Uma metodologia em cascata de quatro etapas para classificar códigos NCM usando técnicas de PLN

Pedro Pinheiro, Marcos Amaris

[1]Universidade Federal do Pará
Faculdade de Engenharia de Computação, Tucuruí - Pará

pedrobraga85@gmail.com, amaris@ufpa.br

Resumo. *Esse trabalho tem como objetivo desenvolver um processo para classificar as descrições dos produtos presentes nas Notas Ficais eletrônicas (NF-e). Essa classificação é feita sobre os 8 dígitos da Nomenclatura Comum do Mercosul (NCM), separado em 4 partes, Capítulo, Posição, Subposição e item/Subitem. A classificação foi realizada utilizando o algoritmo de Máquina de Vetores de Suporte (SVM) e o algoritmo de Naïve Bayes em conjunto com as técnicas de Processamento Natural de Linguagem (PNL), para o processamento de uma base de dados de 340.000 produtos distintos. Os dados foram divididos em 80% treinamento e 20% teste e Obteve-se um acurácia de 90% para um total de 98 classes.*
Palavras–chave*: Processamento de Linguagem Natural; Aprendizagem de máquina; Classificação de Texto; Nomenclatura Comum do Mercosul;*

Abstract. *This work aims to develop a process to classify the descriptions of products present in electronic invoices (NF-e). This classification is based on the 8 digits of the Common Mercosur Nomenclature (NCM), separated into 4 parts, Chapter, Position, Subheading and item/Sub-item. The classification was performed using the Support Vector Machine (SVM) algorithm and the Naïve Bayess algorithm together with Natural Language Processing (NLP) techniques, for processing a database of 340,000 different products. The data were divided into 80% training and 20% testing and an accuracy of 90% was obtained for a total of 98 classes.*
Keywords*: Natural Processing Language, Machine Learning, Text Classification and Mercosul Common Nomenclature.*

1. INTRODUÇÃO

Com a globalização, aumentou-se em larga escala a importação e exportação de mercadorias, diante disso, surgiu-se um grande obstáculo, que desafia tanto as alfandegas quanto aos lojistas, que é categorização das descrições das mercadorias. Em 1988 a Organização Mundial das Alfândegas (WCO) criou a Nomenclatura chamada de código do Sistema Harmonizado (HS) que serviu de base para a criação da Nomenclatura Comum do Mercosul (NCM) que é um sistema ordenado que permite, pela aplicação de regras e procedimentos próprios, determinar um único código numérico para uma dada mercadoria. O NCM é uma Nomenclatura regional para categorização de mercadorias adotada pelo Brasil, Argentina, Paraguai e Uruguai desde 1995, sendo utilizada em todas as operações de comércio exterior dos países do Mercosul [Roberto Scalco et al. 2015].

Cada descrição de produto está relacionado a um especifico código NCM. Os estabelecimento comerciais ou aduaneiro tem a responsabilidade de informar corretamente o NCM de cada produto. Porém, esse processo é suscetível ao erro humano devido à grande diversidade dos produtos e falta de conhecimento da Nomenclatura ou a interpretação errada das regras do NCM.

Na cidade de Paraíba aconteceu um caso [do Bomfim et al. 2020] sobre a má informação desses códigos em estabelecimentos comerciais, seja por um equívoco ou desonestidade do contribuinte. O Estado é afetado diretamente, deixando de arrecadar o imposto sobre a mercadoria ou arrecadando de uma mercadoria que deveria ser isenta de imposto, assim prejudicando o próprio comerciante, devido a que cada código NCM tem um valor de alíquota atrelado, PIS/COFINS, ICMS e TIPI correspondente ao impostos que incidirá sobre a venda do produto.

A criação da Nota Fiscal Eletrônica em 2004 tem como objetivo a implantação de um modelo nacional de documento fiscal eletrônico, que venha substituir a sistemática atual de emissão do documento fiscal em papel, proporcionando validade jurídica garantida pela assinatura digital do remetente, diminuição da sonegação, aumento da arrecadação e crescimento na confiabilidade da Nota Fiscal [SEFAZ 2021]. Com isso fica a cargo da Secretaria do Estado da Fazenda a realização da verificação dos códigos de NCM informados nas Notas fiscais, segundo [Ding et al. 2015] mostra que cerca de 30% das mercadorias enviadas globalmente estão com os códigos errados, tornando-se uma busca global por solução que tragam um melhor desempenho na rotulação dos código NCM dos produtos.

É expressivo a evolução nos últimos anos do aprendizado de maquina juntamente com as técnicas de Processamento de Linguagem Natural (PNL), que é uma área da Inteligência Artificial que estuda a geração e compreensão automática de linguagens humanas naturais tanto em texto, quanto em voz. Essa tecnologia é utilizada para desenvolver aplicações como traduções entre idiomas, chatbots, sumarização de textos, análise de sentimentos e muitas outras [Sebastiani 2002], dentro desse universo exitem trabalhos nas áreas de classificação de texto curto [Wang et al. 2017] e classificação de produtos [Yu et al. 2012].

Através da grande quantidade de dados sobre os produtos que são gerados com as NF-e, onde foi realizado uma rotina de extração de dados das Notas fiscais e inserção no banco de dados. O *dataset* ficou com aproximadamente 340.000 descrições textuais distintas. Com esse expressivo montante de informações e as técnicas de aprendizado de máquina em conjunto com o PNL, é possível a criação de modelos preditivos para auxiliar a classificação do código NCM.

Portanto, a finalidade dessa tese é criar e avaliar dois modelos de aprendizado de maquina distintos para o código NCM através da descrição dos produtos presentes nas NF-e, usando como base a grande quantidade de dados que são gerados através das Notas fiscais eletrônicas e auxiliado pelas técnicas de processamento natural de linguagem. Dessa forma foram usados 2 algorítimos de aprendizado supervisionado com o intuito de encontrar o resultado ideal embasado nos dados. Os resultados evidenciaram que o algoritmo LinarSVC tem um melhor desempenho para esta análise, obtendo uma acurácia de 90%.

Este trabalho está estruturado como: A Seção 2 descreve conceitos fundamentais para o entendimento desse trabalho. A Seção 3 demonstra alguns trabalhos relacionados. A Seção 4 apresenta a metodologia usada nessa pesquisa, depois os resultados são apresentados na Seção 5 e por fim as conclusões na Seção 6.

2. CONCEITOS E BACKGROUND TEÓRICO

2.1. Nota Fiscal Eletrônica

A [Brasil 2003] prevê obrigações acessórias, nomeadamente, a transmissão das informações fiscais e econômicas dos contribuintes às autoridades fiscais para efeitos de fiscalização, bem como o registo fiscal e os dados fiscais. Os três ramos do governo agindo como um. A troca de informações fiscais proporcionada por essa política ajudará cada Secretaria da Fazenda (Sefaz) a combater a sonegação fiscal e reduzir a inadimplência por meio de mecanismos de cruzamento de dados e verificação eletrônica. Isso facilita a identificação de contribuintes em situações irregulares [Sousa 2010].

A implementação da Nota Fiscal Eletrônica (NF-e) pelo Sistema Público de Escrituração Digital (SPED) visa simplificar as obrigações de garantia e economizar no armazenamento de documentos de papel, ao mesmo tempo que contribui para o combate à evasão fiscais. O SPED cria um ambiente no qual o Tesouro nacional e as autoridades fiscais federais podem combinar informações contábeis e fiscais, identificar fraudes e evasões fiscais e abranger toda a cadeia produtiva. Também define novos processos de controle e gestão, confiabilidade das informações, sincronização de registros, consistência e integração entre os sistemas corporativos e fiscais [Bonfim et al. 2012].

