UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Aprendizado de máquina para classificação e direcionamento de reclamações de consumidores

Hussein Hassem Sampaio el Messmar

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Hussein Hassem Sampaio el Messmar

Aprendizado de máquina para classificação e direcionamento de reclamações de consumidores

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

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ronaldo Cristiano Prati

Versão original

São Carlos

2022

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

Sampaio Messmar, Hussein Hassem
S972a Aprendizado de máquina para classificação e direcionamento de reclamações de consumidores / Hussein Hassem Sampaio Messmar; orientador Ronaldo Cristiano Prati. -- São Carlos, 2022.
54 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. INTELIGÊNCIA ARTIFICIAL. 2. APRENDIZADO COMPUTACIONAL. I. Prati, Ronaldo Cristiano, orient. II. Título.

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

Hussein Hassem Sampaio el Messmar

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Ronaldo Cristano Prati

Original version

São Carlos

2022

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus.

Aos meus pais, Elizabeth e Hossem que me forneceram a base para mais essa conquista.

À todos os professores deste curso que conduziram com maestria o papel de passar os conhecimentos necessários para a nossa formação.

Ao meu orientador Ronaldo Prati pelo seu apoio, direcionamento e ideias que foram essenciais para a conclusão deste projeto.

E em especial à minha esposa Elizandreia e minha filha Sophia pelo amor, paciência e apoio em todos os momentos que necessitei.

"The best way to predict the future is to create it."
Alan Kay

RESUMO

Messmar, H.H.S **Aprendizado de máquina para classificação e direcionamento de reclamações de consumidores**. 2022. 54p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

O presente trabalho propõe a utilização de técnicas de aprendizado de máquina em conjunto com mineração de texto para criação de um sistema capaz de classificar e direcionar de maneira autônoma as reclamações abertas por clientes de serviços financeiros. A metodologia utilizada foi exploratória e descritiva em conjunto com a metodologia CRISP que auxiliou nas seguintes etapas: compreensão do problema, interpretação dos dados, modelagem e avaliação dos resultados. Foram aplicadas 2 arquiteturas distintas com intuito de compararação, sendo elas: *Naive Bayes* e *BERT*. As avaliações demonstraram que para a técnica *BERT* foi possível obter os melhores resultados com uma acurácia média de 65,00% na classificação das reclamações entre 23 classes.

Palavras-chave: Aprendizado de Máquina. Aprendizagem Supervisionada. Reclamações de clientes. Classificação de texto.

ABSTRACT

Messmar, H.H.S . 2022. 54p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

The present work proposes the use of machine learning techniques with text mining to create a system capable of classifying and autonomously directing the financial services customer's complaints. The methodology used was exploratory and descriptive in conjunction with the CRISP methodology that helped in the stages of the project: understanding the problem, data interpretation, modeling, and evaluation of results. Two different architectures were applied to compare the *Naive Bayes* and *BERT* models. The evaluations showed that the *BERT* technique obtained the best results with an average accuracy of 65.00% in the classification of complaints among 23 classes.

Palavras-chave: Machine Learning. Supervised Learning. Customers complaints. Text classifying.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CFPB	Consumer Financial Protection Bureau
NLP	Natural Language Processing
KNN	K-Nearest Neighbors
TF	Term Frequency
IDF	Inverse Document Frequency
BERT	Bidirectional Encoder Representations from Transformers
MLM	Masked-Language Modeling
USP	Universidade de São Paulo
AM	Aprendizado de Máquina

SUMÁRIO

1 INTRODUÇÃO

Produtos financeiros são hoje um dos pontos chave para uma economia sólida [Halvorsen and Møkkelgård 2018]. Em muitos países, a regulação desses produtos é feita por agências governamentais como o Banco Central (BACEN) no Brasil [Banco Central do Brasil 2021] e o *Consumer Financial Protection Bureau* (CFPB) nos Estados Unidos [Consumer Financial Protection Bureau 2021]. Nesse contexto, consumidores que se sentem lesados na contratação desses produtos podem abrir uma reclamação por meio dessas agências, caso não consigam resolver seu problema diretamente com a empresa. Com a popularização da Internet, o caminho para realizar uma reclamação ficou mais fácil [Coussement and Poel 2008]. Segundo [BUREAU OF CONSUMER FINANCIAL PROTECTION 2021], o volume de reclamações vem crescendo ao longo dos anos, o que torna a gestão dessas reivindicações algo desafiador para as empresas do ramo.

Grande parte das informações referente às reclamações está contida em dados não estruturados, na forma de textos, em que os consumidores relatam os problemas ocorridos e quais são as suas solicitações e reclamações. Utilizar uma abordagem de classificação automática pode trazer ganhos financeiros, pois reduzimos a necessidade de pessoas realizando o trabalho de leitura inicial desses textos, reduzindo também os erros operacionais que podem ocorrer na classificação das reivindicações e melhorando o tempo entre recebimento do caso e o encaminhamento para área resolvedora [Coussement and Poel 2008].

A implementação de um algoritmo de aprendizado de máquina para classificação de textos se inicia no direcionamento das reclamações, onde os casos são separados e distribuídos entre as áreas encarregadas. Posteriormente, esse mesmo sistema pode ser utilizado na análise das réplicas dos consumidores, podendo identificar casos em que o problema reclamado ainda não foi resolvido. [Joung *et al.* 2019] explicam que a implementação de um sistema de classificação automática de reclamações pode ser um desafio, visto que vários textos recebidos são ambíguos. Essa ambiguidade está relacionada ao fato de que muitos clientes não tem conhecimento suficiente para distinguir o que de fato é o seu real problema. A utilização de uma base de dados rotulada grande e diversificada pode ser uma das formas de tentar melhorar a precisão dessas classificações. Outros atributos, além dos extraídos do texto, também podem ser utilizados e, se bem selecionados, tendem a ser de extrema importância no treinamento desses algoritmos [Yang *et al.* 2018].

1.1 Questões de Pesquisa

Neste projeto, pretendemos utilizar algoritmos de aprendizado de máquina para a geração de um classificador de reclamação de consumidores. Dessa maneira, formulamos a seguinte questões de pesquisa:

Q1 *"Qual a abordagem mais adequada para utilização de dados estruturados e não estruturados na classificação de textos?"*

Os objetivos definidos para solução da questão são:

- Mapear e aplicar técnicas de pre-processamento e aprendizado de máquina para textos envolvendo múltiplas classes utilizando-se de modelos presentes na literatura.
- Avaliar essas técnicas aplicadas, afim de comparar os resultados de assertividade dos algoritmos e também o custo computacional de cada um.

1.2 Organização da Monografia

Os próximos capítulos deste trabalhado serão estruturados da seguinte forma:

- No capítulo 2, serão apresentados os fundamentos teóricos e os trabalhos relacionados.
- No capítulo 3, serão apresentadas a metodologia e as técnicas propostas pelo trabalho, será demonstrado também como foi o processo de desenvolvimento do experimento.
- No capítulo 4 são apresentados os resultados das técnicas implementadas com a análise das métricas de avaliação estabelecidas.
- Por fim, no Capítulo 5 são apresentadas as conclusões discutindo quais foram os avanços e limitações do experimento e também as sugestões para trabalhos futuros.

2 TRABALHOS RELACIONADOS E FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados conceitos importantes relacionados a este trabalho. Dentre os assuntos abordados, estão os sistemas de proteção aos consumidores de serviços financeiros no Brasil e nos Estados unidos, e as principais técnicas de aprendizado de máquina para classificação de textos. Ao final, serão evidenciados os trabalhos encontrados por meio da revisão bibliográfica sistemática.

