Estratégias Computacionais Baseadas em Similaridade de Textos e Visualização Exploratória para a Identificação de Inconsistências em Notas Fiscais Eletrônicas

Mayara C. Marinho[1], **Li Weigang**[1], **Vinícius Di Oliveira**[2], **Maria Cristina F. Oliveira**[3], **Vinicius R. P. Borges**[1]

[1]Universidade de Brasília (UnB), Brasília, DF, Brazil

[2]Secretaria de Economia do Distrito Federal, Brasília, DF, Brazil

[3]Universidade de São Paulo (USP), São Carlos, SP, Brazil

{mayaracm,weigang,viniciusrpb}@unb.br, vinicius.oliveira@economia.df.gov.br, macristina@icmc.usp.br

Abstract. *Examining fiscal invoices for fraud detection by regulatory bodies is challenging in view of the expressive volume of electronic invoices issued daily. To address this challenge, this research introduces a novel approach that combines automatic labeling through text similarity with exploratory visualization techniques. Our aim is to assist specialists in the fiscal auditing process, while balancing between automation and human autonomy. This approach is operationalized through an interactive web visualization system called NFViz. NFViz enables specialists to explore the invoices to identify inconsistencies that may signal potential instances of fraud. By doing so, it streamlines the fraud identification process, thereby contributing to enhance routine tax auditing activities.*

Resumo. *A detecção de fraudes em notas fiscais é um desafio considerável para os órgãos de controle, dada a quantidade expressiva de documentos gerados diariamente. Buscando soluções para esse desafio, esta pesquisa propõe uma abordagem inovadora que combina rotulação automática de notas fiscais baseada em similaridade de texto e técnicas de visualização exploratória. O objetivo é auxiliar especialistas na tarefa de auditoria fiscal, trazendo equilíbrio entre automatização e autonomia por meio de um sistema de visualização web interativo denominado NFViz. O NFViz permite investigar as notas fiscais em busca de incompatibilidades que podem sinalizar possíveis fraudes, otimizando as atividades de auditoria fiscal ao reduzir o escopo de busca por notas suspeitas.*

1. Visão Geral

A fraude fiscal é regulamentada pela Lei Brasileira nº 4.729/65, que exige a declaração da nota fiscal após a venda de um produto ou serviço. A despeito da finalidade de evitar omissões, criação de documentos falsos, dentre outras práticas fraudulentas, a fiscalização da aplicação desta Lei ainda é um processo desafiador, uma vez que são geradas aproximadamente 5,8 milhões de notas fiscais por dia no Brasil, segundo o Ministério da Fazenda[1]. A inspeção individual de parcela significativa das notas fiscais pelos especialistas é inviável, porque demandaria uma análise minuciosa de um grande volume de dados.

[1]http://www.nfe.fazenda.gov.br/portal/infoEstatisticas.aspx, acessado em 12 de abril de 2023.

No que concerne à detecção de fraudes em documentos fiscais, diversos trabalhos da literatura propõem estratégias baseadas em tarefas supervisionadas de classificação [Hajek and Henriques 2017]. No entanto, não existem conjuntos de dados de notas fiscais eletrônicas com rótulos de fraude. Diante desse cenário, uma alternativa é a criação de um procedimento prévio à análise manual dos casos, visando identificar subconjuntos de notas fiscais com indícios de fraude, as quais podem ser encaminhadas para uma auditoria minuciosa posterior [Zha 2020]. Existem trabalhos que consideram essa linha de raciocínio e propõem ferramentas visuais com base em técnicas não supervisionadas de agrupamento, objetivando reduzir o tempo de decisão [Zha 2020] [Resck et al. 2023].