DANFe é um acrônimo para ”Documento Auxiliar de Nota Fiscal Eletrônica”. Este documento é uma apresentação legível e simplificada da Nota Fiscal. Em poucas palavras, o DANFe é um documento impresso que contém as principais informações de uma Nota Fiscal Eletrônica (NFe). Para a pesquisa foram extraídos da NF-e as seguintes informações, descrição do produto e o código NCM na Figura 1, que será introduzido na próxima seção.

2.2. Nomenclatura Comum do Mercosul

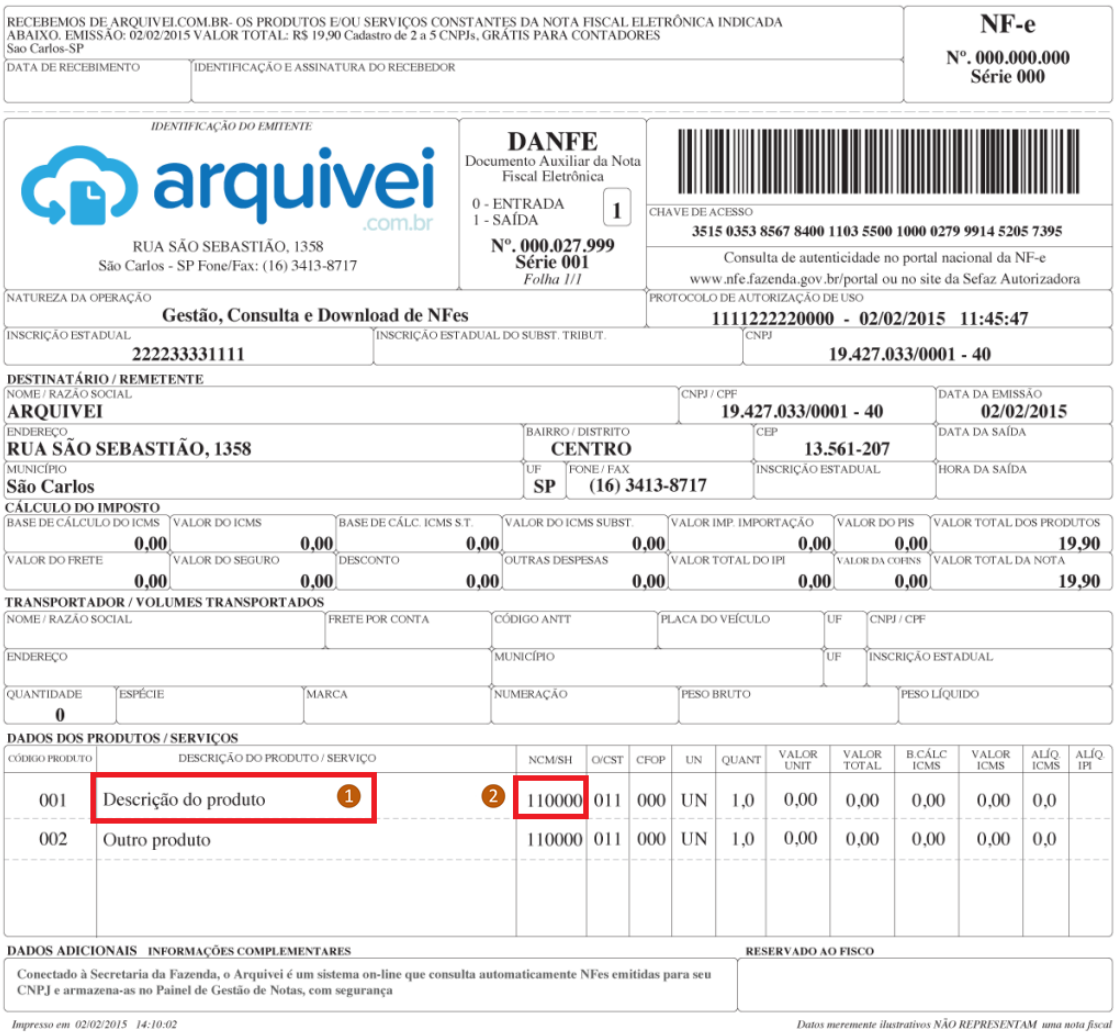
O código NCM, é uma convenção de categorização de mercadorias. Os códigos são compostos por oito dígitos como mostra a Figura 2, os dois primeiros dígitos especificam o capítulo onde correspondem as características de cada produto. Os dois dígitos seguintes se referem a posição na qual demonstra características especificadas no capítulo. O quinto e sexto dígitos definem a subposição de um determinado produto. O sétimo classifica o produto, e por fim, o oitavo dígito descreve especificamente do que se trata o produto, respectivamente como mostra a Tabela 1.

2.3. Processamento de Linguagem Natural

O Processamento de linguagem natural (PLN), tem como peça fundamental o processamento de texto, que é basicamente a conversão de texto puro em uma sequência de números.

Esse processo inicia com a *tokenização*, que é o processo de separação das palavras do texto, quebrando as frases nos espaços em branco e extraindo as palavras,

Figura 1. Nota Fiscal Eletrônica (NF-e)



RECEBEMOS DE ARQUIVEI.COM.BR- OS PRODUTOS E/OU SERVIÇOS CONSTANTES DA NOTA FISCAL ELETRÔNICA INDICADA ABAIXO. EMISSÃO: 02/02/2015 VALOR TOTAL: R$ 19,90 Cadastro de 2 a 5 CNPJs, GRÁTIS PARA CONTADORES
Sao Carlos-SP

DATA DE RECEBIMENTO | IDENTIFICAÇÃO E ASSINATURA DO RECEBEDOR

NF-e
Nº. 000.000.000
Série 000

IDENTIFICAÇÃO DO EMITENTE

arquivei .com.br

RUA SÃO SEBASTIÃO, 1358
São Carlos - SP Fone/Fax: (16) 3413-8717

DANFE
Documento Auxiliar da Nota Fiscal Eletrônica

0 - ENTRADA
1 - SAÍDA
1

Nº. 000.027.999
Série 001
Folha 1/1

CHAVE DE ACESSO
3515 0353 8567 8400 1103 5500 1000 0279 9914 5205 7395

Consulta de autenticidade no portal nacional da NF-e www.nfe.fazenda.gov.br/portal ou no site da Sefaz Autorizadora

NATUREZA DA OPERAÇÃO	PROTOCOLO DE AUTORIZAÇÃO DE USO
Gestão, Consulta e Download de NFes	1111222220000 - 02/02/2015 11:45:47

INSCRIÇÃO ESTADUAL	INSCRIÇÃO ESTADUAL DO SUBST. TRIBUT.	CNPJ
222233331111		19.427.033/0001 - 40

DESTINATÁRIO / REMETENTE

NOME / RAZÃO SOCIAL	CNPJ / CPF	DATA DA EMISSÃO
ARQUIVEI	19.427.033/0001 - 40	02/02/2015

ENDEREÇO	BAIRRO / DISTRITO	CEP	DATA DA SAÍDA
RUA SÃO SEBASTIÃO, 1358	CENTRO	13.561-207	

MUNICÍPIO	UF	FONE / FAX	INSCRIÇÃO ESTADUAL	HORA DA SAÍDA
São Carlos	SP	(16) 3413-8717		

CÁLCULO DO IMPOSTO

BASE DE CÁLCULO DO ICMS	VALOR DO ICMS	BASE DE CÁLC. ICMS S.T.	VALOR DO ICMS SUBST.	VALOR IMP. IMPORTAÇÃO	VALOR DO PIS	VALOR TOTAL DOS PRODUTOS
0,00	0,00	0,00	0,00	0,00	0,00	19,90