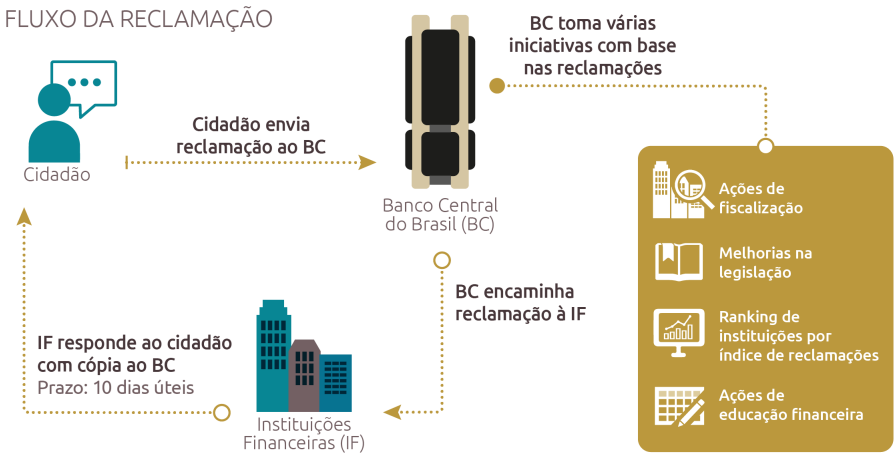
2.1 Sistema de proteção à consumidores de serviços financeiros

Nesta seção será apresentada uma revisão do sistemas responsáveis pela proteção dos consumidores de serviços financeiros.

2.1.1 Sistemas existentes no Brasil

Segundo [Banco Central do Brasil 2021], o responsável por prestar assistência à consumidores de serviços financeiros no Brasil é o Banco Central. Em casos de problemas, a instituição orienta que os consumidores devem primeiro tentar resolver o caso diretamente com a instituição, e, se houver insucesso, pode ser feita uma reclamação formal através das plataformas e canais de atendimento da agência. Conforme mostrado na Figura 1, além de ações na esfera individual, que é a mediação da reclamação do consumidor, outras ações na esfera coletiva são tomadas com o intuito de melhorar os serviços prestados pelas instituições financeiras.

Figura 1 – Fluxo de Reclamação. Fonte: Banco Central do Brasil



2.1.2 Sistema existente nos Estados Unidos

Nos Estados Unidos, conforme [Consumer Financial Protection Bureau 2021], a agência responsável pelo intermédio de interesses entre instituições financeiras e consumi-

dores, é a Consumer Financial Protection Bureau, também conhecida pela sigla CFPB. Ao contrário do Brasil, em que o Banco Central possui diversas atribuições, o departamento americano tem como única finalidade realizar a proteção dos direitos dos consumidores de produtos financeiros.

2.2 Gestão de Reclamações

Atualmente, a experiência do cliente é um dos pilares de qualquer empresa que quer obter sucesso e longevidade. Ter uma gestão eficaz no tratamento de reclamações é uma parte importante na jornada do cliente. Por isso, é necessário que o retorno dessas reclamações seja feita de maneira rápida e com uma solução satisfatória. Segundo [Fornell and Wernerfelt 1988], o tratamento de uma reclamação é uma oportunidade que a empresa tem em recuperar sua imagem perante o cliente. [Álvarez, Casielles and Martín 2010] define a gestão das reclamações como a maneira como as empresas lidam com as reivindicações de seus clientes insatisfeitos com o produto/serviço adquirido. O trabalho de [Johnston and Mehra 2002] orienta que, para uma gestão de reclamação eficaz, deve haver apenas um ponto de contato entre o cliente e o representante da empresa. Além disso, este contato deve estar apto para solucionar o problema do cliente de forma a não gerar experiências negativas para o consumidor. Dessa maneira, ter uma classificação e um direcionamento assertivo é importante para a empresa proporcionar uma resolução rápida e eficaz.

2.3 Mineração de Textos

A mineração de textos tem como propósito principal a descoberta de conhecimento a partir de textos. Esse processo é feito por meio da modelagem dos dados não estruturados, de forma a possibilitar a aplicação de algoritmos de aprendizado de máquina nessas informações. O principal objetivo da mineração de texto, é identificar padrões nos dados textuais analisados, possibilitando a extração de informação do conteúdo [Aggarwal and Zhai 2012], esse processo de mineração tem etapas bem definidas, conforme descreve [Rezende 2003] na Figura 2.

Figura 2 – Processo de Mineração de texto [Rezende 2003].

2.4 Pré Processamento

Quando trabalhamos com aprendizado de máquina, o processo que antecede a implementação dos algoritmos é chamado de pré processamento. Essa etapa é importante para que tenhamos bons resultados. Nos modelos trabalhados na mineração de textos, essa etapa é fundamental pois precisamos extrair atributos de dados não estruturados, que não podem ser aplicados diretamente à um algoritmo de *AM*. [Ozbay and Alatas 2020].

2.4.1 Espaço Vetorial Léxico

Também conhecido como *bag of words*, esse modelo de representação transforma o texto em dados estruturados por meio de um vetor de palavras. Ele é um modelo esparso e multidimensional, visto que cada nova palavra no texto será associada a uma nova dimensão [Aggarwal 2018]. Para melhorar o desempenho, é necessário a implementação em conjunto de técnicas para redução de dimensionalidade [Martins 2003]. [Alves 2010] ilustra a implementação do modelo através da Figura 3, em que d_j, representa os documentos analisados, w_j as palavras extraídas desse documento e a_{ij} o peso atribuído a cada palavra. [Aggarwal 2018] menciona que existem duas técnicas comumente utilizadas dentro deste modelo de representação que são os modelos binário e os modelos que contabilizam a frequência das palavras que são chamados de *Term Frequency-inverse Document Frequency* (TF-IDF).

Figura 3 – Modelo Bag Of Words [Alves 2010].

	w_1	w_2	w_3	⋯	w_n
d_1	a_{11}	a_{12}	a_{12}	⋯	a_{1n}
d_2	a_{21}	a_{22}	a_{23}	⋯	a_{2n}
⋯	⋮	⋯	⋯	⋱	⋮
d_n	a_{n1}	a_{n2}	a_{n3}	⋯	a_{nn}

2.4.1.1 Modelo Binário

No modelo binário, as palavras são extraídas do texto e o valor atribuído a cada uma delas na representação é 0 ou 1, ou seja, se apalavra estiver presente no texto será atribuído o valor 1 caso esteja ausente será atribuído o valor 0 [Aggarwal 2018].

2.4.1.2 Term Frequency - Inverse Document Frequency (TF-IDF)

Nesta técnica, a frequência das palavras influencia no peso. Uma normalização de valores é aplicada para que nenhum *outlier* prevaleça sobre as demais palavras. A equação 2.1 demonstra como é a implementação da técnica em que $freq(t_i, d_j)$ é o número de vezes que o termo t_i aparece no documento d_j, e $idf(t_j)$ é a frequência inversa do termo em relação aos documentos, ou seja, quanto mais frequente a palavra é, menor será o valor retornado [Aggarwal 2018] [Matsubara, Martins and Monard 2003].

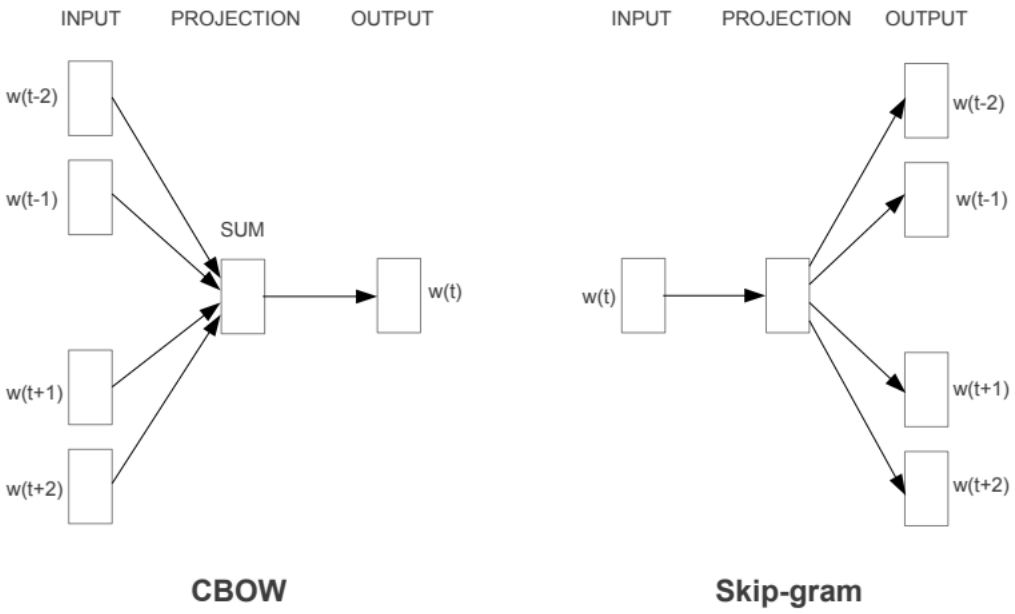
$$tfidf(t_j, d_i) = freq(t_i, d_j) * idf(t_j) \tag{2.1}$$

