme exposto na Tabela 1, que ilustra o exemplo de uma representação de *encoding*.

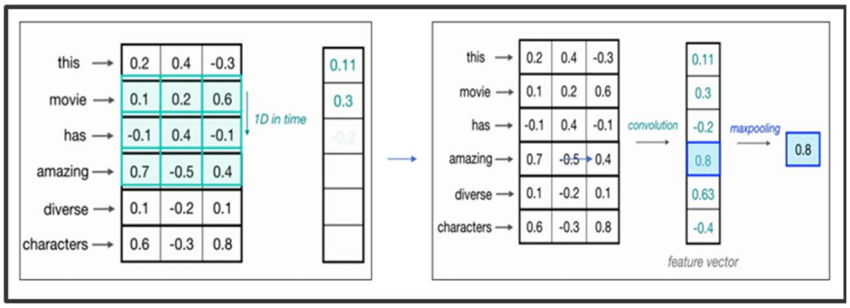
Tabela 1 – Representação de encoding

	VERMELHO	AMARELO	PRETO
VERMELHO	1	0	0
AMARELO	0	1	0
PRETO	0	0	1

Fonte: (Carneiro, 2020).

Como exposto na Figura 4, a matriz em azul que desliza de cima para baixo na imagem é chamada de kernel (filtro), que é basicamente uma matriz menor que o conjunto com valores iniciais aleatórios.

Figura 4 – Matriz de palavras e o processo de convolução



Fonte: Carneiro (2020).

Para Carneiro (2020), os *kernels* permitem que a rede neural atualize seus valores para extrair melhor os recursos da matriz de entrada, assim como as redes neurais convolucionais clássicas para imagens. A camada de convolução na qual o aplicativo do *kernel* é executado é chamada de *1D-conv*, porque a saída é uma matriz unidimensional, em oposição à convolução para imagens, em que a saída é uma matriz *2D-conv* (CARNEIRO, 2020). Sharma (2020) indica que o processo de saída (*output*) com o mesmo número de linhas que a matriz de entrada (*input*) pode ser chamado de vetor de características, pois a aplicação dos *kernels* visa extrair e sumarizar as principais características de encontradas em uma matriz. Segundo o autor, o *Max Pooling* visa extrair o maior valor do vetor de traços, que é basicamente a representação da palavra mais proeminente na frase.

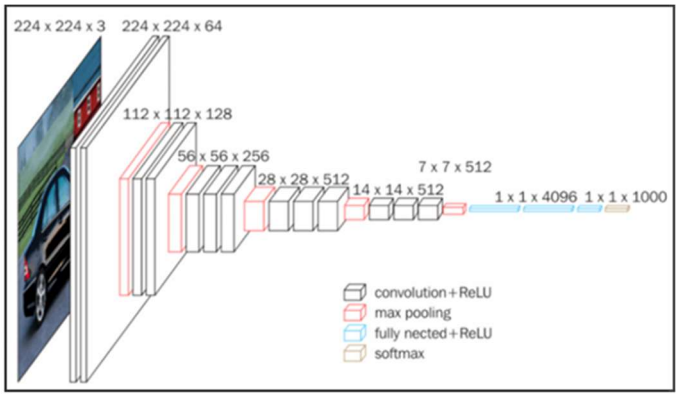
Carneiro (2020) afirma que uma das vantagens das redes neurais convolucionais é justamente que os *kernels* se adaptam durante as iterações para identificar automaticamente as principais características das sentenças e assim resumir toda a matriz de um *corpus* em um único valor. A seguir veja como funciona uma arquitetura de Redes Neurais Convolucionais.

As CNNs têm a propriedade de ter altos requisitos computacionais e de memória devido às suas convoluções sucessivas e, portanto, muitas arquiteturas de CNN foram propostas desde Lenet (LECUN *et al.*, 1998) e AlexNet (KRIZHEVSKY *et*

al., 2012) para pesquisar otimizações a serem buscadas, como maior precisão, menor tempo de processamento e menos requisitos de armazenamento. As principais arquiteturas de CNNs são descritas a seguir, cada uma com suas características:

- **VGG**: é uma rede neural convolucional com variações no número de camadas, com profundidade de 16-19 camadas, com 3 camadas densas no final da rede. (SIMONYAN; ZISSERMAN, 2014).

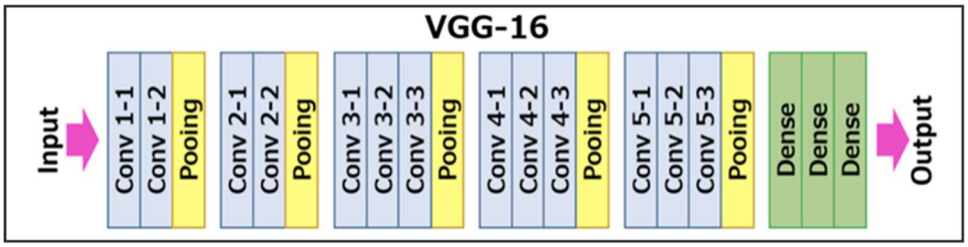
Figura 5 – VGG



Fonte: SAXENA (2017).

A arquitetura do VGG16 com uma imagem RGB de tamanho fixo 224 x 224 como entrada. A imagem é passada por uma pilha de filtros de convolução de tamanho 3x3 e algumas camadas de agrupamento, conforme mostrado na Figura 5.

Figura 6 – VGG16



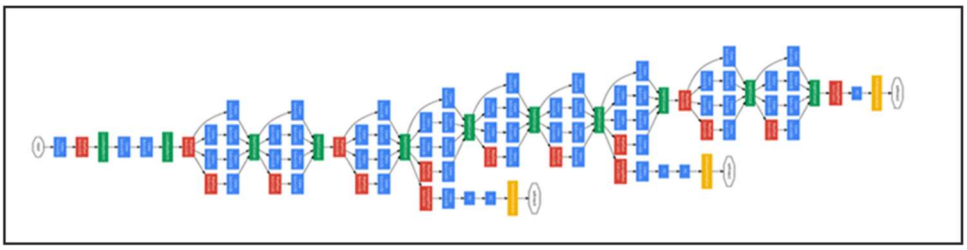
Fonte: NEUROHIVE (2018).

A característica do VGG é estudar filtros com dimensões 3x3 com pequenos campos receptores para uma arquitetura profunda e mostrar que os filtros 7x7 usados

no Alexnet podem ser substituídos por uma sequência de 3 filtros 3x3, reduzindo assim o conjunto de parâmetros. (NEUROHIVE, 2018).

- **Inception**: é uma rede neural convolucional que pode reduzir o número de parâmetros usando o bloco de *inception* sem perder eficiência. (CHRISTIAN SZEGEDY et al., 2015). A principal característica do Inception é quebrar o padrão de redes neurais sequenciais que representam ramificações dentro da rede e usar mais de um classificador, agilizando o treinamento e podendo ser usado para aprender características de diferentes visões.

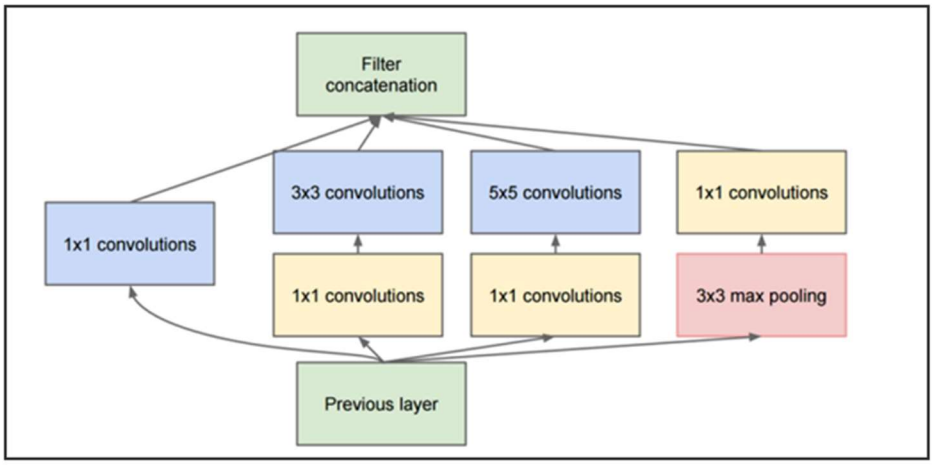
Figura 7 – Bloco de Inception



Fonte: SZEGEDY et al. (2014).

Ao contrário do VGG16, que é uma pilha de camadas, a rede de *inception* consiste em 9 blocos chamados "*Inceptions Blocks*", que propagam as informações de uma camada em 4 fluxos de dados. A figura 7 acima apresenta um exemplo de bloco *inception*.

Figura 8 – Filter Concetenation

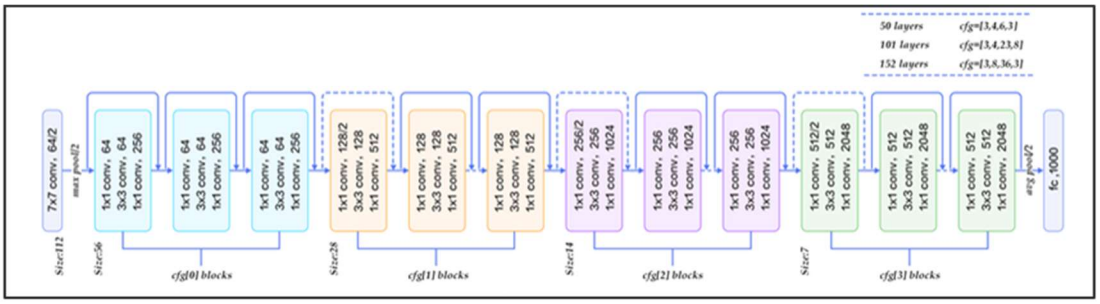


Fonte: SZEGEDY et al. (2014).

A Figura 8 mostra, de baixo para cima, uma camada que se propaga em quatro fluxos de dados, onde cada fluxo é responsável por extrair informações em diferentes níveis de abstração. Ao final, os resultados dos quatro fluxos são concatenados e passados para o próximo cabeçalho.

- **ResNet**: é uma rede neural convolucional com mais de 150 camadas, que apresenta uma estrutura de aprendizado para redes neurais profundas, reformulando as camadas com funções residuais para referências às entradas das camadas. (HE et al., 2016). Uma das características mais importantes do ResNet é utilizar intensa camadas de BatchNormalization e não possui camadas densas além da camada de saída. A seguir, a Figura 9 apresenta a arquitetura ResNet da esquerda para a direita.

Figura 9 – Arquitetura ResNet



Fonte: UTRERA (2018).

A imagem de entrada passa por uma camada de convolução de filtro 7x7, e depois por uma sequência de blocos residuais, adicionando os dados residuais a cada terceira camada de convolução. (UTRERA, 2018).

- **MobileNet**: baseado em convoluções de profundidade separáveis, que é uma forma de convolução que é fatorada em uma convolução de profundidade e uma convolução de ponto 1x1 para combinas as saídas de convolução de profundidade (HE et al., 2016). A seguir, a Figura 10 apresenta um exemplo do modelo MobileNet.

Figura 10 – MobileNet

Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1	3 × 3 × 512 dw	14 × 14 × 512
5× Conv / s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000

Fonte: ANDREW G. HOWARD et al. (2017).

A rede MobileNet representa uma grande redução nos parâmetros (Inception) para os parâmetros MobileNet, apresentando uma melhoria na precisão do inception para mobileNet.

Na próxima seção 2.6.2, é apresentado o estudo das métricas de avaliação e desempenho.

2.6.2 Métricas de avaliação de desempenho

Considerando-se as técnicas de classificação expostas, faz-se necessário avaliar o desempenho do classificador empregado num experimento. Para tanto, algumas métricas são utilizadas para distinguir entre a classe real e a classe prevista (rótulo), empregando os rótulos (‘P’ - positivo e ‘N’ - negativo) para as previsões de classe produzidas por um modelo aplicado ao fenômeno em estudo (RODRIGUES, 2019).

Segundo Mariano (2022), um modelo de classificação de dados visa fazer uma previsão, tendo como base eventos passados. Para tanto, o modelo utiliza como entradas instâncias e atributos. Todas essas informações são usadas para treinar um modelo que é usado para prever os resultados esperados para novos dados que

surgirem no futuro. Ao treinar esse modelo, deve-se usar um conjunto de dados (não usados no treinamento) para testar se o modelo está correto.

a) Matriz de Confusão

A matriz de confusão é uma matriz quadrada que compara os valores verdadeiros de uma classificação com os valores previstos de alguns modelos. A diagonal principal desta matriz quadrada contém os valores corretos, enquanto a matriz secundária contém os erros cometidos pelo modelo (BITTAR, 2020).

Segundo Kunumi (2022), a matriz de confusão permite visualizar facilmente quantos exemplos foram classificados corretamente e incorretamente em cada classe, o que auxilia no entendimento se o modelo favorece uma classe em detrimento de outra. Nesse sentido, Bittar (2020) afirma que uma métrica também pode ser utilizada para comparar o desempenho entre os diferentes modelos aplicados.

Assim, as métricas selecionadas apresentam o desempenho do modelo ou da técnica. Na Tabela 2 são expostos os resultados possíveis de uma matriz de confusão aplicada para a classificação dos dados.

Tabela 2 – Definição de matriz de confusão

VP - Verdadeiro Positivo	O rótulo avaliado é verdadeiro e o modelo aplicado retornou um valor positivo, indicando assim que o modelo estava correto.
FN - Falso Negativo	O rótulo avaliado é positivo e o modelo aplicado retornou um valor negativo, indicando assim um erro de modelo
VN - Verdadeiro Negativo	O rótulo avaliado é negativo e o modelo aplicado retornou um valor negativo, indicando assim que o modelo estava correto.
FP - Falso Positivo	O rótulo avaliado é negativo e o modelo aplicado retornou um valor positivo, indicando assim um erro de modelo.

Fonte: Rodrigues (2019).

Uma matriz de confusão é uma tabela que mostra os sucessos e falhas de seu modelo *versus* os resultados esperados (ou rótulos) (RODRIGUES, 2019). Na Tabela 3, é exposto um exemplo de matriz de confusão.

Tabela 3 - Matriz de confusão

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Rodrigues (2019).

Segundo Rodrigues (2019), a aplicação da matriz de confusão permite calcular as métricas de avaliação de desempenho dos resultados dos experimentos. As métricas de avaliação de desempenho consideradas neste estudo são listadas a seguir:

- **Acurácia:** indica o desempenho geral do modelo, apresentando quantas classificações o modelo classificou corretamente em relação ao total. No entanto, pode haver situações em que é falso, conforme exposto pela fórmula abaixo:

$$(VP+VN) /(VP+FV+FP+FN)$$

- **Precisão:** pode ser usada em uma situação em que os falsos positivos são considerados mais prejudiciais do que os falsos negativos, conforme expressado pela fórmula abaixo.

$$VP/(VP+FP)$$

- **Recall:** pode ser usada em uma situação em que falsos negativos são considerados mais prejudiciais do que falsos positivos, conforme expressado pela fórmula abaixo.

$$VP/(VP+FN)$$

- **F1 Score:** é uma maneira de olhar para apenas uma métrica, ao invés de duas (precisão e recall) em determinadas situações. É uma média harmônica entre os dois indicadores, que está muito mais próxima dos menores valores

do que uma simples média aritmética, conforme expressado pela fórmula abaixo.

(2* Precisão*Recall) /(Precisão+Recall)

As métricas de avaliação de desempenho podem ser usadas em vários classificadores, como aqueles indicados em seções anteriores.

2.7 Principais autores e obras considerados na plataforma teórica da pesquisa

No Quadro 1 são expostos os temas componentes da temática desta pesquisa, bem como o conceito base considerado e os principais autores da plataforma teórica estabelecida nesta dissertação.

Quadro 1 - Sinóptico com os principais autores e obras considerados na plataforma teórica da pesquisa

Tópico	Conceito base	Principais autores considerados
Conhecimento de cliente	O conhecimento há muito é reconhecido como um fator chave no valor das empresas inovadoras, enquanto o conhecimento do cliente é visto como um elemento importante na vantagem competitiva das empresas e como um fator chave na economia deste século. Khosravi, *et al.*, (2017).	Aacinol (2019) Jaziri (2019) Hollanda (2019)
Mídias sociais	As mídias sociais tornaram-se uma nova parte das ferramentas de marketing disponíveis para as empresas, permitindo que elas construam novas formas de se relacionar com seus clientes. Seller; Laurindo (2018).	Tarn (2019) Hazarika; Mousavizadeh; Statista (2020) Drude (2021)

Descoberta de conhecimento – (KDD)	O processo de descoberta de conhecimento em dados de texto, conhecido como Text Mining (MT), muitas vezes combina técnicas de Information Retrieval (IR), Machine Learning (AM) e Natural Language Processing (NLP) em suas fases Martins, *et al.,* (2003).	Fayyad, U. M.; Piatetsky-Shapiro, G.; Smith, P (1996) Phridviraj; Gururao (2014) Ristoski; Paulheim (2016) Peres;
Inteligência artificial	Inteligência Artificial é desenvolver sistemas para executar tarefas que são melhor executadas por humanos do que por máquinas ou tarefas para as quais nenhuma solução algorítmica por computadores tradicionais seja possível. Rich e Knight (1991).	Hariri (2019) Kenji (2019) Ludermir (2021)
Análise de sentimentos	A análise de sentimentos, também conhecida como Mineração de Opinião, é o campo de estudo que analisas as opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, pessoas, eventos e seus aspectos. Liu (2012).	Ohashi (2019) Tebaldi (2019) Tibco (2022)
Redes Neurais	As camadas são compostas de nós. Um nó é apenas um lugar onde os cálculos, vagamente padronizados em um neurônio no cérebro humano que dispara quando recebe estímulos suficientes. Um nó combina os dados de entrada com uma série de coeficientes (ou pesos) que reforçam ou atenuam essa entrada, atribuindo assim significado às entradas em relação à tarefa que o algoritmo está tentando aprender. Pathmind (2019).	Pathmind (2019) Baroni; Itanimá (2021) Herhold (2022)
Métricas de avaliação de desempenho	As métricas são utilizadas para distinguir entre a classe real e a classe prevista. Os rótulos (‘P’ – positivo e ‘N’ - negativo) são usados para as previsões de classe produzidas por um modelo. (Rodrigues, 2019).	Rodrigues (2019) Bittar (2020) Mariano (2021) Kunumi (2022)

Fonte: autora (2023).

3. Métodos e instrumentos de pesquisa

Neste capítulo são apresentados o método e materiais a serem empregados nos experimentos a serem conduzidos nesta dissertação, bem como uma descrição detalhada da escolha da metodologia para a realização dos experimentos. Segundo (Prodanov; Freitas, 2013, p. 14), metodologia é a aplicação de procedimentos e técnicas que precisam ser desenvolvidas para a construção do conhecimento, com o objetivo de verificar sua validade e utilidade em diferentes campos.

3.1 Tipologia da pesquisa

A metodologia de pesquisa utilizada nesta dissertação será bibliográfica, exploratória e experimental. Segundo (Gil, 2002), a pesquisa bibliográfica envolve a leitura, análise e interpretação de livros, periódicos, documentos etc. O material escolhido para estudo deve passar por triagem para que seja aplicado um plano de leitura, no qual a leitura deve ser cuidadosa, atenta e sistemática, o pesquisador deve fazer anotações que sirvam de base teórica para o desenvolvimento do estudo.

A pesquisa bibliográfica foi baseada em consultas a fontes de referência bibliográfica e teórica, tais como: artigos, livros, teses, dissertações e sites com conteúdo sobre mineração de opinião, análise de sentimentos e descoberta de conhecimento do cliente. Foram consultadas as seguintes bases de dados: *SCIELO*, *IEEE XPLORE*, *SCOPUS* e *GOOGLE SCHOLAR* onde foram selecionados esses materiais de estudo e selecionados os mais relevantes para serem desenvolvidos no decorrer deste trabalho.

(Yin, 2006), afirma que a pesquisa exploratória permite maior intimidade entre o pesquisador e o sujeito estudado, é pouco explorada e pouco conhecida. A pesquisa exploratória visa conhecer melhor o problema para torná-lo explícito ou levantar hipóteses visando aprimorar ideias ou descobrir percepção.

Segundo (Gil, 2002), um experimento identifica um objeto de estudo selecionando as variáveis encontradas para definir formas de controle e efeitos das variáveis que produz sobre o objeto. A pesquisa experimental tem como foco a aquisição de conhecimento, neste caso a base documental composta por opiniões de clientes. Nesta pesquisa foram realizados experimentos com técnicas de análise e

classificação aplicadas aos comentários extraídos de mídias sociais e redes sociais para a descoberta de conhecimento do cliente em empresa varejista.

3.2 Universo, amostragem e amostra

Nesta pesquisa, o universo considerado é formado por empresas varejistas. Segundo SAFETEC (2022), os dados dos clientes no varejo são informações importantes que ajudam a analisar as estratégias de vendas criadas e os resultados alcançados. O varejo é um setor bastante analógico e muitas lojas ainda não acompanham o desempenho de suas vendas por meio de indicadores pertinentes. Além disso, uma gestão de dados organizada permite avaliar o desempenho das operações diárias, extraindo informações sobre todos os processos da organização.

Conquistar um cliente é o ponto culminante do processo de vendas. Qualquer pessoa que compra alguma coisa é uma grande fonte de informação para a empresa. Assim, a empresa objetiva entender melhor o relacionamento com os clientes ao longo do tempo, para ser mais ágil e assertiva na tomada de decisões sobre estratégias voltadas ao atendimento ao cliente (SAFETEC, 2022).

A amostragem não probabilística aplicada nesta pesquisa é respaldada por critérios de acessibilidade e conveniência da pesquisadora. Assim, este é um estudo de caso único. Segundo FIA (2020), os estudos de caso são um método de pesquisa abrangente sobre um determinado tema, permitindo aprofundar o conhecimento sobre o mesmo e, assim, incentivar novas pesquisas sobre o mesmo tema. Yin (2001) afirma que um estudo de caso é uma estratégia de pesquisa que responde a perguntas do tipo ‘como’ e ‘por que’ e se concentra nos contextos reais dos casos atuais.

A amostra do objeto enfocado nesta pesquisa recaiu sobre uma grande empresa varejista, dada a sua expressiva representatividade neste segmento de atuação, uma vez que tal empresa é um dos maiores varejistas atuantes no país, conforme a importância do estudo de caso único a ser selecionado indicada por Gil (2022). A empresa enfocada nesta pesquisa de caso único é paradigmática, por possuir muitas lojas e milhões de clientes no país, além de vários canais de comunicação com o cliente para a coleta de *feedback* de seus produtos e serviços.

A empresa varejista enfocada nesta pesquisa possui um sistema interno que contém diferentes bases de dados para *download* em formato MSExcel. Essas bases

de dados têm diversos atributos, dentre os quais se destacam registros com comentários sobre produtos e serviços adquiridos pelos clientes. Tais registros/comentários são continuamente importados das mídias sociais mantidas pela empresa e armazenados nesse sistema interno. O sistema da empresa possui apenas relatórios de dados estruturados e não apresenta a capacidade de analisar o *feedback* expressado pelos clientes nos comentários por eles elaborados.

Ressalte-se ainda que a empresa demonstrou anuência à pesquisa ora proposta, indicando ainda interesse a respeito dos experimentos e aplicações previstos, com especial foco nos possíveis resultados vislumbrados nesta pesquisa. Portanto, o acesso e uso dessas bases de dados foram autorizados e liberados pela empresa foco do estudo de caso desta pesquisa.

A aplicação dos dados utilizados nos experimentos desta pesquisa estão de acordo com a LGPD (Lei Geral de Proteção de Dados), que aplica-se a todas as operações de processamento realizadas por pessoa física ou jurídica, pública ou privada, independentemente do meio, do país de sua sede ou do país em que os dados estejam localizados, desde que a operação de processamento de dados seja realizada no Brasil; a atividade de tratamento se destine à oferta de bens, serviços ou tratamento de dados de residentes no país; ou ainda, que os dados pessoais objeto do tratamento foram recolhidos internamente (STJ, 2020). Assim, os dados utilizados nos experimentos desta dissertação voltam-se tão somente ao desenvolvimento, aplicação e validação de técnicas inteligentes para a análise de sentimentos em comentários extraídos de mídias sociais para a descoberta de conhecimento do cliente.

A empresa em questão possui múltiplos sistemas e canais de comunicação com seus consumidores e busca sempre melhorar seus produtos e serviços por meio do *feedback* de seus clientes. Possui um grande banco de dados de comentários de clientes sobre os serviços e produtos. Esses comentários são provenientes de mídias sociais, central de atendimento, estandes em lojas físicas, aplicativos e portais na internet, sendo que todos esses canais usam a metodologia NPS (*Net Promoter Score*), para que o cliente expresse sua opinião em relação produtos e serviços prestados pela empresa.

Os comentários externados pelos clientes podem ser positivos, indicando que o cliente está satisfeito, ou negativos, indicando que o cliente não está satisfeito com os serviços e produtos que adquiriu da empresa em questão. E ainda pode acontecer

que o cliente se posicione de forma neutra quanto aos seus comentários. Em todos os casos, o cliente pode indicar uma nota NPS, mas não tecer comentários sobre os serviços e produtos da empresa. Os comentários manifestados pelos clientes têm como característica apresentarem baixa estruturação, com ocorrência de erros de ortografia e erros de concordância, tornando complexa a compreensão para a leitura e entendimento de seu significado.

Por esta razão, nesta dissertação são aplicadas técnicas inteligentes para a análise de sentimentos em comentários de clientes para a descoberta de conhecimento do cliente, considerando-se os dados disponíveis no banco de dados da empresa varejista, conforme já explicado anteriormente.

Experimentos computacionais baseados em métodos e técnicas de análise de sentimentos foram realizados com os dados disponibilizados. Os experimentos realizados visam desenvolver uma solução eficiente para descobrir conhecimentos úteis para a tomada de decisões relacionadas aos produtos e serviços oferecidos aos clientes.

Na Tabela 4 são expostas as características da base de dados considerada para os experimentos realizados nesta pesquisa, quais sejam: nome da base de dados, breve descrição da base de dados, quantidade de registros disponíveis para aplicação nos experimentos. Esta base de dados foi extraída do sistema da empresa alvo desta pesquisa, que contém registros de dados com opiniões e *feedbacks* de clientes sobre os produtos e serviços da empresa.

Tabela 4 – Base de dados com registros de opiniões de clientes e feedbacks sobre produtos e serviços da empresa varejista

Nome da Base	Descrição da Base	Quantidade de Registros
Atendimento	Atendimento aos clientes (central de atendimento, estandes, app e portal)	17.547

Fonte: autora (2023).