VALOR DO FRETE	VALOR DO SEGURO	DESCONTO	OUTRAS DESPESAS	VALOR TOTAL DO IPI	VALOR DA COFINS	VALOR TOTAL DA NOTA
0,00	0,00	0,00	0,00	0,00	0,00	19,90

TRANSPORTADOR / VOLUMES TRANSPORTADOS

NOME / RAZÃO SOCIAL	FRETE POR CONTA	CÓDIGO ANTT	PLACA DO VEÍCULO	UF	CNPJ / CPF

ENDEREÇO	MUNICÍPIO	UF	INSCRIÇÃO ESTADUAL

QUANTIDADE	ESPÉCIE	MARCA	NUMERAÇÃO	PESO BRUTO	PESO LÍQUIDO
0					

DADOS DOS PRODUTOS / SERVIÇOS

CÓDIGO PRODUTO	DESCRIÇÃO DO PRODUTO / SERVIÇO	NCM/SH	O/CST	CFOP	UN	QUANT	VALOR UNIT	VALOR TOTAL	B.CÁLC ICMS	VALOR ICMS	ALÍQ ICMS	ALÍQ IPI
001	Descrição do produto (1)	(2) 110000	011	000	UN	1,0	0,00	0,00	0,00	0,00	0,0	
002	Outro produto	110000	011	000	UN	1,0	0,00	0,00	0,00	0,00	0,0	

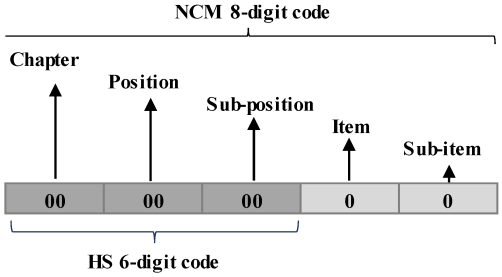
DADOS ADICIONAIS INFORMAÇÕES COMPLEMENTARES | RESERVADO AO FISCO

Conectado à Secretaria da Fazenda, o Arquivei é um sistema on-line que consulta automaticamente NFes emitidas para seu CNPJ e armazena-as no Painel de Gestão de Notas, com segurança

Impresso em 02/02/2015 14:10:02 — *Datos meremente ilustrativos NÃO REPRESENTAM uma nota fiscal*

Fonte: https://arquivei.com.br/o-que-e-danfe-entenda-sua-funcao-e-importancia/

Figura 2. Composição do código NCM (Nomenclatura Comum do Mercosul)



Fonte: https://www.mdpi.com/BDCC/BDCC-06-00008/article_deploy/html/images/BDCC-06-00008-g001.png

Tabela 1. Capítulo 21 na tabela NCM.

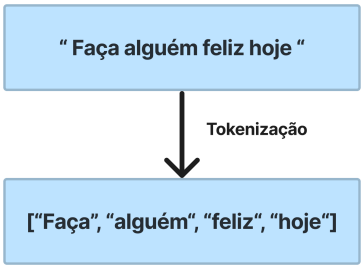
NCM	Descrição
21.01	Extratos, essências e concentrados de café, chá ou mate e preparações à base destes produtos ou à base de café, chá ou mate; chicória torrada e outros sucedâneos torrados do café e respectivos extratos, essências e concentrados.
2101.1	Extratos, essências e concentrados de café e preparações à base destes extratos, essências ou concentrados ou à base de café:
2101.11	Extratos, essências e concentrados.
2101.11.10	Café solúvel, mesmo descafeinado.
2101.11.90	Outros.
2101.12.00	Preparações à base de extratos, essências ou concentrados ou à base de café.
2101.20	Extratos, essências e concentrados de chá ou de mate e preparações à base destes extratos, essências ou concentrados ou à base de chá ou de mate.
2101.20.10	De chá.
2101.20.20	De mate.
2101.30.00	Chicória torrada e outros sucedâneos torrados do café e respectivos extratos, essências e concentrados.

gerando um vetor de tokens. Seguindo para a remoção das palavras vazias ou *Stopwords*, que se refere às palavras mais comuns na língua ou palavras de conexão, que estão presentes na maioria dos documentos e não tem grande significância para frase e acaba posteriormente atrapalhando a classificação. Por fim é feito o processo de *stemming* ou stemização [Orengo and Huyck 2001] que é o nome dado ao processo de extração de raiz da palavra, que a mesma é comum para as diversas variações da palavra.

2.3.1. Tokenização

Tokenização é o mecanismo de fragmentar o texto em palavras, termos ou símbolos, conhecidos como tokens. A lista de tokens se transforma em entrada para processamento, análise ou mineração de texto. Normalmente, o processo de tokenização ocorre no nível da palavra. Mas, às vezes é difícil definir o que se entende por uma ”palavra”. O principal uso da tokenização é identificar as palavras-chaves significativas. A desvantagem da tokenização é a dificuldade de tokenizar o documento sem espaços em branco, caracteres especiais ou outras marcas. Figura 3, mostra um exemplo de Tokenização.

Figura 3. Tokenização



Fonte: Próprio Autor

2.3.2. *Stop-Words*

Nos documentos, as palavras de maior frequência são mais importantes para representar o conteúdo do que as de menor frequência. No entanto, algumas palavras de alta frequência como ”o”, ”para”, ”em”com baixo conteúdo, podem enviesar os resultados, essas palavras estão presentes na lista de *Stop-words*. Portanto, precisamos ignorá-las, sendo excluidas para reduzir a quantidade de dimensões e aumentar a relevância entre palavras e documentos ou categorias, Tabela 2.

Tabela 2. Lista de Stop Words

Stop Words ENG	Stop Words PT-BR
a	uma
about	sobre
above	acima
across	através
after	depois de
again	novamente
against	contra
all	todos
almost	quase
alone	sozinho
along	ao longo
already	já
also	além disso
although	apesar
always	sempre
among	entre
an	a
and	e
another	outro

2.3.3. *Stemming*

A raiz é o elemento que contém o significado fundamental de uma palavra. Por exemplo, as palavras vidro, vidraça, vidraceiro, envidraçar têm a mesma raiz ”vidr”, como mostra a Tabela 3. Essa métrica é usada para reduzir a dimensão do vetor de palavras e aumentar a relevância das mesmas, mantendo somente a raiz.

Tabela 3. Stemming

Palavra	Raiz
Vidro	Vidr
Vidraça	Vidr
Vidraceiro	Vidr
Envidraçar	Vidr
Casa	Cas
Casona	Cas
Casinha	Cas
Casebre	Cas

2.3.4. TF-IDF

TF-IDF (*Term frequency – inverse document frequency*) é um processo estatístico utilizado para medir a relevância das palavras perante o texto, ou seja, esta técnica assume que a importância de uma palavra é diretamente proporcional à sua frequência no texto e inversamente proporcional à sua frequência no corpus, sendo bastante popular na área de processamento de linguagem natural (PLN) [Trstenjak et al. 2014]. O termo de frequência de uma palavra w no documento d, denotado por $TF(w, d)$, é o número de vezes que a palavra w aparece no documento d. Quanto maior, mais representativa é a palavra w do documento d. A frequência de documento de uma palavra w, denotada por $DF(w)$, é o número de documentos em que w ocorre. A frequência inversa do documento de uma palavra w, denotada por $IDF(w)$. Portanto, uma palavra w terá relevância baixa se a palavra aparecer em mais de um documento [Neto et al. 2000].

$$IDF(w) = 1 + log(|D|/DF(w)) \tag{1}$$

Para encontrar as palavras mais bem ranqueadas. Podemos expressar em uma única fórmula TFIDF (w, d):

$$TFIDF(w, d) = TF(w, d) * IDF(w) \tag{2}$$

Após realizar o pré-processamento nas descrições e a transformação para representação vetorial, está pronto para ser a entrada do classificador.