Os principais desafios relacionados à análise de Notas Fiscais do Consumidor Eletrônicas (NFC-es) são a quantidade expressiva de documentos, a presença de atributos categóricos e textuais curtos e pouco padronizados, e a necessidade de participação de um especialista em auditoria fiscal na análise das diversas inconsistências que podem sinalizar uma fraude. Além disso, um ponto chave desse problema é a combinação do conhecimento e experiência prática dos especialistas e de tecnologias atuais para identificar casos suspeitos de fraude, sinalizados pela ocorrência de inconsistências no preenchimento dos campos das notas fiscais. A utilização de ferramentas que facilitem a identificação dessas inconsistências pode proporcionar aos especialistas um suporte para o aprimoramento das atividades de auditoria e também maior eficiência no processo. Entretanto, essas soluções precisam buscar um equilíbrio entre eficiência e autonomia do especialista.

Esta pesquisa teve como principal objetivo desenvolver um método que facilitasse a detecção de inconsistências, utilizando técnicas de Aprendizado de Máquina e de Visualização de dados. Foi desenvolvido um sistema *web* interativo que possibilita fazer uma análise visual das similaridades das notas fiscais, bem como a visualização de agrupamentos de notas similares para auxiliar a identificação de inconsistências. Espera-se que essa abordagem visual, analítica e inteligente proporcione aos especialistas uma nova perspectiva para a tarefa de detecção de fraudes fiscais, contribuindo para a eficiência do esforço rotineiro de fiscalização. Assim, as principais contribuições deste projeto são:

- A proposta de uma medida de similaridade para comparar notas fiscais eletrônicas descritas por atributos textuais e categóricos;
- A criação de rótulos de inconsistência para a identificação automática de notas fiscais suspeitas de fraude;
- Uma avaliação de técnicas de projeção multidimensional, como a *MDS* [Cox and Cox 2008], a *t-SNE* [Van der Maaten and Hinton 2008] e a *UMAP* [McInnes et al. 2018], para a visualização de notas fiscais;
- O desenvolvimento e validação preliminar de um sistema *web* interativo que incorpora algoritmos de cálculo de similaridade, de visualização e de agrupamento para viabilizar a identificação, guiada pelo especialista, de notas fiscais inconsistentes.

2. Revisão Bibliográfica

A detecção de fraudes tem sido amplamente explorada na literatura, embora poucos trabalhos abordem técnicas específicas para as notas fiscais eletrônicas. No que concerne auditoria fiscal, pode-se citar um método para mensurar e ranquear o potencial de fraude de contribuintes baseando-se na análise dos principais tipos de indicadores de fraude [Matos et al. 2015], um estudo de caso de auditoria [Kieckbusch et al. 2020]

e uma classificação baseada em Redes Neurais Convolucionais para definir a categoria do produto comercializado [Kieckbusch et al. 2021]. Ademais, foi publicado um conjunto de dados de notas fiscais com descrições rotuladas em categorias específicas [Di Oliveira et al. 2022], no entanto, não existem rótulos quanto à ocorrência de fraude.

Métodos não supervisionados são frequentemente utilizados em conjuntos de dados não rotulados, visto que podem auxiliar a compreensão do comportamento do sistema a ser estudado e, consequentemente, auxiliar a detecção anomalias nos dados [Bolton and Hand 2001], o que pode ser essencial em contextos de detecção de fraude. Ao considerar o problema de busca por transações suspeitas como uma tarefa de classificação semi-supervisionada, Zha [Zha 2020] propôs um assistente de auditoria fiscal capaz de analisar visualmente as informações e detectar evidências de fraude. De maneira semelhante, Resck et al. desenvolveu um sistema para especialistas do direito utilizando técnicas de Aprendizado de Máquina e análise visual interativa [Resck et al. 2023].

As NFC-es evidenciam alguns desafios em aberto nas abordagens computacionais da literatura, como mencionado anteriormente. Esta pesquisa busca preencher essas lacunas propondo um método de visualização exploratória apoiado por técnicas de Aprendizado de Máquina e de similaridade de textos para identificar notas fiscais inconsistentes.