2.4.2 Espaço Vetorial Semântico - Word Embeddings

Apesar de uma técnica amplamente utilizada, a representação léxica não é capaz de tratar questões de semântica do texto como a polissemia, em que palavras de mesma grafia podem ter sentidos diferentes. Com o intuito de atender essa necessidade foi criada técnica conhecida pelo termo *word embedding*. Neste método a sequência das palavras é considerada na modelagem do espaço vetorial [Aggarwal 2018]. Duas das grandes empresas do cenário da tecnologia criaram suas próprias representações utilizando essa técnica, o *Word2Vec* (Google) e o *FastText* (Facebook).

A técnica *Word2Vec* consiste em uma rede neural de duas camadas, que pode predizer palavras a partir do contexto. Essa técnica é conhecida como *continuous bag of words*. Também é possível predizer contextos através das palavras presentes no texto, em que essa abordagem é chamada de *skip-gram* [Mikolov *et al.* 2013]. Na Figura 4 podemos observar a ilustração dessas duas variações da técnica Word2Vec.

Figura 4 – Representação das variações da técnica Word2Vec [Mikolov *et al.* 2013]



Complementando o modelo *Word2Vec*, o *FastText* acrescenta informações referente aos n-gramas de caracteres, incluindo referências de sufixos, prefixos e radicais, permitindo que a representação semântica de uma palavra seja expandida com mais atributos.

2.4.3 *Bidirectional Encoder Representations from Transformers* (BERT)

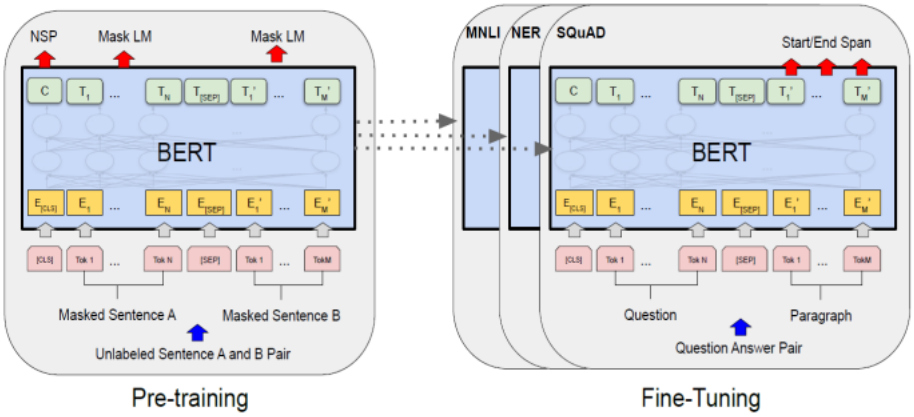
Conforme [Devlin and Chang 2018], hoje um dos maiores desafios quando trabalhamos com processamento de língua natural (*NLP*, do inglês *Natural Language Processing*) é a obtenção de dados rotulados para treinamento dos modelos. Como forma de reduzir necessidade de dados classificados para treinamento, o modelo *BERT* foi criado. BERT é um algoritmo que utiliza a abordagem de transferência de aprendizado em um modelo que já passou por um pré treinamento utilizando bilhões de dados. Além da enorme quantidade de dados utilizados na geração deste modelo, o *BERT* se destaca por armazenar a relevância das palavras no texto e por utilizar uma técnica que permite ter vetores de atenção bidirecionais, que são mecanismos que analisam o documento nos dois sentidos, esquerda para direita e direita para esquerda. [Devlin *et al.* 2018] [Vaswani *et al.* 2017].

[Devlin *et al.* 2018] explica que o treinamento do modelo *BERT* é constituído de duas partes. Na primeira, também conhecida como *pre-training*, são utilizados dados não rotulados em um pré-treinamento aplicando as técnicas *Masked Language Model*(MLM), que irá ocultar determinada palavra e tentar prever ela levando em consideração as demais palavras de forma aleatória. Na outra tarefa, chamada de *Next Sentence Prediction*(NSP), são recebidos dois pares de sentença, e o modelo tenta prever se uma está ligada à outra.

A segunda parte do treinamento é conhecido como *fine-tuning*, que é feita após a conclusão da primeira etapa. Nela, são utilizados dados rotulados com um conjunto

específico para a tarefa desejada. Esse pré treinamento tem um custo computacional muito menor, o que o torna viável para ser realizado em máquinas comuns [Kenton and Toutanova 2019]. Na Figura 5 podemos observar esse fluxo implementado no *BERT*.

Figura 5 – Etapas de treinamento BERT. [Devlin *et al.* 2018]



2.5 Aprendizado de Máquina

Após a construção dos atributos, precisamos submeter o conjunto para treinamento a um algoritmo de aprendizado de máquina. Como trabalharemos com dados rotulados, serão utilizados apenas algoritmos de aprendizado de máquina supervisionados. Nesta seção apresentaremos os algoritmos utilizados no desenvolvimento deste trabalho.

2.5.1 Naive Bayes

Naive Bayes é um algoritmo probabilístico baseado no teorema de Bayes, dado pela equação 2.2 que descreve a probabilidade de um evento ocorrer.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{2.2}$$

Segundo [McCallum, Nigam *et al.* 1998], dois modelos deste algoritmo bastante aplicados na classificação de textos são os modelos Multinomial e o Bernouli.

O modelo multinomial modela os atributos como uma distribuição normal multivariada, e se baseia na frequência dos termos e é dado pela equação 2.3, em que a distribuição é parametrizada por $\theta_y = (\theta_{y1}....\theta_{yn})$ para cada classe y, n é o numero de atributos e θ_{yi} é a probabilidade $P(x_i|y)$ do atributo i aparecer na classe, N_{yi} é o número de vezes que o atributo x_i aparece no conjunto de treinamento de determinada classe e N_y é a contagem total de todos os atributos da classe [Scikit Learn - Naive Bayes 2022].

$$\hat{\theta}_i = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{2.3}$$

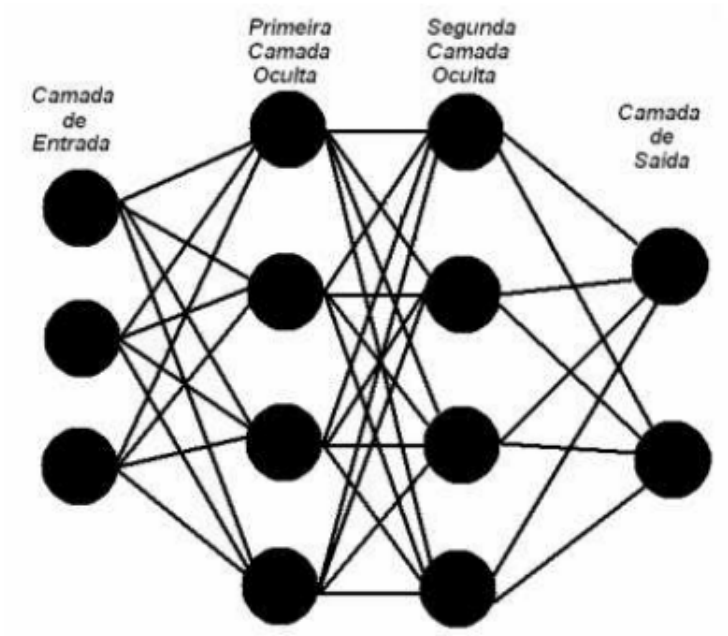
O modelo de Bernouli, dado pela equação 2.4, é baseado na Distribuição Multivariada de Bernouli. Neste modelo, os atributos assumem valores booleanos para cada classe. A principal diferença entre o modelo de Bernouli é que ele irá penalizar atributos ausentes enquanto no modelo Multinomial ele é apenas ignorado [Scikit Learn - Naive Bayes 2022].