Com base nas opiniões/*feedbacks* de clientes sobre produtos e serviços foram realizados experimentos aplicando-se técnicas de IA para realizar a análise de sentimentos dos comentários extraídos da base de dados.

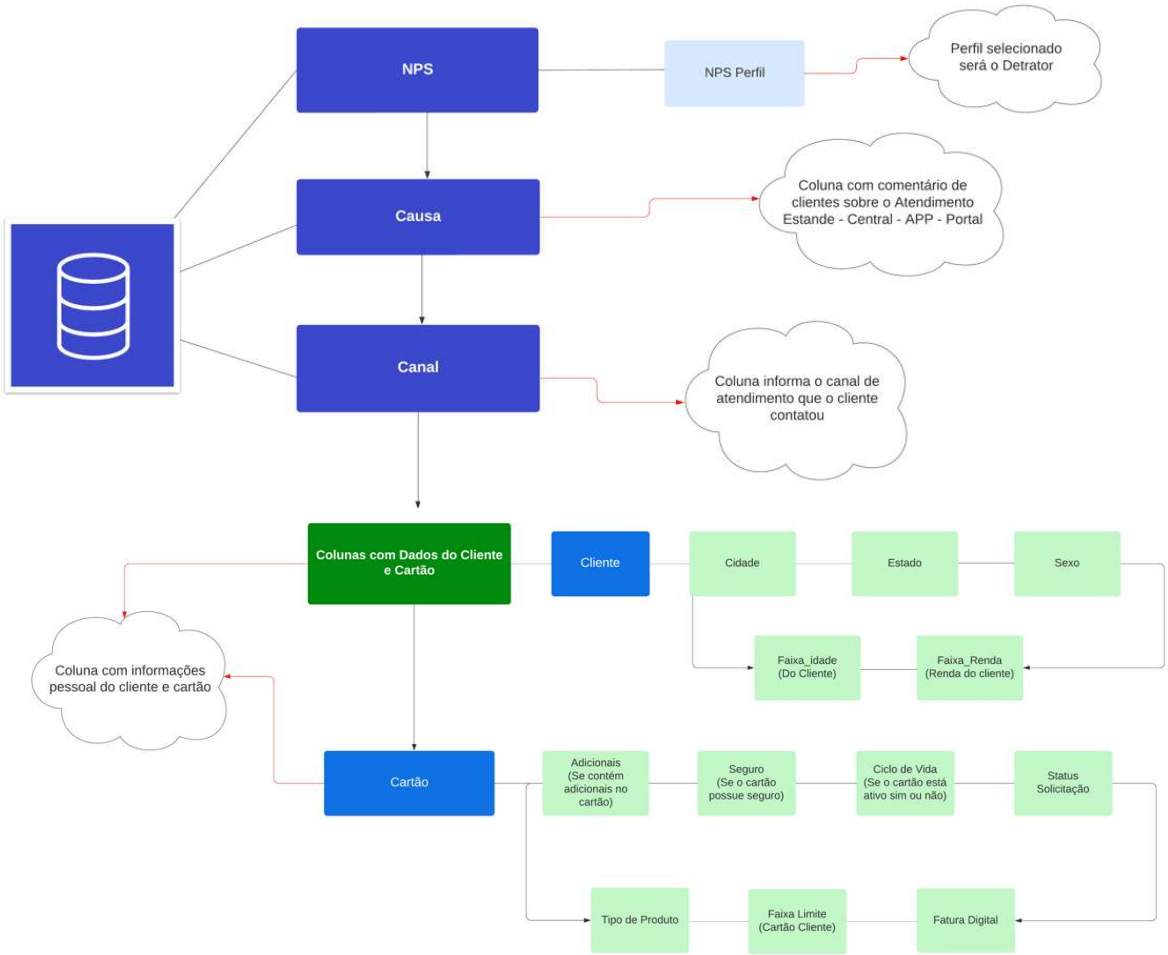
3.3 Estrutura de Base de Dados de Atendimento, Arquitetura Computacional e Metodologia Experimental

Nesta seção são apresentadas a estrutura da base de dados de atendimento, a arquitetura computacional e a metodologia experimental desenvolvida nesta dissertação.

3.3.1 Estrutura da Base de Dados de Atendimento

A estrutura da base de dados de atendimento considerada para os experimentos computacionais nesta pesquisa é apresentada na Figura 11.

Figura 11 – Estrutura da base de dados de atendimento



Fonte: autora (2023).

Na figura 11 são expostas todas as categorias de dados consideradas e suas correlações para a aplicação dos dados da base ‘Atendimento’ nos experimentos delineados. As correlações e cruzamentos foram determinados com a participação de um especialista da empresa relatada nesta dissertação.

O especialista acompanhou todas as etapas do processo de KDD delineado nesta pesquisa. Foram realizadas reuniões via Google Meet e reuniões presenciais, além de trocas de mensagens com o especialista da empresa responsável pela área de soluções de TI com foco no atendimento ao cliente. Tais interações visaram esclarecer e entender a estrutura e características da base de dados ‘Atendimento’, bem como sanar dúvidas relevantes que surgiram durante o trabalho de pesquisa e aplicação dos experimentos. O profissional responsável esteve sempre disponível para esclarecer dúvidas, tendo participado da seleção dos atributos e auxiliado na correlação das categorias dos dados para o cruzamento nos experimentos. O especialista também participou das discussões dos resultados sobre necessidades e aplicações da solução idealizada para a empresa/negócio. A seguir são apresentadas as correlações das categorias de dados para os cruzamentos, conforme validação realizada pelo especialista da empresa.

Correlações de categorias de dados para cruzamentos visando a aplicação nos experimentos para a descoberta de conhecimento do cliente

As correlações (cruzamentos) dos campos/atributos da base de dados selecionada foram efetuadas com o auxílio do especialista da empresa. Portanto, as correlações delineadas a seguir têm apelo prático para o negócio/empresa visando proporcionar atendimento mais eficaz ao cliente, com base no conhecimento descoberto para os perfis específicos estabelecidos. Abaixo são apresentadas as correlações indicadas pelo especialista da empresa como as mais expressivas.

1) Correlação entre a ‘Causa’ (comentário do cliente) X ‘Perfil do Cliente’, cruzando os atributos ‘sexo’, ‘idade’, ‘renda’ e ‘estado’.

2) Correlação entre a ‘Causa’ (comentário do cliente) X ‘Produto cartão’, cruzando com os atributos ‘faixa limite’, ‘tipo de produto’, ‘ciclo de vida’ e ‘adicional’.
3) Correlação entre a ‘Causa’ (comentário do cliente) X ‘Canais de atendimento’, cruzando com os atributos ‘central de atendimento’ ‘estandes/lojas’, ‘aplicativos’, ‘portais’ e ‘status da solicitação’.
4) Correlação entre ‘Causa’ (comentário de cliente) X ‘Estratificação de notas detratoras do NPS’ (notas de 0 a 6), segregadas em dois substratos: notas de 0 a 3 (detratores inferiores) e notas de 4 a 6 (detratores superiores).

A título da condução dos experimentos e para a validação da solução delineada, nesta pesquisa optou-se por apresentar apenas os resultados dos atributos da categoria de ‘perfil do cliente’ (‘sexo’, ‘idade’, ‘renda’ e ‘estado’) em cruzamento com os resultados de NPS detratores. Seria possível considerar também as correlações com as demais categorias (‘produto cartão’ e ‘canais de atendimento’) e seus respectivos atributos, mas em termos práticos teríamos como resultado apenas a aplicação da solução delineada em outros conjuntos de dados. Ou seja, obteria-se tão somente a mesma indicação de que a solução apresenta resultados satisfatórios para o objetivo para o qual foi concebida.

3.3.2 Arquitetura computacional

Os softwares utilizados para os experimentos computacionais desta dissertação foram o Python 3.10.7 e Colaboratory, ferramenta do Google para leitura da base de dados. As bibliotecas Scikit-learn e spaCy também foram utilizadas para segregar os dados da base selecionada, e assim realizar o treinamento e teste dos dados.

Em termos de hardware foi utilizado um computador para processar os experimentos, composto por uma placa de vídeo Graphics 620, processador Intel(R) Core (TM) i7-8565U CPU @ 1.80GHz 1.99 GHz, SSD 512GB e 16GB de memória, sistema operacional Windows 11 Home Single Language. A seguir são apresentadas no Quadro 2 as principais bibliotecas e ferramentas utilizadas para realização dos

experimentos nesta dissertação.

Quadro 2 – Bibliotecas e ferramentas utilizadas nos experimentos

Bibliotecas	**Descrição**	**URL**
Colaboratory	ferramenta do Google para escrever e escutar de código em Python.	https://colab.research.google.com/notebooks/welcome.ipynb?hl=pt-br
Pandas	para leitura e análise de dados.	https://pandas.pydata.org/
NLTK	para suporte à normalização de texto.	https://www.nltk.org/
Numpy	para operações de álgebra linear e matriz ao gerar vetores para entrada em redes neurais artificiais.	https://numpy.org/
Scikit-learn	para extração de atributos e algoritmos de aprendizado de máquina para criar treinamentos e testes.	https://scikit-learn.org/stable/
SpaCy	para extração de informações, entender linguagem natural ou pré-processar textos para posterior uso em modelos de Deep Learning.	https://spacy.io/
WordCloud	para a criação da nuvem de palavras para a análise de sentimentos.	https://pypi.org/project/wordcloud/
Tokenização	para divisão de uma frase em palavras ou tokens individuais e para a remoção de pontuação e caracteres especiais.	https://spacy.io/api/tokenizer
Seaborn	biblioteca do Python voltada para visualização de dados estatísticos, em um nível mais alto que o matplotlid, criando gráficos de Pareto das palavras.	https://seaborn.pydata.org/
Stop Words	utilizados para o pré-processamento dos textos, visando a remoção de pontuação, caracteres especiais para o tratamento de palavras que não agregam significado, ao menos em termos semânticos e, portanto, são irrelevantes.	http://python.w3.pt/?p=234
Lematização	‘Lema’ será utilizado para vocabulário e análise morfológica de uma palavra de acordo com seu significado no dicionário.	https://spacy.io/api/lemmatizer
Stemização	para extração do radical das palavras.	https://www.nltk.org/_modules/nltk/stem/rslp.html

Fonte: autora (2023).

Optou-se por prosseguir com a biblioteca TF-IDF (abreviação da frequência do termo – frequência inversa do documento), visando indicar a importância de uma palavra em um documento em relação a uma coleção de documentos ou corpus de linguagem. (SCIKIT-LEARN, 2022).

Descrevendo brevemente o passo a passo da configuração da rede neural convolucional, como mostra a Figura 12, a linha 2 corresponde ao pré-processamento e na linha 5 foi aplicado a biblioteca rake para encontrar as principais palavras-chaves. Em seguida na linha 8 foi iniciado o desenvolvimento do modelo de redes neurais convolucionais, com a consequente configuração para criação do modelo. Os testes foram executados em outras configurações para redes neurais convolucionais e a rede com os melhores resultados foi a de Redes Neurais Convolucionais, com um número total de cinco camadas de convolução, *batch* de tamanho 128, com 26 parâmetros e utilização da função de ativação Soft Max:

- camada de entrada com 28 x 28 neurônios;
- camada de saída com 5 x 5 neurônios;
- camada oculta com 2 camadas 24 x 24 neurônios por camada;
- filtros 3x3.

Na Figura 12 é apresentado o código-fonte da configuração das redes neurais convolucionais.

Figura 12 - código-fonte da configuração das redes

```
#Definição do pré-processamento
for n in text.split():
    semstop = [p for p in text.split() if p not in stopwords]
bagofwords = semstop

#Análise do Texto - Principais Palavras-Chave (Keywords)
r = Rake(include_repeated_phrases=False, min_length=1, max_length=5) text_to_rake = text
r.extract_keywords_from_text(text_to_rake)

#Modelo a CNN
def cluster_text(text):
    vectorizer = TfidfVectorizer(stop_words=stopwords)
    X = vectorizer.fit_transform(bagofwords)

    import matplotlib.pyplot as plt
    from sklearn.cluster import KMeans
    Sum_of_squared_distances = []
    K = range(2,10)
    for k in K:
      km = KMeans(n_clusters=k, max_iter=200, n_init=10)
      km = km.fit(X)
      Sum_of_squared_distances.append(km.inertia_)
    plt.plot(K, Sum_of_squared_distances, 'bx-')
    plt.xlabel('k')
    plt.ylabel('Sum_of_squared_distances')
    plt.title('Elbow Method For Optimal k')
    plt.show()

    print('Quantidade de Clusters: ')
    true_k = int(input())
    model = KMeans(n_clusters=true_k, init='k-means++', max_iter=200, n_init=10)
    model.fit(X)
    labels=model.labels_
    clusters=pd.DataFrame(list(zip(text,labels)),columns=['title','cluster'])
    #print(clusters.sort_values(by=['cluster']))

    for i in range(true_k):
      print(clusters[clusters['cluster'] == i])

    return
```

Fonte: autora (2023).

A estrutura concebida foi treinada e o modelo foi executado. Para tanto, foram utilizadas duas bibliotecas para identificar os principais tópicos: Gensim e Bertopic. A seguir, na Figura 13 é apresentada a configuração das bibliotecas para a realização dos experimentos.

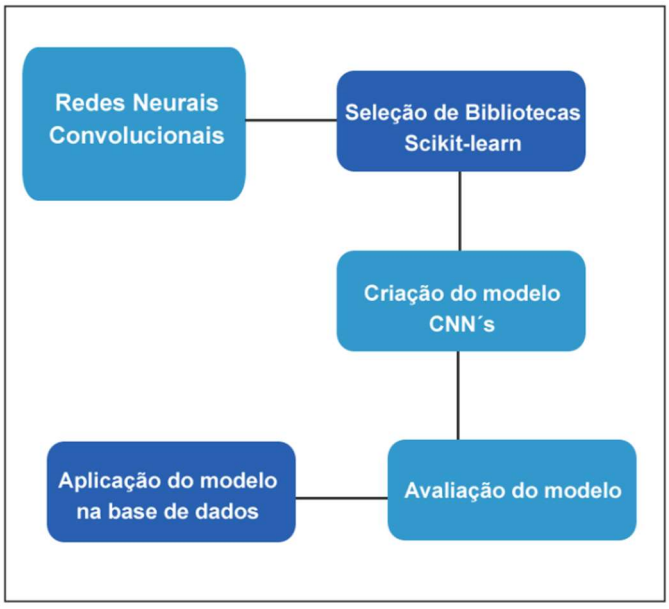
Figura 13 – Codificação de bibliotecas bertopic e Gensim

```
#Bibliotecas para Encontrar os principais tópicos

import gensim

from gensim import corpora
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize


# Treinando e Executando o modelo LDA na Matriz de Termos.

Lda = gensim.models.ldamodel.LdaModel
ldamodel = Lda(doc_term_matrix, num_topics=10, id2word = dictionary, passes=50,
random_state=4)


#Bertopic

from bertopic import BERTopic

topic_model = BERTopic()
topics, probs = topic_model.fit_transform(newsentences)
```

Fonte: autora (2023).

Na Figura 14 é apresentado o desenho da arquitetura idealizada para a criação do modelo de redes neurais convolucionais considerada nos experimentos.

Figura 14 – Arquitetura do modelo.



Fonte: autora (2023).

Após a construção do modelo da Rede Neural Convolucional, primeiro decidiu-se dividir a base de dados em treinamento e teste, com 80% dos dados usados para treinamento e 20% para teste. Com a base de dados foi particionada desta forma não

se verificou um bom resultado com o modelo usado neste estudo. Desta forma, foi realizada uma segunda divisão da base de dados, com 60% dos dados destinados ao treinamento e 40% dos dados destinados ao teste. Ainda assim não obteve-se bom desempenho do modelo, notadamente em razão de que não obteve-se bom desempenho do modelo CNNs aplicados, especialmente por se tratar de uma base de dados com muito ruídos, o que apresentou muitas dificuldades para a aplicação do modelo treinado.

Foi então realizada uma terceira divisão na base, com o intuito de que o modelo obtivesse melhor desempenho. Assim, a base de dados foi segmentada em 70% para treinamento e 30% para testes do modelo aplicado. Na Tabela 5 são indicados os percentuais de dados destinados ao treinamento e testes, conforme as três divisões realizadas na base de dados.

Tabela 5 – Divisão da base de dados atendimento.

	Base Treinamento 80%	**Base Teste 20%**
1 – Divisão da base	5.252	845
	Base Treinamento 60%	Base Teste 40%
2 – Divisão da base	3.939	1.126
	Base Treinamento 70%	Base Teste 30%
3 – Divisão da base	4.593	845,1

Fonte: autora (2023).

3.3.3 Metodologia Experimental

A metodologia dos experimentos delineada foi realizada em seis fases, tendo por base o processo de KDD de Fayyad *et al.* (1996), porém com adaptação para viabilizar o atingimento dos objetivos desta pesquisa. Foi aplicada a técnica inteligente de Redes Neurais Convolucionais para a extração de conhecimento da base de atendimento considerada. Cada fase do processo de KDD adaptado é explicada a seguir:

- Fase 1: Seleção da base de dados de ‘Atendimento’. A base de dados contém 17.547 registros referentes ao ano de 2021, dentre os quais se destaca o *feedback* dos clientes expressados em forma de comentário em texto livre que foram coletados por meio da central de atendimento, estandes (lojas), aplicativos e portal da empresa varejista. Além de comentários dos clientes, a base contém informações importantes acerca do cliente, produtos e serviços. Esta base de dados contém originalmente 59 colunas com e sem registros. Foi realizada uma análise em conjunto com o profissional especialista da empresa responsável pela base de dados considerada nesta dissertação. A pesquisadora realizou várias reuniões com o especialista para o entendimento dos campos/atributos da base de dados. Como fruto dessas reuniões como o especialista da empresa foram encontradas algumas colunas (categorias de atributos) que não faziam sentido para esta pesquisa. O especialista chegou à conclusão de que não faria sentido processar todo o conjunto de dados originais nos experimentos da presente pesquisa. Mais adiante será apesentada tabela com os nomes das colunas (atributos) que foram retiradas em consenso com o especialista da empresa varejista.

- Fase 2: Pré-processamento da base de dados ‘Atendimento’. Nesta fase se inicia o pré-processamento da base selecionada para a extração/redução de atributos aplicando-se técnicas de inteligência artificial. As técnicas utilizadas para o pré-processamento foram: i) ‘stop words’ – utilizadas para o pré-processamento dos textos com objetivo de remover as pontuações, caracteres especiais; ii) ‘lematização’ – também usada para vocabulário e análise morfológica de uma palavra, de acordo com seu significado no dicionário; iii) ‘stemização’ – para extração do radical das palavras e, por fim; iv) ‘tokenização’ – para divisão de uma frase em palavras ou tokens individuais. Foram utilizadas as bibliotecas ‘Seaborn’ e ‘Word Cloud’ para a visualização de dados estáticos/gráficos e nuvens de palavras, respectivamente.

- Fase 3: Transformação dos dados da base ‘Atendimento’. Esta fase visa transformar os dados originais em formatos mais adequados para o processo de mineração de dados considerado neste estudo. Segundo Moura (2019), esta fase consiste na aplicação de técnicas de transformação, tais como

normalização, agregação, criação de novos atributos, redução e síntese dos dados. Os dados são agrupados em um mesmo local para aplicação dos modelos de análise. Esses dados de entrada devem ser transformados antes de serem usados para treinamento do modelo.

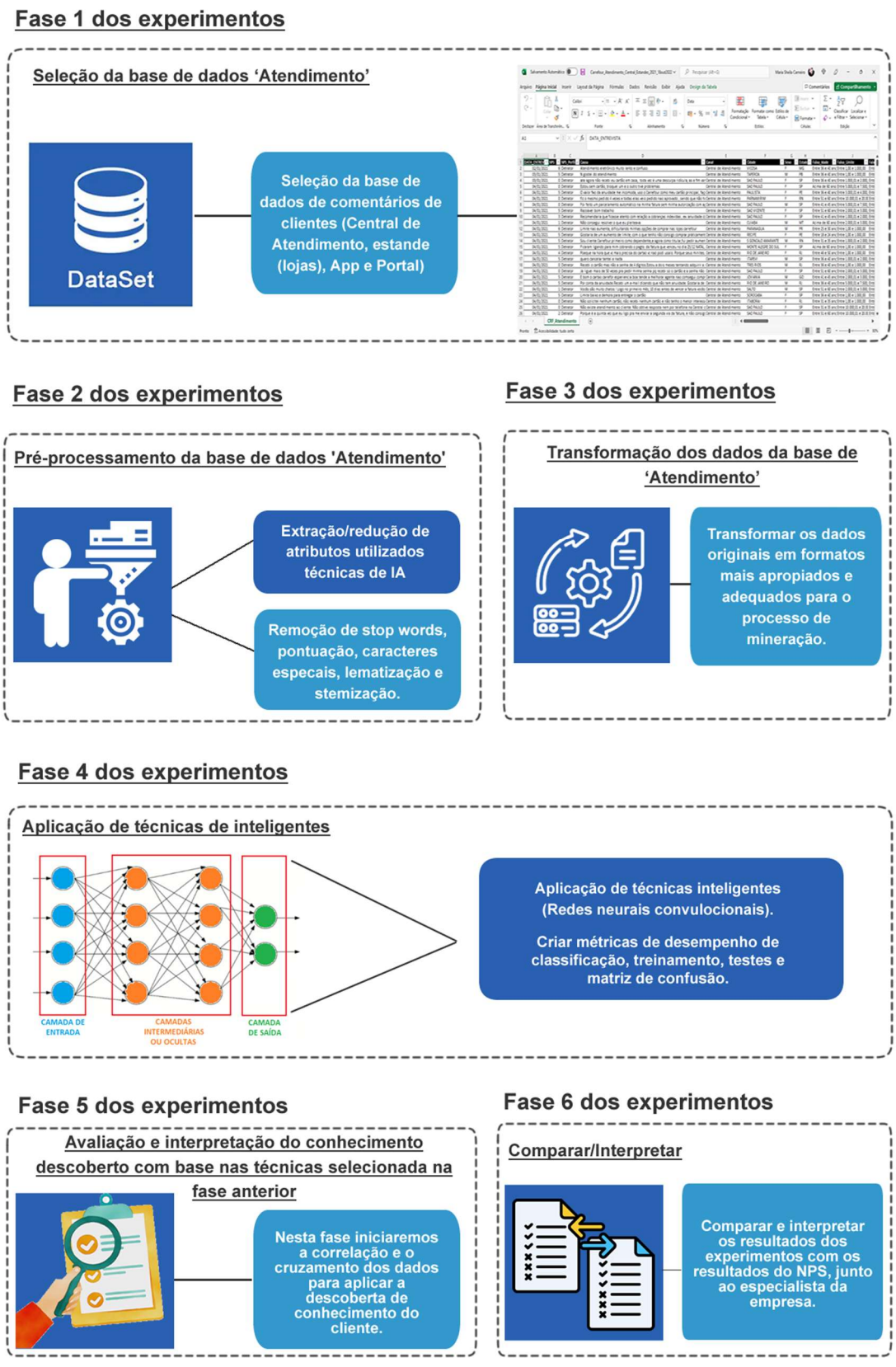
- Fase 4: Aplicação de técnicas inteligentes de Redes Neurais Convolucionais. Esta fase consiste em construir modelos ou aplicar de técnicas de mineração de dados. Essas técnicas visam testar a hipótese formulada e descobrir novos padrões de forma autônoma, conforme indicado por Moura (2019). Além disso, a descoberta pode ser dividida em preditiva e descritiva. Primeiramente, cria-se um modelo de classificador para treinamento e então iniciam-se os primeiros testes que produzem uma matriz de confusão. Durante os experimentos, métricas de avaliação de desempenho como acurácia, precisão, recall e F1 score são utilizadas para avaliar a performance do modelo.

- Fase 5: Interpretação e avaliação do conhecimento descoberto. Esta fase consiste em avaliar o desempenho do modelo (MOURA, 2019). Foram realizados as correlações determinadas e o cruzamento dos dados para a descoberta de conhecimento do cliente. A validação pode ser feita de diferentes formas, dentre as quais destaca-se o uso de medidas estatísticas e avaliação por especialistas do negócio.

- Fase 6: Comparação dos resultados dos experimentos com os resultados dos indicadores detratores do NPS (*Net Promoter Score*) da empresa. Nesta fase é realizada a comparação em conjunto com os especialistas da empresa.

Considerando-se as indicações expostas acima, na Figura 15 são detalhados os passos e respectivas ações executadas em cada uma das seis fases do processo de KDD adaptado de Fayyad *et al.* (1996) para os fins desta pesquisa:

Figura 15 – Desenho das fases dos experimentos.



Fonte: autora (2023).

A seguir é exposto um descritivo detalhado da operacionalização de cada uma das seis fases dos experimentos previstos nesta pesquisa.

- Fase 1: Seleção da base de dados ‘Atendimento’.

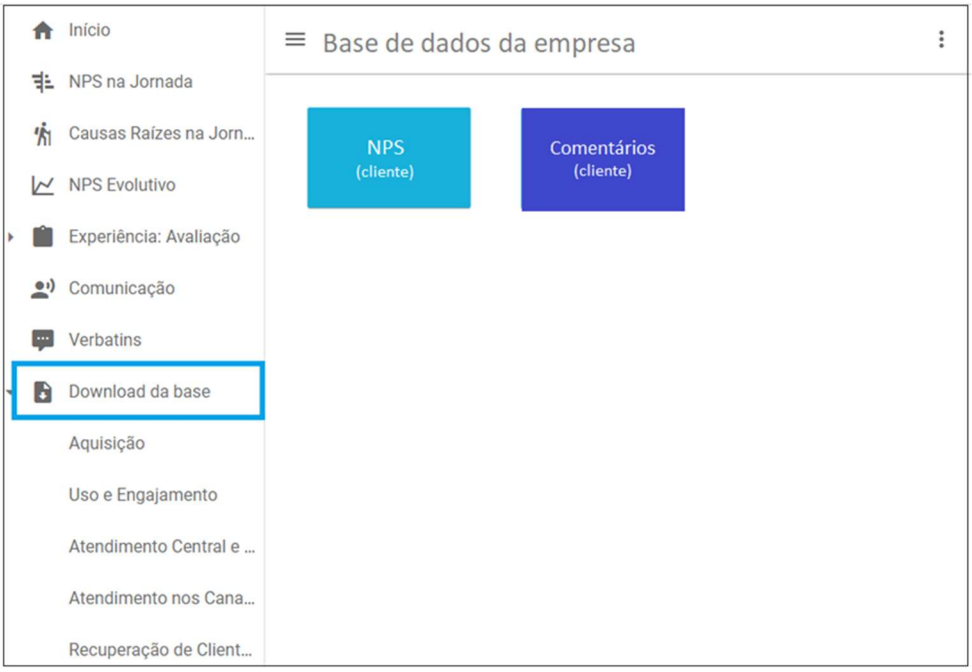
A base selecionada contém dados e *feedbacks* (comentários) dos clientes sobre produtos e serviços oferecidos pela empresa. Os dados foram coletados pela central de atendimento, estandes (lojas físicas), aplicativos e portais, bem como os resultados do NPS (*Net Promoter Score*) do varejista considerado nesta pesquisa.