3. Metodologia

Essa seção apresenta dois métodos complementares para a detecção de inconsistências em NFC-es. A Subseção 3.1 descreve informações relacionadas ao conjunto de dados NFC-e. A Subseção 3.2 detalha a estratégia automática baseada em similaridade de textos, enquanto a Subseção 3.3 apresenta o processo de visualização exploratória, que fornece mais autonomia aos especialistas na identificação de inconsistências em NFC-es.

3.1. Notas Fiscais Eletrônicas do Distrito Federal

Nessa pesquisa foi utilizado o conjunto privado de NFC-es do Distrito Federal, Brasil. Criado em 2019, possui formato tabular e contém 359.494 instâncias compostas por três atributos descritos na Tabela 1. Além do conjunto de dados principal, é utilizada a tabela de descrição oficial concatenada do NCM de 2019 como uma forma de gabarito (*ground-truth*) em relação à descrição informada pelo comerciante na nota fiscal do produto. É esperado que a descrição informada na nota fiscal seja compatível com as nomenclaturas do CST e NCM correspondentes à categoria do produto. Assim, se a descrição informada for compatível, segundo o padrão NCM e de acordo com a categoria de imposto CST, a nota é considerada consistente, caso contrário, ela é considerada inconsistente.

Tabela 1. Descrição dos atributos do conjunto de dados NFC-e.

Atributo	Tipo	Significado
DESC (Descrição informada)	Texto	Descrição do produto pelo vendedor. Esse texto é livre, portanto sem um padrão previamente definido.
NCM	Categórico	Nomenclatura Comum do Mercosul. Esse código é padronizado para cada categoria de produto comercializado.
CST	Categórico	Código de Situação Tributária, base para o cálculo do ICMS. Pode assumir os valores “00” (completamente taxado), “20” (possui uma taxa base), “40” (isento), “41” (não taxado) e “60” (ICMS taxado por substituição).

3.2. Identificação Automática de Inconsistências

O fluxograma da Figura 1 ilustra as etapas do método proposto para a identificação automática de inconsistências, desde o pré-processamento do conjunto de dados até a rotulação de cada nota fiscal fornecida como entrada como consistente ou inconsistente.

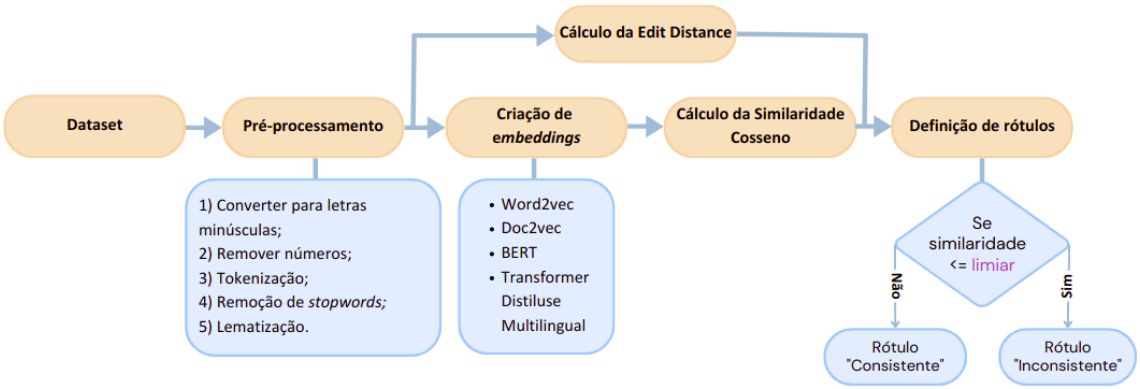


Figura 1. Fluxograma da metodologia de criação de rótulos de consistência.

Primeiramente, foi aplicado um pré-processamento aos textos correspondentes às descrições informadas e às descrições oficiais obtidas do gabarito, que consistiu na conversão para minúsculo, remoção de números, *tokenização*, remoção de *stopwords* e lematização. Em seguida, foi realizado o cálculo de distância entre ambas as descrições, com o objetivo de definir o quão similares elas são e, assim, detectar descrições informadas pouco compatíveis com o esperado, o que pode caracterizar uma inconsistência. Para que essa comparação pudesse ser feita, assumiu-se que o NCM informado é o correto.