2.4. Aprendizado de Máquina para Classificação de texto

O aprendizado de máquina é um campo da inteligência artificial que visa desenvolver técnicas de computação para aprender e construir sistemas que possam adquirir conhecimento de forma automática. Visando a automatização de procedimentos manuais, um importando ramo de pesquisa é a Classificação de texto, devido ser uma tarefa exaustiva e por consumir muito tempo esse tema tem ganhado notoriedade [Kadhim 2019]. A Classificação de Texto é um processo de aprendizagem supervisionada no qual um corpo de texto, precisa ser atribuído a um conjunto de rótulos de classes. Algumas das aplicações comuns de classificação de texto são as análise de sentimentos, categorização de artigos e detecção de spam e os Algoritmos de *Machine Learning* mais usados são *Naïve Bayes*, *Support vector machine* [Luppes et al. 2019].

Esta seção discute as diferentes técnicas de aprendizado de máquina supervisionado, como classificadores Naïve Bayes (NB), Support vector machine (SVM) que são

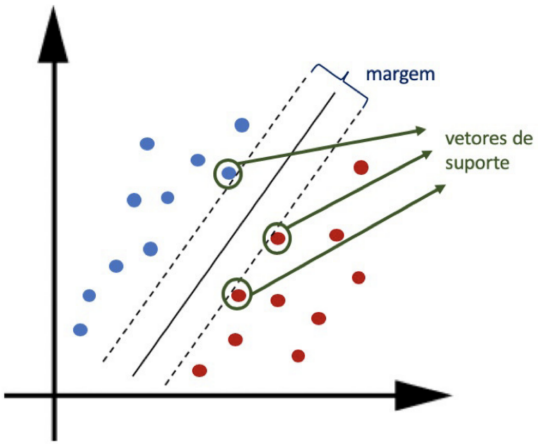
usadas no presente trabalho de pesquisa e analisa o efeito de cada técnica na classificação de texto usando algoritmos de aprendizado de máquina.

2.4.1. Máquina de vetores de suporte

Iniciado os estudos por Vapnik (1995), As Máquinas de Vetores de Suporte (SMV) utilizam métodos de classificação supervisionada baseados em particionamento, cujo objetivo é construir hiperplanos com a melhor separação possível entre as classes. Esta técnica tem sido usada com sucesso em uma série de aplicações de reconhecimento de padrões, tais como: classificação de texto, classificação de SPAM, reconhecimento de caligrafia, análise de texto, expressão gênica, etc.

O funcionamento do SVM é definido da seguinte forma: Dado um conjunto de pontos pertencentes a duas classes, o SVM define uma melhor forma de separar os dados lineares, através do conceito de hiperplanos que visa colocar a maior quantidade de pontos possíveis da mesma classe no mesmo lado, amparado pela margem, que tenta encontrar os vetores de suporte a uma distância máxima do hiperplano de cada camada Figura 4. Quando os dados não são linearmente separáveis o SVM aplica a técnica de *kernel trick* como mostra a Figura 5, mapear os recursos originais em um grande espaço dimensional e, assim, conseguir uma melhor separação entre as classes.

Figura 4. Exemplo de classificador linear, margem e vetores de suporte.



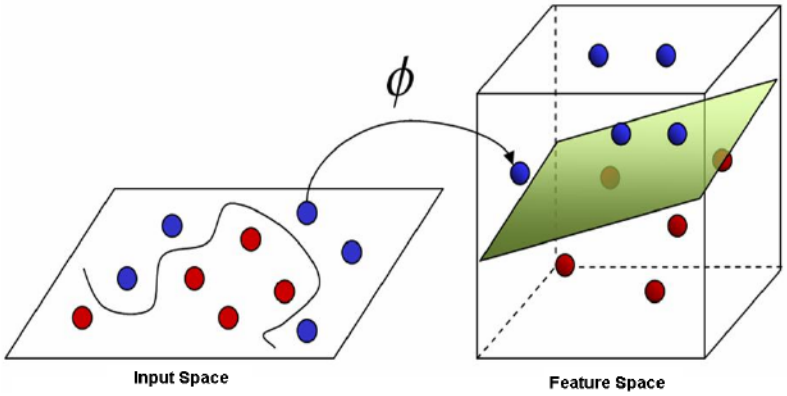
Fonte: [Escovedo and Koshiyama 2020]

2.4.2. Naive Bayes

Conhecido por ser um dos modelos mais populares na literatura de aprendizado de máquina e por aplicar conceitos de probabilidade, o modelo *Naive Bayes* é um conjunto de classificadores probabilísticos de observações baseados na aplicação do teorema de *Bayes* que define que a probabilidade condicional de um evento é a probabilidade obtida pela informação adicional de outro evento que já ocorreu [Andre Dieb Martins 2013].

Apesar de ser um modelo simples apresenta bons resultados e é muito útil para grande volume de dados. Em resumo, o termo *"Naive"* (ingênuo) refere-se à forma como

Figura 5. O classificador SVM não linear com o Kernel trick



Fonte: [Delanerolle et al. 2021]

o algoritmo analisa as características no conjunto de dados, pressupondo que as características sejam independentes umas das outras, ou seja, ele assume que a existência de uma característica particular em uma classe não está relacionada com a existência de qualquer outro recurso[Russell and Norvig 2003].

Para o cálculo da probabilidade de um evento c dado que um evento x ocorreu $P(C/X)$, pelo Teorema de Bayes temos

$$P(c/x) = \frac{P(x/c)P(c)}{P(x)} \tag{3}$$

onde,

- $P(c/x)$ é a probabilidade posterior da classe (c, alvo) dada preditor (x, atributos).
- $P(c)$ é a probabilidade original da classe.
- $P(x/c)$ é a probabilidade que representa a probabilidade de preditor dada a classe.
- $P(x)$ é a probabilidade original do preditor

3. TRABALHOS RELACIONADOS

Existem poucos trabalhos que discutem o problema de predição de códigos HS ou NCM, realizamos uma busca rápida no Google Scholar e IEEExplore e encontramos 79 trabalhos e 1 artigo, respectivamente, utilizando a string de busca a seguir em ambas as bibliotecas eletrônicas (*"supervised learning"AND classification AND ("Harmonized System"OR "Mercosur Common Nomenclature")*). Abaixo, mencionamos os trabalhos que achamos mais relevantes para esta pesquisa.

O Trabalho de [de Abreu Batista et al. 2018] consiste no desenvolvimento de um classificador para a categorização automática de descrições de produtos em seus códigos de NCM, o objetivo é extrair dados da Nota Fiscal Eletrônica ao Consumidor (NFC-e), para realizar um aprendizado supervisionado utilizando o algorítimo de *Naïve Bayes*, os resultados mostraram uma acurácia média de 86.5% para 2 classes.

[Luppes et al. 2019] Propôs uma arquitetura de Rede Neural Convolucional (CNN) para rotular as descrições com base em descrições de texto curtas, utilizaram as técnicas

de *embeddings word* com diversas bases de dados onlines, como o **DBpédia**, obtiveram resultados de 92% para os 2 dígitos inicias, também capitulo do **HS-2**.

