Classificação dos Códigos de NCM Usando Processamento de Linguagem Natural

Pedro Pinheiro, Marcos Amaris

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

pedrobraga85@gmail.com, amaris@ufpa.br

Resumo. *Esse artigo 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 Capítulos (primeiros dois dígitos) da Nomenclatura Comum do Mercosul (NCM). A classificação foi realizada utilizando o algorítimo de Máquina de vetores de suporte (SVM), com uma base de dados de 340.000 produtos distintos, que foram tratados usando as técnicas de Processamento natural de linguagem. Obteve-se um acurácia de 87% para um total de 50 classes.*
Palavras–chave*: Processamento de Linguagem Natural; Aprendizagem de máquina; Classificação de Texto and Nomenclatura Comum do Mercosul;*

Abstract. *This article aims to develop a process to classify the descriptions of products in electronic invoices (abbreviated NF-e in portuguese). This classification is done on the Chapters (first two digits) of the Mercosul Common Nomenclature (NCM). The classification was performed using the Support Vector Machine algorithm, with a database of 340,000 distinct products, which were treated using Natural Language Processing techniques. An accuracy of 84% was obtained for a total of 50 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, em 1983 a Organização Mundial das Alfândegas (WCO) criou uma nomenclatura comum dos produtos em todo o mundo, conhecido como Sistema Harmonizado de Descrição e Codificação de Mercadorias, Chamado de Código **HS**. Baseado no código HS foi criada a Nomenclatura Comum do Mercosul, conhecida pelo acrônimo NCM, que utiliza como base os 6 dígitos do Código HS, é uma convenção de categorização de mercadorias adotada desde 1995 pelos países que compõem o bloco econômico do Mercosul.

Normalmente, a verificação das descrições dos produtos assegura que as mercadorias estejam em conformidade com os regulamentos governamentais para evitar a entrada indevida ou ilegal no país de destinado [Che et al. 2018]. Contudo, a classificação correta dos códigos de NCM continua sendo uma tarefa desafiadora, conforme listado em [Luppes et al. 2019].

Sendo assim, a classificação incorreta de mercadorias além de impactar nos tributos como PIS e Cofins, que fere os cofres públicos, acarreta em sanções tributárias para o contribuinte ou terceiro envolvido na classificação. É expressivo a evolução nos últimos anos do aprendizado de maquina juntamente com as técnicas de Processamento de Linguagem Natural (PNL), e através da grande quantidade dados que são gerados com a NF-e, esse trabalho tem como objetivo a criação de um modulo de aprendizado de maquina, utilizando o processamento natural de linguagem predizer o Codigo NCM através da descrição dos produtos.

Este trabalho está estruturado como segue: 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 finalmente as conclusões na Seção 6.

2. CONCEPTOS E BACKGROUND TEÓRICO

2.1. 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 em um texto. Seguindo para a remoção de palavras de paradas ou *Stop-words*, que se refere às palavras de conexão que tem pouca contribuição para classificação ou análise. Por fim é feito o processo de *stemming*, que é em resumo remover os sufixos das palavras, é ficando com o radical da palavra que representa o significado. Finalmente é usada uma métrica estatística, que tem como propósito demonstrar o grau de importância de uma palavra no texto, levando em consideração toda as palavras do texto sob análise.

3. TRABALHOS RELACIONADOS

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% de 2 classes só. [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 tecnicas 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**.

4. METODOLOGIA

Foi usada a Linguagem de Programação Python para os experimentos e processo de classificação. O dataset usado para esse estudo é composto por 340.000 descrições distintas separadas em 98 classes (Levando em consideração apenas o capitulo do NCM-2). Nota-se que os dados estão claramente desbalanceados, seguindo o pesamento de [Prati 2006] foi usado a ideia de remoção de exemplos das classes majoritárias, chamado de *Under-sampling*, que é uma abordagem bem direta para solucionar o desbalanceamento, também foi usado a técnica de inclusão de exemplos nas classes minoritárias, chamado de *Over-sampling*, bem similar ao anterior, só que esse replica os exemplos para solução do problema, foi feita uma tradução para o inglês e a retradução para o português dos textos,

afim de tentar modifica-los, e assim ficou o dataset após a analise e balanceamento figura 2 com 402.634 amostras.