Esta base de dados foi fornecida pela empresa, sendo que os dados foram extraídos de seu sistema interno. Os dados estão em acordo com a Lei Geral de Proteção de Dados (LGPD), que estabelece as regras sobre o uso de dados pessoais em todas as transações online.

Tal legislação proíbe o uso arbitrário de dados pessoais. Esta dissertação segue todas as normas de acordo com a legislação que define que a base geral de utilização dos dados deve indicar a finalidade específica dos dados utilizados (Neoenergia, 2020). Nesta etapa inicial a base de dados foi extraída por meio do sistema utilizado pela empresa *(Quest Manager - QWST).*

Este sistema abrange bases de dados para *downloads* de arquivos contendo comentários sobre serviços e produtos adquiridos pelos clientes. Foi realizada a exportação desta base de dados para planilha no formato *MSExcel* com extensão *.CSV*. Na Figura 16 é exposta a tela do sistema para extração de banco de dados de comentários dos clientes.

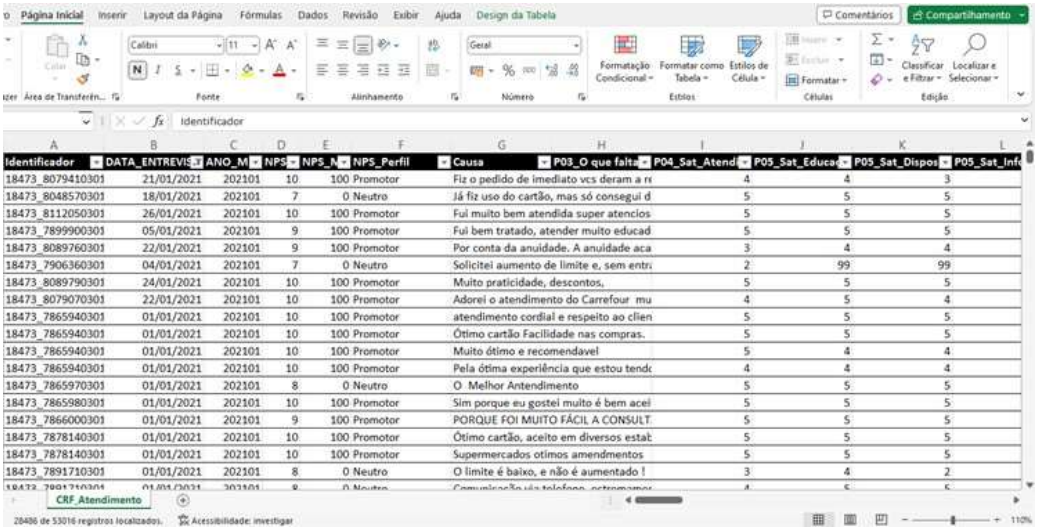
Figura 16 – Sistema para extração de banco de dados de comentários.



Fonte: autora (2023).

Na Figura 17 é exposta uma imagem geral da base de dados ‘Atendimento’ original, ou seja, sem nenhum pré-processamento e transformação dos dados.

Figura 17 – Base de dados ‘Atendimento’ original extraída do sistema da empresa.



Identificador	DATA_ENTREVIS	ANO_M	NPS	NPS_N	NPS_Perfil	Causa	P03_O que falta	P04_Sat_Atend	P05_Sat_Educa	P05_Sat_Dispos	P05_Sat_Inf
18473_8079410301	21/01/2021	202101	10	100	Promotor	Fiz o pedido de imediato vcs deram a r		4	4	3	
18473_8048570301	18/01/2021	202101	7	0	Neutro	Já fiz uso do cartão, mas só consegui d		5	5	5	
18473_8112050301	26/01/2021	202101	10	100	Promotor	Fui muito bem atendida super atencios		5	5	5	
18473_7899900301	05/01/2021	202101	9	100	Promotor	Fui bem tratado, atender muito educad		5	5	5	
18473_8089760301	22/01/2021	202101	9	100	Promotor	Por conta da anuidade. A anuidade aca		3	4	4	
18473_7906360301	04/01/2021	202101	7	0	Neutro	Solicitei aumento de limite e, sem entr		2	99	99	
18473_8089790301	24/01/2021	202101	10	100	Promotor	Muito praticidade, descontos,		5	5	5	
18473_8079070301	22/01/2021	202101	10	100	Promotor	Adorei o atendimento do Carrefour mu		4	5	4	
18473_7865940301	01/01/2021	202101	10	100	Promotor	atendimento cordial e respeito ao clien		5	5	5	
18473_7865940301	01/01/2021	202101	10	100	Promotor	Ótimo cartão Facilidade nas compras.		5	5	5	
18473_7865940301	01/01/2021	202101	10	100	Promotor	Muito ótimo e recomendavel		5	4	4	
18473_7865940301	01/01/2021	202101	10	100	Promotor	Pela ótima experiência que estou tend		4	4	4	
18473_7865970301	01/01/2021	202101	8	0	Neutro	O Melhor Antendimento		5	5	5	
18473_7865980301	01/01/2021	202101	10	100	Promotor	Sim porque eu gostei muito é bem acei		5	5	5	
18473_7866000301	01/01/2021	202101	9	100	Promotor	PORQUE FOI MUITO FÁCIL A CONSULT		5	5	5	
18473_7878140301	01/01/2021	202101	10	100	Promotor	Ótimo cartão, aceito em diversos estab		5	5	5	
18473_7878140301	01/01/2021	202101	10	100	Promotor	Supermercados otimos amendmentos		5	5	5	
18473_7891710301	01/01/2021	202101	8	0	Neutro	O limite é baixo, e não é aumentado !		3	4	2	

Fonte: autora (2023).

A base de dados ‘Atendimento’ contém 59 colunas com e sem registros. A partir de análises e algumas reuniões com o especialista da empresa foram encontradas algumas colunas que não faziam sentido para esta pesquisa, tendo sido removidas da base de dados ‘Atendimento’. Na Tabela 6, a seguir são indicados os nomes das colunas (atributos) que foram removidas em consenso com o especialista da empresa.

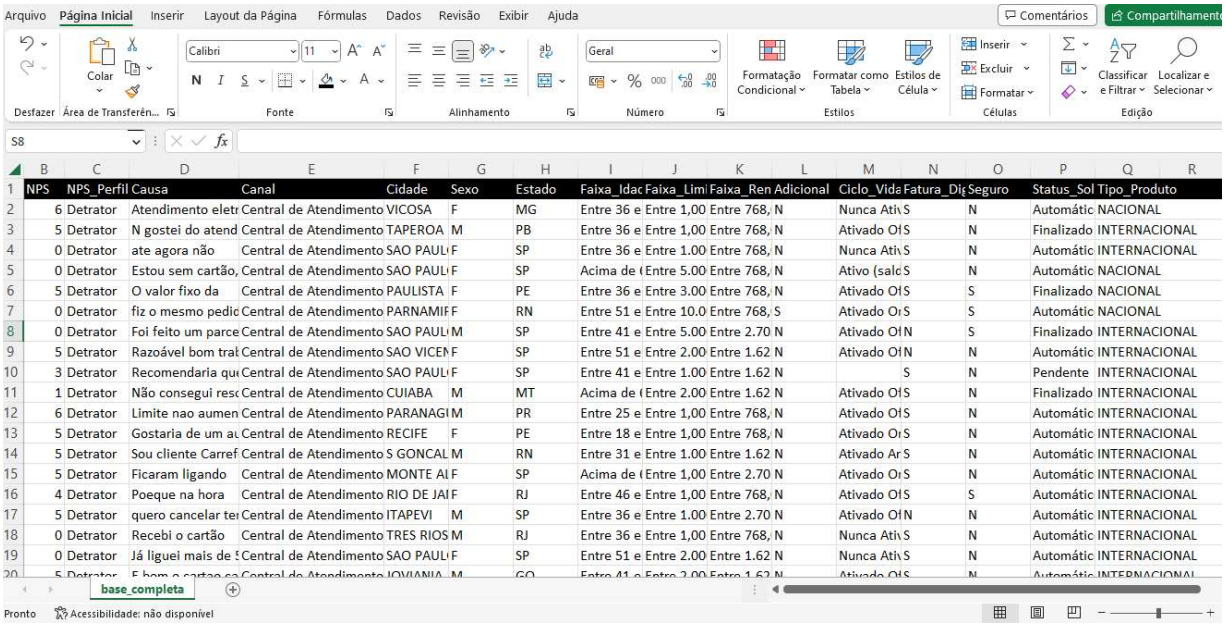
Tabela 6 – Nomes de colunas (atributos) removidas da base.

Nome das colunas removidas da Base de Atendimento	
Identificador	Data da entrevista
Ano_mes_entidade	NPS Média
P03_O que falta pra 10	P04_Sat_Atendimento recebido
P05_Sat_Educação e cortesia do atendente	P05_Sat_Disposição do atendente em resolver
P05_Sat_Informações passadas pelo atendente	P05_Sat_Tempo de espera pelo atendimento
P05_Sat_Temp total do atendimento	P05_Sat_Solução dada
P06_principal motivo do contato	P07_Detalhe do motivo
P08_Status da solicitação	P09_1_Tentou resolver por outro canal
P09_10_Tentou resolver por outro canal	P10_1_Opinião URA
Causa 1	Subcausa 1
Confiança 1	Causa 2
Subcausa 2	Confiança 2
NR_Conta	Numberx
Data_atendimento	Regional
Mob	FX_Mob
Ano_Mes_Ultima_Compra	CD_Portfolio
Classe_Social	Mindset
Nivel_Engajamento	Referencia
Bandeira	Canal_Adesao
CD_Posicao	Loja_Adesao
Nome_Loja	Nome_Loja_Central
Atendente	

Fonte: autora (2023).

As 44 colunas removidas da base de dados foram analisadas e avaliadas em conjunto com o especialista da empresa varejista. Muitas dessas colunas não tinham registros, o que foi considerado um critério para sua remoção. Outras colunas apresentavam registros, mas o especialista avaliou e concluiu que tais registros não tinham sentido conforme o objetivo e desenho de pesquisa delineados. Na Figura 18 é exposto como ficou a base ‘Atendimento’ após a remoção realizada.

Figura 18 – Base de dados ‘Atendimento’ após a remoção de colunas.



	NPS	NPS_Perfil	Causa	Canal	Cidade	Sexo	Estado	Faixa_Idac	Faixa_Limi	Faixa_Ren	Adicional	Ciclo_Vida	Fatura_Di	Seguro	Status_Sol	Tipo_Produto
2	6	Detrator	Atendimento eletr	Central de Atendimento	VICOSA	F	MG	Entre 36 e	Entre 1,00	Entre 768,	N	Nunca Ativ	S	N	Automátic	NACIONAL
3	5	Detrator	N gostei do atend	Central de Atendimento	TAPEROA	M	PB	Entre 36 e	Entre 1,00	Entre 768,	N	Ativado Ol	S	N	Finalizado	INTERNACIONAL
4	0	Detrator	ate agora não	Central de Atendimento	SAO PAULI	F	SP	Entre 36 e	Entre 1.00	Entre 768,	N	Nunca Ativ	S	N	Automátic	INTERNACIONAL
5	0	Detrator	Estou sem cartão,	Central de Atendimento	SAO PAULI	F	SP	Acima de (	Entre 5.00	Entre 768,	N	Ativo (salc	S	N	Automátic	NACIONAL
6	5	Detrator	O valor fixo da	Central de Atendimento	PAULISTA	F	PE	Entre 36 e	Entre 3.00	Entre 768,	N	Ativado Ol	S	S	Finalizado	NACIONAL
7	0	Detrator	fiz o mesmo pedic	Central de Atendimento	PARNAMIF	F	RN	Entre 51 e	Entre 10.0	Entre 768,	S	Ativado Ol	S	S	Automátic	NACIONAL
8	0	Detrator	Foi feito um parce	Central de Atendimento	SAO PAULI	M	SP	Entre 41 e	Entre 5.00	Entre 2.70	N	Ativado Ol	N	S	Finalizado	INTERNACIONAL
9	5	Detrator	Razoável bom trat	Central de Atendimento	SAO VICEN	F	SP	Entre 51 e	Entre 2.00	Entre 1.62	N	Ativado Ol	N	N	Automátic	INTERNACIONAL
10	3	Detrator	Recomendaria qu	Central de Atendimento	SAO PAULI	F	SP	Entre 41 e	Entre 1.00	Entre 1.62	N		S	N	Pendente	INTERNACIONAL
11	1	Detrator	Não consegui resc	Central de Atendimento	CUIABA	M	MT	Acima de (	Entre 2.00	Entre 1.62	N	Ativado Ol	S	N	Finalizado	INTERNACIONAL
12	6	Detrator	Limite nao aumen	Central de Atendimento	PARANAGI	M	PR	Entre 25 e	Entre 1,00	Entre 768,	N	Ativado Ol	S	N	Automátic	INTERNACIONAL
13	5	Detrator	Gostaria de um au	Central de Atendimento	RECIFE	F	PE	Entre 18 e	Entre 1,00	Entre 768,	N	Ativado Ol	S	N	Automátic	INTERNACIONAL
14	5	Detrator	Sou cliente Carref	Central de Atendimento	S GONCAL	M	RN	Entre 31 e	Entre 1.00	Entre 1.62	N	Ativado Ar	S	N	Automátic	INTERNACIONAL
15	5	Detrator	Ficaram ligando	Central de Atendimento	MONTE AL	F	SP	Acima de (	Entre 1,00	Entre 2.70	N	Ativado Ol	S	N	Automátic	INTERNACIONAL
16	4	Detrator	Poeque na hora	Central de Atendimento	RIO DE JAI	F	RJ	Entre 46 e	Entre 1,00	Entre 768,	N	Ativado Ol	S	S	Automátic	INTERNACIONAL
17	5	Detrator	quero cancelar te	Central de Atendimento	ITAPEVI	M	SP	Entre 36 e	Entre 1.00	Entre 2.70	N	Ativado Ol	N	N	Automátic	INTERNACIONAL
18	0	Detrator	Recebi o cartão	Central de Atendimento	TRES RIOS	M	RJ	Entre 36 e	Entre 1,00	Entre 768,	N	Nunca Ativ	S	N	Automátic	INTERNACIONAL
19	0	Detrator	Já liguei mais de	Central de Atendimento	SAO PAULI	F	SP	Entre 51 e	Entre 2.00	Entre 1.62	N	Nunca Ativ	S	N	Automátic	INTERNACIONAL

Fonte: autora (2023).

O critério utilizado foi a seleção da base de dados com registros de 2021, com foco nos comentários dos clientes. Na base de dados, os comentários dos clientes estão na coluna rotulada como (Causa), doravante denominado como o nome do atributo na base que contém todos os comentários de clientes, conforme exposto na Figura 18.

Esses comentários de clientes vêm das mídias sociais. Ressalta-se que foram selecionados apenas os comentários associados aos detratores do NPS (Net Promoter Score) para realização dos experimentos nesta dissertação. Assim, somente os comentários associados às notas atribuídas pelos clientes que variavam de 0 (zero) a 6 (seis), conforme a metodologia NPS. Importante ressaltar que todas as perguntas/respostas que fazem parte desta base de dados são respondidas pelos

clientes diretamente nos canais de comunicação da empresa (central de atendimento, estandes – lojas, aplicativos e portais).

Segundo Magalhães (2019), o NPS é um método de satisfação do cliente desenvolvido para avaliar o nível de fidelização do cliente como a empresa, facilmente aplicável por qualquer empresa, independentemente de seu tamanho e setor. O NPS pode medir desde a satisfação global, até cada ponto de contato com o cliente da empresa.

Para calcular o NPS, considera-se a porcentagem de clientes promotores e subtrai-se dele a porcentagem de clientes detratores. De acordo com Reichheld (2023), calcular o NPS é bem simples, por exemplo, se 60% dos respondentes são promotores, 10% são detratrores e 30% são neutros, seu NPS seria 60-10=50. A pontuação é número inteiro que varia de -100 a 100 e indica a satisfação com a empresa. Magalhães (2019) indica que os clientes são classificados em três níveis distintos, quais sejam:

- Promotores (notas 9 ou 10): são clientes genuinamente satisfeitos com produtos e serviços e com maior probabilidade de recomendar a empresa;
- Neutros (notas 7 ou 8): são clientes que não estão engajados com a empresa e satisfeitos passivamente. Eles normalmente não recomendam os produtos ou serviços da empresa;
- Detratores (notas de 0 a 6): são clientes insatisfeitos com os produtos ou serviços que a empresa oferece.

Nesta dissertação, apenas os detratores (notas de 0 e 6) foram considerados para realização dos experimentos e descoberta de conhecimento. Os neutros (notas de 7 e 8) não foram considerados neste trabalho por serem comentários ambíguos, e os positivos (notas de 9 a 10) também não foram considerados por serem comentários positivos sobre produtos e serviços da empresa.

A base de dados contém 17.547 linhas com registros de comentários de clientes. Após a remoção de alguns atributos, em concordância com o especialista da empresa varejista, a base de dados ficou com 16 colunas (atributos) com os respectivos registros validados pelo especialista da empresa quanto aos objetivos de pesquisa desta dissertação. Na tabela 7 a seguir são apresentados os nomes das colunas (atributos) e suas respectivas descrições.

Tabela 7 – Nome e descrição dos atributos.

Nome da coluna	Descrição
NPS	coluna de notas NPS atribuídas de 0 a 6.
Perfil do NPS	o perfil selecionado ‘Detratores’.
Causa	coluna com comentários dos clientes.
Canal	coluna com informações sobre o canal de atendimento se o contato foi feito por meio ‘central de atendimento’, ‘estandes (lojas)’, ‘aplicativo’ ou ‘portais’ da empresa.
Cidade Estado Sexo Faixa de idade Faixa de renda	colunas com os dados pessoais dos clientes.
Faixa limite	coluna com informações de limite de cartão de cliente.
Adicional	coluna com informações se o cliente possui cartões adicionais.
Seguro	coluna com informações se o cartão do cliente possui algum tipo de seguro oferecido pela empresa.
Ciclo de vida	coluna com informações se o cartão fidelidade está ativo, inativo ou nunca ativo.
Fatura digital	coluna com informação se o cliente ativo possui fatura digital.
Tipo de produto	coluna com informações se o cartão é nacional ou internacional.
Status solicitação	coluna contém informar o status do atendimento ‘Automático’, ‘Finalizado’ ou ‘Pendente’.

Fonte: autora (2023).

- Fase 2: Pré-processamento da base de dados ‘Atendimento’.

Nesta fase foi iniciado o pré-processamento de textos (comentários), com a extração/redução de atributos aplicando-se técnicas de inteligência artificial. Para o processo de normalização foi realizada a remoção de ‘stop words’, remoção de caracteres especiais, stemização, lematização, tokenização e padding de palavras.

Segundo Carneiro (2020), a etapa de pré-processamento volta-se a remoção de *stop words* (remoção de palavras) que não agreguem valor semântico ao texto. Quando aplicadas, as palavras chamadas stop words são consideradas dispensáveis em uma frase porque, embora contribuam para a compreensão do que está escrito,

não contribuem muito para o resultado final da semântica da frase. Exemplos de *stop words* são artigos, conjunções e preposições.

Outro método para transformar palavras em texto é a *tokenização*, ação em que cada palavra única recebe um ID de identificação. Um pré-processamento importante a ser executado é a *padding* de palavras, que possibilita o preenchimento (conclusão) da matriz de palavras, uma vez que ela é tokenizada. Portanto, as palavras semelhantes assumem dimensões próximas umas das outras (CARNEIRO, 2020).

Nesta dissertação, as bibliotecas Seaborn e Word Cloud foram aplicadas para visualizar gráficos e palavras em nuvens. Segundo Lago (2022), Seaborn é uma ferramenta de gráficos em Python, ideal para análise exploratória de dados e para aprender em profundidade como os dados se comportam em um banco de dados. E a Word Cloud (nuvens de palavras) é uma ferramenta voltada à representação visual de palavras, destacando as palavras que aparecem com mais frequência.

- Fase 3: Transformação dos dados da base de 'Atendimento'.

Esta fase é executada para transformar os dados originais em formatos mais apropriados e adequados ao processo de mineração. Ela envolve as seguintes atividades:

- Normalização para dimensionar os valores de dados em um intervalo específico;
- Seleção de atributos para a geração de novos atributos a partir de um conjunto de atributos fornecido para ajudar no processo de mineração;
- Discretizarão para a transferência de funções contínuas, modelos, variáveis e equações em contrapartes discretas.
- Redução e sintetização dos dados.

Todos os tratamentos dos dados foram realizados para dar início ao processo de mineração de dados. Os dados ficaram agrupados em um mesmo local para a aplicação dos modelos de análise. Ainda nesta fase, pegou-se todo o texto pré-processado e criou-se BagOfWords.

O BagOfWords é uma representação simplificada das palavras de um texto usado para tarefas de Processamento de Linguagem Natural, com a mineração de

texto. Segundo Fonseca (2020), BagOfWords é uma forma de representar o texto de acordo com as ocorrências de palavras. Quando traduzida para o português, o ‘saco de palavras’ recebe esse nome porque não leva em consideração a ordem ou a estrutura das palavras no texto, mas tão somente se elas ocorrem ou a frequência com que ocorrem no texto.

- Fase 4: Aplicação de técnicas de Redes Neurais Convolucionais.

Nesta fase, primeiramente foram construídos modelos e aplicação de técnicas de mineração de dados com o objetivo de extrair conhecimento. Usando redes neurais convolucionais para classificar os resultados e aplicação de métricas de avaliação de desempenho como acurácia, precisão, recall e F1 Score para avaliar o modelo.

Após o treinamento foram realizados os primeiros testes gerando-se assim a matriz de confusão com os resultados dos experimentos. Logo após todo o treinamento e teste foi avaliada e selecionada a técnica com melhor resultado, conforme as métricas de avaliação de desempenho consideradas (acurácia, precisão, recall e F1 Score) da técnica de IA aplicada.

- Fase 5: Interpretação e avaliação do conhecimento descoberto com base nas técnicas selecionadas na fase anterior.

Nesta fase, com base no modelo de classificação utilizado, foi avaliado o desempenho do modelo criado anteriormente, considerando-se as correlações determinadas e o cruzamento dos dados para a descoberta de conhecimento do cliente. A validação pôde ser feita de diferentes formas, com uso de medidas estatísticas e avaliação do especialista da empresa.

No Quadro 3 são apresentadas as correlações das categorias de dados para o cruzamento realizado de acordo tópico 3.3.1, visando a descoberta de conhecimentos sobre os clientes da empresa varejista enfocada nesta pesquisa.

Quadro 3 – Correlações de categorias de dados para cruzamentos.

Causa (comentário do cliente)	Perfil do cliente (sexo, idade, renda e estado)
	Produto cartão (faixa limite, tipo de produto, ciclo de vida e adicional)
	Canais de atendimento (central de atendimento, estandes/lojas, aplicativos, status da solicitação)
	Estratificação de notas do NPS (notas de 0 a 6, segregadas em dois substratos: notas de 0 a 3 e notas de 4 a 6).

Fonte: autora (2023).

- Fase 6: Comparação dos resultados dos experimentos com os resultados dos indicadores detratores do NPS (*Net Promoter Score*) da empresa.

A comparação foi realizada em conjunto com os profissionais especialistas da empresa varejista analisada. Por fim, a descoberta de conhecimentos será compartilhada com outros profissionais e áreas afins da organização, para que esta pesquisa possa colaborar com a empresa varejista e fornecer informações/resultados que suportem o processo de tomada de decisão.

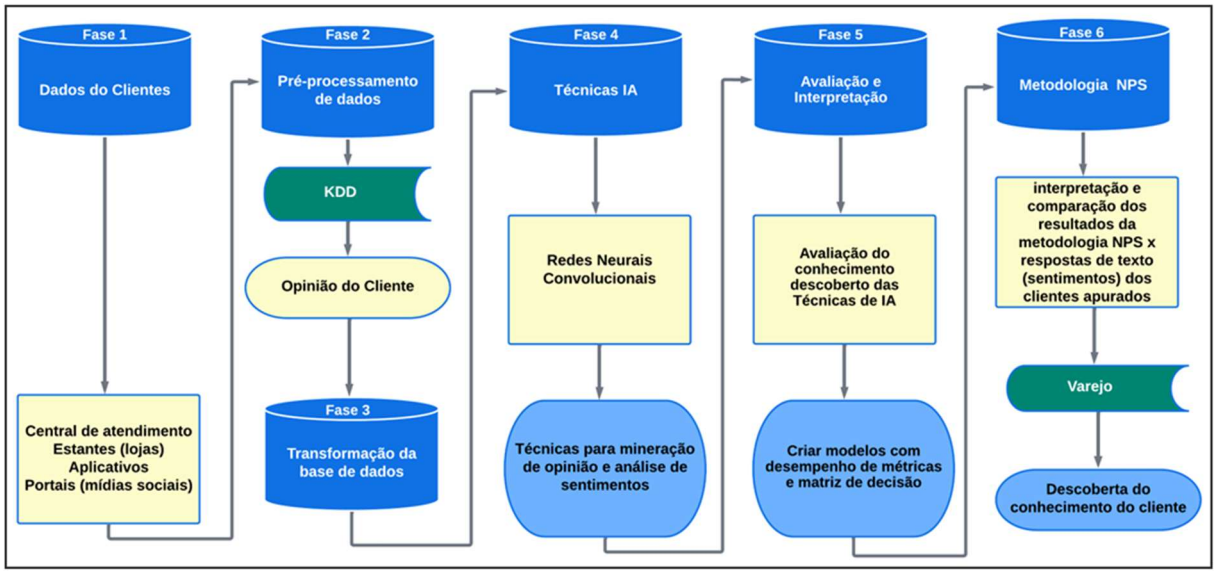
3.4 Modelo teórico-empírico

O modelo teórico-empírico desta pesquisa considera a descoberta de conhecimento de clientes a partir de dados registrados em opiniões manifestadas em redes sociais e mídias sociais de um varejista atuante no país. A temática abordada neste estudo inclui técnicas/métodos de inteligência artificial e redes neurais convolucionais voltadas à análise de sentimentos em comentários postados em mídias sociais e redes sociais, visando a descoberta de conhecimento em bancos de dados (KDD), bem como o cruzamento desse conhecimento com a metodologia NPS (*Net Promoter Score*) já implementada na empresa foco desta investigação.