[Ding et al. 2015] Usa uma abordagem de espaço vetorial. Seu conjunto de dados foi obtido na Alfândega de Cingapura. Obteve ótimo desempenho em um dos capítulo, com 98% de acurácia, a principal parte das suas conclusões foi que quanto menor é as descrições disponíveis, mais difícil se tornava classificá-las em um capítulo: ao classificar uma descrição com até três tokens, eles só podem atingir uma precisão de apenas 15%.

[Li and Li 2019] Aborda a o problema de forma diferente. Construíram duas CNNs simples, uma para texto e outra para imagens, para classificar as descrições de calçados em quatro classes com base em seis tipos de códigos HS. Eles obtiveram umas acurácia de 93%. Seu conjunto de dados inclui 10.000 imagens dos sapatos junto com uma descrição de texto.

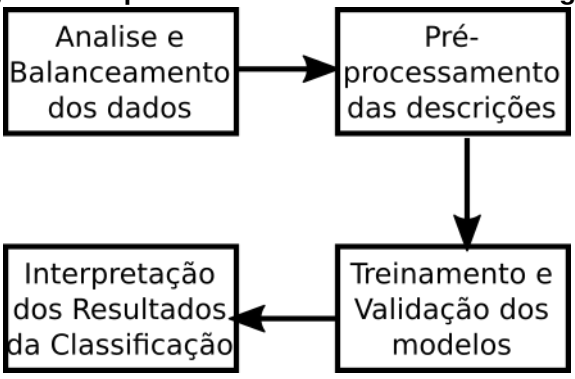
Os trabalhos de [de Abreu Batista et al. 2018] e [**?**] apresentam a classificação completa do NCM, mas em capítulos específicos, já [Luppes et al. 2019] apresenta a classificação dos primeiros 4 dígitos, diferente do presente trabalho que tem uma proposta de classificar 98 capítulos diferentes e o código NCM completo, usando a metodologia de cascata proposta no trabalho de [Granitto et al. 2005] relacionando 4 modelos de classificação usando usando Máquina de vetor de suporte, diferente dos trabalhos de [Ding et al. 2015] e [Li and Li 2019] que utilizam Redes Neurais.

4. METODOLOGIA

O objetivo desta seção é apresentar um método para classificar as descrições textuais dos produtos contidos nas Notas Fiscais Eletrônicas em seus respectivos códigos da Nomenclatura Comum do Mercosul. Este método consiste na configuração dos dados de texto em estruturas de dados lógicas baseadas na frequência dos termos, os dados foram rotulados e organizados de acordo com o NCM presente na NFe. Posteriormente, foi utilizados diferentes algoritmos de classificação baseados em técnicas de aprendizado supervisionado para finalmente avaliar seus desempenhos.

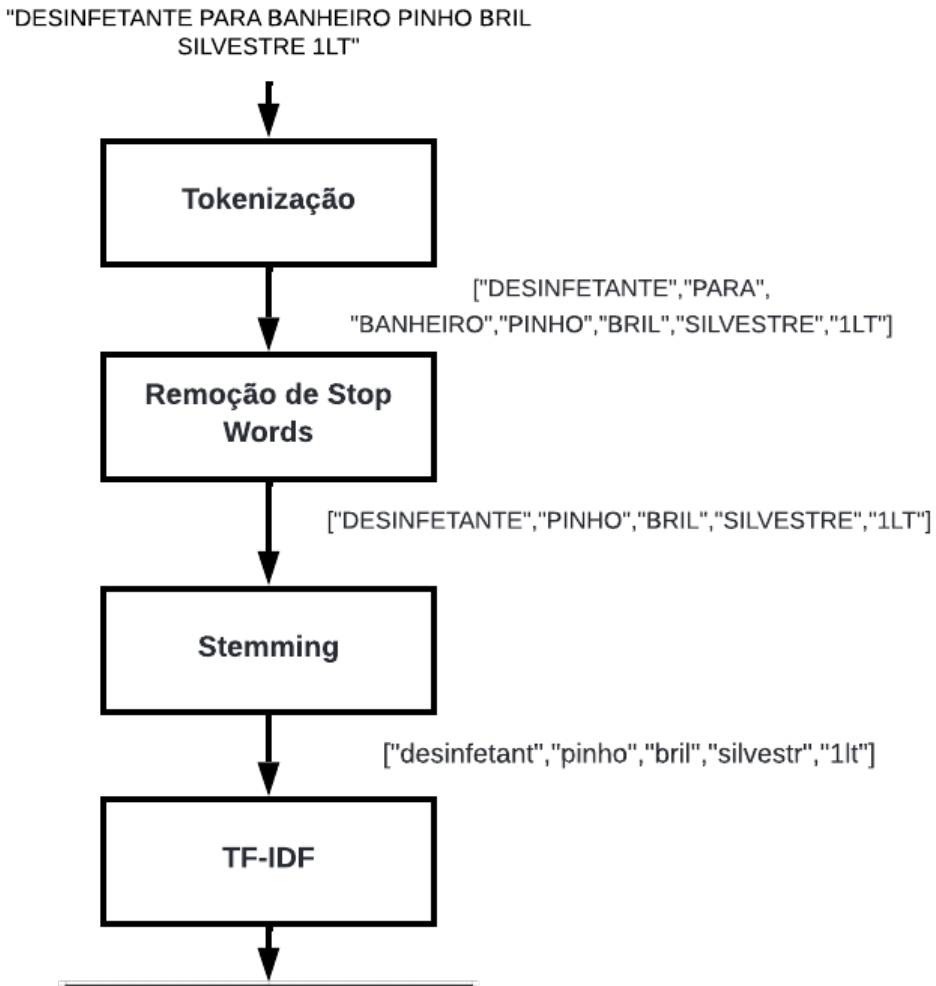
Para isso, construímos uma metodologia dividida em três etapas, sendo essas: (1) Analise e balanceamento dos dados; (2) pré-processamento das descrições dos produtos; (3) treinamento e validação dos classificadores, veja Figura 6.

Figura 6. Etapas realizadas durante a metodologia desse trabalho



Fonte: Autor

Figura 7. Etapas realizadas durante o PNL



	TF-IDF
desinfetant	1
pinho	1
bril	1
silvestr	1
1lt	1

Fonte: Autor

4.1. Analise e balanceamento do dados

Os dados usados no presente trabalho foram cedidos por um escritório de contabilidade do município de Tucuruí-PA, onde estavam armazenados localmente em um Sistema de Gerenciamento de Banco de dados (SGDB) *PostgreSQL*. As informações presentes no DB são as descrições dos produtos e os seus respectivos códigos NCM como mostra a Tabela 5 que foram extraidas das Notas Ficais Eletrônicas (NF-e), totalizando 340.000 descrições distintas. Conforme proposto no trabalho de [Luppes et al. 2019] o código NCM foi divido em 4 partes: capítulo, posição, subposição e item/subitem, como mostra a Tabela 5.

Devido a grande variedade de códigos de NCM e diversidade de produtos por código, alguns código tendem a aparecer mais e outros menos, gerando um desbalanceamento nos dados como mostram as Figuras 8 e 9.