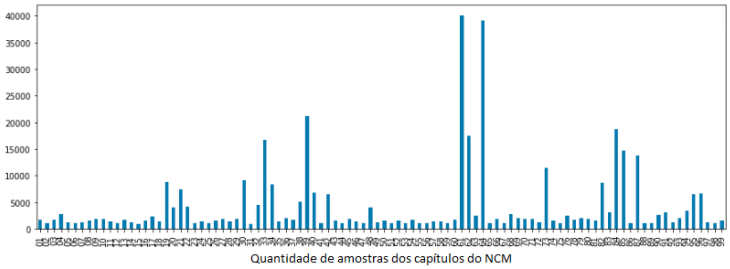

Figura 1. Quantidade de amostras por Capitulo de ncm

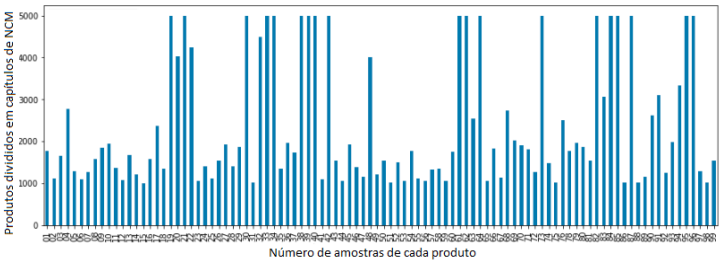

Figura 2. Quantidade de amostras por Capitulo de ncm

Após o balanceamento dos dados, foi utilizado a biblioteca NLTK para a retirada de Pontuações, Caracteres Especias, *Stop Words* e o método de *Stemming*, que extrai a raiz da palavra. Com a normalização dos dados foi aplicado a técnica TF-IDF (*Term Frequency–Inverse Document Frequency*). Que é uma medida estatística que mostra a importância de cada uma das palavras em um texto.

Por fim, para a criação do modelo foi utilizado Máquina de Vetores de Suporte usando a classe `LinearSVM` do *sckitlearn*, tendo como entrada do modelo as descrições do produtos tratadas e no formato TF-IDF, e o Capítulo do NCM como rótulo. O modelo foi divido em 80% para Treino e 20% para Teste.

5. RESULTADOS

Após a criação do modelo foi disposto um conjunto de dados de teste para validação, com 46.120 descrições para um total de 98 rótulos, foi escolhido a métrica de acurácia que informa em geral o quão o modelo está acertando, onde obteve-se uma media total de acurácia de 87%, para fins de comparação foi criado um novo modelo utilizando o algorítimo de *Multinomal Naive Bayes* (MNB) e foi usado os mesmo processos de validação para ambos, como pode se observar na Tabela 1, podemos notar que os resultados obtidos com o *Linear SVM* foram superiores devido a abordagem utilizada para a interação com as *Features*, pois o MNB trata as mesmas como independentes.

A Tabela 2 foi criada com 10 rótulos aleatórios, para expor os resultados com mais precisão foram utilizados algumas métricas de avaliação, como o *Precision* que é a razão entre os verdadeiros positivos e todos os positivos, *Recall* é a medida do modelo identificando corretamente os verdadeiros positivos, e o *F1-Score* é a combinação dos valores. Observa-se que algumas classes pode ter suas valores de métricas aumentadas se houver um tratamento especifico para cada classe. Como é o caso do Capítulo 39 que tem como descrição ”PLÁSTICO E SUAS OBRAS”, que pode ter diversa palavras com

grande expressão que não representa necessariamente o capítulo e acaba atrapalhando a classificação.

Tabela 1. Resultados para diferentes algorítimos de classificação

Modelo	Acurácia
Linear SVM	87%
Multinominal Naive Bayes	79%

Tabela 2. Métricas para alguns rótulos

label	16	21	27	33	39	62	65	72	82	96
precision	0.91	0.88	0.89	0.78	0.73	0.81	0.90	0.84	0.84	0.88
recall	0.93	0.88	0.83	0.59	0.72	0.81	0.97	0.76	0.90	0.93
f1-score	0.92	0.88	0.86	0.68	0.73	0.81	0.93	0.80	0.87	0.91
sample	338	998	398	1018	1012	1032	203	241	1024	983

6. CONCLUSÕES E TRABALHOS FUTUROS

A predição dos dígitos do NCM é realmente uma tarefa muito complexa, requer dados muito bem normalizados. Através de técnicas de processamento de linguagem natural conseguimos construir um modelo com uma acurácia de 87%, tendo como base as descrições textuais tiradas das NF-e.

Para um Próximo passo, um tratamento específico das descrição para cada classe, com o objetivo de retirar as palavras invesadas, e também temos como objetivo para trabalhos futuros a criação de mais modelos de aprendizado, tendo como entrada a saída do modelo anterior junto com a descrição do produto, assim criando 4 modelos para predizer por completo os 8 dígitos do Código.

Referências

[Che et al. 2018] Che, J., Xing, Y., and Zhang, L. (2018). A comprehensive solution for deep-learning based cargo inspection to discriminate goods in containers. In *Proceedings of the CVPR IEEE Conference*, pages 1206–1213.

[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.

[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*.

[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.