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i) \tag{2.4}$$

2.5.2 Rede Perceptron Multicamadas

A rede perceptron de múltiplas camadas é utilizada quando precisamos resolver problemas não linearmente separáveis, em que não seria possível a aplicação de redes de neurônio de camada única. Essa rede é composta de uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A Figura 6 ilustra a arquitetura desta rede com uma camada de entrada com três neurônios, duas camadas ocultas com 4 neurônios cada e uma e a saída com dois neurônios [Falcão *et al.* 2013].

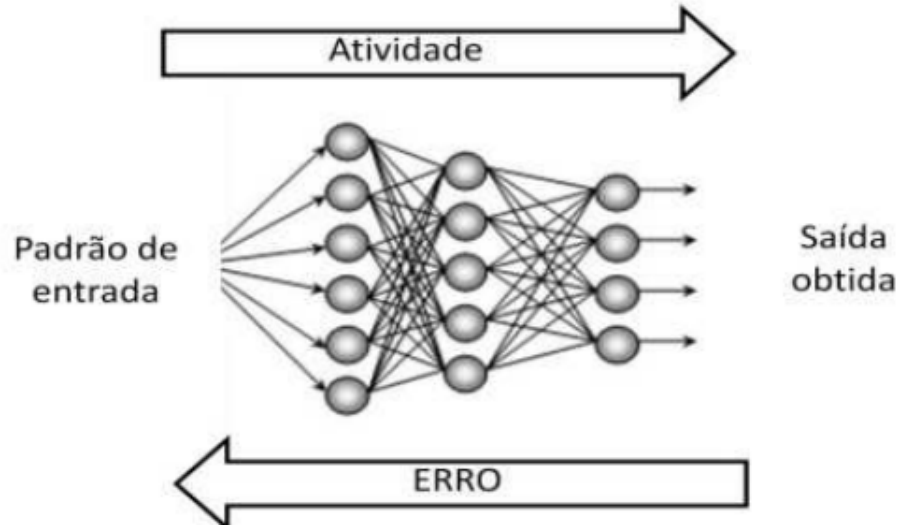
Figura 6 – Arquitetura de RNA MLP com duas camadas ocultas [Falcão *et al.* 2013]



.

Para treinar as redes de Perceptron de múltiplas camadas o algoritmo *backpropagation* é normalmente utilizado. Esse algoritmo tem duas etapas onde a saída da rede é melhorada a cada interação. Na Figura 7 podemos ver essas duas etapas. A primeira conhecida como propagação (*forward*), a aquisição dos dados é feita através da camada de entrada e propagado através da rede gerando uma saída, através da saída um erro é calculado e em seguida a retropropagação se inicia ajustando os pesos dos neurônios de forma a reduzir o erro na saída [Falcão *et al.* 2013].

Figura 7 – Algoritmo Backpropagation [Falcão *et al.* 2013]



.

2.6 Trabalhos Relacionados

Em [Joung *et al.* 2019], foram explorados dados de reclamações de clientes de uma empresa de venda de ar condicionados. Nesse trabalho, foi aplicada a técnica *Outcome-Drive Innovation*(ODI). O intuito do trabalho não era classificar o texto, mas gerar *insights* através de agrupamento não supervisionado das reclamações. Ao fim do experimento foram identificados duas necessidades dos clientes que os especialistas da área consideraram relevantes.

[Yang *et al.* 2018] implementaram uma abordagem para classificação de reclamações de consumidores de uma empresa de telecomunicações. Além dos atributos extraídos do texto da reclamação, também foram utilizados outros seis atributos numéricos como a força do sinal e a qualidade da chamada do cliente. Os algoritmos classificadores trabalharam com rótulos binários e obtiveram um bom resultado.

Em [Khedkar and Shinde 2020] os autores realizaram a análise de sentimento em avaliações de consumidores de um hotel para identificação de casos de elogio e reclamação. Uma das abordagens para classificação do texto adotado, foi uma técnica que mistura diversos algoritmos conhecida como *ensemble*. No final, os autores conseguiram um desempenho satisfatório em praticamente todos os sistemas de avaliação.

Em [Yang *et al.* 2019] os autores analisaram textos de reclamações no chat de uma empresa de *e-commerce* com intuito de prever se o cliente é propenso a escalar sua reivindicação à órgãos reguladores. Como produto final da pesquisa foi desenvolvido uma funcionalidade de monitoramento das reclamações que sinaliza essa propensão dos clientes que estão reclamando. Como algoritmo de classificação foram utilizadas Redes Neurais LSTM.

Em [Krishna *et al.* 2019] os autores utilizaram reclamações de bancos indianos.

Nesse trabalho, foram aplicadas diversas técnicas de pre processamento como a *Liguistic Inquiry and Word Count*(LIWC), TF-IDF, *Word2Vec* e *document-term matrix*(DTM). Uma gama grande de algoritmos de aprendizado de máquina também foi aplicado e o que melhor performou foi o algoritmo Floresta Aleatória.

Em [Leem and Eum 2021] os autores analisaram avaliações de consumidores de um banco digital. Eles tinham a proposta de realizar a análise de sentimento e também classificar como está a qualidade do atendimento da empresa em determinadas áreas. Para classificação de sentimento foi utilizado um software chamado *TEXTOM*. Após a análise de sentimento, foram agrupadas palavras chaves de determinados assuntos para classificação dos textos e posteriormente realizada a análise de quais assuntos tinham o maior percentual de avaliações negativas.

Em [Peixoto 2021], o autor implementou técnicas de aprendizado de máquina para triagem de reclamações de consumidores no site Reclame Aqui, foram implementadas técnicas de pre processamento baseada em transformers (BERT) e *embeddings* FastText, os melhores resultados vieram da combinação do *Naive Bayes SVM* (NBSVM) e *FastText*.

No Quadro 1 são apresentados os trabalhos citados , realizamos a divisão das técnicas aplicadas para pré processamento e aprendizado de máquina.

Quadro 1 – Comparativo de Trabalhos

	Pré Processamento			Aprendizado de Máquina				
Trabalho	Espaço Vetorial Léxico	Espaço Vetorial Semântico	BERT	SVM	Floresta Aleatória	Naive Bayes	RNA	Outros
[Joung *et al.* 2019]	x							
[Yang *et al.* 2018]	x			x	x	x		x
[Khedkar and Shinde 2020]	x			x	x	x	x	
[Yang *et al.* 2019]	x	x						
[Krishna *et al.* 2019]	x	x		x	x	x	x	x
[Leem and Eum 2021]	x							x
[Peixoto 2021]		x	x					x

3 METODOLOGIA

Neste capítulo é apresentada a metodologia proposta para o desenvolvimento do trabalho relacionado à classificação de reclamações de clientes de produtos financeiros. Os dados utilizados no desenvolvimento foram coletados do site da instituição americana Consumer Financial Protection Bureau[1]. O processo de limpeza inicial removeu campos não utilizados, seleção de uma categoria dos produtos reclamados e também a remoção de classes com baixa recorrência. Como metodologia base no desenvolvimento foi utilizado o *Cross Industry Standard Processo for Data Mining* (CRISP-DM) que significa Processo Padrão Inter-Indústrias para Mineração de Dados, representado na Figura 8. Essa técnica é amplamente utilizada em diversos trabalhos envolvendo aprendizado de máquina e se destaca pela simplicidade de implementação [Wirth and Hipp 2000]. A seguir, um resumo do das etapas da metodologia:

1. **Entendimento do problema**: Nessa fase os objetivos do projeto são identificados e detalhados.