A determinação da similaridade entre os atributos descritivos das notas é um componente fundamental da solução. Foram consideradas duas estratégias distintas, visando identificar a que melhor se adaptasse aos padrões textuais das notas fiscais: a aplicação da *Edit Distance* nos textos; e alternativamente a geração de *embeddings* a partir dos textos, sobre os quais foi computada a similaridade cosseno. Para gerar os *embeddings*, foram comparadas múltiplas abordagens, como *Word2vec* com a arquitetura *Continuous Bag of Words*, *Doc2vec* [Le and Mikolov 2014], *Transformer Distiluse Multilingual* e *Bidirectional Encoder Representations from Transformers (BERT)* [Khurana et al. 2023].

Com base na similaridade calculada anteriormente, é preciso estabelecer um limiar que determine o valor de similaridade mínima entre as descrições para uma nota ser considerada consistente. Para isso, foram conduzidos experimentos na Seção 4 analisando o comportamento de distintos valores de limiar na determinação da similaridade.

3.3. Visualização Exploratória

O fluxograma da Figura 2 ilustra as etapas que compõem o método proposto, desde o pré-processamento do conjunto de dados até a geração da visualização interativa. Além disso, foi criado o *NFViz*, um sistema *web* que implementa a metodologia proposta.

Novos atributos foram criados a partir da frequência dos dados originais na etapa de pré-processamento, além de ser aplicada a técnica *One-Hot-Encoding* aos atributos categóricos e o *TF-IDF* aos atributos textuais para a extração de vetores de características. Em seguida, foram comparados dois cálculos de distância: a distância de

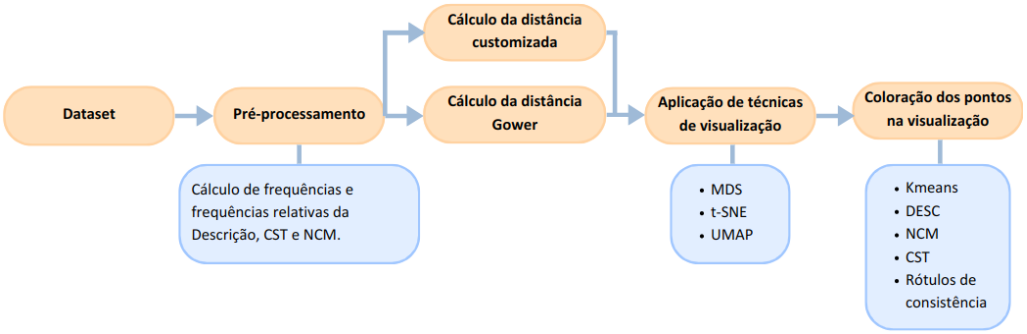


Figura 2. Etapas da metodologia do processo de visualização de notas fiscais.

Gower [Gower 1971] e uma função de distância customizada, definida pela somatória da distância euclideana aplicada aos atributos numéricos e a distância cosseno aplicada aos atributos textuais [Marinho et al. 2022]. Essa última foi proposta como uma alternativa à distância de Gower para tratar o conjunto de dados pós-processado, que inclui atributos textuais, numéricos e categóricos.

Foram consideradas três técnicas de redução de dimensionalidade para possibilitar a visualização das notas fiscais: *Multidimensional Scaling (MDS)*, *t-Distributed Stochastic Neighbor Embedding (t-SNE)* e *Uniform Manifold Approximation and Projection (UMAP)*. Por meio delas, cada instância de dados associada a uma NFC-e é representada por um ponto bidimensional no espaço visual (*layout*), de modo que instâncias semelhantes tendem a estar mais próximas, enquanto as dissimilares aparecem distanciadas. Os pontos, que representam as instâncias, podem ser coloridos para refletir atributos, rótulos de consistência, ou a pertinência a diferentes agrupamentos identificados pelo *K-Means*.