De acordo com Juristo e Moreno (2013), o desenho do modelo teórico-empírico deve ser consistente, ou seja, deve ser definido de tal forma que a pesquisa e

experimentos possam ser operacionalizados. O modelo teórico-empírico desta pesquisa é apresentado na Figura 19.

Figura 19 – Modelo teórico-empírico.



Fonte: autora (2023).

Conforme exposto na Figura 19, primeiramente houve a importação dos comentários de clientes extraídos de redes sociais e mídias sociais da empresa, que contêm opiniões sobre os serviços e produtos oferecidos. Após a coleta desses dados dos clientes, iniciou-se o processo do KDD (*Knowledge Discovery in Databases*) pela seleção dos dados.

O pré-processamento (tratamentos dos dados), foi iniciado logo após a seleção dos dados, sendo esses dados limpos e preparados, com exclusão de colunas que não continham registro e exclusão de colunas que não contêm dados de comentários do cliente e, portanto, são desnecessárias para a abordagem da análise de sentimentos enfocada nesta pesquisa.

O próximo passo a ser realizado nesta dissertação foi a aplicação de técnicas inteligentes de Redes Neurais Convolucionais para realizar a classificação e análise de sentimentos dos comentários dos clientes extraídos das mídias sociais e redes sociais, visando assim a descoberta de conhecimento acerca do cliente.

Em seguida, avalia-se e compara-se o desempenho das técnicas inteligentes aplicadas no processo de análise e classificação de sentimentos. Para tanto, será

desenvolvida uma abordagem para extração de conhecimento a partir de comentários armazenados em bancos de dados.

Em seguida, após avaliação e comparação do desempenho das técnicas aplicadas, os resultados foram comparados com os resultados do NPS disponíveis e fornecidos pela empresa. Por fim, os resultados foram interpretados e avaliados acerca do conhecimento sobre o cliente para proporcionar apoio à tomada de decisão da empresa. Estima-se que a partir desses resultados a empresa varejista possa conhecer mais detalhadamente seus clientes e assim oferecer serviços e produtos mais adequados aos clientes.

4. Apresentação, Análise e Discussão dos Resultados

Neste capítulo são apresentados, analisados e discutidos os resultados dos experimentos computacionais realizados. Conforme descrito no capítulo 3, os experimentos foram divididos em seis fases previstas para a descoberta de conhecimento do cliente, conforme os tópicos apresentados a seguir.

4.1 Fase 1 – Seleção da base de dados de comentários de clientes (base ‘Atendimento’)

Nesta fase foi selecionada a base de dados original utilizada para a realização dos experimentos. A base contém 17.547 linhas com registros de comentários de clientes e catorze colunas de atributos, quais sejam: sexo, idade, renda, estado, faixa de limite, tipo de produto, ciclo de vida e adicional, status da solicitação, canal de atendimento (central de atendimento, estande/lojas, aplicativos e portal), notas do NPS de 0 a 3 e, por fim, notas do NPS de 4 a 6.

Primeiramente, com o apoio do especialista da empresa, foi realizada a análise da base e retiradas todas as linhas e colunas que não continham registros, uma vez que havia muitos campos sem registros por ser uma base não estruturada. Apenas os resultados detratores do NPS (notas de 0 a 6, numa escala de 0 a 10) foram considerados para a realização dos experimentos.

Após essa análise em conjunto com o especialista, restaram *9.382* registros no banco de dados com NPS detratores, que foram considerados para os experimentos delineados nesta pesquisa a fim de obter *insights* para a tomada de decisão da empresa. Foram criados conjuntos de atributos para correlações de categorias de dados visando os cruzamentos entre eles para a execução dos experimentos computacionais.

A primeira a correlação foi entre o atributo Causa (comentários de clientes), vinculado ao conjunto de atributos do Perfil do Cliente (sexo, idade, renda, estado). Em complemento foi realizada a correlação entre os atributos comentário do cliente associando-o aos atributos notas do NPS detratores (‘0’ a ‘6’), segregadas em dois conjuntos (‘0 a 3’ e ‘4 a 6’).

Neste estudo optou-se por apresentar apenas as categorias do atributo perfil do cliente. Embora fosse possível demonstrar as demais correlações com os outros atributos (‘produto cartão’ e ‘canais de atendimento’) e suas respectivas categorias, entende-se que a solução desenvolvida seja passível de aplicação e validação tão somente a partir dos resultados dos experimentos com o atributo perfil do cliente. Ou seja, experimentos realizados com a categoria do perfil do cliente X causa (comentário de cliente) já são capazes de demonstrar que as técnicas inteligentes e experimentos aplicados nesta pesquisa são viáveis e apresentam resultados satisfatórios para a descoberta de conhecimento do cliente a fim de se obter *insights* para os gestores da empresa.

Na próxima seção será iniciada a etapa de pré-processamento da base de dados ‘Atendimento’ com a aplicação de técnicas inteligentes para a condução dos experimentos.

4.2 Fase 2 – Pré-processamento da base de dados

Nesta fase inicia-se o pré-processamento dos dados selecionados na fase 1, para torná-los aptos para a execução das próximas fases do processo de descoberta de conhecimento delineado nesta pesquisa. Para a leitura da base de dados foi utilizado o *Google Collaboratory Notebook* (*Colab*), produto do *Google Research*, que é uma ferramenta amplamente utilizada na área científica para a realização de experimentos computacionais. Esta ferramenta permite que qualquer pessoa escreva e execute código *Python* no navegador, sendo particularmente útil para aprendizado de máquina, análise de dados e educação (COLABORATORY, 2022).

Para a realização dos experimentos, as bibliotecas NLTK, *Spacy, Pandas, Randon, Numpy, Re, Seaborn, Matplotlib, PIL* e *WordCloud* foram instaladas e importadas no *notebook colab* para a execução dos experimentos. A seguir é descrita a ação de carregamento da base de dados.

4.2.1 Carregamento da base de dados

Nesta fase, a base de dados foi carregada para execução do pré-processamento dos textos comentários de clientes. A seguir, inicia-se a extração/redução de atributos aplicando as técnicas inteligentes. Na Figura 20 é exposta a base de dados selecionada.

Figura 20 – Carregamento da base de dados.

```
[5]  base_dados = pd.read_csv('/content/drive/MyDrive/Experimentos 2023/base 2023/base_dados_completa.csv'
                              encoding='ISO-8859-1', delimiter=';')

[6]  base_dados.shape

     (9382, 5)
```

	NPS	NPS_Perfil	Causa	Canal	Cidade	Sexo	Estado	Faixa_Idade	Faixa_Limite	Faixa_Renda	Adicional	Ciclo_Vida	Fatura_Digita
0	6	Detrator	Atendimento eletrônico muito lento e confuso.	Central de Atendimento	VICOSA	F	MG	Entre 36 e 40 anos	Entre 1,00 e 1.000,00	Entre 768,01 e 1.625,00	N	Nunca Ativo	
1	5	Detrator	N gostei do atendimento	Central de Atendimento	TAPEROA	M	PB	Entre 36 e 40 anos	Entre 1,00 e 1.000,00	Entre 768,01 e 1.625,00	N	Ativado Off	
2	0	Detrator	ate agora não recebi eu cartão em casa, toda ...	Central de Atendimento	SAO PAULO	F	SP	Entre 36 e 40 anos	Entre 1.000,01 e 2.000,00	Entre 768,01 e 1.625,00	N	Nunca Ativo	
3	0	Detrator	Estou sem cartão, bloquei um e o outro tive pr...	Central de Atendimento	SAO PAULO	F	SP	Acima de 60 anos	Entre 5.000,01 e 7.500,00	Entre 768,01 e 1.625,00	N	Ativo (saldo)	
4	5	Detrator	O valor fixo da anuidade me	Central de Atendimento	PAULISTA	F	PE	Entre 36 e 40 anos	Entre 3.000,01 e 4.000,00	Entre 768,01 e 1.625,00	N	Ativado Off	

✓ 0s conclusão: 00:00

Fonte: autora (2023).

4.2.2 Funções para pré-processamento de textos

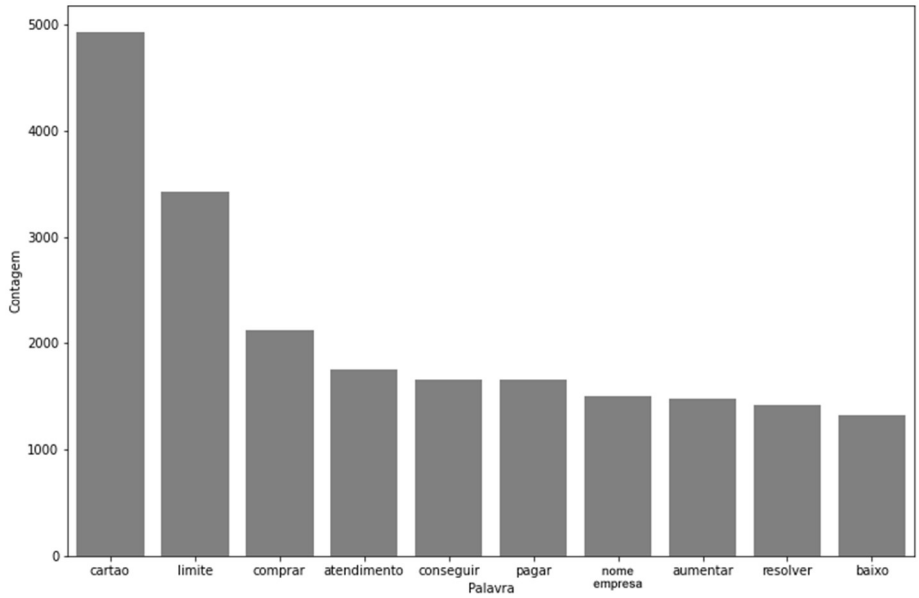
Nesta fase, os dados de causa (comentários de clientes) selecionados na fase 1 foram pré-processados para torná-los adequados para a aplicação de ferramentas de mineração de texto. Primeiramente foram definidas as *stopwords* e *caracteres especiais*, que em seguida foram retirados do texto.

As *stopwords* e caracteres especiais extraídos foram artigos, pronomes, sinais de pontuação e outros elementos textuais não interessantes para a extração do conhecimento proposto neste estudo. Finalmente, todas as palavras foram deixadas em minúsculas. A função desenvolvida para o pré-processamento de textos encontra-se no apêndice desta dissertação.

Na sequência foi criada uma *Bag of Words* (saco de palavras) que foi executada na base de dados para o atributo ‘causa’ (comentários de clientes), conforme código exposto no apêndice. A *Bag of Words* contendo o texto pré-processado possuía o tamanho *91.119 palavras*.

Logo, após a criação dos *Bag of Words*, foi aplicada a técnica de tokenização na base de dados para a geração das palavras com ocorrência de maior frequência. A Figura 21 expõe o gráfico de Pareto com as dez palavras com maior frequência em toda a base de dados de comentários do cliente.

Figura 21 – Gráfico Pareto com as 10 palavras com mais frequências.



Fonte: autora (2023).

As palavras com maior frequência identificadas foram cartão, limite, comprar, atendimento, conseguir, pagar, nome da empresa, aumentar, resolver e baixo. A partir da aplicação da ferramenta *Wordcloud* foi possível gerar uma nuvem de palavras após o pré-processamento dos textos. A nuvem de palavras é apresentada na Figura 22.

Figura 22 – Nuvem de palavras.

Fonte: autora (2023).

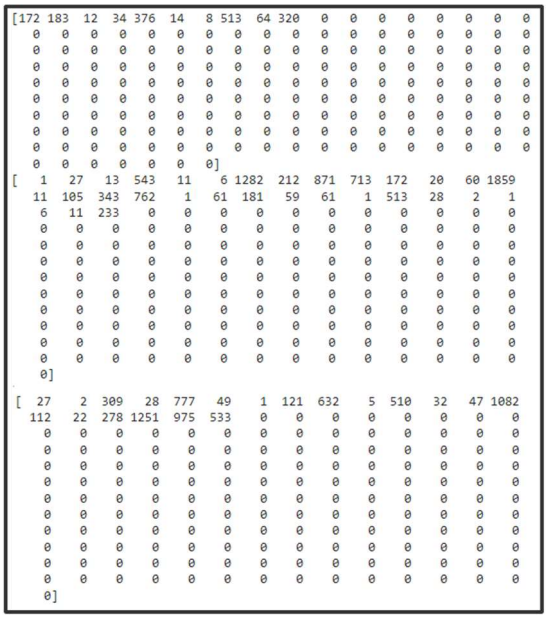
Em consonância às dez palavras com maior ocorrência nos comentários dos clientes, na nuvem de palavras gerada é possível identificar com maior destaque o mesmo conjunto de palavras indicado na figura 22, a exemplo de cartão, comprar e limite, dentre outras ocorrências.

4.2.3 *Padding* e vetorização dos textos

Logo após o processo de normalização dos comentários dos clientes da base de dados foi necessário adequar os textos aos classificadores considerados neste experimento, visto que estes não reconhecem valores (textos) discretos. Portanto, foi necessário aplicar o processo de *padding* e vetorização dos textos antes de iniciar o treinamento do modelo.

A Figura 23 expõe os textos convertidos em números. Cada processo assume uma codificação, sendo que todos os classificadores considerados nos experimentos usaram essa técnica para treinar seus modelos e, em seguida, realizar a classificação do texto.

Figura 23 – Padding e vetorização dos textos.



Fonte: autora (2023).

4.3 Fase 3 – Transformação dos dados da base de ‘Atendimento’

Nesta fase dos experimentos, após realização de todo o processo de pré-processamento do texto através de normalizações, remoção de *stopwords*, *stemming* e *padding*, conforme explicado no capítulo 3, foi iniciada a transformação da base de dados, com a aplicação de técnicas inteligentes para a descoberta de conhecimento do cliente. Para fins ilustrativos foi utilizada a primeira correção descrita anteriormente no atributo ‘causa’ (comentário do cliente), de modo a vinculá-lo ao conjunto de atributos do perfil do cliente (sexo, idade, renda, estado), conforme ilustrado na Figura 24.

Figura 24 – Correção causa x perfil do cliente.

	causa	sexo	idade	renda	estado
0	atendimento eletronico lento confuso	F	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	MG
1	atendimento	M	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	PB
2	receber cartao ridicula receber receber pesso...	F	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	SP
3	cartao bloquei problema	F	Acima de 60 anos	Entre 768,01 e 1.625,00	SP
4	fixar anuidade incomodar carrefour cartao prin...	F	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	PE

Fonte: autora (2023).

Cada registro do atributo causa (comentário do cliente) foi relacionado aos respectivos os atributos sexo, idade, renda e estado do perfil do cliente que realizou o comentário. A título de exemplo, uma cliente do sexo feminino, com idade entre 36 e 40 anos, renda entre R$ 768,01 e R$ 1.625,00 e moradora no estado de Minas Gerais foi associada ao comentário ‘atendimento eletrônico lento confuso’.

4.3.1 Divisão de base em treinamento e teste

A base de dados original foi dividida em Treinamento com 6.567 registros e Teste com 2.815 registros. A Figura 25 mostra como a base de dados foi dividida.

Figura 25 – Divisão de base em treinamento e teste.

```
[232] train_inputs, test_inputs, train_labels, test_labels = train_test_split(data_inputs,
                                                                              data_labels,
                                                                              test_size=0.3,
                                                                              stratify = data_labels)
```

```
[234] train_inputs.shape
      (6567, 147)
[235] train_labels.shape
      (6567,)
[236] test_inputs.shape
      (2815, 147)
[237] test_labels.shape
      (2815,)
```

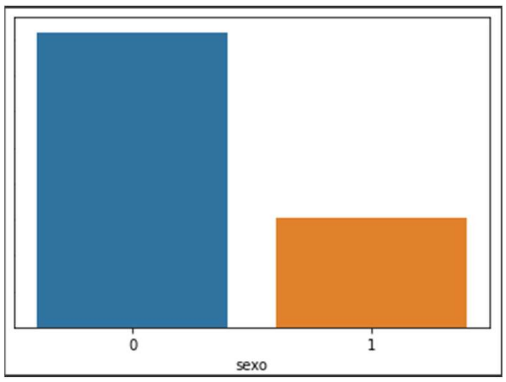
Fonte: autora (2023).

4.3.2 Tratamento das classes (atributos)

Nesta etapa dos experimentos foram aplicadas técnicas inteligentes para modificar os atributos da base de dados em formato numérico de 0 e 1. Na seção Apêndice está disponível a função desenvolvida para alterar os atributos da base de dados. A Figura 26 mostra um exemplo de como o atributo de gênero (M - masculino, F - feminino) foi convertido em um formato numérico.

Figura 26 – Modificação de atributos textuais para o formato numérico.

	causa	sexo
0	atendimento eletronico lento confuso	0
1	atendimento	1
2	receber cartao ridicula receber receber pesso...	0
3	cartao bloquei problema	0
4	fixar anuidade incomodar carrefour cartao prin...	0
5	aprovar motivo cancelar comprar	0
6	parcelamento automatico fatura autorizacao at...	1
7	razoavel trabalhar	0
8	recomendar atentar relacao cobrancas indevido ...	0
9	conseguir resolver pleitear	1
10	limite aumentar dificultar opcoes comprar loj...	1



Fonte: autora (2023).

A seguir inicia-se a fase 4 do processo de descoberta de conhecimento KDD adaptado para esta pesquisa. Nela foram aplicadas técnicas inteligentes para a

criação de modelos para classificação de texto, visando a descoberta de conhecimento do cliente.

4.4 Fase 4 – Modelo Redes Neurais Convolucionais

Nesta fase, após a transformação da base de dados, aplicou-se ferramentas para a classificação de texto com o objetivo de extrair conhecimento. Primeiramente, verificou-se quais eram as principais palavras chaves (*keywords*). Para isso, a ferramenta Rake foi utilizada para construir um ranking de palavras-chaves. Na Tabela 8 são apresentadas as dez primeiras posições do ranking.

Tabela 8 – *Ranking* de palavras.

Ranking	**Palavras**	**Score**
1	atendimento / cartão / vencimento/ cancelamento	25.0
2	cobrar/ juro/ cliente	24.5
3	cancelamento / assinatura/ sistema	23.0
4	atendimento / cartao / cobrar/ anuidade	16.8
5	parcelar / cartão / alimentacao	15.8
6	aumentar / limite / cancelar	15.8
7	limite / aumentar/ parcelar	15.5
8	senha / cadastro / cartao	9.5
9	diminuíram / atraso	9.5
10	bloquear / cartao	9.0

Fonte: autora (2023).

Ao efetuar a análise dos resultados expostos na tabela 8, percebe-se quais são as palavras-chave mais importantes identificadas nos textos de comentários dos clientes. O parâmetro *score* foi utilizado para formar o *ranking* das palavras. O parâmetro de avaliação é a soma dos valores das métricas frequências da palavra, grau da palavra e relação grau/frequência.

Frequência da palavra é o valor que indica a frequência com que a palavra ocorre no texto. O grau da palavra é o valor que representa as palavras que se repetem com mais frequência com outras palavras no *ranking*. E a relação grau-frequência é a divisão entre o grau da palavra e a frequência da palavra. Sendo assim,

as palavras atendimento, cartão, vencimento e cancelamento foram as palavras com maior *score*, pois além de serem as palavras mais citadas no texto, também sempre foram citadas junto com as demais palavras que aparecem na classificação do texto. Após a construção do *ranking* de palavras-chave apresentado no quadro, a seguir é exposta a aplicação das técnicas inteligentes para criação da modelagem de tópicos.

4.4.1 Criando modelagem de tópicos com a ferramenta Gensim

Nesta etapa foi criado um modelo para gerar os principais tópicos da base de dados. A ferramenta Gensim foi utilizada para realizar a tarefa de modelagem de tópicos. Para tanto, foi realizada a tokenização do texto e a criação de um dicionário de termos das palavras. Dessa forma, cada termo será um índice. Em seguida, foi realizada a conversão da lista em uma matriz de texto, tendo como parâmetro o dicionário criado anteriormente.

Logo após foi realizado o treinamento e, na sequência, a execução do modelo LDA na matriz de termos, visando a geração dos principais tópicos. Os dez principais tópicos encontrados na base de dados pela ferramenta Gensim são apresentados na tabela 9.

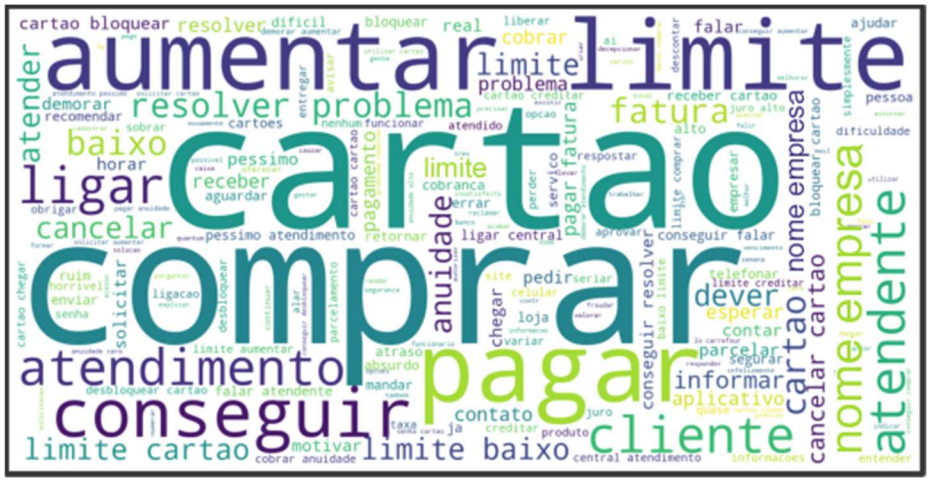
Tabela 9 – Principais tópicos.

	Principais tópicos
1	cartao / limite / atendimento / comprar / nao conseguir
2	cartao / limite / comprar / nao conseguir/ pagar
3	cartao / limite / atendimento / comprar / aumentar
4	cartao / limite / atendimento / comprar / nome empresa
5	cartao / limite / comprar / pagar / atendimento
6	cartao / limite / comprar / atendimento / nao conseguir
7	cartao / limite / não conseguir / pagar / atendimento
8	cartão / limite / comprar / nao conseguir / atendimento
9	cartao / limite / comprar / nao conseguir / pagar
10	cartao / limite / atendimento / comprar / aumentar

Fonte: autora (2023).

Nesta etapa foi analisado todo o conteúdo da *BagOfWords* criada, com a consequente geração dos principais tópicos. Os principais tópicos (problemas) identificados foram: *cartão*, *limite*, *atendimento*, *compra* e *pagamento*. Na Figura 27 é possível visualizar a nuvem de palavras com os principais temas indicados anteriormente.

Figura 27 – Nuvem de palavras dos principais tópicos.



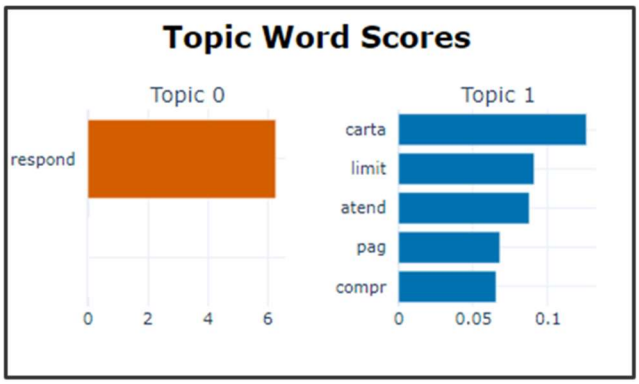
Fonte: autora (2023).

4.4.2 Modelagem de tópicos com a ferramenta Bertopic

Nesta etapa a ferramenta Bertopic foi aplicada para gerar representações de tópicos. Segundo Grootendorst (2022), primeiro cada documento é convertido em sua representação incorporada usando para tanto um modelo de linguagem pré-treinado. Em seguida, antes de agrupar essas incorporações, a dimensionalidade das incorporações resultantes é reduzida para agilizar o processo de agrupamento. Por fim, as representações de tópicos são extraídas dos clusters de documentos usando uma variante personalizada baseada em classe do TF-IDF.

Assim, o Bertopic foi aplicado para identificar os principais tópicos previamente identificados na base de dados. A Figura 28 mostra os tópicos encontrados e as palavras associadas.

Figura 28 – Modelagem de Tópicos e palavras correlacionadas.



Fonte: autora (2023).

A análise dos resultados expostos na Figura 28 mostra que persistem os cincos temas descobertos com a ferramenta Gensim, quais sejam: cartão, limite, atendimento, pagamento e compra. Após a aplicação do modelo Bertopic, a Figura 29 mostra os principais tópicos encontrados na biblioteca.

Figura 29 – Principais tópicos encontrados utilizando-se a biblioteca Bertopic.

```
topic_model.get_topic(1)

[('cartao', 0.17331227942109445),
 ('limite', 0.12165412396967884),
 ('comprar', 0.08549579782046478),
 ('atendimento', 0.07414527457233243),
 ('conseguir', 0.07114273796535003),
 ('pagar', 0.07094722618274876),
 ('nome empresa', 0.06595112219333853),
 ('aumentar', 0.06524430272849434),
 ('resolver', 0.06313947919404458),
 ('baixo', 0.05989469346332091)]
```

Fonte: autora (2023).

Para validar a classificação do *ranking* de palavras-chave e da modelagem de tópicos, calculou-se a frequência de todas as palavras, exceto os *stopwords*, que foram calculadas aplicando-se a biblioteca de probabilidade. Na Tabela 10 são apresentadas as palavras mais encontradas nos comentários dos clientes.

Tabela 10 – Palavras mais frequentes.

	Palavras frequentes	
1	cartao	4.931
2	limite	3.432
3	comprar	2.115
4	atendimento	1.749
5	conseguir	1.656
6	pagar	1.650
7	nome empresa	1.499
8	aumentar	1.478
9	resolver	1.416
10	baixo	1.322

Fonte: autora (2023).