2. **Interpretação dos dados**: Nessa etapa a coleta dos dados é realizada bem como a exploração e análise dos dados de forma a garantir que eles tenham qualidade e atendam os requisitos para desenvolvimento do projeto.

3. **Preparação dos dados**: Após a coleta dos dados, é realizada a preparação da base de forma a deixá-la pronta para a próxima etapa de modelagem, alguma das atividades aplicadas nesta fase são: remover atributos que não serão utilizados, retirada de valores nulos, adequação do formato dos atributos, geração de representações que sejam interpretáveis pelos algoritmos de AM.

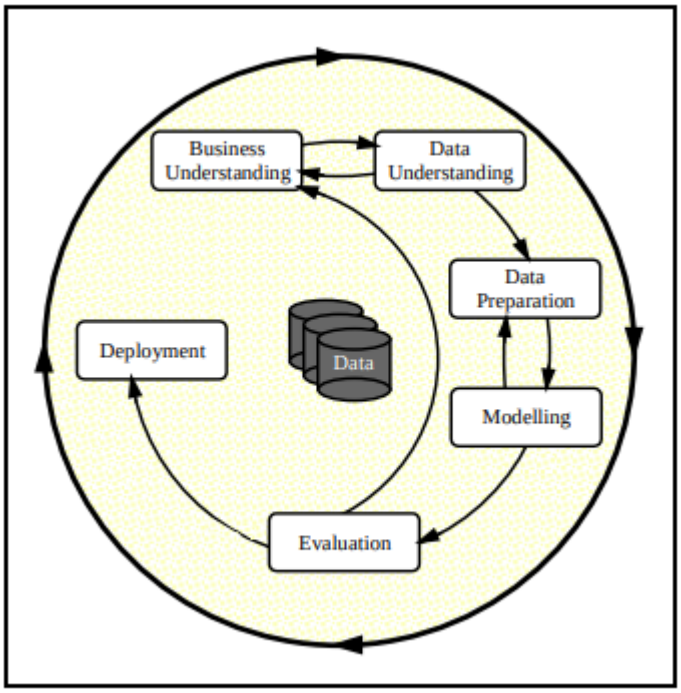
4. **Modelagem**: Nesta etapa o modelo é construído, podendo ser implementado com diversas variações de algoritmos e representações dos dados geradas na etapa anterior.

5. **Avaliação**: O modelo é avaliado verificando se os objetivos estabelecidos na etapa de entendimento do problema foram alcançados, caso contrário, a implementação do projeto retorna para o estágio inicial da metodologia para que sejam realizadas alterações necessárias com intuito alcançar os resultados esperados.

6. **Implementação**: O modelo final desenvolvido e validado é publicado.

[1] https://www.consumerfinance.gov/

Figura 8 – Fases Metodologia CRISP-DM [Wirth and Hipp 2000]



.

3.1 Entendimento do Problema

A gestão de reclamações de clientes sempre foi um desafio para as empresas ao redor do mundo, com o advento da internet, o volume dessas revindicações através de canais digitais vem crescendo exponencialmente. Antes, era necessário um longo deslocamento até a sede da empresa ou horas no telefone para submeter uma reclamação. Hoje em dia, é algo simples bastando apenas o preenchimento de um formulário online através de um computador ou celular [Coussement and Poel 2008]. Ter um sistema que classifique automaticamente e direcione essas reclamações é de grande relevância para agilidade do processo,para eliminação de erros humanos e também pode ser um diferencial estratégico para empresas que querem fidelizar seus clientes.

3.2 Interpretação dos dados

A coleta dos dados foi realizada a através do arquivo disponibilizado pelo *Consumer Financial Protection Bureau*[2], essa base compreende todas as reclamações recebidas desde o ano de 2011 até o final de 2021. Após obtenção dos dados, foi utilizada a biblioteca *Pandas* [team 2020] para análise inicial, foi verificado que o dataset tem reclamações de diversos produtos como podemos ver no Quadro 2. Para o desenvolvimento deste projeto foram utilizadas as reclamações referente ao produto cartão de crédito.

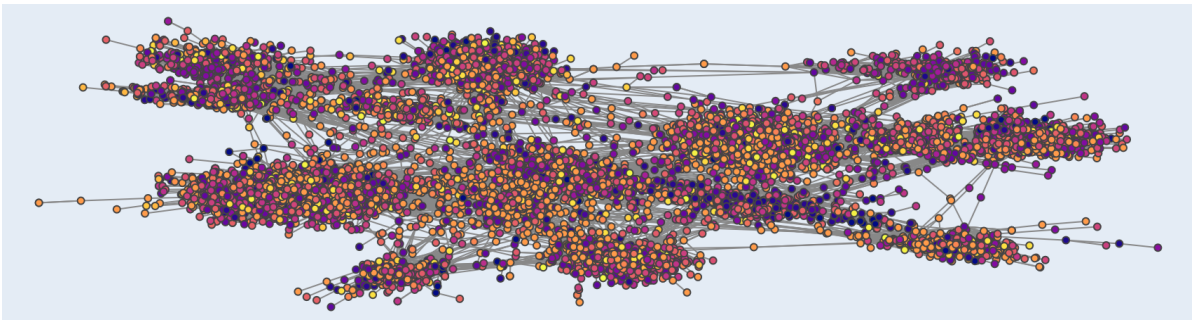
[2] https://www.consumerfinance.gov/data-research/consumer-complaints/

Quadro 2 – Reclamações por produto

Product	Quantity
Credit reporting, credit repair services, or other personal consumer reports	870908
Debt collection	390254
Mortgage	340776
Credit reporting	140431
Credit card or prepaid card	**125886**
Checking or savings account	104463
Credit card	**89190**
Bank account or service	86206
Student loan	63707
Money transfer, virtual currency, or money service	34147
Consumer Loan	31600
Vehicle loan or lease	28709
Payday loan, title loan, or personal loan	19416
Payday loan	5543
Money transfers	5354
Prepaid card	3819
Other financial service	1058
Virtual currency	18

Para o produto selecionado, inicialmente existem 51 classes distintas na base. Como análise preliminar do agrupamento dessas classes, foi gerada uma representação com a técnica *Bag of Words* e posteriormente aplicado o algoritmo *KNN-G* [Eppstein, Paterson and Yao 1997], utilizando a distância do cosseno (equação 3.1) como métrica para identificar as similaridades. Para representação gráfica da técnica aplicada, foi gerado o grafo de k-vizinhança [Silva and Zhao 2016] da figura 9 com as 15 classes com maior ocorrência. Para criação das arestas foram utilizados os 3 vizinhos mais próximos considerando a distância do cosseno. As cores dos nós representam as classes e a proximidade indicam a similaridade. Pode-se notar que não há uma separação visível das classes e que será necessária uma boa representação dos dados para aplicação de algoritmos de aprendizado de máquina.

Figura 9 – Grafo 15 das classes mais representativas



.

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}} \tag{3.1}$$

3.3 Preparação dos dados

Após a seleção e obtenção da base de dados, é necessário verificar se eles estão consistentes o suficiente para aplicarmos os algoritmos de aprendizado de máquina, remover impurezas e gerar representações adequadas são etapas vitais para criação de modelos que consigam minerar padrões que gerem classificações precisas.

Essa etapa foi dividida em três fases: limpeza e seleção dos dados, agrupamento e balanceamento das classes e geração de representações para os textos.

Na etapa de limpeza e seleção dos dados, o principal objetivo é a identificação e remoção de ruídos que são inconsistências nas características utilizadas pelos modelos na realização da classificação, na base utilizada, o ruído identificado foram reclamações com valores nulos que foram removidas. Atributos que não serão utilizados também foram removidos restando apenas dois, o texto da reclamação e a sua classificação original. No quadro 3 podemos ver algumas linhas do dataset após a limpeza inicial.