Seguindo o pensamento de [Prati 2006] foram usadas duas técnicas para o balanceamento, *under-sampling* e *over-sampling* para solução do problema. No processo de replicação das classes minoritárias (*over-sampling*), realizou-se a tradução e a retradução dos textos e o inserindo novamente no *dataset*, afim de tentar modifica-los para maior diversificação da classe, limitando cada classe com no máximo 5000 amostras, as classes que ficaram com menos 1000 amostras foram retiradas como mostra a Tabela 4, e as classes com maioritárias foi retirando amostras aleatórias usado a técnica de *under-sampling* limitando em 5000 amostras, o *dataset* antes do processo de balanceamento tinha aproximadamente 340 mil itens, e após o balanceamento 250 mil itens, essa redução se dá pelo corte feito nas classes com mais de 5 mil itens.

A Figura 8 apresentam o *Dataset* antes do balanceamento e a Figura 10 mostra o resultado após a analise e o balanceamento com o experimento do capítulo dos NCM com os dois primeiros dígitos, pode-se observar que existe uma linha vermelha presente nos gráficos, que representa a media de cada classe e nota-se que foi reduzido quase pela metade. Por questão de espaço e por questão de reprodutibilidade outras imagens, scripts dos experimentos e dados podem ser encontrados neste repositório `https://github.com/pedrobragap/products_classification`.

Tabela 4. Número de classes antes de depois do balanceamento

	Desbalanceados	**Balanceados**
2 dígitos	99	98
4 dígitos	85	81
6 dígitos	83	82
8 dígitos	90	88

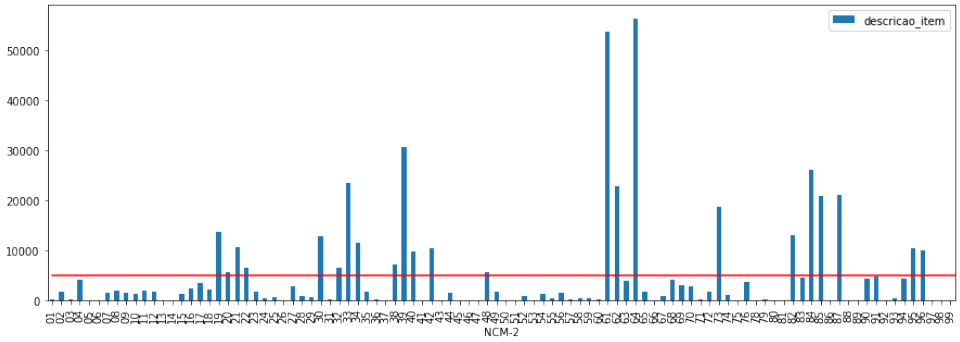
4.2. Pré-processamento das descrições textuais

Esta etapa descreve o pré-processamento realizado sobre as descrições textuais dos produtos contidos na NF-e, e tem como objetivo extrair o vetor de características de cada um dos documentos de entrada como mostra a Figura 7. O algoritmo pega uma série de descrições como entrada e inicia com a tokenização dos mesmos, quebrando o texto e extraindo vetor de *tokens*. A próxima fases é remover os *stopwords* e seguindo para

Tabela 5. Códigos NCM Segregados

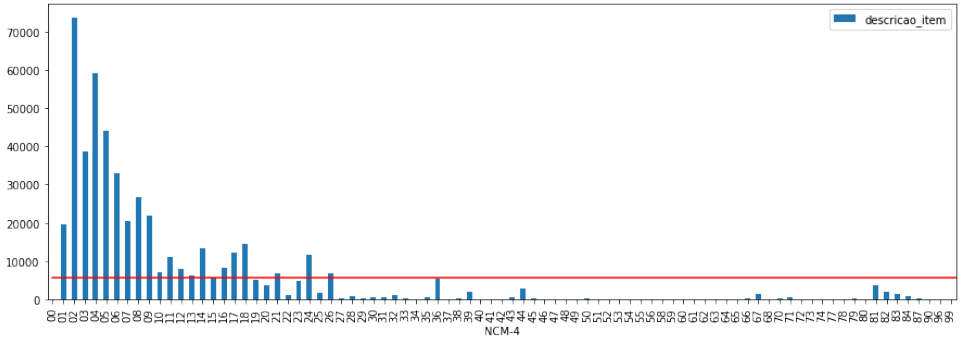
DESCRIÇÃO DOS PRODUTOS	CÓDIGO NCM	NCM-2	NCM-4	NCM-6	NCM-8
Buscofem 400mg 10 Caps Gel	30049029	30	04	90	29
ACHOCOLATADO EM PO TODDY ORIGINAL 200G	90308490	90	30	84	90
TEMPERO SAZON CARNES 60G	21039021	21	03	90	21
VERDURAS DE REAPROVEITAMENTOKG	07020000	07	02	00	00
OLEO DE SOJA COMIGO 900ML	15079011	15	07	90	11
FEIJAO IMPERIAL CARIOCA 1KG	07133399	07	13	33	99
OLEO DE SOJA COMIGO 900ML	15079011	15	07	90	11
SOUTIEN C:41T:EG	62121000	62	12	10	00
DESINFETANTE PINHO BRIL SILVESTRE 1LT	38089419	38	08	94	19
COXA SOBRECOXA FRANGO KG	02071400	01	07	14	00

Figura 8. Quantidade de amostras no Capitulo de NCM-2 antes do balanceamento



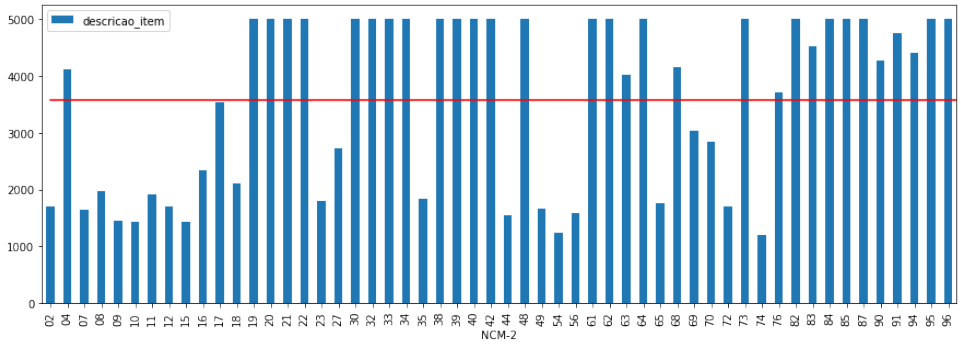
Fonte: Próprio Autor

Figura 9. Quantidade de amostras no Capitulo de NCM-4 antes do balanceamento



Fonte: Próprio Autor

Figura 10. Quantidade de amostras no Capitulo de NCM-2 após o balanceamento

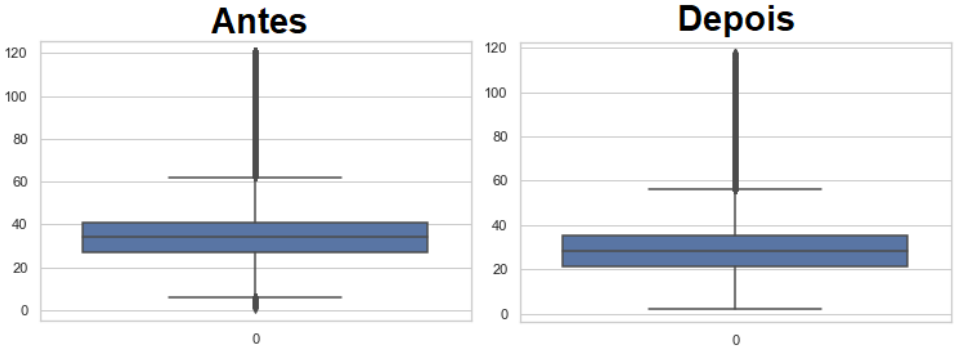


Fonte: Próprio Autor

a aplicação de *stemming* sobre os *tokens* restantes no vetor, conforme mencionado na Subseção 2.3