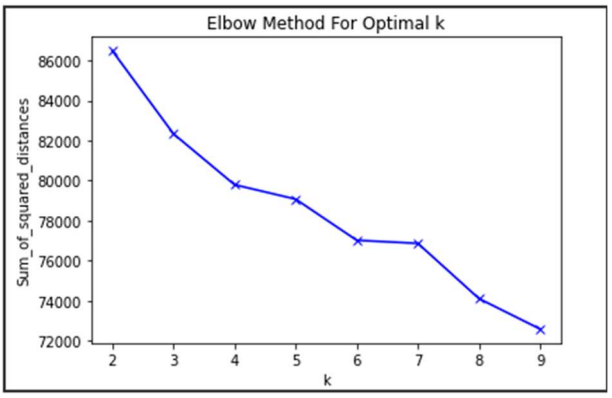
4.4.3 Criação de *clusters* (agrupamentos)

Nesta etapa foi criada um método para aplicação de K-means (agrupamentos), que é um método de segregar em torno de centros diversos dados, criando o *clustering* que produz o efeito de particionar ‘n’ observações dentre ‘k’ grupos, onde cada observação pertence ao grupo mais próximo da média.

A técnica K-means foi escolhida por ser uma técnica de agrupamento aplicável à mineração de textos. É necessário definir K, ou seja, o número de grupos-alvo.

Para encontrar o melhor K foi calculada a soma dos quadrados intra-cluster (soma dos quadrados intra-cluster, comumente abreviada como wcss), levando-se em consideração que quanto mais próximo de ‘0’ o valor estiver, melhor será o K. A Figura 30 expõe a soma dos quadrados intra-cluster.

Figura 30 - Soma dos quadrados intra-clusters.



Fonte: autora (2023).

Analisando-se a Figura 30 faz-se perceptível que os valores de erro para os valores de K de 2 a 9 foram elevados. Portanto, os clusters gerados pela técnica K-means não foram considerados. Deve-se entender que houve muitos erros, pois a *Bag of words* criada contém um número significativo de palavras que se repetem. Na próxima seções são apresentadas as técnicas utilizadas para o agrupamento de palavras, que contribuíram significativamente para os resultados acerca de descoberta de conhecimento sobre o cliente.

4.4.4 Construção de classificador e treinamento

Nesta etapa foi criado um modelo de classificador de textos, com a definição das categorias (classes), adição às classes feminino e masculino e início do treinamento do modelo. Para tanto, utilizou-se o modelo begin_training para iniciar o treinamento. Este modelo utiliza o recurso de *Deep Learning* com redes neurais convolucionais para a classificação dos textos. O gráfico matplotlib exposto na Figura 31 ilustra a progressão dos erros do classificador. O modelo criado apresentou bom desempenho na classificação dos atributos do perfil do cliente.

Figura 31 – Progressão dos erros do classificador.

Fonte: autora (2023).

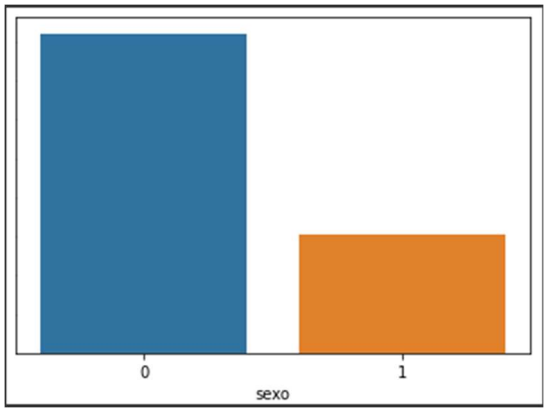
Após a criação do modelo de classificação, o modelo foi carregado e foram executados alguns testes na base de dados. O modelo apresentou resultado satisfatório na classificação de textos e os respectivos atributos de perfil de cliente. O próximo passo foi aplicar as métricas de avaliação de desempenho são elas, (acurácia, precisão, recall e F1 score) à matriz de confusão para pontuar o modelo.

4.4.5 Avaliação do modelo e aplicação de métricas de avaliação de desempenho

Nesta etapa são demonstrados os resultados apresentados nos classificadores. A base de dados foi dividida em treinamento e teste. O conjunto de treinamento representou 80% do *feedback* do cliente, enquanto o conjunto de teste representou 20% do *feedback* do cliente.

Para treinamento e teste o modelo considerou Causa (comentários do cliente) X Perfil do Cliente (atributo sexo). O balanceamento entre o perfil do cliente (sexo: 0 - feminino e 1 – masculino) foi realizado conforme exposto na Figura 32.

Figura 32 – Distribuição da base de teste.

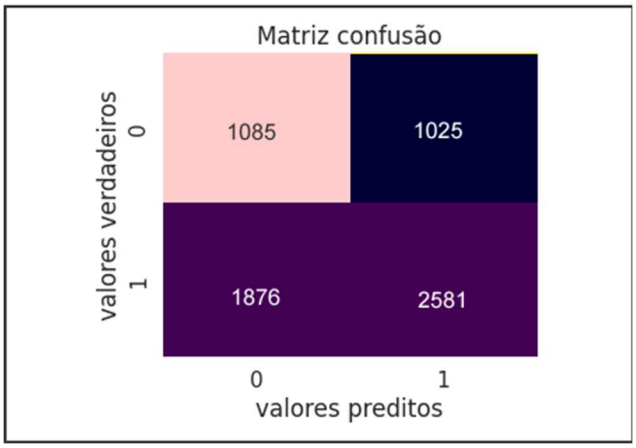


Fonte: autora (2023).

A primeira métrica utilizada para avaliar os resultados foi a matriz de confusão, avaliando-se assim como cômputo a classe real e a classe prevista, com aplicação dos rótulos ‘P – Positivo’ e ‘N – Negativo’. A métrica de acurácia, que mede a porcentagem de acertos em relação ao número de ocorrências, também foi utilizada. As outras métricas avaliadas foram precisão, recall e pontuação F1-score.

Após a aplicação da base de teste ao modelo, os primeiros resultados podem ser vistos na Figura 33, que expõe a matriz de confusão criada mostrando os acertos e fracassos do modelo.

Figura 33 – Matriz de confusão treinamento.



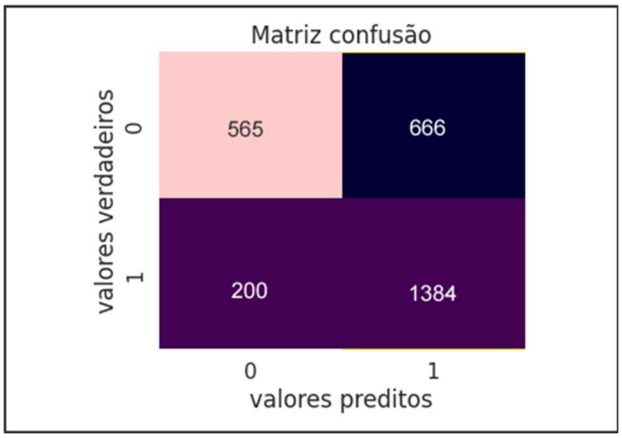
Fonte: autora (2023).

Ao realizar a análise da Figura 33 foi identificado que 6.567 comentários de clientes foram usados em conjunto com o atributo de perfil de cliente. O modelo

classificou 1.025 textos com erros de 15,6%, valor este que representa os falsos positivos. Na segunda linha da matriz de confusão havia 4.457 de textos que, destas, o algoritmo classificou como 2.581 de textos, o que equivale a 39,2% de acerto, este valor representa os verdadeiros positivos.

Um segundo teste foi realizado considerando-se o comentário do cliente e o atributo de perfil do cliente. Esse outro teste foi realizado num conjunto de 2.815 comentários de clientes que não foram usados na fase de treinamento. A Figura 34 é exposta a matriz de confusão gerada no teste realizado com os experimentos.

Figura 34 – Matriz de confusão teste.



Fonte: autora (2023).

De forma diferente dos primeiros testes realizados, neste segundo teste foram utilizados menos registros (apenas 2.815 comentários de clientes), ou seja, aqueles registros que não foram utilizados na fase de treinamento. O modelo classificou 666 comentários com erros de 27,6%, valor este que representa os falsos positivos. Na segunda linha da matriz de confusão o modelo classificou 1.384 comentários, o que corresponde a uma acurácia de 39,2% de acerto, valor este que representa os verdadeiros positivos. Nesta fase, os modelos foram construídos pela primeira vez e técnicas de mineração de dados foram aplicadas para extração de conhecimento dos comentários dos clientes. Aplicou-se redes neurais convolucionais para classificar os resultados e aplicar métricas de avaliação de desempenho, como acurácia, precisão, recall e F1-Score para avaliar o modelo.

Para o processo de treinamento, a base de dados ‘Atendimento’ foi dividida em dois conjuntos: treinamento e teste. O conjunto de treinamento contém 6.567 registros/comentários de clientes, que representa 80% da base. O conjunto de teste

contém 2.815 registros/comentários de clientes, que representa 20% da base. Os seguintes resultados das quatro métricas selecionadas foram usados para as bases de treinamento e teste:

Tabela 11 - Resultados consolidados da Matriz de Confusão – RNC.

Primeira Base de teste com 6.567 comentários de clientes				
Classificador	Acurácia	Precisão	Recall	F1 Score
RNC	39,2%	39,2%	29,1%	56,4%

Fonte: autora (2023).

Como primeiro resultado, apresenta-se o valor de acerto (acurácia) de 39,2%, o que indica uma classificação pouco exitosa das classes de perfis de clientes. Os resultados das métricas de avaliação de desempenho com precisão, recall e F1-score para cada uma das classes analisadas. A precisão do modelo foi de 39,2% de acerto cruzando-se os comentários com os atributos do perfil do cliente. Já o resultado da métrica de recall foi de 29,1%, que não obteve um resultado muito bom. Das métricas utilizadas, a que mais se destacou foi a F1-score que teve um resultado de 56,4%, apresentando-se assim como a métrica que obteve um melhor desempenho nos experimentos realizados nesta pesquisa. Em seguida, a Figura 35 mostra os resultados gerais de precisão dos modelos.

Figura 35 – Resultado da acuraria geral dos modelos (treinamento).

	precision	recall	f1-score	support
0	0.00	0.00	0.00	3986
1	0.39	1.00	0.56	2580
accuracy			0.39	6566
macro avg	0.20	0.50	0.28	6566
weighted avg	0.15	0.39	0.22	6566

Fonte: autora (2023).

Após o treinamento foram realizados os primeiros testes gerando-se assim a matriz de confusão com os resultados dos experimentos. Logo após todo o treinamento e teste a técnica com melhor resultado foi avaliada e selecionada,

conforme as métricas de avaliação de desempenho consideradas (acurácia, precisão, recall e F1 Score) da técnica de IA aplicada, conforme exposto na Tabela 12.

Tabela 12 – Resultados consolidados Precisão, Recall e F1-Score – RNC.

Segunda Base de teste com 2.815 comentários de clientes				
Classificador	Acurácia	Precisão	Recall	F1 Score
RNC	27,6%	27,6%	13%	42,7%

Fonte: autora (2023).

No segundo teste com a base de dados obteve-se resultado de 27,6% de acerto, indicando o sucesso na classificação das classes de perfis de cliente. Os resultados das métricas de avaliação de desempenho precisão, recall e F1-score também foram bem avaliados. A precisão obteve uma classificação do modelo na ordem de 27,6% de acerto ao cruzar os comentários do cliente com os atributos do perfil do cliente. Já o recall obteve uma classificação de 13% acertos, enquanto a métrica F1-score obteve um resultado de 42,7% de resultado. A Figura 36 mostra os resultados gerais de precisão dos modelos.

Figura 36 – Resultado da acuraria geral dos modelos (teste).

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2052
1	0.27	1.00	0.43	765
accuracy			0.27	2817
macro avg	0.14	0.50	0.21	2817
weighted avg	0.07	0.27	0.12	2817

Fonte: autora (2023).

Os resultados obtidos na primeira base de teste não apresentaram uma boa classificação das métricas de avaliação de desempenho. Isto se deve às características inerentes ao conteúdo do banco de dados em si, pois é tipicamente um banco de dados com muito ruído, razão pela qual o modelo teve mais dificuldade em classificar as métricas de avaliação de desempenho. O segundo teste correu bem,

uma vez que a base de dados ficou significativamente menor e o modelo conseguiu classificar melhor as métricas.

Assim, os resultados obtidos subsidiaram algumas decisões que serão úteis para a empresa. Por meio de técnicas inteligentes, descobriu-se conhecimentos que a empresa não possuía sobre a percepção dos clientes sobre seus serviços e produtos. Por meio deste estudo constatou-se que clientes com diferentes perfis estão insatisfeitos com os mesmos serviços e produtos que a empresa oferece. E com resultados obtidos por meio dos experimentos realizados a empresa poderá obter diversos *insights* que poderão auxiliar na tomada de decisão dos gestores quanto às melhorias dos serviços, produtos e atendimento ao cliente.

A partir das análises realizadas e dos resultados alcançados foi possível concluir que os resultados não tiveram maior sucesso por se tratar de um base de dados com muito ruído, o que naturalmente dificulta a obtenção de melhores resultados com a aplicação das técnicas inteligentes consideradas neste trabalho. Como nada foi encontrado na literatura sobre análise de sentimentos e comparação com os resultados de NPS em avaliação de atendimento de empresas, não foi possível comparar os resultados deste estudo com a literatura. Cabe ressaltar que, embora a acurácia tenha sido baixa após a realização dos experimentos e avaliação do modelo, obteve-se algumas descobertas de conhecimento sobre o cliente que até então eram desconhecidas da empresa. Na próxima seção são apresentadas a avaliação e interpretação do conhecimento descoberto utilizando técnicas de Redes Neurais Convolucionais aplicadas nos experimentos realizados.

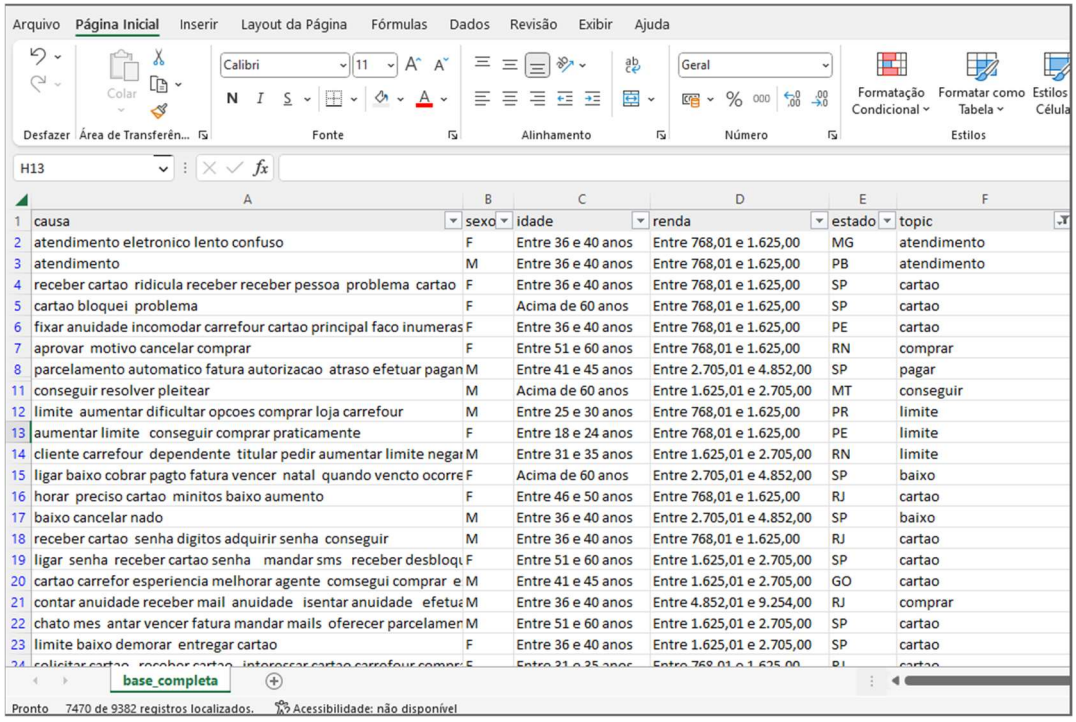
4.5 Fase 5 – Avaliação e interpretação do conhecimento descoberto com a aplicação das técnicas selecionadas

Nesta etapa iniciou-se a avaliação e interpretação do conhecimento descoberto com as técnicas de redes neurais convolucionais aplicadas nos experimentos. Como visto nas seções anteriores, foram aplicados modelos de classificação e filtros nos textos (comentários dos clientes). Assim, foram utilizadas as ferramentas Rake para classificar o ranking das palavras, Gensim e Bertorpic para modelagem dos tópicos, K-Means para 0 agrupamento (clusterização) dos textos. Nesta seção é apresentada a avaliação e interpretação do conhecimento descoberto.

4.5.1 Base de dados com tópicos criados

Logo após a aplicação de técnicas inteligentes para geração dos tópicos utilizando os modelos citados nas seções anteriores foi criada uma coluna chamada “Topic”, com os principais tópicos que o modelo identificou na base de dados, conforme exemplificado na Figura 37.

Figura 37 – Base de dados com coluna topic.



	causa	sexo	idade	renda	estado	topic
2	atendimento eletronico lento confuso	F	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	MG	atendimento
3	atendimento	M	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	PB	atendimento
4	receber cartao ridicula receber receber pessoa problema cartao	F	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	SP	cartao
5	cartao bloquei problema	F	Acima de 60 anos	Entre 768,01 e 1.625,00	SP	cartao
6	fixar anuidade incomodar carrefour cartao principal faco inumeras	F	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	PE	cartao
7	aprovar motivo cancelar comprar	F	Entre 51 e 60 anos	Entre 768,01 e 1.625,00	RN	comprar
8	parcelamento automatico fatura autorizacao atraso efetuar pagan	M	Entre 41 e 45 anos	Entre 2.705,01 e 4.852,00	SP	pagar
11	conseguir resolver pleitear	M	Acima de 60 anos	Entre 1.625,01 e 2.705,00	MT	conseguir
12	limite aumentar dificultar opcoes comprar loja carrefour	M	Entre 25 e 30 anos	Entre 768,01 e 1.625,00	PR	limite
13	aumentar limite conseguir comprar praticamente	F	Entre 18 e 24 anos	Entre 768,01 e 1.625,00	PE	limite
14	cliente carrefour dependente titular pedir aumentar limite negar	M	Entre 31 e 35 anos	Entre 1.625,01 e 2.705,00	RN	limite
15	ligar baixo cobrar pagto fatura vencer natal quando vencto ocorre	F	Acima de 60 anos	Entre 2.705,01 e 4.852,00	SP	baixo
16	horar preciso cartao minitos baixo aumento	F	Entre 46 e 50 anos	Entre 768,01 e 1.625,00	RJ	cartao
17	baixo cancelar nado	M	Entre 36 e 40 anos	Entre 2.705,01 e 4.852,00	SP	baixo
18	receber cartao senha digitos adquirir senha conseguir	M	Entre 36 e 40 anos	Entre 768,01 e 1.625,00	RJ	cartao
19	ligar senha receber cartao senha mandar sms receber desbloq	F	Entre 51 e 60 anos	Entre 1.625,01 e 2.705,00	SP	cartao
20	cartao carrefor esperiencia melhorar agente comsegui comprar e	M	Entre 41 e 45 anos	Entre 1.625,01 e 2.705,00	GO	cartao
21	contar anuidade receber mail anuidade isentar anuidade efetua	M	Entre 36 e 40 anos	Entre 4.852,01 e 9.254,00	RJ	comprar
22	chato mes antar vencer fatura mandar mails oferecer parcelamen	M	Entre 51 e 60 anos	Entre 1.625,01 e 2.705,00	SP	cartao
23	limite baixo demorar entregar cartao	F	Entre 36 e 40 anos	Entre 1.625,01 e 2.705,00	SP	cartao

Fonte: autora (2023).

A coluna ‘Topic’ mostra os principais tópicos gerados pelo modelo com a aplicação de técnicas inteligentes de redes neurais convolucionais, são eles: cartão, limite, atendimento, pagar, comprar, conseguir, nome empresa, aumentar, resolver e baixo. Esses tópicos foram gerados pela solução de inteligência artificial desenvolvida e apresentada nesta dissertação, as técnicas utilizadas classificaram cada linha de (comentário) e logo após, todos os experimentos aplicados geraram os tópicos localizados na coluna “F” da base de dados, conforme Figura 37.

Apesar da baixa acurácia, conforme indicado no tópico 4.4.5, obtida em função da base de dados ter baixa estruturação e ser ruidosa, optou-se por dar prosseguimento aos experimentos. O próximo passo foi a aplicação das correlações

determinadas e o cruzamento dos dados para a descoberta de conhecimento a partir das reclamações elaboradas pelos clientes.

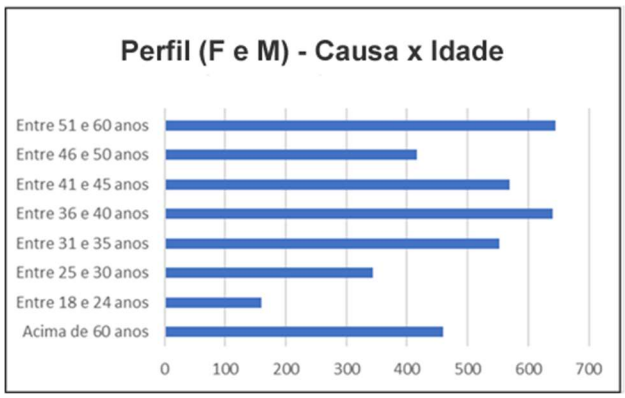
4.5.2 Correlação do atributo Causa (comentário do cliente) x Perfil do cliente (sexo, idade, renda e estado)

Nesta etapa inicia-se o estabelecimento das correlações entre os atributos causa (comentário do cliente) e o atributo ‘sexo’ do perfil do cliente em cruzamento com ‘idade’, ‘renda’ e ‘estado do cliente’. Os resultados alcançados são apresentados segregados nos principais tópicos considerados como mais significativos para fins de análise desta pesquisa, quais sejam: cartão, limite, comprar, atendimento, pagar e aumento).

a) Tópico ‘Cartão’

Na correlação entre o atributo ‘causa’ e o atributo ‘sexo’ (feminino e masculino do perfil do cliente verificou-se que a maioria dos clientes insatisfeitos com os serviços relacionados ao cartão tem idade entre 36 e 40 anos e entre 51 e 60 anos, conforme indicado na Figura 38.

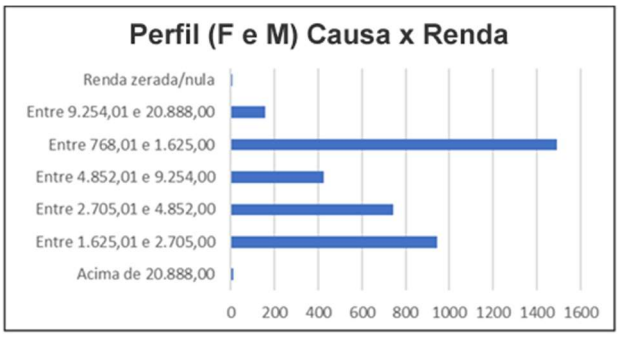
Figura 38 – Cartão: causa x perfil (F e M) x idade.



Fonte: autora (2023).

Foi identificado também que os clientes que possuem renda mensal entre R$ 768,01 e R$ 1.625,00 compõem a maioria dos insatisfeitos com o cartão da empresa, conforme indicado na Figura 39.

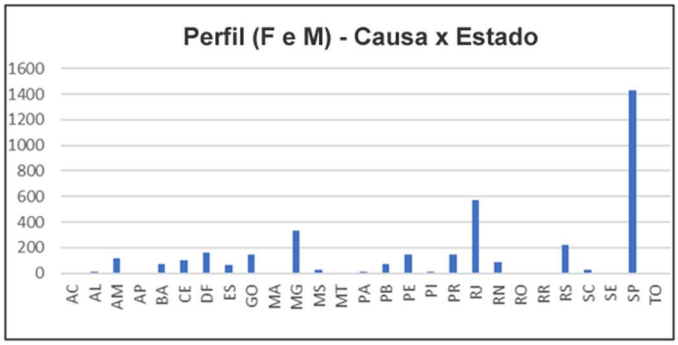
Figura 39 – Cartão: causa x perfil (F e M) x renda.



Fonte: autora (2023).

Ao avaliar a base de dados foi possível identificar que o estado com mais reclamações sobre serviços de cartão é o estado de São Paulo (SP), com aproximadamente 30% das reclamações. Deve-se levar em consideração que o estado de São Paulo tem a maior população do país. Não obstante, foi o estado que se destacou com a maior quantidade de reclamações, tanto para o perfil feminino quanto para o perfil masculino, nos registros da base de dados da empresa. Outros estados identificados com aproximadamente 15% reclamações foram o Rio de Janeiro e Minas Gerais. Os demais estados compuseram em torno de 40% das reclamações dos clientes, conforme indicado na Figura 40.

Figura 40 – Cartão: causa x perfil (F e M) x estado.

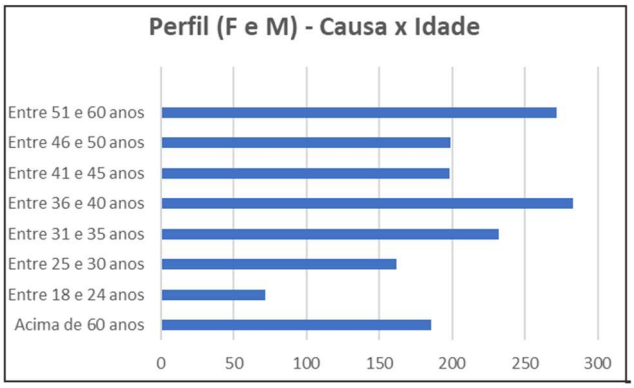


Fonte: autora (2023).

b) Tópico ‘Limite’

Na correlação entre o tópico ‘Limite’ e o atributo ‘sexo’ (feminino e masculino) do perfil do cliente, verificou-se que 40% dos clientes insatisfeitos com o limite de cartão tem idade entre 36 e 40 anos e 25% de clientes com idade entre 51 e 60 anos. Os clientes reclamaram principalmente que o limite é baixo, dificuldade para solicitar aumento de limite do cartão, limite incompatível com a renda e dificuldade para aprovação de aumento de limite. 35% das demais faixas etárias estão insatisfeitos com demais serviços oferecidos pela empresa. A Figura 41 apresenta os resultados encontrados.