Quadro 3 – Dataset após limpeza inicial

Consumer complaint narrative	**Issue**
I have a bank of America small business credit card...	Fees or interest
I opened a citi double cash card the beginning...	Getting a credit card
XXXX Chase offered a card with free points...	Fees or interest

Na etapa de agrupamento e balanceamento das classes, foi reduzida a quantidade excessiva de classes presentes no dataset original com o agrupamento de classes com descrição similar e também a remoção de categorias com frequência menor que 500 e do tipo *outras*, no apêndice A é possível ver as classes pré e pós processamento. Também para o equilíbrio do número de ocorrências entre as classes remanescentes, foram aplicadas as técnicas de *subamostragem*, em que categorias com um alto número de registro tem algumas linhas removidas aleatoriamente e *superamostragem* onde categorias com um baixo volume de dados são aumentadas criando novas linhas com variações dos textos já existentes, para essa técnica foi utilizada a biblioteca *textaugment* [Marivate and Sefara 2020] que a partir de um texto gera variações trocando palavras por sinônimos da base *WordNet* [Fellbaum 1998]. É importante ressaltar que esse equilíbrio é feito somente na base de treinamento, deixando a base de validação e teste sem alterações.

Na fase de geração das representações, foram trocadas as representações textuais das classes por numéricas e para aplicação do algoritmo *Naive Bayes* a representação dos

textos das reclamações foi feita através de uma *Bag of Words* com 1,2 e 3 *gramas* e a frequência dos termos sendo calculada pela técnica TF/IDF.

3.4 Modelagem

Essa etapa consiste na aplicação dos algoritmos de AM na geração do modelo para executar a tarefa desejada. Para esse experimento foram selecionados dois modelos distintos para comparação:

- Naive Bayes - O algoritmo foi utilizado com a representação de *Bag of Words* e *TF/IDF* geradas na etapa de pré processamento, conforme descrito no Capítulo 2. A técnica consiste na aplicação do teorema de Bayes para realização das predições. O termo naive vem do fato do algoritmo considerar que, dada a classe, os atributos selecionados são independentes entre si.

- BERT - Será utilizado o modelo *bert-base-uncased* [Devlin *et al.* 2018] disponibilizado no site do *Hugging Face*[3]. O modelo pré treinado na língua inglesa utilizando a técnica *MLM* que foi descrita no Capítulo 2.

Como existe um alto custo computacional envolvendo o treinamento dos modelos, utilizaremos a técnica *Holdout* na aplicação dos dois algoritmos descritos acima. Essa técnica consiste em dividir a base de dados em 3 partes, uma exclusivamente para treino, outra para validação e outra para teste.

Os experimentos serão construídos na linguagem Python com o auxilio das bibliotecas *Sklearn* [Pedregosa *et al.* 2011] para geração do modelo com o algoritmo *Naive Bayes* e *Flair* [Akbik *et al.* 2019] para geração do modelo com o algoritmo BERT. Para execução dos códigos, foi utilizada a plataforma Google Colaboratory com uma licença PRO que forneceu as seguintes configurações de máquina:

- **Processador**: Intel(R) Xeon(R) CPU @ 2.20GHz
- **Memória**: 25GB
- **GPU**: Tesla P100-PCIE-16GB

3.4.1 Treinamento

Após dez épocas de treinamento com o modelo *BERT* e uma rodada de treinamento com o modelo *Naive Bayes* podemos concluir que o primeiro modelo tem um tempo muito maior quando comparado com o segundo:

[3] https://huggingface.co/

- **Tempo de Treinamento Algoritmo *Naive Bayes***: 1 Segundo
- **Tempo de Treinamento Algoritmo *BERT***: 620 Minutos

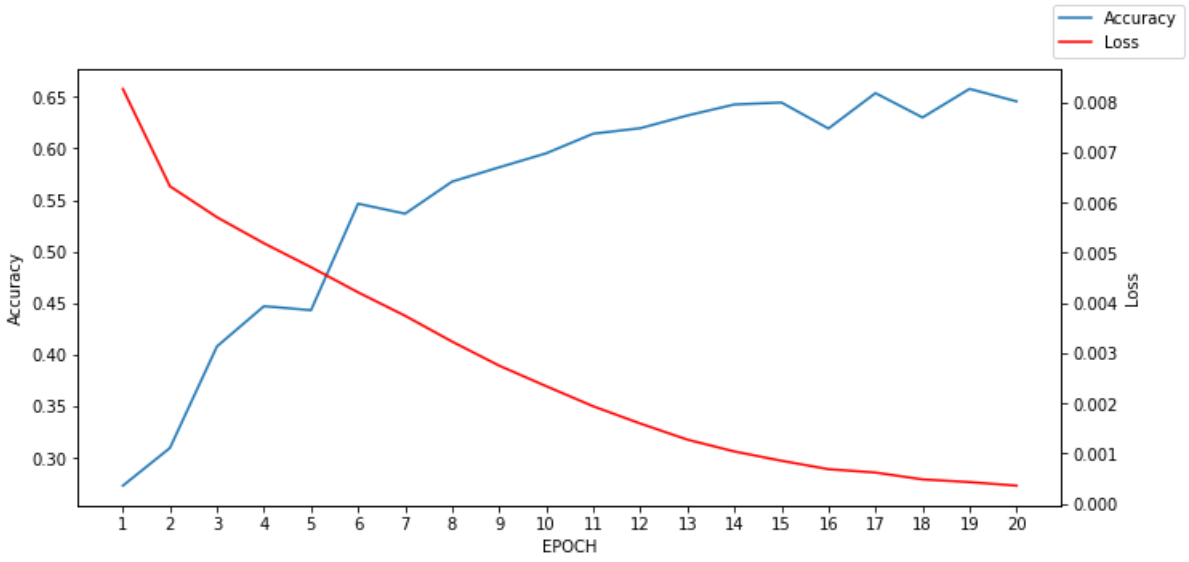
No próximo capitulo serão comentadas as métricas utilizadas para avaliação, os resultados obtidos em cada modelo e qual foi o indicado para implementação da solução proposta.

4 RESULTADOS

4.1 Bert

Após execução de cada época no treinamento do modelo BERT, foram medidos os valores da função de perda e a acurácia para os dados de validação. Na Figura 10 podemos ver esses dados coletados ao longo das épocas, em que podemos averiguar que o melhor valor foi atingido na época 19, com uma acurácia média de 65,76%.

Figura 10 – Resultados de Validação Treinamento BERT



.

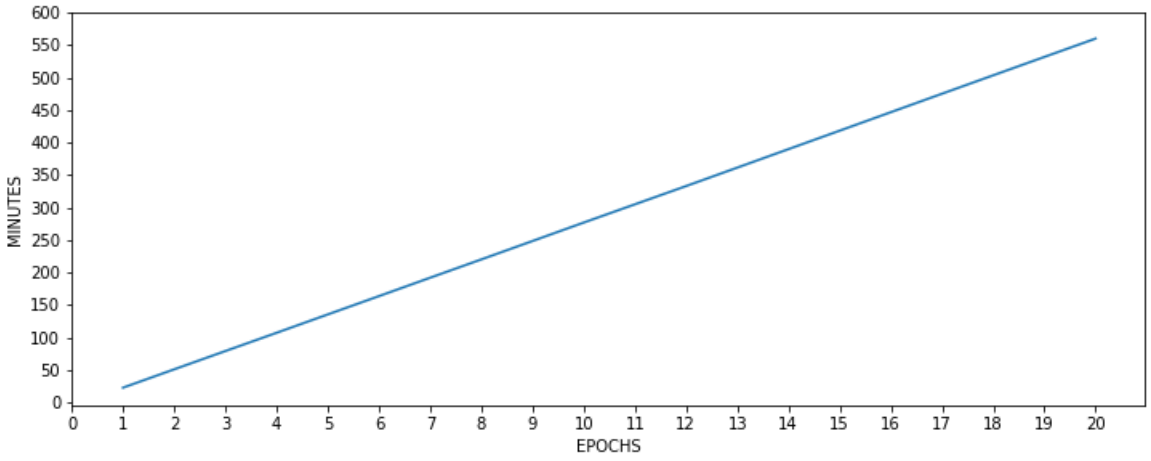
Após a validação do treinamento, foi utilizada a base de teste para obtenção das métricas de avaliação final. No Quadro 4 estão listadas a Precisão, Revocação e F1-Score obtidas através do modelo final para cada classe.