Com a conclusão das etapas anteriores, as descrições textuais tendem a reduzir suas dimensões devido a remoção de *stopwords* e o *stemming*, a média da dimensão das descrições antes do pré-processamento era de aproximadamente 36 caracteres e depois do tratamento essa media cai para aproximadamente 29 caracteres como nota-se na Figura 11

Figura 11. Media da dimensão das descrições textuais antes e depois do Pré-processamento



Fonte: Próprio Autor

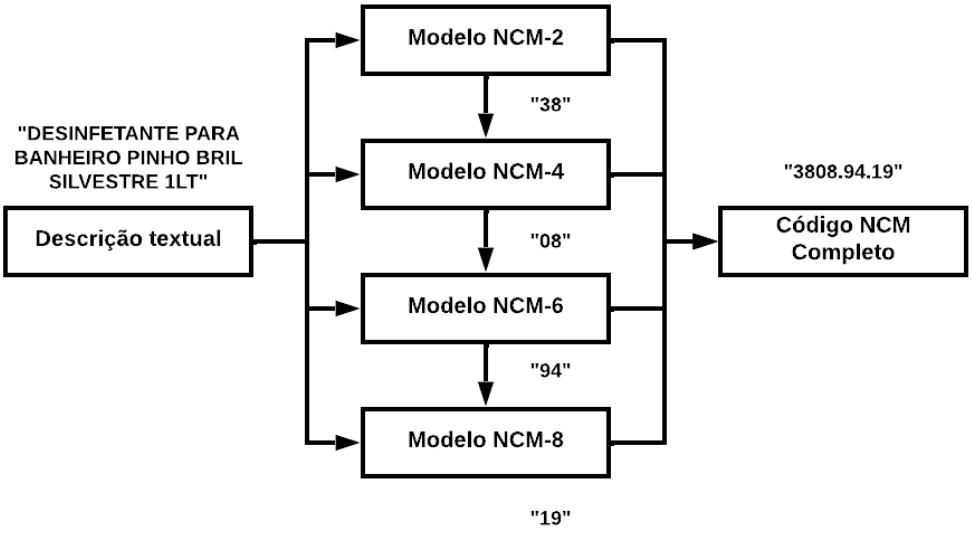
Os passos seguintes é referente à contabilização dos termos para calcular as métricas TF e IDF, que servirão de base para o cálculo de TF-IDF, conforme explicado na Subseção 2.3.4, cujos resultados produzirão os pesos dos vetores de características a serem utilizados para o treinamento do classificador.

4.3. Treinamento do classificador

Nesse trabalho foi utilizado a metodologia de cascata, usando 4 modelos de classificação, afim de, predizer os 8 dígitos do NCM a partir da descrição textual, o fluxograma presente na Figura 12 apresenta o processo de classificação realizado, neste caso, para cada saída dos modelos, o NCM predito é utilizado como entrada em conjunto com a descrição textual tokenizada para predizer os 2 dígitos subsequentes do NCM. Foram selecionados 2 algorítimos de classificação: *Multinomial Naive Bayes*,*Maquinas de Vetores de Suporte*,

todos são baseados no aprendizado de maquina supervisionado. Os dados de entradas foram divididos 80% treino e 20% teste, as *features* de entrada são os as descrições em forma vetorial e o *label* são os dígitos do NCM, Para assim encontrar o algorítimo de classificação mais eficiente para o problema.

Figura 12. Fluxo dos modelos



Fonte: Próprio Autor

5. RESULTADOS

As configurações utilizadas para a criação dos modelos e *datasets* são os mesmos para ambos. a tabela 6 e a Figura 13 apresenta o resultado do teste dos modelos feito separadamente, foi utilizado a metrica de acurácia *F1-score* para medi-las. E por fim os resultados agrupados na tabela 7 e a Figura 14. Observa-se que os resultados obtidos pelo *Linear SVC* nos dois primeiros modelos tem um alto nível de acurácia e no geral foi melhor que o *Naïve Bayes*, assim como no trabalho de [Mishu and Rafiuddin 2016] que o *Linear SVC* foi superior, possivelmente o SVC se saiu melhor devido a técnica de *kernel trick* que opera no espaço dimensional superior sem computar as características originais.