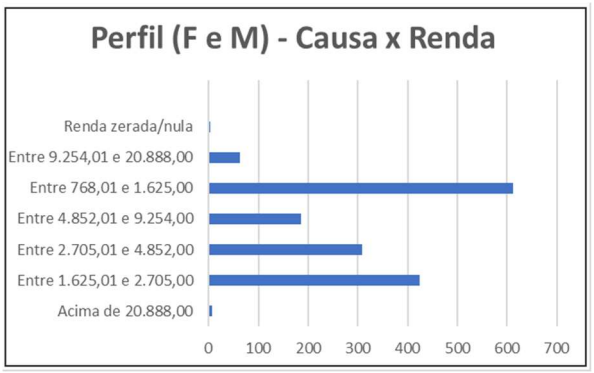
Figura 41 – Limite: causa x perfil (F e M) x idade.



Fonte: autora (2023).

Identificou-se também que a maioria dos clientes insatisfeitos com o limite de seu cartão possui renda mensal entre R$ 768,01 e R$ 1.625,00. A Figura 42 ilustra a correlação entre os atributos analisados.

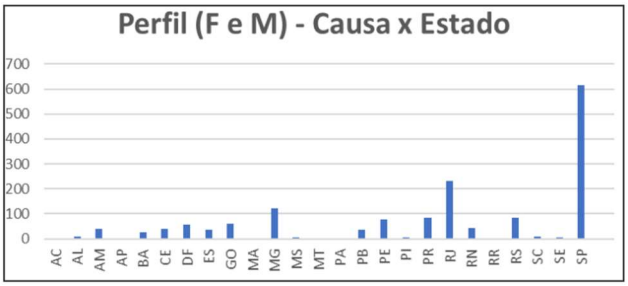
Figura 42 – Limite: causa x perfil (F e M) x renda.



Fonte: autora (2023).

O estado com mais reclamações sobre limite de cartão é o estado de São Paulo (60%), também correlacionado ao fato de ser o estado mais populoso. Os demais estados compuseram 40% das reclamações sobre limites, considerando-se clientes homens e mulheres, conforme evidenciado na Figura 43.

Figura 43 – Limite: causa x perfil (F e M) x estado.



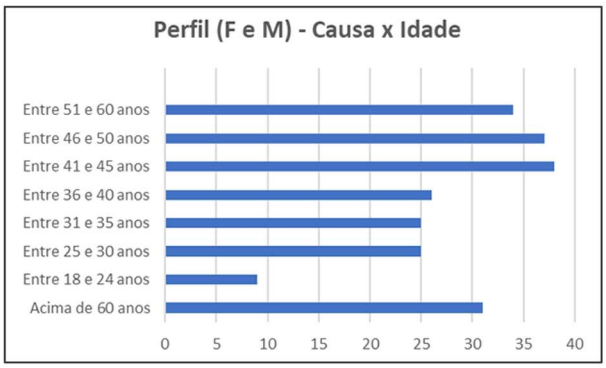
Fonte: autora (2023).

c) Tópico ‘Comprar’

Na correlação entre o tópico ‘Comprar’ e o atributo ‘sexo’ do perfil do cliente (feminino e masculino) verificou-se que a maioria dos clientes insatisfeitos com as compras realizadas na empresa (lojas físicas ou pela internet) são clientes com idade entre 41 e 45 anos. Indica-se ainda que 47% do perfil masculino estão insatisfeitos com os seguintes aspectos: parcelamento de compras, dificuldade em realizar compras no site da empresa e dificuldade em comprar com o cartão. Já 53% do perfil feminino estão insatisfeitos com os seguintes aspectos: aprovação de compras no cartão, análise de crédito para compras no cartão, tarifa alta nas compras parceladas,

reembolso na compra, dificuldade em comprar com o cartão e vencimento das compras efetuadas no cartão. A Figura 44 expõe a distribuição das reclamações encontradas do tópico ‘comprar’.

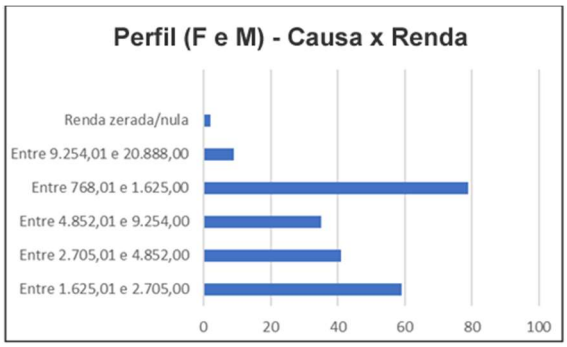
Figura 44 - Comprar: causa x perfil (F e M) x idade.



Fonte: autora (2023).

Continuando a análise da base de dados foi identificado que 60% dos clientes com renda entre R$ 768,01 e R$ 1.625,00 estão insatisfeitos com as compras online e nas lojas físicas da empresa, enquanto 40% dos clientes das demais faixas de renda estão insatisfeitos com outros produtos e serviços, conforme expressado na Figura 45.

Figura 45 – Comprar: causa x perfil (F e M) x renda.

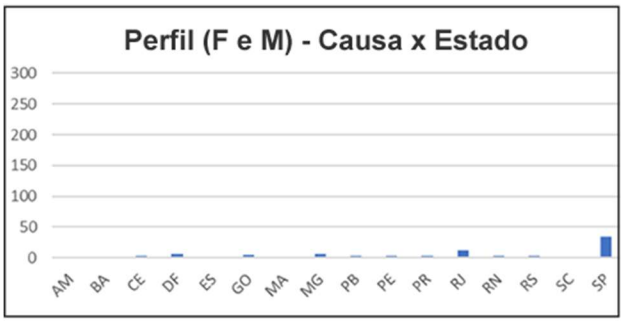


Fonte: autora (2023).

Analisando-se os resultados, constatou-se que 60% dos clientes residem no estado de São Paulo (SP), sendo estes os que mais reclamaram à empresa. Constatou-se que tanto o perfil masculino quanto o feminino de clientes residentes no

estado de São Paulo encontram dificuldades na utilização do cartão para compras online e nas lojas físicas da empresa, conforme demonstrado na Figura 46.

Figura 46 – Comprar: causa x perfil (F e M) x estado.

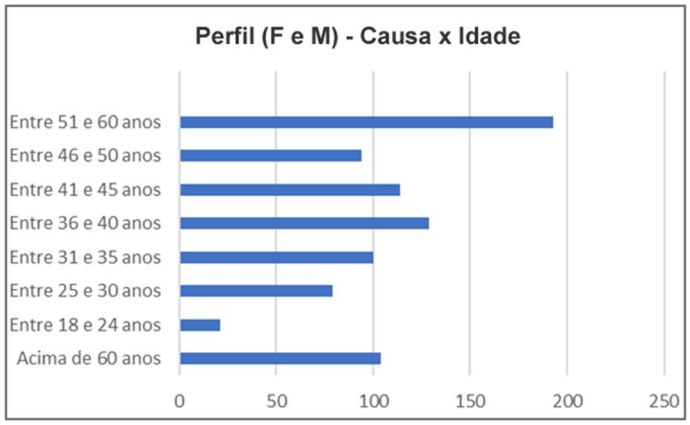


Fonte: autora (2023).

d) Tópico ‘Atendimento’

Na correlação entre o tópico ‘Atendimento’ e o atributo ‘sexo’ (feminino e masculino) do perfil do cliente verificou-se que a maioria dos clientes insatisfeitos com o atendimento da empresa têm entre 51 e 60 anos de idade. 40% do perfil masculino estão insatisfeitos com a prestação de serviço no atendimento oferecido ao cliente por conta das seguintes razões: demora no atendimento em todos os canais de atendimento, central de atendimento demora no atendimento, pessoas desqualificadas no atendimento). Outros 60% do perfil feminino estão insatisfeitos com o atendimento confuso, dificuldade no atendimento via telefone, mentiras no atendimento que o funcionário presta ao passar informações falsas e muita espera no atendimento. A Figura 47 expõe a distribuição dos resultados desta correlação.

Figura 47 – Atendimento: causa x perfil (F e M) x idade.



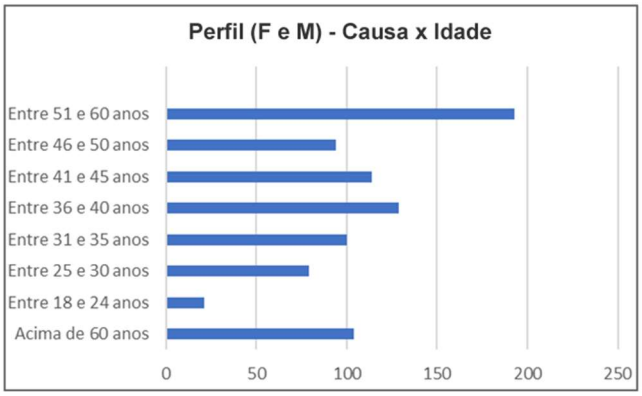
Fonte: autora (2023).

e) Tópico ‘Pagamento’

Foi verificado que 52% do perfil feminino estão insatisfeitos com os serviços prestados pela empresa, pois esses clientes reclamam das seguintes questões: promoções falsas, não conseguir contato na central de atendimento, não conseguir acessar o aplicativo da empresa, péssimo atendimento via chat e seguros e tarifas muito altas. Já no perfil masculino, 48% estão insatisfeitos devido aos seguintes aspectos: cobranças indevidas pela empresa, produtos da empresa, problemas com app, atendimento ruim nos estandes em lojas físicas e problemas com fatura do cartão.

Em relação à faixa etária dos clientes, a maioria das reclamações sobre pagamentos advém de clientes com idade entre 51 e 60 anos. Os clientes das demais faixas etárias estão insatisfeitos com diversos produtos e serviços oferecidos pela empresa. A Figura 48 expõe a distribuição das reclamações encontradas.

Figura 48 – Causa x perfil (F e M) x idade.



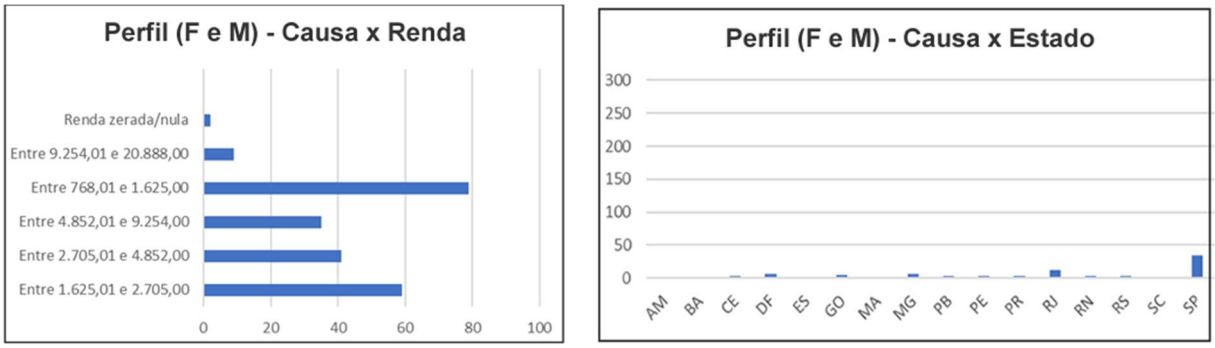
Fonte: autora (2023).

Ainda avaliando-se os resultados descobertos foi possível identificar que 35% do perfil masculino estão insatisfeitos com a empresa devido aos seguintes itens: não conseguem atendimento via chat, não conseguem troca senha pelo aplicativo e quando ligam na central de atendimento ficam horas na espera para resolver o problema de senha, não conseguem desbloquear o cartão e, por fim, não conseguem ao acessar o app do cartão.

Por outro lado, 65% do perfil feminino indicaram os seguintes aspectos: não conseguem resolver problemas na central de atendimento, dificuldade no pagamento de fatura, suporte ruim, não conseguem desbloquear cartão, não conseguem atendimento via chat e e-mail e ainda dificuldades em acessar o app do cartão.

Em relação à faixa de renda mensal dos clientes, verificou-se que 60% dos clientes residentes no estado de São Paulo (SP) recebem entre R$ 768,01 e R$ 1.625,00 e 40% dos clientes reclamaram de múltiplos serviços relacionados ao pagamento, como mostra a Figura 49.

Figura 49 – Causa x perfil (F e M) x renda e estado.



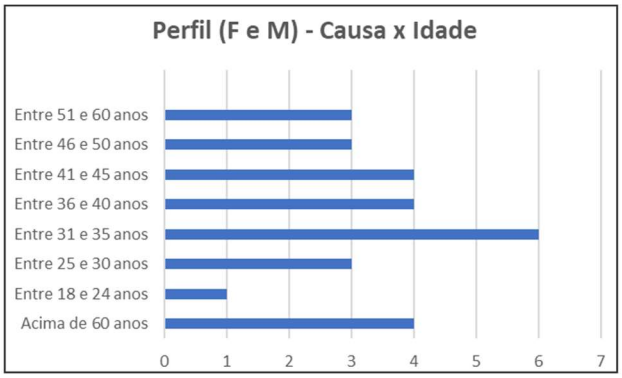
Fonte: autora (2023).

Com as análises realizadas foi possível identificar que os clientes têm dificuldade em resolver diversos problemas relacionados à senha do cartão, acesso ao aplicativo e central de atendimento. Pôde-se ainda apurar que tanto o perfil masculino (50%) quanto o perfil feminino (50%) estão insatisfeitos com questões relacionadas ao tópico ‘pagamento’. Nesse sentido, foi possível encontrar comentários relacionados aos seguintes aspectos: pagamento da anuidade, problemas com pagamento automático, opções de pagamento e taxas de pagamento.

f) Tópico ‘Aumento’

Na correlação entre o tópico ‘aumento’ e o atributo ‘sexo’ (feminino e masculino) do perfil do cliente foi possível verificar que a maioria dos clientes insatisfeita com o aumento de limite têm idade entre 31 e 35 anos, conforme indicado na Figura 50.

Figura 50 – Aumento: causa x perfil (F e M) x idade.

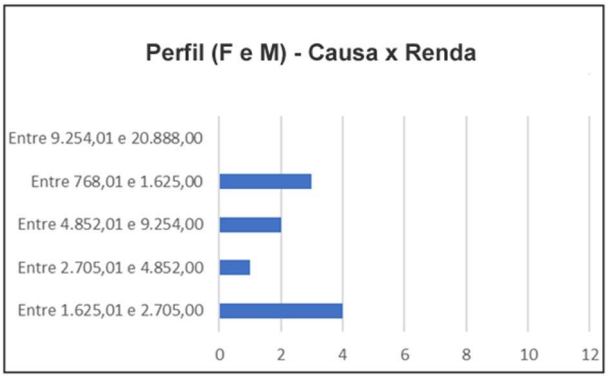


Fonte: autora (2023).

51% do perfil masculino estão insatisfeitos devido à dificuldade em aumentar limite de crédito e aumento de anuidade; enquanto 49% do perfil feminino estão insatisfeitos em razão do aumento de pontuação de compras e dificuldade em aumentar limite de cartão.

Continuando-se a análise dos resultados, identificou-se ainda que os clientes com renda entre R$ 1.625,01 e 2.705,00 (Figura 51) são a maioria dos insatisfeitos quanto ao aumento de limite de crédito e aumento de anuidade de cartão. Isto porque os clientes reclamam que mesmo com uma boa renda, a empresa em questão não aumenta os limites para compras nas lojas físicas e compras online.

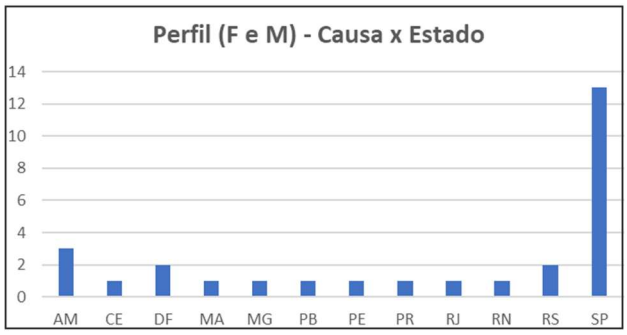
Figura 51 – Aumento: causa x perfil (F e M) x renda.



Fonte: autora (2023).

Conforme exposto na Figura 52, São Paulo e Amazonas são os estados com mais reclamações dos clientes em relação ao limite de crédito fornecido pela empresa.

Figura 52 – Causa x perfil (F e M) x estado.



Fonte: autora (2023).

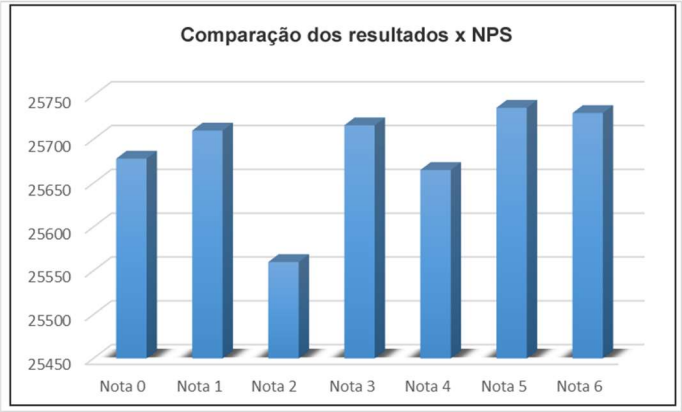
4.6 Fase 6 – Comparação dos resultados dos experimentos com os resultados dos indicadores detratores do NPS (*Net Promoter Score*)

A partir dos resultados auferidos nos experimentos realizados, nesta última etapa do processo de KDD adaptado desta pesquisa foi realizada a comparação com os indicadores detratores do NPS (*Net Promoter Score*). Esta atividade foi realizada em conjunto com o especialista da empresa, de modo a aproveitar sua expertise no fenômeno analisado.

4.6.1 Comparação dos resultados dos experimentos da correlação Causa x perfil do cliente com os resultados dos indicadores detratores do NPS

Foram considerados os resultados da primeira correlação analisada (causa – comentário do cliente) e atributos do perfil do cliente (sexo, idade, renda e estado) em cruzamento com os resultados dos indicadores do NPS. A Figura 53 apresenta a distribuição das notas detratoras (0 a 6) de NPS indicadas pelos clientes que manifestaram reclamações junto à empresa.

Figura 53 – Comparação dos resultados x NPS.



Fonte: autora (2023).

As notas mais indicadas pelos clientes foram, em ordem de ocorrência, 5, 6, 3, 1, 0, 4 e 2. Além disso, os resultados indicam que 60% dos clientes que aplicaram as menores notas de NPS detratores (de 0 a 3) são do perfil feminino, demonstrando elevada insatisfação deste público com os serviços e produtos da empresa. Já no perfil masculino, 40% de clientes deram notas de NPS detratores mais elevadas, ou seja, entre 4 e 6. Isto indica que as clientes do perfil feminino, em geral, tendem a atribuir notas mais baixas aos serviços e produtos da empresa analisada quando manifestam suas reclamações nos canais de atendimento disponibilizados.

Percebeu-se ainda que 70% do perfil com renda entre R$ 768,01 e 1.625,00 foram os que aplicaram a pior faixa de nota de NPS (entre 0 e 3). É um perfil insatisfeito principalmente com os itens aumento de limite de crédito e aumento de anuidade de cartão. Já outro perfil de renda (entre 9.254,01 e 20.888,00) atingiu 30% de clientes que estão insatisfeitos apenas com os itens cartão e limite.

Já 50% dos clientes com faixa etária acima de 60 anos aplicaram nota de NPS de 0 a 3. Esse perfil apresentou problemas relacionados principalmente ao atendimento via Central, lojas e app da empresa. Outros 50% do perfil entre 18 a 50 anos atribuiu notas NPS de 4 e 6, demonstrando insatisfação com os itens limite e anuidade do cartão, demonstrando a ocorrência de clientes com faixa renda de interesse da empresa, mas que esta não tem oferecido um limite adequado para este perfil de cliente.

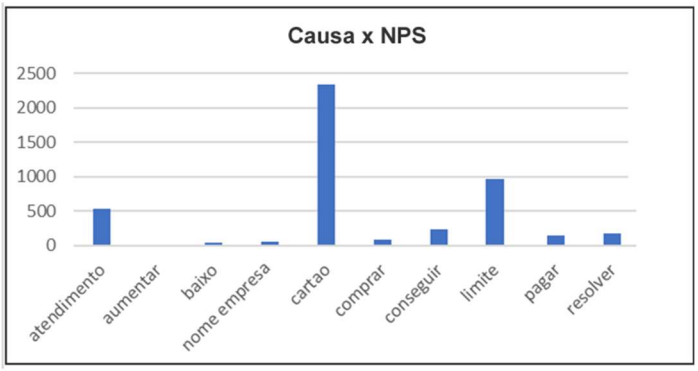
Já analisando-se os perfis de clientes por estado do país chegou-se à conclusão que 60% dos clientes que moram no estado de São Paulo e que aplicaram a nota do NPS entre 0 e 3 estão muito insatisfeitos com vários serviços e produtos

oferecidos pela empresa. Isto porque apenas 40% dos demais estados aplicaram nota NPS entre 4 e 6, sendo a maior parte das insatisfações deste perfil voltadas à falta de atendimento adequado, tanto em estandes de lojas físicas, quanto na central de atendimento.

4.6.2 Comparação dos resultados dos experimentos da correlação Causa (comentário de cliente) x Estratificação de notas detratoras do NPS (notas de 0 a 6) com os resultados dos indicadores detratores do NPS

Comparando-se os resultados apresentados nos tópicos 4.4.1 e 4.4.2 quanto aos principais tópicos (cartão, limite, compras, atendimento, pagamentos e aumento) descobertos com a solução de inteligência artificial aplicada, com os resultados do NPS manifestados pelos clientes, foi possível efetuar a primeira correlação com as notas dos detratores (0 a 3). Como resultados, os tópicos com mais reclamações de clientes foram sobre cartão, limite e atendimento nos canais disponibilizados pela empresa. A Figura 54 ilustra a comparação dos resultados com o NPS detrator 0 a 3.

Figura 54 – Comparação dos resultados x NPS notas 0 a 3.



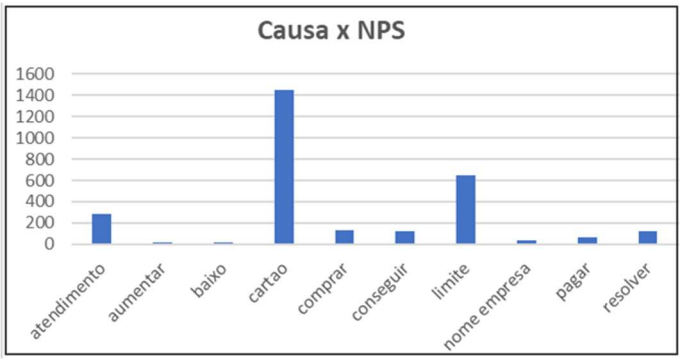
Fonte: autora (2023).

Os tópicos com menos comentários de cliente foram: baixo, nome da empresa, comprar, não conseguir atendimento e resolver problema com a empresa. Ressalte-se que o tópico aumentar não apresentou nenhum comentário de cliente na correlação entre os experimentos analisados e o detrator 0 a 3 do NPS.

Ainda comparando-se os resultados dos experimentos com os resultados do NPS, fez-se a segunda correlação com as notas 4 a 6 de NPS detratores. Os tópicos

com mais insatisfações relatadas pelos clientes foram sobre cartão, limite e atendimento. Já os tópicos com menos reclamações foram: aumentar, baixo, comprar, pagar, nome empresa e resolver problemas. A Figura 55 ilustra a comparação dos resultados com o NPS de detratores com notas entre 4 e 6.

Figura 55 – Comparação dos resultados x NPS nota 4 a 6.



Fonte: autora (2023).

A partir da análise e descoberta de conhecimento do cliente foi possível chegar à conclusão de que muitos comentários dos clientes estão relacionados a problemas com cartão, atendimento, limite, pagamento e problemas não resolvidos pelos canais de atendimentos da empresa. Também foi possível verificar que em determinados comentários, o mesmo cliente demonstrou insatisfação com mais de um problema (tópico identificado) indicado pela solução desenvolvida nesta pesquisa. A título de exemplo, em alguns comentários o cliente aborda dois ou até três dos principais tópicos evidenciados nos resultados desta pesquisa.

4.7 Principais conhecimentos sobre o cliente identificados nos resultados da pesquisa

Na Tabela 13 são apresentados os principais conhecimentos descobertos a respeito das manifestações de insatisfação do cliente a partir da aplicação de técnicas de rede neurais convolucionais na solução desenvolvida nesta dissertação.

Tabela 13 – Principais conhecimentos descobertos sobre o cliente.

Tópico	Conhecimento descoberto
Cartão	Clientes com perfis diferentes estão insatisfeitos com os serviços oferecidos pela empresa, sendo os principais motivos: • Falta de aumento de limite; • Dificuldades de contato com a central de atendimento; • Desbloqueio e cancelamento de cartões; • Obtenção de informações referentes aos cartões.
Limite	Existência de clientes com perfis diferentes com dificuldade em aumentar o limite do cartão, mesmo quando o cliente tem renda alta.
Compras	Clientes com diversos problemas relacionados a: • Compras efetuadas nas lojas próprias da empresa; • Problemas com entregas, pois muitos produtos chegam danificados e com muito tempo de entrega atrasado em relação ao prometido pela empresa.
Atendimento	Clientes reclamaram dos seguintes problemas: • Atendimento na central, lojas e aplicativos, pois há muita demora no tempo de espera para ser atendido; • Estado com mais reclamação no atendimento é o estado de São Paulo, provavelmente em razão do volume de clientes.
Pagamento	Problemas relacionados a: • Pagamento de fatura de cartão; • Multas muito altas para pagamentos efetuados em atraso.
Aumento	Clientes de diferentes perfis, estados e idades estão insatisfeito com: • Falta de aumento do limite de crédito; • Aumento de anuidade do cartão.

Fonte: autora (2023).

5. Conclusão

O atendimento ao cliente sempre foi um aspecto de suma importância para as empresas. Nos últimos anos o volume de dados gerado diariamente nos canais de atendimento de grandes empresas com dezenas de milhões de clientes tem se expandido cada vez mais, tornando-se um desafio para a empresa extrair informações úteis para a geração de conhecimento acerca dos clientes. A Inteligência Artificial tem métodos e técnicas capazes de analisar grandes massas de dados visando a extração de conhecimentos acerca do fenômeno abordado. Assim, esta pesquisa teve como objetivo aplicar técnicas de Redes Neurais Convolucionais para a análise e classificação de sentimentos em comentários de clientes visando a descoberta de conhecimento do cliente em empresa varejista.