4. Experimentos

Essa seção detalha os experimentos e os resultados obtidos. A Subseção 4.1 apresenta os experimentos para a avaliação das estratégias propostas na identificação de inconsistências. As Subseções 4.2 e 4.3 apresentam os experimentos de visualização e o *NFViz*.

4.1. Identificação Automática de Inconsistências

Nessa tarefa, os experimentos avaliaram a qualidade dos rótulos gerados, visto que não existem dados rotulados quanto à consistência. Para isso, foi utilizado um subconjunto de 10.000 instâncias aleatórias da NFC-e.

Foi aplicada uma tarefa supervisionada de classificação, utilizando o modelo *Long Short-Term Memory (LSTM)*. A entrada consistiu em um texto único composto pela descrição informada, pelo NCM e pelo CST, e o rótulo de inconsistência como alvo. A arquitetura utilizada nos experimentos é composta pelas camadas (na ordem): *LSTM*, *Dropout* e Totalmente Conexa com função de ativação sigmóide. Foram utilizadas a função de perda *Binary Crossentropy*, a otimização *Adam* e a técnica de *Early Stopping*. Por fim, também foi aplicado o *Synthetic Minority Oversampling Technique (SMOTE)* [Bowyer et al. 2002] para lidar com dados desbalanceados e o método *Holdout* 70-10-20. Os valores de *F1-score* obtidos para diferentes configurações do cálculo de distância e método de *embedding* são mostrados na Tabela 2.

Ao analisar os valores de *F1-score* observa-se que a *Edit Distance* produziu rótulos com um bom desempenho preditivo na etapa de avaliação automática. Essa

Tabela 2. Resultado da classificação de um subconjunto aleatório da NFC-e.

Embedding	Similaridade	Limiar=0,1	Limiar=0,2	Limiar=0,3
-	**Edit Distance**	**0,94**	**0,94**	**0,94**
Doc2vec	Cosine distance	0,46	0,51	0,49
BERT	Cosine distance	0,49	0,54	0,59
Transformer Distiluse	Cosine distance	0,30	0,51	0,56
Word2vec	Cosine distance	-	0,33	0,37

configuração resultou em um *F1-score* de 94%, enquanto todas as demais configurações utilizando as técnicas de *embedding* obtiveram resultados inferiores a 60% no subconjunto aleatório. Este resultado implica que a rede *LSTM* conseguiu reconhecer os padrões nas NFC-es com base nos rótulos de consistência criados pela *Edit Distance*, mas que não foi possível reproduzir os rótulos gerados ao utilizar as demais representações, o que pode ser explicado pela sua simplicidade e pelo fato das descrições dos produtos serem curtas, com uma média de 4,5 palavras.

4.2. Visualização Exploratória

Os experimentos para validar o processo de visualização exploratória proposto foram realizados em um subconjunto da NFC-e cujas descrições informadas continham os termos "gin", "vodka" e "cigarro", totalizando 5.000 instâncias. Para definir as técnicas de visualização exploratória, foram realizadas uma análise visual e uma avaliação de qualidade por meio das métricas *Neighborhood Preservation*, que quantifica a preservação das vizinhanças após a redução do espaço dimensional, e o Coeficiente de Silhueta, que mede a qualidade dos grupos em termos de coesão interna e de separabilidade.

As visualizações por similaridade de pontos geradas pelas técnicas *t-SNE* e *MDS* produziram *layouts* em que os agrupamentos apresentaram mais separabilidade quando comparados à *UMAP*. Observou-se um valor de preservação de vizinhança mais alto e estável ao utilizar a *t-SNE* com a distância customizada, aproximadamente $0,83$, e o resultado da métrica Coeficiente de Silhueta foi satisfatório para o espaço de alta dimensão utilizando a distância customizada e rótulos do *K-Means*, obtendo o valor aproximado de $0,90$. Com isso, conclui-se que a configuração que utiliza a *t-SNE* com a distância customizada mostrou-se mais adequada a esse subconjunto de dados.