O treinamento durante 20 épocas do modelo BERT durou aproximadamente 9 horas conforme mostrado na Figura 11. Essa figura ilustra a evolução do tempo ao longo das épocas.

Quadro 4 – Resultado por classes do modelo BERT

Classe	Precisão	Revocação	F1-Score
0	0,44	1,00	0,61
1	0,71	0,45	0,55
2	0,56	1,00	0,72
3	0,66	1,00	0,80
4	0,52	1,00	0,69
5	0,51	1,00	0,68
6	0,78	0,99	0,87
7	0,81	0,56	0,66
8	0,79	0,62	0,69
9	0,60	1,00	0,75
10	0,47	1,00	0,64
11	0,66	0,93	0,77
12	0,79	0,64	0,71
13	0,54	0,86	0,67
14	0,93	0,60	0,73
15	0,50	1,00	0,67
16	0,60	0,95	0,74
17	0,57	1,00	0,72
18	0,86	0,52	0,65
19	0,87	0,57	0,70
20	0,73	0,65	0,69
21	0,43	0,83	0,57
22	0,66	0,73	0,70
Média Global	0,65	0,81	0,68

Figura 11 – Tempo de treinamento ao longo das épocas do modelo BERT



.

4.2 Naive Bayes

Para o modelo Naive Bayes, o tempo de treinamento foi praticamente instantâneo, como mostrado na Figura 12.

Figura 12 – Tempo de treinamento ao longo das épocas do modelo Naive

```
[ ] from sklearn.naive_bayes import MultinomialNB
    from time import time

    t = time()

    naive_bayes_classifier = MultinomialNB()
    naive_bayes_classifier.fit(x_train, y_train)

    training_time = time() - t
    print("train time: %0.3fs" % training_time)

    train time: 0.876s
```

.

No Quadro 4.2 podemos ver as métricas de precisão, revocação e f1-score na base de teste.

Classe	Precisão	Revocação	F1-Score
0	0,27	0,95	0,42
1	0,83	0,23	0,36
2	0,39	0,86	0,54
3	0,45	0,77	0,57
4	0,22	0,95	0,36
5	0,27	0,68	0,38
6	0,33	0,88	0,49
7	0,69	0,34	0,46
8	0,82	0,42	0,55
9	0,16	0,94	0,28
10	0,28	0,91	0,43
11	0,50	0,81	0,62
12	0,78	0,41	0,54
13	0,35	0,93	0,51
14	0,83	0,69	0,75
15	0,31	1,00	0,48
16	0,61	0,87	0,72
17	0,47	0,76	0,58
18	0,80	0,36	0,50
19	0,79	0,66	0,72
20	0,81	0,44	0,57
21	0,35	0,68	0,46
22	0,71	0,41	0,52
Média Global	0,52	0,69	0,61

4.3 Considerações finais e dificuldades e limitações

Entre as limitações encontradas é importante citar a presença de ruído nas bases, na qual diversos textos tinham censura devido à presença de dados pessoais e também

caracteres desconhecidos. Foi notado também que os textos classificados não tinham um padrão fixo, em que muitas vezes eram classificadas de maneira diferente o que pode indicar que essa tarefa foi feita por diversas instituições com métodos próprios.

Conforme podemos ver nas seções anteriores o modelo BERT obteve uma assertividade maior na classificação da base de teste. Mesmo com métricas melhores, acreditamos que ele ainda pode ser melhorado com técnicas de otimização de hiper-parâmetros e uma abordagem de pré-processamento diferente do que foi apresentado neste trabalho. Para implementação em produção do modelo desenvolvido é importante que esses resultados sejam ligeiramente melhores para que se evite o retrabalho de reclassificação manual, para que o projeto fique viável financeiramente.

5 CONCLUSÃO

O presente trabalho teve como motivação a implementação de uma solução autônoma, a partir de aprendizado de máquina, para classificação de reclamações de clientes do sistema financeiro. A gestão dessas reclamações pode ser um diferencial competitivo para as empresas que pretendem aumentar sua carteira , retendo possíveis saída de clientes insatisfeitos com o tratamento de suas reivindicações.

Inicialmente, foi realizada uma análise e tratamento do conjunto de dados disponibilizado no site do *Consumer Financial Protection Bureau* (CFPB) [Consumer Financial Protection Bureau 2021]. Após essa etapa, foi criado um grafo em conjunto com o algoritmo K-NNG [Eppstein, Paterson and Yao 1997] para criação das arestas, o intuito era analisar o agrupamento das classes sem um pré processamento. Posteriormente foram utilizadas duas técnicas de aprendizado supervisionados para implementação do classificador: *BERT* e *Naive Bayes*. O treinamento do modelo *BERT* por 20 épocas durou aproximadamente 9 horas, em que o melhor modelo foi obtido na época 19. A técnica de validação utilizada foi a *holdout*. Durante o treinamento, os dados eram validados com a base de validação e na etapa final foi utilizada a base de testes para obtenção das métricas de avaliação final. O treinamento do modelo *Naive Bayes* foi feito em uma única vez sem otimização de parâmetros. No final, as métricas foram obtidas a partir da base de teste.

Dentre as técnicas aplicadas, o *BERT* foi o que obteve os melhores resultados em termos das métricas investigadas. É importante verificar antes da implementação do modelo se os custos com o retratamento de reclamações classificadas de maneira errada e o aumento no tempo para finalização dessas reinvocações não irão inviabilizar o projeto com um aumento nos custos operacionais e na quantidade de disputas abertas pelos clientes.

Como trabalhos futuros, podem ser utilizados os demais atributos do conjunto de dados para auxilio na classificação das reclamações. A utilização de outras técnicas de pré processamento e também de algoritmos supervisionados pode resultar em métricas de avaliação melhores, deixando o modelo proposto mais confiável para utilização em produção. Por fim, pode-se também com a mesma base tentar desenvolver uma classificação de propensão de clientes que tendem a abrir disputas após o tratamento da reivindicação.

REFERÊNCIAS

AGGARWAL, C. C. **Machine Learning for Text**. 1st. ed. [*S.l.: s.n.*]: Springer Publishing Company, Incorporated, 2018. ISBN 3319735306.

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

AKBIK, A.; BERGMANN, T.; BLYTHE, D.; RASUL, K.; SCHWETER, S.; VOLLGRAF, R. FLAIR: An easy-to-use framework for state-of-the-art NLP. *In*: **NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**. [*S.l.: s.n.*], 2019. p. 54–59.

ÁLVAREZ, L. S.; CASIELLES, R. V.; MARTÍN, A. M. D. Analysis of the role of complaint management in the context of relationship marketing. **Journal of Marketing Management**, Taylor & Francis, v. 27, n. 1-2, p. 143–164, 2010.

ALVES, A. I. M. **Modelo de representação de texto mais adequado à classificação**. 2010. Tese (Doutorado) — Instituto Politécnico do Porto. Instituto Superior de Engenharia do Porto, 2010.

BANCO Central do Brasil. 2021. https://www.bcb.gov.br/acessoinformacao/registrar_reclamacao. Accessed: 2021-09-26.

BUREAU OF CONSUMER FINANCIAL PROTECTION. **2020 Consumer Response annual report**. [*S.l.*], 2021.

CONSUMER Financial Protection Bureau. 2021. https://https://www.consumerfinance.gov/. Accessed: 2021-09-26.

COUSSEMENT, K.; POEL, D. V. den. Improving customer complaint management by automatic email classification using linguistic style features as predictors. **Decision Support Systems**, v. 44, n. 4, p. 870–882, 2008. ISSN 0167-9236. Available at: https://www.sciencedirect.com/science/article/pii/S0167923607001820.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, abs/1810.04805, 2018. Available at: http://arxiv.org/abs/1810.04805.