A aplicação da técnica de redes neurais convolucionais buscou analisar e classificar os sentimentos expressados pelos clientes da empresa varejista enfocada nesta dissertação. Para tanto foi realizada a avaliação e interpretação do conhecimento descoberto, além do cruzamento deste conhecimento produzido com os indicadores detratores do NPS (*Net Promoter Score*) da empresa varejista. Para atingir os objetivos indicados foi considerada uma base de dados com registros provenientes da central de atendimento aos clientes, bem como os resultados do NPS. Os dados foram utilizados para a realização de experimentos computacionais baseados em métodos e técnicas voltados à análise e classificação de sentimentos, visando assim constituir uma solução eficiente de descoberta de conhecimento útil para a tomada de decisões a respeito de produtos e serviços oferecidos aos clientes.

O método de redes neurais convolucionais foi aplicado por meio de modelos de classificação de texto com uso de ferramentas como Rake, Gensim e Bertopic para modelagem de tópicos, além da ferramenta K-means para agrupamento dos textos. Após a aplicação das técnicas inteligentes, o modelo criado pôde identificar os principais tópicos dos sentimentos extraídos dos comentários dos clientes na base de dados analisada, quais sejam: cartão, limite, comprar, atendimento, pagamento e aumento. Logo após a descoberta dos principais tópicos na base de dados foram aplicadas as correlações entre os atributos causa (comentários de clientes) com os atributos da base de dados. A título da condução dos experimentos e para a validação da solução delineada, optou-se por apresentar apenas os resultados dos atributos da

categoria de perfil de cliente (sexo, idade, renda e estado) em cruzamento com os resultados do NPS detratores.

Foi possível identificar que clientes com renda mensal entre R$ 768,01 e R$ 1.625,00 e provenientes do estado de São Paulo estão insatisfeitos com serviços relacionados a questões envolvendo o cartão, limite, atendimento *online* e atendimento em lojas físicas. Verificou-se que 47% do perfil masculino estavam insatisfeitos com os seguintes aspectos: parcelamento de compras, dificuldade para comprar no site da empresa e dificuldade para fazer compras com cartão nas lojas da empresa. E ainda que 53% do perfil feminino tem dificuldade com análise de crédito, reclamações sobre altas taxas para compras parceladas e dificuldades para pedir reembolso de uma compra.

Em complemento, a solução desenvolvida indicou que 47% de clientes do perfil masculino estão insatisfeitos com o atendimento prestado ao cliente pelos seguintes motivos: demora no atendimento em todos os canais de atendimento, demora no atendimento da central de atendimento e funcionários desqualificados no atendimento. Outros 48% de clientes do perfil feminino estão insatisfeitas com atendimento confuso, dificuldade para atender o telefone, mentir no atendimento, funcionários dando informações erradas e muita espera no atendimento. Quanto à questão dos pagamentos, constatou-se que 52% do perfil feminino estão insatisfeitos com os serviços da empresa, uma vez que essas clientes reclamam dos seguintes problemas: promoções incorretas, indisponibilidade da central de atendimento, não acesso ao aplicativo da empresa, péssimo atendimento via chat e seguros e taxas muito altas. Quanto ao perfil masculino, 48% estão insatisfeitos devido aos seguintes aspectos: cobrança indevida da empresa, produtos da empresa, problemas com o app, atendimento ruim nos estandes das lojas físicas e problemas com a fatura do cartão.

Constatou-se também que a maioria dos clientes insatisfeitos com questões de pagamento e que atestou dificuldades em solucionar problemas com a empresa tem idade entre 51 e 60 anos. Nessa faixa etária, identificou-se que 46% do perfil masculino estão insatisfeitos com a empresa com base nos seguintes pontos: não recebem ajuda pelo chat, não conseguem alterar a senha pelo aplicativo, muito tempo de espera na central de atendimento, problemas com a senha, desbloqueio de cartão e dificuldade de acesso ao aplicativo do cartão. Quanto ao perfil feminino, 49% relataram os seguintes aspectos: não conseguem resolver problemas no call center,

dificuldades para pagar a conta, suporte ruim, não conseguem desbloqueio do cartão, não recebem ajuda via chat e e-mail e ainda relataram dificuldades para acessar o app.

Constatou-se também que que 51% dos clientes do sexo masculino estão insatisfeitos por causa dos seguintes aspectos: dificuldade em aumentar o limite de crédito e aumento da previdência; enquanto 49% do perfil feminino estão insatisfeitos devido aos seguintes pontos: aumento dos pontos de compra e dificuldade de aumentar o limite do cartão. Analisando-se os resultados, constatou-se também que os clientes com renda entre R$ 1.625,01 e 2.705,00 são a maioria dos insatisfeitos com o aumento de limite de crédito e aumento de previdência do cartão. Isto porque na opinião destes perfis de clientes, mesmo apresentando um rendimento compatível, a empresa não aumenta os limites para compras em lojas físicas e compras online. Quanto ao estado de domicílio indicado pelos clientes, São Paulo e Amazonas são os estados com mais reclamações de clientes quanto ao limite de crédito concedido pela empresa.

A partir dos resultados dos experimentos executados nesta pesquisa, na última etapa do processo KDD adaptado foi realizada a comparação com os indicadores críticos do NPS (*Net Promoter Score*) detratores. Desta forma, forma considerados os resultados da correlação analisada de causa (comentários de clientes) e atributos do perfil do cliente (sexo, idade, renda e estado) em cruzamento com os resultados dos indicadores de NPS detratores. Foi possível constatar que 60% dos clientes que aplicaram as notas críticas mais baixas do NPS detrator (entre 0 e 3) são do sexo feminino, demonstrando a elevada insatisfação desse público com os serviços e produtos da empresa. No perfil masculino, 40% dos clientes relataram escores mais altos de NPS crítico, ou seja, atribuíram notas entre 4 e 6. Isso indica que as clientes do perfil feminino geralmente tendem a atribuir notas mais baixas aos serviços e produtos da empresa analisada, expressando assim suas reclamações nos canais de atendimento disponíveis.

Verificou-se também que 70% dos perfis de clientes com renda entre R$ 768,01 e R$ 1.625,00 foram os que aplicaram a pior faixa de pontuação do NPS detrator (entre 0 e 3). Ou seja, é um perfil particularmente insatisfeito com o aumento de pontos de limite de crédito e aumento de previdência do cartão. Outro perfil de renda (entre R$ 9.254,01 e R$ 20.888,00) atinge 30% dos clientes insatisfeitos apenas com o cartão e itens relacionados ao limite estabelecido. Como 50% dos clientes tinham mais

de 60 anos, eles deram uma nota de 0 a 3 para o NPS detrator. Esse perfil apresentou principalmente questões relacionadas ao atendimento ao cliente por meio da central da empresa, lojas e aplicativo. Outros 50% de perfil entre 18 e 50 anos atribuíram notas NPS detrator de 4 a 6, refletindo sua insatisfação com os itens tais como limite e anuidade do cartão. Tais conhecimentos gerados a partir da base de dados analisada indicam diferentes perfis de clientes de interesse da empresa, no que tangem à faixa de renda, idade e sexo dos clientes, notadamente quanto ao fato de a empresa não ter oferecido um limite de cartão adequado para esses perfis de clientes.

A análise do perfil dos clientes por estado do país revelou que 60% deles são residentes no estado de São Paulo e com NPS detrator entre 0 e 3, tendo alegado estarem muito insatisfeitos com os diversos serviços e produtos da empresa. Isso porque apenas 40% dos demais estados aplicaram uma nota de NPS entre 4 e 6, sendo que a maior parte da insatisfação desse perfil está centrada na falta de atendimento adequado, tanto nas bancas das lojas quanto no central de atendimento.

Comparando-se os resultados dos experimentos com os resultados de NPS detratores manifestados pelos clientes, a primeira correlação foi estabelecida com as notas detratoras mais críticas (0 a 3). Observou-se que os assuntos com mais reclamações dos clientes foram sobre limite e atendimento nos canais disponibilizados pela empresa. Os tópicos com menos comentários de clientes foram: limite baixo, comprar, não conseguem atendimento e dificuldade a resolver problemas. Ressalte-se que o tópico aumentar não apresentou nenhum comentário de cliente na correlação entre os experimentos analisados e a faixa de NPS detrator entre 0 e 3.

A partir da análise realizada e consequente descoberta do conhecimento do cliente, pode-se concluir que muitos comentários dos clientes dizem respeito a problemas com cartão, atendimento, limite, pagamento e problemas que não são solucionados pelos canais de atendimento da empresa. Também foi possível verificar que tanto o perfil feminino, quanto o perfil masculino demostram insatisfação com mais de um problema (tópicos identificados) indicado pela solução aplicada nesta pesquisa.

Como contribuições desta pesquisa para a Academia indica-se a aplicação de técnicas de redes neurais convolucionais voltadas à análise de sentimentos de comentários de clientes para descoberta de conhecimento, que se demonstra um tema de pesquisa atual e relevante. Isto porque as técnicas de análise de sentimentos de clientes aplicadas nesta dissertação denotam sua capacidade de produzir conhecimentos novos acerca do cliente e, portanto, merecem a atenção dos

pesquisadores. Portanto, esta pesquisa contribui para avançar o estudo e aplicação de métodos e técnicas inteligentes de modo a ampliar o conhecimento acerca destas por parte dos pesquisadores, que poderão aplicar a solução aqui desenvolvida em outros fenômenos e objetos de pesquisa.

Também se indica contribuições desta pesquisa para os profissionais, gestores e organizações de mercado. Isto porque a solução desenvolvida nesta dissertação mostrou-se relevante para a viabilização da aplicação de métodos e técnicas inteligentes em demandas e problemas reais das empresas atuais, no que concerne à descoberta de conhecimento de clientes a partir da análise de sentimentos em comentários realizados por clientes de empresas. Dessa forma, a solução desenvolvida poderá ser replicada em outros tipos de empresas (porte, setor de atuação, perfil de clientes) para a descoberta de conhecimento dos clientes.

Segundo o autor (Nonaka, 2006), adquirir e descobrir o conhecimento do cliente é vital, principalmente para os negócios, além de ajudar a empresa a aprender mais sobre o cliente que compra seus serviços e produtos. Portanto, ao realizar experimentos aplicando-se técnicas inteligentes de redes neurais convolucionais (RNC) para a análise de sentimentos do cliente em comparação à metodologia NPS (*Net Promoter Score*), prevê-se importante contribuição para o campo científico da Inteligência Artificial e seus métodos e técnicas. O desenho de pesquisa aqui delineado poderá permitir o desenvolvimento e validação de diferentes tipos de soluções inteligentes para a descoberta de conhecimento do cliente, subsidiando assim uma melhor tomada de decisão dos gestores nas empresas acerca dos clientes.

Como limitações desta pesquisa indica-se a seleção da técnica inteligente inicialmente aplicada para a realização dos experimentos realizados, que se restringiu às redes neurais convolucionais. Essa técnica foi escolhida em razão da plataforma teórica estabelecida, uma vez que trabalhos já publicados indicavam que essa técnica é muito aplicada para à análise de sentimentos. Portanto, nesta pesquisa de dissertação foram utilizadas técnicas de redes neurais convolucionais para classificação e clusterização (agrupamento) de textos de forma que os registros semelhantes fossem agrupados e diferenciados dos registros de dados de outros subconjuntos. Outra limitação diz respeito a base de dados analisada. Não obstante sua relevância em função da diversidade e volume dos dados analisados, a base em questão retrata um fenômeno restrito: a insatisfação de clientes em relação ao atendimento prestado por empresa varejista. Há ainda a ser mencionado o fato de que

os experimentos foram realizados em uma única empresa (estudo de caso único). Embora a empresa enfocada nesta pesquisa seja uma multinacional do setor de varejo, há de se considerar as especificidades do negócio de atuação da empresa, bem como suas características únicas.

Não obstante, estima-se que em função do volume e características dos dados disponíveis na empresa enfocada, os resultados dos experimentos ora realizados possam ajudar a resolver problemas que ocorrem costumeiramente em empresas varejistas, não sendo possível, entretanto, generalizar os resultados para toda e qualquer empresa varejista. Assim, os resultados alcançados nesta pesquisa indicam que a solução desenvolvida é capaz de proporcionar descoberta de conhecimento do cliente que auxilie a gestão dos clientes em empresas varejistas.

Por fim, como indicação de pesquisas futuras sugere-se realizar a aplicação desta solução em outras bases de dados, bem como em empresas de diferentes setores de atuação. Indica-se ainda a aplicação de outros métodos e técnicas inteligentes além das redes neurais convolucionais, realizando-se assim um comparativo de suas métricas e resultados.

Referências

AACINOL. **A importância de conhecer o cliente e ações que podem ser feitas**. Disponível em: <https://acinol.com.br/importancia-de-conhecer-seu-cliente-e-acoes-que-podem-ser-feitas/>. Acesso em: 27 dez. 2022.

ALVES GISELY. **Entendendo Redes Convolucionais (CNNs).** Publicado em: 08 out. 2018. Disponível em: < https://medium.com/neuronio-br/entendendo-redes-convolucionais-cnns-d10359f21184>. Acesso em: 31 dez. 2022.

AGGARAVAL, C. C.; ZHAI, C. Mining text data. [S.l.]: **Springer Science & Business Media, 2012**.

AGGARWAL, C. C.; ZHAI, C. (EDS.). **Mining Text Data**. Boston, MA: Springer US, 2012.

ANDREW G. HOWARD et al. **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.** 2017. ArXiv:1704.04861. Disponível em: https:// arxiv.org/abs/1704.04861. Acesso em: 04 mai. 2023.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais: Teoria e Aplicações (2 ed.)**. Rio de Janeiro: LTC, 2011.

BARONI, ITANIMÁ. **A matemática por trás das redes neurais**. (2021). Disponível em: <https://repositorio.uninter.com/handle/1/1023/>. Acesso em: 04 set. 2022.

B. KENJI, 2019. **Machine Learning para Leigos**. Disponível em: <https://www.venturus.org.br/machine-learning-para-leigos/>. Acesso em: 01 set. de 2022.

BIANCHI, ALEXANDRE**. As classificações dos algoritmos de Machine Learning**. Publicado em 27 mai. 2020. Disponível em: <https://www.viceri.com.br/insights/as-classificacoes-dos-algoritmos-de-machine-learning>. Acesso em: 02 abr. 2022.

BING LI. Sentiment Analysis: **Mining Opinions, Sentiments, and Emotions**. Cambridge University Press, 2015. doi: 10.1017/CBO9781139084789.

BITTAR, MARCEL. **Métricas para avaliação de modelos de Machine Learning**. Publicado em: 25 ag. 2020. Disponível em: <https://mabittar.github.io/Metricas/>. Acesso em: 22 jan. 2023.

BO, Z., YING-JIAO, Kou. **Research of Customer Knowledge Management and Realizing Process**. In: 2009 International Conference on Management and Service Science. 2009.

BORDIN JUNIOR, A. **Aplicação de programação genética na análise de sentimentos.** Dissertação (Mestrado) — Universidade Federal de Goiás, 2018.

BOBBIO, A.; PORTINALE, L.; MINICHINO, M.; CIANCAMERLA, E. **Improving the Analysis of Dependable Systems by Mapping Fault Trees into Bayesian Networks. Reliability.** ENGINEERING & SYSSTEM SAFETY, Vol. 71, p.249-260, 2001 BUCHNOWSKA, D. (2011), **Customer Knowledge Management Models: Assessment and Proposal. In Research in Systems Analysis and Design: Models and Methods.** Springer Berlin Heidelberg, pp. 25-38.

BREIMAN L, 2001. **"Florestas Aleatórias". Aprendizado de máquina**. 45 (1): 5-32. Bibcode: 2001 Machine Learning. doi:10.1023/A:1010933404324.

CASALI RUDINEY, 2021. **Árvore de Decisão: como se aplica no aprendizado de máquina?** Disponível em: <https://www.digitalhouse.com/br/blog/arvore-de-decisao/>. Acesso em: 08 set. 2022.

CASTRO, L. N. FERRARi, D. G. **Introdução a Mineração de dados: conceitos bás icos, algoritmos e aplicações**. São Paulo: Saraiva, 2016.

CARNEIRO, ALVARO LEANDRO CAVALCANTE. **Redes Neurais Convolucionais para processamento de linguagem natural.** Publicado em 07 jul. 2020. Disponível em:<https://medium.com/data-hackers/redes-neurais-convolucionais-para-processamento-de-linguagem-natural-935488d6901b>. Acesso em: 31 dez. 2022.

CHEN, Y.; SU, C. **A Kano-CKM model for customer knowledge discovery**. Total Quality Management and Business Excellence, v. 17, n. 5, p. 589-608, 2006.

CHRISTIAN SZEGEDY et al. **Rethinking the Inception Architecture for Computer Vision**. 2015. ArXiv:1512.00567. Disponível em: https://arxiv.org/abs/1512.00567. Acesso em: 05 mai. 2023.

DANESHGAR, F.; PARIROKH, M. **An integrated customer knowledge management framework for academic libraries**. The Library Quarterly, v. 82, n. 1, p. 7-28, 2012.

CHEN, Y.; TSAI, S.; HOU, S.; CHEN, L. **Measuring customer innovativeness via FuzzyArt network modeling**. In: IEEM 2009 - IEEE International Conference on Industrial Engineering and Engineering Management, p. 1558-1562, 2009.

DUDA, R. O.; HART, P. E.; Stork, D. G. **Pattern Classification. USA: Wiley Interscience**, 2000. ISBN 0471056693.

DRUDE, K. P. (2021). **Introduction to the Special Edition on social media.** Journal of Technology in Behavioral Science, June. 1–4. https://doi.org/10.1007/s41347-021-

00217-3.

EGGERS WILLIAM, D. Schatsky, P. Viechnicki (2017). **AI-augmented government using cognitive technologies to redesign public sector work**. Deloitte University Press (2017). Disponível em: <https://www2.deloitte.com/content/dam/insights/us/articles/3832_AI-augmented-government/DUP_AI-augmented-government.pdf>. Acesso em: 25 dez. 2022.

FAYYAD, U. M.; Piatetsky-Shapiro, G.; Smith, P. From **Data Mining to Knowledge Discovery: An Overview**. Knowledge Discovery and Data Mining, Menlo Park, AAAI Press, 1996.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMITH, P. **The KDD process for extracting useful knowledge from volumes of data**. Comunications of the ACM, v.39, p.27-34, 1996.

FERNANDES, ANITA MARIA DA ROCHA**. Inteligência artificial: noções gerais**. Florianópolis: Visual Books, 2003.

FREITAS, A. A. **Uma Introdução a Data Mining. Informática Brasileira em Análise, Centro de Estudos e Sistemas Avançados do Recife (C.E.S.A.R.),** Recife, Pe, ano II, n. 32, 2000.

FIA, BUSINESS. **Estudos de Caso: O que são, Exemplos e Como Fazer para TCC.** Publicado em: 20 set. 2020.

Disponível em: <https://fia.com.br/blog/estudos-de-caso/>. Acesso em: 24 jan. 2023.

FIDEL, P.; SCHLESINGER, W.; CERVERA, A. Collaborating to innovate: Effects on customer knowledge management and performance. Journal of Business Research, v. 68, n. 7, p. 1426-1428, 2015.

FUNCHAL, JOÃO P.; MADSEN, Carlos A.; ADAMATTI, Diana F. **Classificação automática de dados para a descoberta de conhecimento: um estudo de caso para a classificação de risco na área da saúde.** Passo Fundo-RS. Revista Brasileira de Computação Aplicada, 2015.

GEBERT, H., GEIB, M., KOLBE, L., RIEMPP, G. (2002), **Towards customer knowledge management: Integrating customer relationship management and knowledge management concepts**. In: The Second International Conference on Electronic Business (ICEB), pp. 296-298.

GIL, A. C**. Como Elaborar Projetos de Pesquisa**. 4ª ed. São Paulo: Atlas, 2002.

GOLDSCHMIDT, R. R.; PASSOS, E.; BEZERRA E. **Data Mining: conceitos, técnicas, algoritmos, orientações e aplicações**. (2 ed.). Campus: Rio de Janeiro,

2015.

Guilherme Defreitas Juraszek. **Reconhecimento de produtos por imagem utilizando palavras visuais e redes neurais convolucionais**. Dissertação, Centro de Ciências Tecnológicas, Universidade do Estado de Santa Catarina, Joinville, Brasil, 2014. URL http://tede.udesc.br/tede/tede/1761.

GONÇALVES, P., Dores, W., BENEVENUTO, F., and Preto-MG-Brasil, O. (2012). Panas-t: **Uma escala psicometrica para medic¸ ao de sentimentos no twitter.**

GROOTENDORST MAARTERN. **BERTopic: Neural topic modeling with a class-based TF-IDF procedure.** Publicado: 11 mar. 2022.

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques, third edition**. Waltham, Mass.: Morgan Kaufmann Publishers, 2012.

HARIRI, R. H., Fredericks, E. M., & Bowers, K. M. (2019). **Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data***, 6(1).* doi:10.1186/s40537-019-0206-3.

HAZARIKA, BIDYUT BIKASH; MOUSAVIZADEH, MOHAMMADREZA; TARN, MIKE. **A Comparison of Hedonic and Utilitarian Digital Products Based on Consumer Evaluation and Technology Frustration. Journal of Information Systems and Technology Management**, [S.l.], v. 16, dez. 2019. ISSN 1807-1775. Disponível em: <http://www.jistem.tecsi.org/index.php/jistem/article/view/3094/719>. Acesso em: 13 mai. 2022. doi: http://dx.doi.org/10.4301/S1807-1775201916008.

HEMMATIAN, F.; SOHRABI, M. K. **A survey on classification techniques for opinion mining and sentiment analysis**. Artificial Intelligence Review, Springer, p. 1–51, 2017.

HERHOLD, LEIGH ANN. **Redes Neurais. O que são e qual sua importância?** Disponível em: <https://www.sas.com/pt_br/insights/analytics/neural-networks.html/>. Acesso em: 06 set. 2022.

HOLLANDA, T. **Conhecimento do cliente em customer experience.** Publicado em: 17 abr. 2019. Disponível em: <https://www.metricx.blog/conhecimento-do-cliente-em-customer-experience/>. Acesso em: 27 dez. 2022.

HUB. REALIZE. **Análise de Sentimentos e a Estratégia de sua empresa**. Publicado em: 09 jun. 2022. Disponível em: <https://realizehub.com/analise-de-sentimentos-e-a-estrategia-de-sua-empresa/>. Acesso em: 31dez. 2022.

HE, K. et al. **Deep Residual Learning for Image Recognition**. 2016.

IAN GOODFELLOW, YOSHUA BENGIO, and Aaron Courville (2016). **Deep Learning**.

ICHI.PRO. **Máquina de vetores de suporte (SVM) explicada.** 2020. <https://ichi.pro/pt/maquina-de-vetores-de-suporte-svm-explicada-97743104690915>. Acesso em: 14 mai.2022.

IVONE PENQUE MATSUNO YUGOSHI. **Mineração de opiniões baseada em aspectos para revisões de produtos e serviços.** PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2018.

JAZIRI, D. **The advent of customer experiential knowledge management approach (CEKM): The integration of offline & online experiential knowledge**. Journal of Business Research, v. 94, n. May, p. 241-256, 2019.

JIEBING, W.; BIN, G.; YONGJIANG, S. **Customer knowledge management and IT-enabled business model innovation: A conceptual framework and a case study from China.** European Management Journal, v. 31, n. 4, p. 359-372, 2013.

JOTHI, N.; RASHID, N. A. A.; e HUSAIN, W. **Data mining in healthcare–a review**. Procedia Computer Science, v.72, p.306-313, 2015.

JURISTO, N.; MORENO, A. M. **Basics of software engineering experimentation**. [s.l.] Springer Science & Business Media, 2013.

KAREN SIMONYAN; ANDREW G. HOWARD. **Very Deep Convolutional Networks for Large-Scale Image Recognition.** 2014. ArXiv:1409.1556. Disponível em: https: //arxiv.org/abs/1409.1556. Acesso em: 05 mai. 2023.

KAUFFMANN, E.; GIL, D.; FERRÁNDEZ, A.; SELLERS, R.; MORA, H. **A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making**. Industrial Marketing Management, v. 90, n. 10, p. 523-537, 2020.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. **Supervised machine learning**: **A review of classification techniques. Emerging artificial intelligence applications in computer engineering**, v.160, p.3-24, 2007.

KHOSRAVI, A.; HUSSIN, A. R. C. **Customer Knowledge Management: Development Stages and Challenges**. Journal of Theoretical and Applied Information Technology, v. 91, n. 2, p. 264-274, 2016.

KHOSRAVI, A.; HUSSIN, A. R. C.; DAHLAN, H. M. **Toward a survey instrument for investigating customer knowledge management in software companies**. Journal of Theoretical and Applied Information Technology, v. 95, n. 23, p. 6494-6509, 2017.

KUNUMI. **Métricas de avaliação em Machine Learning: Classificação**. Publicado em: 18 mai. 2022. Disponível:<https://www.kunumi.com/2022/05/18/metricas-de-

avaliacao-em-machine-learning-classificacao/>. Acesso em: 22 jan.2023.

LAK, B.; REZAEENOUR, J. **Effective Factors of Social Customer Knowledge Management (SCKM) in Organisations: Study of Electronic Service Providers in Iran.** Journal of Information & Knowledge Management, v. 16, n. 02, p. 1750014-1750014, 2017.

LEITE, TIAGO M. **Redes Neurais, Perceptron Multicamadas e o Algoritmo Backpropagation**. Publicado em: 10 mai. 2018. Disponível em: <https://medium.com/ensina-ai/redes-neurais-perceptron-multicamadas-e-o-algoritmo-backpropagation-eaf89778f5b8>. Acesso em: 10 mai. 2021.

LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A.; CHENG, T. **Geospatial big data handling theory and methods: A review and research challenges**. ISPRS Journal of Photogrammetry and Remote Sensing, v.115, p.119-133, 2016.

LINOFF, G. S.; BERRY, M. J. **Data mining techniques: for marketing, sales, and customer relationship management**. John Wiley & Sons, 3ª ed., 2011.

LIU, B. **Sentiment analysis and opinion mining**. Synthesis lectures on human language technologies, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

LIU P, CHOO KKR, WANG L, HUANG F (2016) **SVM ou aprendizado profundo? Um estudo comparativo sobre sensoriamento remote classificação de imagem.** Computação Suave. https://doi.org/10.1007/s00500-016-2247-2

LU, H.; SETIONO, R.; LIU, H. **Effective Data Mining using Neural Networks. IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 957-961, 1996.