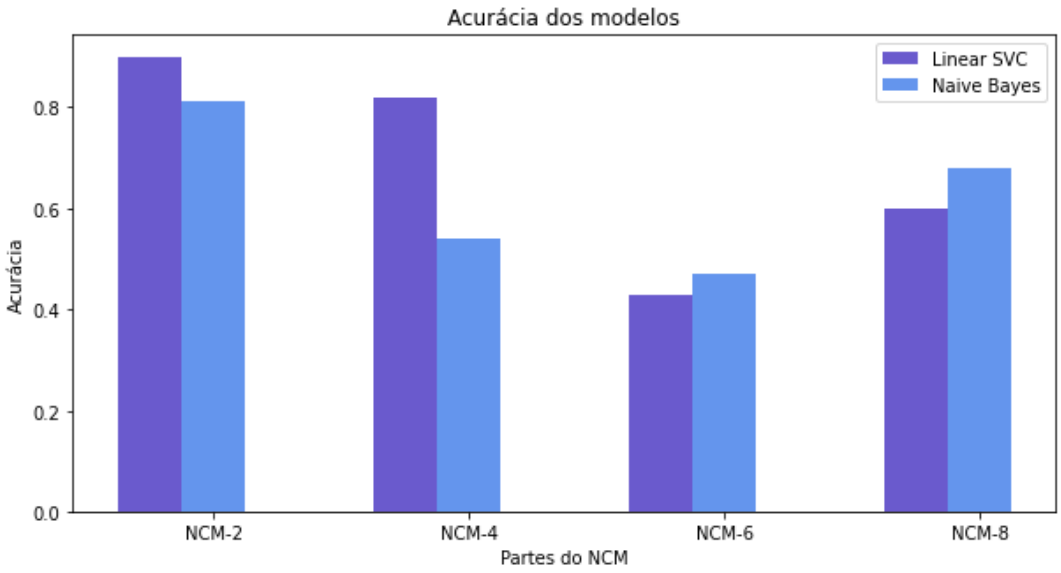
Tabela 6. Acurácia dos Modelos

Partes do NCM	LINEAR SVC	NAIVE BAYES
NCM-2	0.90	0.81
NCM-4	0.82	0.54
NCM-6	0.43	0.47
NCM-8	0.60	0.68

Tabela 7. Acurácia dos Modelos utilizando a tecnica de cascata

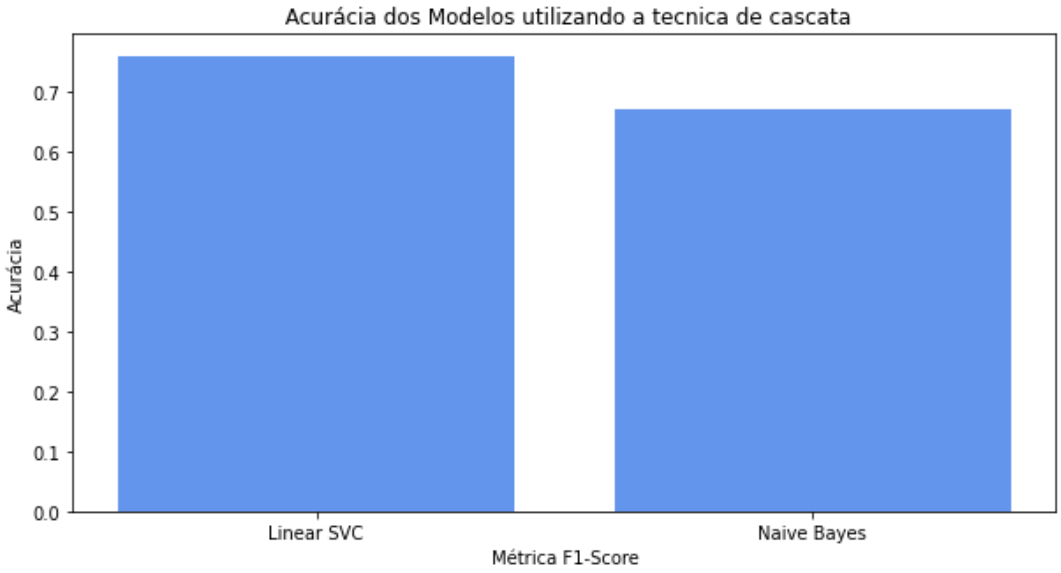
Métricas	LINEAR SVC	NAIVE BAYES
F1-score	0.76	0.67

Figura 13. Acurácia dos Modelos



Fonte: Próprio Autor

Figura 14. Acurácia dos Modelos



Fonte: Próprio Autor

6. CONCLUSÕES

Neste estudo foi desenvolvida e avaliada uma aplicação de classificação de texto, com o objetivo de analisar três técnicas de aprendizado de máquina supervisionado, a saber: 1) Multinomial Naive-Bayes; 2) Suporte Vector Machine; foi utilizado a mesma configurações de pré-processamento de texto, como técnicas de balacemento como under-sampling e over-sampling, vetorização Term Frequency–Inverse Document Frequency (TF-IDF), remoção de stop words e processamento de linguagem natural stemming. Utilizou-se também descrições extraídas das NF-e. Após a análise dos resultados dos dois classificadores utilizados, observou-se que o linear SVC foi superior. Assim, de acordo com os resultados dos experimentos realizados neste estudo, conclui-se que o classificador SVC é mais eficaz na classificação dos códigos NCM do que o classificador Naive-Bayes.

Para trabalhos futuros, gostaríamos de trabalhar com redes neurais convolucionais (CNN), baseado nos resultados de trabalho de [Luppes et al. 2019] e com o modelo BERT (*Bidirectional Encoder Representations from Transformers*), que é um rede neural capaz de aprender as formas de expressão da linguagem humana, o trabalho de [**?**] demostra ótimos resultados utilizado essa ferramenta.

Referências

[Andre Dieb Martins 2013] Andre Dieb Martins, Bruno B. Albert, E. C. G. (2013). Classificador de textos otimizado utilizando lei de potencia para palavras raras. *XXXI SIMPOSIO BRASILEIRO DE TELECOMUNICAÇÕES*.

[Bonfim et al. 2012] Bonfim, D. P., Moraes, D., Machado, H., Amorim, M. O., and Raimundini, S. L. (2012). Nota fiscal eletrônica: uma mudança de paradigma sob a perspectiva do fisco estadual. *ConTexto*, 12(21):17–28.

[Brasil 2003] Brasil (2003). Emenda constitucional n. 42.

[de Abreu Batista et al. 2018] de Abreu Batista, R., Bagatini, D. D., and Frozza, R. (2018). Classificação automática de códigos ncm utilizando o algoritmo naïve bayes. *iSys-Revista Brasileira de Sistemas de Informação*, 11(2):4–29.

[Delanerolle et al. 2021] Delanerolle, G., Yang, X., Shetty, S., Raymont, V., Shetty, A., Phiri, P., Hapangama, D., Tempest, N., Majumder, K., and Shi, J. (2021). Artificial intelligence: A rapid case for advancement in the personalization of gynaecology/obstetric and mental health care. *Women's Health*, 17:174550652110181.

[Ding et al. 2015] Ding, L., Fan, Z., and Chen, D. (2015). Auto-categorization of hs code using background net approach. *Procedia Computer Science*, 60:1462–1471.

[do Bomfim et al. 2020] do Bomfim, E. T., da Silva, R. B., and da Silva, J. A. M. (2020). Reflexos tributários causados pela classificação incorreta da ncm no valor pis/cofins devido por um supermercado paraibano. *Revista de Contabilidade e Gestão Contemporânea UFF*, 3(1):49–62.

[Escovedo and Koshiyama 2020] Escovedo, T. and Koshiyama, A. (2020). *Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise*. Casa do Código.

[Granitto et al. 2005] Granitto, P. M., Rébola, A., Cerviño, U., Gasperi, F., Biasoli, F., Ceccatto, H., et al. (2005). Cascade classifiers for multiclass problems. In *Proceedings*

of the 7-th Argentine Symposium on Artificial Intelligence (ASAI), Rosario, Argentina, pages 29–30. Citeseer.

[Kadhim 2019] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292.

[Li and Li 2019] Li, G. and Li, N. (2019). Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network. *Electronic Commerce Research*, 19(4):779–800.

[Luppes et al. 2019] Luppes, J., de Vries, A. P., and Hasibi, F. (2019). Classifying short text for the harmonized system with convolutional neural networks. *Radboud University*.

[Mishu and Rafiuddin 2016] Mishu, S. Z. and Rafiuddin, S. (2016). Performance analysis of supervised machine learning algorithms for text classification. pages 409–413.

[Neto et al. 2000] Neto, J. L., Santos, A. D., Kaestner, C. A., Alexandre, N., Santos, D., A, C. A., Alex, K., Freitas, A. A., and Parana, C. (2000). Document clustering and text summarization.

[Orengo and Huyck 2001] Orengo, V. M. and Huyck, C. R. (2001). A stemming algorithmm for the portuguese language. In *spire*, volume 8, pages 186–193.

[Prati 2006] Prati, R. C. (2006). *Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos*. PhD thesis, Universidade de São Paulo.

[Roberto Scalco et al. 2015] Roberto Scalco, P., Klaold Lippi, M., and de Almeida, M. I. S. (2015). Preço e renda como determinantes da demanda por bens de luxo no brasil: Um estudo econométrico com produtos importados da nomenclatura comum do mercosul. *Brazilian Journal of Management/Revista de Administração da UFSM*, 8(3).

[Russell and Norvig 2003] Russell, S. J. and Norvig, P. (2003). Instructor's solution manual for artificial intelligence: a modern approach.

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

[SEFAZ 2021] SEFAZ (2021). Sobre a nf-e.

[Sousa 2010] Sousa, J. P. R. d. (2010). Impactos da utilização da nota fiscal eletrônica nas atividades de monitoramento e fiscalização do icms: um estudo na secretaria da fazenda do estado do ceará. Master's thesis, Universidade Federal do Ceará,.

[Trstenjak et al. 2014] Trstenjak, B., Mikac, S., and Donko, D. (2014). Knn with tf-idf based framework for text categorization. *Procedia Engineering*, 69:1356–1364.

[Wang et al. 2017] Wang, J., Wang, Z., Zhang, D., and Yan, J. (2017). Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI*, volume 350.

[Yu et al. 2012] Yu, H.-F., Ho, C.-H., Arunachalam, P., Somaiya, M., and Lin, C.-J. (2012). Product title classification versus text classification. *Csie. Ntu. Edu. Tw*, pages 1–25.