4.3. Visão Geral do Sistema *NFViz*

A metodologia descrita serviu de base para a implementação do *NFViz*, um sistema *web* interativo que tem como objetivo auxiliar os especialistas na detecção de ocorrências de notas ficais inconsistentes. A interface principal do *NFViz* é apresentada na Figura 3.

São apresentadas a seguir algumas questões de investigação que motivaram o *design* do *NFViz*, identificadas a partir de interações com um especialista em auditoria fiscal:

- Q1. É possível identificar notas fiscais que apresentam valores incompatíveis nos campos relativos ao NCM e ao CST?
- Q2. É possível identificar notas fiscais em que as descrições informadas são distintas das descrições oficiais provenientes do NCM?
- Q3. Existem tipos específicos de produtos em que são comuns as ocorrências de divergências nos valores de descrição informada ou de CST?

Figura 3. Painel principal do *NFViz*.

- Q4. Dado agrupamentos de notas similares, é possível identificar ocorrências de divergência nos valores de descrição informada, NCM ou CST em relação aos demais do grupo?

Para ilustrar como a metodologia proposta e sua implementação no *NFViz* permitem buscar respostas para as questões citadas, são apresentados exemplos ilustrativos com base em visualizações obtidas com técnicas de projeção multidimensional. Essa visualização apresenta um *layout* das notas fiscais em que a proximidade indica similaridade, e os pontos, que representam as notas fiscais, podem ser coloridos para auxiliar a inspeção pelo especialista. As cores podem ser mapeadas para refletir propriedades como similaridade, pertinência a agrupamentos, atributos, ou descrições informadas. Ademais, o *NFViz* inclui funcionalidades de interação, como filtros e opções de seleção de instâncias na visualização. Devido aos melhores resultados de qualidade do *layout* apresentados na Seção 4.2, o *NFViz* emprega por padrão a técnica *t-SNE* para gerar os *layouts*.

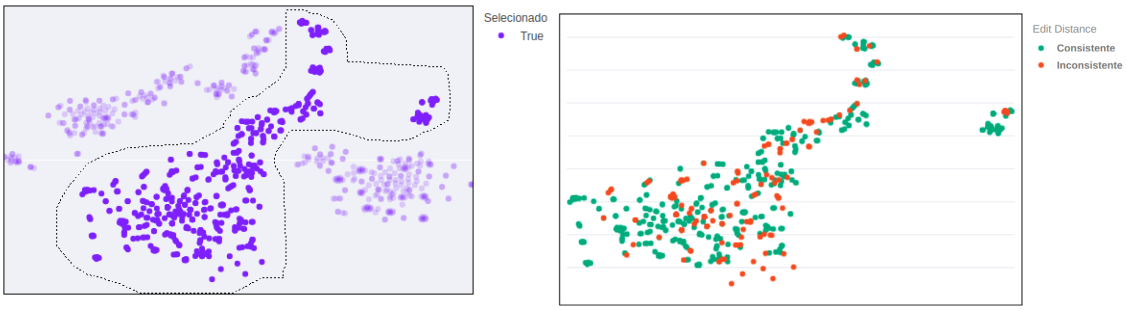
A investigação da questão Q1 pode ser exemplificada pela Figura 4, em que o usuário inicia o processo de busca por inconsistências em notas fiscais a partir da seleção de um conjunto de notas cujas descrições informadas continham a palavra “cigarro”. Após a seleção, foram observadas notas com o mesmo valor de NCM e distintos valores de CST, sendo apresentadas pelas opções “Não informado” e “60 - ICMS taxado por substituição”.