LUDERMIR, TERESA BERNARDA. **Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências**. São Paulo, v. 35, n. 101, p. 85-94, abr. 2021. Publicado abr. 19, 2021. <https://doi.org/10.1590/s0103-4014.2021.35101.007>.

LUCIDCHART, 2021. **O que você quer fazer com árvore de decisão?** Disponível em: <https://www.lucidchart.com/pages/pt/o-que-e-arvore-de-decisao/>. Acesso em: 09 set.2022.

LYU, J.; YANG, S.; CHEN, C. **Transform customer knowledge into company value - case of a global retailer.** In: 2009 6th International Conference on Service Systems and Service Management. IEEE, 2009. p. 959-964, 2009.

L. ZHANG, L. Jiang, C. Li, G. Kong, **Duas abordagens de ponderação de recursos para classificadores de texto naive bayes, Sistema Baseado em Conhecimento.** 100 (2016) 137– 144, <http: //dx.doi.org/10.1016/j.knosys.2016.02.017>.

MA, Z.; QI, L. **Toward an Integrated Customer Knowledge Management Model: A Process Based Approach.** In: 2009 International Conference on Management and Service Science. IEEE, 2009. p. 1-4.

MARIANO, DIEGO. **Métricas de avaliação em Machine Learning**. Publicado em: 30 nov. 2019. Disponível em: <https://diegomariano.com/metricas-de-avaliacao-em-machine-learning/>. Acesso em: 22 jan. 2023.

MARTINS, CLAUDIA APARECIDA et al. (2003). **Uma Experiência em Mineração de Textos Utilizando Clustering Probabilístico Clustering Hierárquico**. 2003. Instituto de Ciências Matemáticas e de Computação. São Carlos: Universidade de São Paulo.

MITCHEL, T.M., 1997, **"Machine Learning"**, Pittsburgh: McGraw Hill.

MJV, 2021. **Redes Neurais: o que são e porque podem representar a evolução dos negócios**. Disponível em: <https://www.mjvinnovation.com/pt-br/blog/redes-neurais-o-que-sao-e-porque-podem-representar-a-evolucao-dos-negocios/>. Acesso em: 05 set. 2022.

MOREIRA, SANDRO. **Rede Neural Perceptron Multicamadas.** Publicado em: 24 dez. 2018. Disponível em: <https://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9>. Acesso em: 10 mai. 2021.

MOURA, KARINA. **Ciclo de vida dos dados.** Publicado em: 08 jan. 2019. Disponível em: <https://medium.com/@kvmoura/kdd-process-9b8e3062142>. Acesso em: 10 jan. 2023.

NEOENERGIA - **A lei Geral de Proteção de Dados - LGDP e a segurança online**. Disponível em: <https://www.neoenergia.com/pt-br/te-interessa/tecnologia/Paginas/lei-geral-protecao-dados-lgpd.aspx/>. Acesso em: 10 set. 2022.

NONAKA, I. **Organizational Knowledge Creation Theory: Evolutionary Paths and Future Advances**. Organization Studies, v. 27, n. 8, p. 1179–1208, 6 jun. 2006.

NS HARZEVILI, SH ALIZADEH, **Mistura de classificador multinomial latente naive bayes, Appl**. Computação Suave. 69 (2018) 516–527.

NHACUONGUE, J. A. **O campo da ciência da informação: contribuições, desafios e perspectivas da mineração de dados para o conhecimento pós-moderno**. Universidade Estadual Paulista (UNESP), 2015.

OLIVEIRA; DANIEL JOSÉ SILVA; BERMEJO, PAULO HENRIQUE DE SOUZA. **Mídias sociais e administração pública: análise do sentimento social perante a**

atuação do Governo Federal brasileiro. Organ. Soc. Salvador, v. 24, n. 82, p. 491-508, set. 2017.

OHASHI, MASARU RODRIGO. **Da Análise de Sentimentos para o Reconhecimento de Emoções: Uma história PLN**. Disponível em: <https://medium.com/neuronio-br/da-an%C3%A1lise-de-sentimentos-para-o-reconhecimento-de-emo%C3%A7%C3%B5es-uma-hist%C3%B3ria-pln-171f27734c56>. Acesso em: 29 dez. 2022.

PANDEY, S. C.; SHUKLA, M. K.; MAURYA, U. K. **Managing Customer Knowledge in Service Economy**. Banking, Finance, and Accounting, p. 907-918, 2014.

PANWAR, Shailesh; RAIWANI, Y. **Data reduction techniques to analyze NSL-KDD dataset**. International Journal of computer engineering & technology (IJCET). Vol 5. Pag. 21-31.2014.

TEBALDI, PEDRO CÉSAR. **Análise de Sentimentos com Machine Learning**. Publicado em: 12 mai. 2019. Disponível em: <https://www.datageeks.com.br/analise-de-sentimentos/>. Acesso em: 28 dez. 2022.

PATHMIND. A Beginner's Guide to LSTMs and Recurrent Neural Networks. Disponível em: <https://pathmind.com/wiki/lstm>. Acesso em: 05 set. 2022.

PEARL, J., 2000, **"Causality: Models, Reasoning, and Inference", Cambridge University Press.**

PEREIRA, LUÍS MONIZ. **Inteligência Articial Mito e Ciência**. São Paulo, 2005. Disponível em: <http://centria.fct.unl.pt/~lmp/publications/online-papers/ia-mito.pdf>. Acesso em: 13 mai. 2022.

PIATETSKY-SHAPIRO, G.; MATHEUS, C. J.; CHAN, P. K. **Systems for Knowledge Discovery in Data bases**. IEEE Transactions on Knowledge and Data Engineering, 5, p.903-912, 1993.

POZZI, F. A.; FERSINI, E.; MESSINA, E.; LIU, B. **Sentiment analysis in social networks**. [S.l.]: Morgan Kaufmann, 2016.

PHRIDVIRAJ, M. S. B.; GURURAO, C. V**. Data mining–past, present and future–a typical survey on data streams.** Procedia Technology, v.12, p.255-263, 2014.

PRODANOV, C. C.; FREITAS, E. C. DE. **Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico-2a Edição**. [s.l.] Editora Feevale, 2013.

RAMESH WADWADAGI AND VEERAPPA PAGI. **Sentiment analysis with deep neural networks: comparative study and performance assessment**. Artificial

Intelligence Review, 2020.

RAVI, K.; RAVI, V. **A survey on opinion mining and sentiment analysis: tasks, approaches and applications.** Knowledge-Based Systems, Elsevier, v. 89, p. 14–46, 2015.

RECUERO, RAQUEL. **Redes sociais na Internet**. Porto Alegre: sulina, 2009.

RINA DECHTER (1986). **Learning while searching in constraint-satisfaction problems.** University of California, Computer Science Department, Cognitive Systems Laboratory.

REICHHELD FREDERICK F. **The One Number You Need to Grow.** Disponível em: Acesso em: 26 abril. 2022.

RISTOSKI, P.; PAULHEIM, H. **Semantic Web in data mining and knowledge discovery: A comprehensive survey.** Web semantics: science, services and agents on the World Wide Web, v. 36, p.1-22, 2016.

ROCKCONTEN. **Entenda o que é NPS (Net Promoter Score) e como implementar essa metodologia na sua empresa**. Publicado em: 5 mai. 2019. Disponível em: <https://rockcontent.com/br/blog/nps>. Acesso em: 03 mai. 2023.

SACRAMENTO GABRIEL, 2021. **Árvore de Decisão: entenda esse algoritmo de Machine Learning**. Publicado em: 12 jul. 2021. Disponível em: <https://blog.somostera.com/data-science/arvores-de-decisao/>. Acesso em: 08 set. 2022.

SAFETEC, INFORMÁTICA. **Dados do cliente no varejo: entenda a importância da gestão e análise.** Disponível em: <https://safetec.com.br/cloud-computing/dados-cliente-varejo/>. Acesso em: 10 jan. 2023.

SACHAMANOROM, W.; SENOO, D. **Voice of the Customer Through Customer Cocreation: The Case of Fuji Xerox Japan**. Association for Information Systems AIS Electronic Library (AISeL), 2016.

SANTOS, F. L. d. **Mineração de opinião em textos opinativos utilizando algoritmos de classificação**. Monografia (Graduação) - Trabalho de Conclusão de Curso, Universidade Brasília, 2013. Ceará, 2014.

SANTOS ALAN; AIRES KELSON; VERAS RODRIGO; UCHÕA VALESKA; SANTOS LUÍS. **Uma abordagem de classificação de imagens dermatoscópicas utilizando aprendizado profundo com redes neurais convolucionais.** In Anais do XVII Workshop de Informática Médica, Porto Alegre, RS, Brasil, 2017. SBC. URL <https://sol.sbc.org.br/index.php/sbcas/article/view/3717>.

SHARMA, NIKITA. **How to create custom NER in Spacy.** Publicado em 30 nov. 2019. Disponível em: <https://nikkisharma536.medium.com/how-to-create-custom-ner-in-spacy-cfcd531f8773>. Acesso em: 31 dez. 2022.

SEDIGHI, M. M.; MOKFI, T.; GOLRIZGASHTI, S. **Proposing a customer knowledge management model for customer value augmentation: A home appliances case study.** Journal of Database Marketing and Customer Strategy Management, v. 19, n. 4, p. 321-347, 2012.

SELLER, Michel Lens; LAURINDO, Fernando José Barbin. **Comunidade de marca ou boca a boca eletrônico: qual o objetivo da presença de empresas em mídias sociais?** Gest. Prod., São Carlos, v. 25, n. 1, p. 191-203, mar. 2018.

SEMMELROCK-PICEJ, M. T.; KANDUTSCH, H. **Information Technology based Customer Knowledge Management Externalisation Techniques for Requirements Analysis.** Proceedings of the European Conference on Information Management & Evaluation, p. 353-364, 2010.

SERRANOGUERRENO, JESUS et al. **Sentiment analysis: A review and compara tive analysis of web services**. Information Sciences, v. 311, p. 1838, 2015.

SILVA, L. A. Da; PERES, S. M.; BOSCARIOLI, C. Introdução à Mineração de Dados: Com Aplicações em R. [s.l.] Elsevier Brasil, 2017.

SIRKEMAA, S. J. (2010), **Information and Knowledge Sharing: Involving Customers in Developing Services**. In: Thinkmind / Service Computation 2010, The Second International Conferences On Advanced Service Computing. Anais. Lisboa, Portugal: Iaria, pp. 116-120.

SONG H.A.; LEE, S. Y. (2013). **Hierarchical Representation Using NMF. Neural Information Processing. Col: Lectures Notes in Computer Sciences.** 8226. [S.l.]: Springer Berlin Heidelberg. pp. 466–473. ISBN 978-3-642-42053-5. doi:10.1007/978-3-642-42054-2_58.

SOUZA, ALEX. (2019). **Algoritmo SVM (Máquina De Vetores De Suporte).** Publicado em 10 abr. 2019. Disponível em: <https://blogdozouza.wordpress.com/2019/04/10/algoritmo-svm-maquina-de-vetores-de-suporte-a-partir-de-exemplos-e-codigo-python-e-r/>. Acesso em: 04 abr. 2022.

SCHMIDHUBER, JURGEN (2015). **Deep Learning. Scholarpedia**: doi:10.4249/scholarpedia.32832.

SCIKIT LEARN – SVM (2021). **Support Vector Machines**. Disponível em: <https://scikit-learn.org/stable/modules/svm.html#svm-classification>. Acesso em: 04

abr. 2022.

STATISTA. **Twitter - Statistics & Facts**. Publicado em: 02 nov. 2022. Disponível em: < https://www.statista.com/topics/737/twitter/ >Acesso em: 23 dez. 2022.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de dados: com aplicações em R**. Elsevier. 1ª. ed. Rio de Janeiro, 2016.

SILVA, T. L. DA; SOUSA, F. R.; MACÊDO, J. A. F. de; Machado, J. C.; Cavalcante, A. A. **Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem.** [S.l.]: Quixadá, 2013.

STJ, 2020 - **Lei Geral de Proteção de Dados (LGPD)**. Disponível em: <https://www.stj.jus.br/sites/portalp/Leis-e-normas/lei-geral-de-protecao-de-dados-pessoais-lgpd./ >. Acesso em: 08 set. 2022.

TANG, S. H.; HOMAYOUNI, S. M.; ALAEI, H. **The role of intelligent agents in customer knowledge management**. Journal of Business, v. 5, n. 16, p. 7042-7049, 2011.

TAGHIZADEH, S. K.; RAHMAN, S. A.; HOSSAIN, M. M. **Knowledge from customer, for customer or about customer: which triggers innovation capability the most?** Journal of Knowledge Management, v. 22, n. 1, p. 162-182, 2017

TENFEN, EMERSON. **A técnica de Knowledge Discovery In Databases (KDD) aplicada nas ocorrências atendidas pela polícia militar**. 2003. Dissertação (Graduação), Universidade Regional de Blumenau. Blumenau, 2003.

TELLES, ANDRÉ. **A revolução das mídias sociais: cases, conceitos, dicas e ferramentas.** São Paulo:M. Books, 2010.

TIBCO. 2022. **O que é análise de sentimentos?** Disponível em: <https://www.tibco.com/pt-br/reference-center/what-is-sentiment-analysis>. Acesso em: 29 dez. 2022.

TORTELLA, P. L.; COELLO, J. M. A. **Análise de sentimentos em mídias sociais. Laboratório de Banco de Dados - Departamento de Ciência da Computação da Universidade Federal de Minas Gerais-UFMG**, 2015. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2013/0047.pdf>. Acesso em: 04 abr. 2022.

TWITTER. **Q1' 2020 Shareholder Letter**. Disponível em: <http://q4live.s22.clientfiles.s3-website-us-east 1.amazonaws.com/826641620/files/doc_financials/2020/q1/Q1-2020-Shareholder-Letter.pdf >. Acesso em: 07 mai. 2022.

VALACHERRY, A. K.; PAKKEERAPPA, P. **Customer Knowledge Management via**

social media: A Case Study of an Indian Retailer. Journal of Human Values, v. 24, n. 1, p. 39-55, 2018.

VIERA FILHO, V.; ALBUQUERQUE, M. T. C. F. **Abordagem Bayesiana para Simulação de Jogos Complexos**. In: SB Games, 2007, São Paulo. Proceedings of SB Games 2007.

WAFA MOUSSER; SALIMA OUADFEL. **Deep feature extraction for pap-smear image classification: A comparative study.** In Proceedings of the 2019 5th International Conference on Computer and Technology Applications, ICCTA 2019, pages 6–10, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-7181-0. doi: 10.1145/3323933. 3324060. URL <http://doi.acm.org/10.1145/3323933.3324060>.

WALTRICK, Camila (2020). **Machine Learning — O que é, tipos de aprendizagem de máquina, algoritmos e aplicações**. Publicado em: 07 mai. 2021. Disponível em: <https://medium.com/camilawaltrick/introducao-machine-learning-o-que-e-tipos-de-aprendizado-de-maquina-445dcfb708f0>. Acesso em: 02 set. 2022.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Practical machine learning tools and techniques.** Morgan Kaufmann, 2016.

XU, G. **Research on customer knowledge management based on CRM**. Proceedings – 2014 6th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2014, v. 1, p. 210-213, 2014.

XU, L. Managing customer services: **Customer knowledge management in service innovation. 8th International Conference on Service Systems and Service Management** - Proceedings of ICSSSM'11, 2011.

XUELIAN, L.; CHAKPITAK, N.; YODMONGKOL, P. **A novel two-dimension' customer knowledge analysis model.** Asian Social Science, v. 11, n. 16, p. 257-266, 2015.

XUELIAN, L. **Local Hospitality on European Market in Western.** v. 49, n. Icemse, p. 25-33, 2017.

X. WU, V. KUMAR, JR QUINLAN, J. GHOSH, Q. YANGANG, H. MOTODA, GJ MCLACHLAN, A. NG, B. LIU, PS YU. **Top 10 algoritmos em mineração de dados, Knowl. Inf**. Sistema 14 (1) (2008).

YIYI, Y.; RONGQIU, C. **Customer participation: Co-creating knowledge with customers. In: 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing**. IEEE, 2008. p. 1-6, 2008.

YIN, R. K. **Estudo de Caso: planejamento e métodos.** Trad. Daniel Grassi e Cláudio

Damacena. 2.ed. Porto Alegre: Bookman, 2006. p. 205.

YANO, T.; SMITH, N. A. **What's worthy of comment? content and comment volume in political blogs**. In: Fourth International AAAI Conference on Weblogs and social media. [S.l.: s.n.], 2010.

ZAFARANI, R.; ABBASI, M. A.; LIU, H. **Social Media Mining**. 1. ed. New York: Cambride University Press, 2014.

ZHONG, NING; LIU, CHUNNIAN; KAKEMOTO, YOSHITSUGU; OHSUGA, SETSUO. **KDD Process Planning. Association for the Advancement of Artificial Intelligence**. **KDD-97**. p. 291-294.1997.

Apêndices

Apêndice A – Importação e instalação de bibliotecas

```
#Import das Bibliotecas
import pandas as pd
import string
import re
import spacy
import math
import matplotlib.pyplot as plt
import numpy as np
import random
import seaborn as sns
import zipfile
import nltk
nltk.download('stopwords')
nltk.download('punkt')

from bs4 import BeautifulSoup
from google.colab import drive

from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
```

```
!pip install spacy==2.2.3
```

```
[2] import spacy
    spacy.__version__
```

```
'2.2.3'
```

```
[3] !python3 -m spacy download pt
```

```
!pip install Unidecode
```

```
[ ] !pip install rake-nltk
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev
Requirement already satisfied: rake-nltk in /usr/local/lib/python3.8/d
Requirement already satisfied: nltk<4.0.0,>=3.6.2 in /usr/local/lib/py
Requirement already satisfied: tqdm in /usr/local/lib/python3.8/dist-p
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/pytho
Requirement already satisfied: click in /usr/local/lib/python3.8/dist-
Requirement already satisfied: joblib in /usr/local/lib/python3.8/dist
```

Apêndice B – Remoção dos Stop Words e caracteres especiais na base de dados

```
[19] stop_words = spacy.lang.pt.stop_words.STOP_WORDS
```

```
[20] print(stop_words)
```

```
{'contra', 'mesmo', 'ontem', 'põem', 'ambos', 'ter', 'me', 'onde', 'estar', 'sem', 'iniciar', 'nove', 'aí', 'te', 'dezasseis', 'outra', 'novo', 'é',
```

```
[21] string.punctuation
```

```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

```
[22] from bs4 import BeautifulSoup
     import unicodedata
```

```
def preprocessamento(texto):
  texto = BeautifulSoup(texto, 'lxml').get_text()
  # Letras minúsculas
  texto = texto.lower()

  # Nome do usuário
  texto = re.sub(r"@[A-Za-z0-9$-_@.&+]+", ' ', texto)

  # URLs
  texto = re.sub(" xx",' ', texto)
  texto = re.sub(" xxx",' ', texto)
  texto = re.sub("\r\n",' ', texto)
  texto = re.sub(r"@[A-Za-z0-9]+", ' ', texto)
  texto = re.sub(r"https?://[A-Za-z0-9./]+", ' ', texto)
  texto = re.sub(r"\.(?=[0-9]|[a-z]|[A-Z])", ' ', texto)
  #texto = re.sub(r"\.(?=[0-9]|[a-z]|[A-Z])", ".$$$", texto)

# Espaços em branco
texto = re.sub(r" +", ' ', texto)
texto = re.sub(r'\d', '', texto)
texto = unicodedata.normalize('NFD', texto)
texto = texto.encode('ascii', 'ignore')
texto = texto.decode("utf-8")

# Lematização
documento = pln(texto)

lista = []
for token in documento:
  lista.append(token.lemma_)

# Stop words e pontuações
lista = [palavra for palavra in lista if palavra not in stop_words and palavra not in string.punctuation]
lista = ' '.join([str(elemento) for elemento in lista if not elemento.isdigit()])

return lista
```

Apêndice C – Tokenização

```
[ ] from nltk import tokenize

    frase = "Bem vindo ao mundo do PLN"
    token_espaco = tokenize.WhitespaceTokenizer()
    token_frase = token_espaco.tokenize(frase)
    print(token_frase)
```

```
['Bem', 'vindo', 'ao', 'mundo', 'do', 'PLN']
```

```
[ ] import nltk

    todas_palavras = ' '.join([texto for texto in bagofwords])
    frequencia = nltk.FreqDist(token_espaco.tokenize(todas_palavras))
    df_frequencia = pd.DataFrame({"Palavra": list(frequencia.keys()),
                                  "Frequência": list(frequencia.values())})
```

```
df_frequencia.head(10)
```

Apêndice D – Função para visualização de WordCloud

```
[34] %matplotlib inline

     from wordcloud import WordCloud

     todos_palavras = ' '.join([texto for texto in base_dados["causa"]])
```

```
[36] nuvem_palavras = WordCloud(width = 800, height = 500, max_font_size = 110,
                                collocations = False).generate(todos_palavras)
```

Apêndice E – Função para modificar atributos da base de dados

```
[ ]
    r = Rake(include_repeated_phrases=False, min_length=1, max_length=5)
    text_to_rake = text
    r.extract_keywords_from_text(text_to_rake)
```

```
[ ]
    words_ranks = [keyword for keyword in r.get_ranked_phrases_with_scores() if keyword[0] > 5]

    print("Principais palavras-chave")
    words_ranks
```

Apêndice F – Função para filtrar as principais palavras-chave, utilizado a biblioteca rake

```
[ ] #Análise do Texto - Principais Palavras-Chave (Keywords)
    r = Rake(include_repeated_phrases=False, min_length=1, max_length=5)
    text_to_rake = text
    r.extract_keywords_from_text(text_to_rake)
```

```
[ ] # Filtro para as principais Palavras-Chave
    words_ranks = [keyword for keyword in r.get_ranked_phrases_with_scores() if keyword[0] > 5]

    print("Principais palavras-chave")
    words_ranks
```

Apêndice G – Função para encontrar os principais tópicos na base de dados

```
import gensim

from gensim import corpora
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
```

```
[349]
doc_complete = text
docs = word_tokenize(doc_complete)

docs_out = []
docs_out.append(docs)
```

```
[350]
dictionary = corpora.Dictionary(docs_out)
```

```
[351]
doc_term_matrix = [dictionary.doc2bow(doc) for doc in docs_out]
```

```
[30]
Lda = gensim.models.ldamodel.LdaModel
ldamodel = Lda(doc_term_matrix, num_topics=10, id2word = dictionary, passes=50, random_state=4)
```

Apêndice H – Função para identificar as 10 palavras mais frequentes

```
[358] #Calculando a frequência das Palavras (Com exceção das Stopwords)
      from nltk.probability import FreqDist

      freq = FreqDist(bagofwords)
      freq = freq.most_common(10)

      print("As 10 palavras mais frequentes\n")
      print(freq)
```

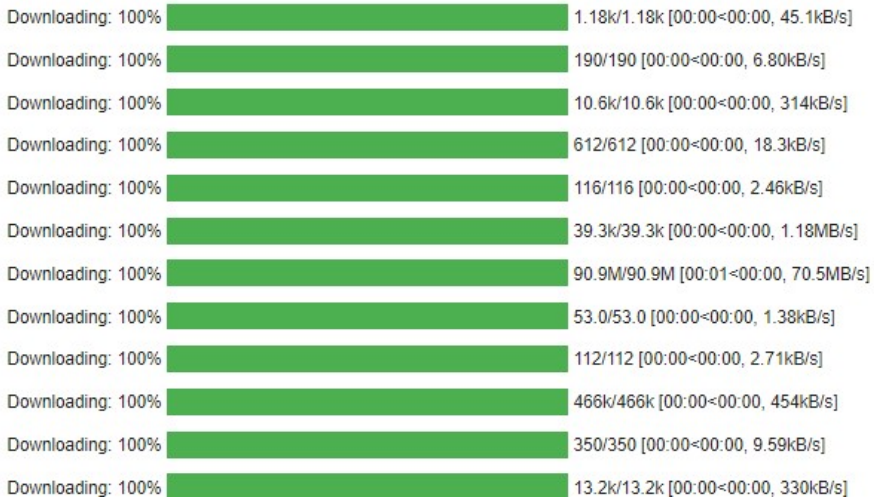
```
As 10 palavras mais frequentes

[('carta', 4398), ('atend', 2401), ('limit', 2396), ('compr', 1912), ('carrefour', 1334),
```

Apêndice I – Método clustering aplicado no texto

```
def cluster_text(text):
    vectorizer = TfidfVectorizer(stop_words=stopwords)
    X = vectorizer.fit_transform(bagofwords)

    import matplotlib.pyplot as plt
    from sklearn.cluster import KMeans
    Sum_of_squared_distances = []
    K = range(2,10)
    for k in K:
       km = KMeans(n_clusters=k, max_iter=200, n_init=10)
       km = km.fit(X)
       Sum_of_squared_distances.append(km.inertia_)
    plt.plot(K, Sum_of_squared_distances, 'bx-')
    plt.xlabel('k')
    plt.ylabel('Sum_of_squared_distances')
    plt.title('Elbow Method For Optimal k')
    plt.show()

    print('Quantidade de Clusters: ')
    true_k = int(input())
    model = KMeans(n_clusters=true_k, init='k-means++', max_iter=200, n_init=10)
    model.fit(X)

    labels=model.labels_
    clusters=pd.DataFrame(list(zip(text,labels)),columns=['title','cluster'])
    #print(clusters.sort_values(by=['cluster']))

    for i in range(true_k):
        print(clusters[clusters['cluster'] == i])

    return
```

Apêndice J – Função Bertopic

```
from bertopic import BERTopic

topic_model = BERTopic()
topics, probs = topic_model.fit_transform(newsentences)
```



Apêndice L – Modelo classificador de texto

```
[ ] modelo = spacy.blank('pt')
    categorias = modelo.create_pipe("textcat")
    categorias.add_label("F")
    categorias.add_label("M")
    modelo.add_pipe(categorias)
    historico = []

[ ] modelo.begin_training()
    for epoca in range(10):
      random.shuffle(base_dados_treinamento_final)
      losses = {}
      for batch in spacy.util.minibatch(base_dados_treinamento_final, 256):
        textos = [modelo(texto) for texto, entities in batch]
        annotations = [{'cats': entities} for texto, entities in batch]
        modelo.update(textos, annotations, losses=losses)
        historico.append(losses)
      if epoca % 5 == 0:
        print(losses)

    {'textcat': 1.669861997444927e-05}
    {'textcat': 4.8259638701052265e-08}
```