DEVLIN, J.; CHANG, M.-W. Open sourcing bert: state-of-the-art pre-training for natural language processing. **Google AI Blog**, v. 2, 2018.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

EPPSTEIN, D.; PATERSON, M. S.; YAO, F. F. On nearest-neighbor graphs. **Discrete & Computational Geometry**, Springer, v. 17, n. 3, p. 263–282, 1997.

FALCÃO, H. S.; LOVATO, A. V.; SANTOS, A.; OLIVEIRA, L.; MANIÇOBA, R. H. C.; GUIMARÃES, M. A.; SANTANA, M. S. Classificação de vagas de estacionamento com utilização de rede perceptron multicamadas. **Revista de Sistemas de Informação da FSMA, Visconde de Araújo**, v. 12, p. 41–48, 2013.

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. [*S.l.: s.n.*]: Bradford Books, 1998.

FORNELL, C.; WERNERFELT, B. A model for customer complaint management. **Marketing Science**, INFORMS, v. 7, n. 3, p. 287–298, 1988.

HALVORSEN, J.; MøKKELGåRD, P. **Financial Literacy and Other Determinants of Consumer Complaint Behavior**. 2018. Dissertação (Mestrado) — NORWEGIAN SCHOOL OF ECONOMICS, Bergen NO, 2018.

JOHNSTON, R.; MEHRA, S. Best-practice complaint management. **The Academy of Management Executive (1993-2005)**, JSTOR, p. 145–154, 2002.

JOUNG, J.; JUNG, K.; KO, S.; KIM, K. Customer complaints analysis using text mining and outcome-driven innovation method for market-oriented product development. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 11, n. 1, p. 40, 2019.

KENTON, J. D. M.-W. C.; TOUTANOVA, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. *In*: **Proceedings of NAACL-HLT**. [*S.l.: s.n.*], 2019. p. 4171–4186.

KHEDKAR, S.; SHINDE, S. Deep learning and ensemble approach for praise or complaint classification. **Procedia Computer Science**, Elsevier, v. 167, p. 449–458, 2020.

KRISHNA, G. J.; RAVI, V.; REDDY, B. V.; ZAHEERUDDIN, M.; JAISWAL, H.; TEJA, P. S. R.; GAVVAL, R. Sentiment classification of indian banks' customer complaints. *In*: IEEE. **TENCON 2019-2019 IEEE Region 10 Conference (TENCON)**. [*S.l.: s.n.*], 2019. p. 429–434.

LEEM, B.-H.; EUM, S.-W. Using text mining to measure mobile banking service quality. **Industrial Management & Data Systems**, Emerald Publishing Limited, 2021.

MARIVATE, V.; SEFARA, T. Improving short text classification through global augmentation methods. *In*: SPRINGER. **International Cross-Domain Conference for Machine Learning and Knowledge Extraction**. [*S.l.: s.n.*], 2020. p. 385–399.

MARTINS, C. A. **Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. 2003. Tese (Doutorado) — Universidade de São Paulo, 2003.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. **Techinical Report**, v. 209, n. 4, p. 10–11, 2003.

MCCALLUM, A.; NIGAM, K. *et al.* A comparison of event models for naive bayes text classification. *In*: CITESEER. **AAAI-98 workshop on learning for text categorization**. [*S.l.: s.n.*], 1998. v. 752, n. 1, p. 41–48.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

OZBAY, F. A.; ALATAS, B. Fake news detection within online social media using supervised artificial intelligence algorithms. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 540, p. 123174, 2020.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEIXOTO, L. H. R. **Aprendizado de Máquina Aplicado no Atendimento de Reclamações de Clientes**. 2021. 73 p. Dissertação (Mestrado) — Universidade de São Paulo, São Paulo, 2021.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. [*S.l.: s.n.*]: Editora Manole Ltda, 2003.

SCIKIT Learn - Naive Bayes. 2022. https://scikit-learn.org/stable/modules/naive_bayes.html. Accessed: 2022-01-09.

SILVA, T. C.; ZHAO, L. **Machine Learning in Complex Networks**. Springer, 2016. ISBN 978-3-319-17289-7. Available at: https://doi.org/10.1007/978-3-319-17290-3.

TEAM, T. pandas development. **pandas-dev/pandas: Pandas**. Zenodo, 2020. Available at: https://doi.org/10.5281/zenodo.3509134.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *In*: **Advances in neural information processing systems**. [*S.l.: s.n.*], 2017. p. 5998–6008.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. *In*: MANCHESTER. **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. [*S.l.: s.n.*], 2000. v. 1, p. 29–39.

YANG, W.; TAN, L.; LU, C.; CUI, A.; LI, H.; CHEN, X.; XIONG, K.; WANG, M.; LI, M.; PEI, J. *et al.* Detecting customer complaint escalation with recurrent neural networks and manually-engineered features. *In*: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)**. [*S.l.: s.n.*], 2019. p. 56–63.

YANG, Y.; XU, D.-L.; YANG, J.-B.; CHEN, Y.-W. An evidential reasoning-based decision support system for handling customer complaints in mobile telecommunications. **Knowledge-Based Systems**, Elsevier, v. 162, p. 202–210, 2018.

APÊNDICES

APÊNDICE A – CLASSES

Classe Pré Processamento	Frequência	Classes Pós Processamento
Problem With A Purchase Shown On...	14791	Problem With A Purchase Shown On...
Other Features, Terms, Or Problems	7702	Removida
Fees Or Interest	7223	Fees Or Interest
Problem When Making Payments	5691	Problem When Making Payments
Getting A Credit Card	4760	Getting A Credit Card
Closing Your Account	4559	Closing Account (Agrupada)
Advertising And Marketing...	3819	Advertising(Agrupada)
Billing Disputes	3102	Billing Disputes
Trouble Using Your Card	2662	Trouble Using Card (Agrupada)
Other	1940	Removida
Identity Theft / Fraud / Embezzlement	1723	Identity Theft / Fraud / Embezzlement
Problem With A Purchase Or Transfer	1705	Problem With A Purchase Or Transfer
Problem With A Credit Reporting	1472	Credit Report (Agrupada)
Trouble Using The Card	1465	Trouble Using Card (Agrupada)
Incorrect Information On Your Report	1462	Credit Report (Agrupada)
Closing/Cancelling Account	1440	Closing Account (Agrupada)
Problem Getting A Card Or Closing...	1215	Problem Getting A Card Or Closing...
Struggling To Pay Your Bill	1168	Struggling To Pay Your Bill
Unexpected Or Other Fees	1080	Removida
Customer Service / Customer Relations	973	Customer Service / Customer Relations
Rewards	900	Rewards
Delinquent Account	834	Delinquent Account
Advertising And Marketing	818	Advertising(Agrupada)
APR Or Interest Rate	785	APR Or Interest Rate
Late Fee	771	Late Fee
Credit Card Protection / Debt Protection	712	Credit Card Protection / Debt Protection

Transaction Issue	681	Transaction Issue
Billing Statement	620	Billing Statement
Payoff Process	513	Payoff Process
Credit Determination	506	Credit Determination
Other Fee	502	Removida
Unsolicited Issuance Of Credit Card	491	Removida
Credit Line Increase/decrease	451	Removida
Balance Transfer	298	Removida
Improper Use Of Your Report	233	Removida
Advertising	160	Removida
Application Processing Delay	107	Removida
Privacy	105	Removida
Sale Of Account	98	Removida
Bankruptcy	96	Removida
Credit Monitoring Or...	88	Removida
Forbearance / Workout Plans	85	Removida
Arbitration	83	Removida
Balance Transfer Fee	51	Removida
Cash Advance	44	Removida
Cash Advance Fee	41	Removida
Convenience Checks	37	Removida
Problem With Overdraft	32	Removida
Overlimit Fee	31	Removida
Unable To Get Your Credit...	29	Removida
Problem With Fraud Alerts Or Security Freezes	20	Removida
Problem With An Overdraft	3	Removida