A investigação da questão Q2 pode ser exemplificada pela Figura 5, que apresenta uma visualização de um subconjunto aleatório de notas fiscais. Após a visualização e seleção de notas pelo usuário, o ponto correspondente a cada nota selecionada foi colorido de modo a refletir o seu rótulo de consistência. Nesse caso, o especialista em auditoria pode ajustar o valor do limiar a partir do qual a nota será considerada consistente.

A investigação da questão Q3 pode ser exemplificada pela Figura 6, que apresenta notas fiscais com valor específico de NCM (NCM = “22011000”). Esse valor de NCM possui a seguinte descrição oficial resumida: “Bebidas, líquidos alcoólicos e vinagres. Águas, incluindo as águas minerais, naturais ou artificiais, e as águas gaseificadas; gelo e neve”. Entretanto, pode-se notar que existem notas fiscais com esse NCM cujas descrições informadas não correspondem à descrição oficial citada, por exemplo, “guardanapo naxim

(a) (b)

Figura 4. Questão Q1: visualização das notas cujos pontos correspondentes são coloridos (a) pelo valor do NCM; (b) pelo valor do CST.



(a) (b)

Figura 5. Respondendo a questão Q2; (a) Seleção de pontos que representam notas fiscais com valor CST "Não informado"; (b) Visualização das notas selecionadas, cujos pontos são coloridos pelo rótulo de consistência.

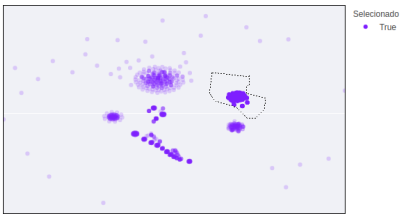
50 guardanapos" e "coifa lado roda". Esses casos são considerados inconsistências.

Finalmente, a investigação da questão Q4 pode ser exemplificada por meio da execução do algoritmo *K-Means* com distintos valores de K, a serem definidos pelo especialista, visando identificar um agrupamento com descrições relacionadas a distintos produtos ou a subcategorias de um produto nas visualizações. A Figura 7 exemplifica o caso de um agrupamento que possui notas fiscais cujas descrições informadas possuem termos que fazem referência a mais de um produto, caracterizando uma inconsistência.

5. Conclusão

Esta pesquisa apresentou uma abordagem inovadora para a detecção de casos de inconsistência em notas fiscais, enfrentando o desafio de lidar com o grande volume de notas geradas diariamente. Visando auxiliar os especialistas na análise das NFC-es, o sistema *web NFViz* combinou duas abordagens: rotulação automática a partir da similaridade de textos, e visualização exploratória. Assim, integrando automatização e recursos interativos para fornecer mais efetividade e controle dos especialistas no processo de auditoria.

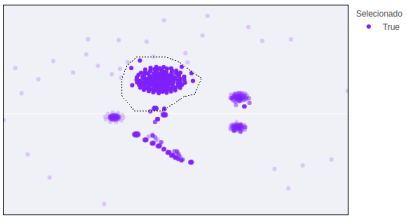
Como trabalhos futuros, há oportunidades para aprimorar o desempenho do sistema antes de disponibilizá-lo e explorar conjuntos de dados abertos relacionados à fraude fiscal. Também é fundamental ampliar o esforço de validação do *NFViz*, por meio da avaliação qualitativa da execução das tarefas no sistema pelos especialistas.



(a)

DESCRICAO	NCM	CST
AGUA MINERAL CRYSTAL C GAS 500ML	22011000	60 - ICMS taxado por substituição
AGUA MINERAL (IFOOD)	22011000	60 - ICMS taxado por substituição
AGUA MINERAL INDAIA C/GAS 1.5L	22011000	60 - ICMS taxado por substituição
AGUA MINERAL LIA COPO 200MLDE:1,63	22011000	60 - ICMS taxado por substituição
AGUA MINERAL LIA COPO 200MLDE:1,63	22011000	60 - ICMS taxado por substituição

(b)



(c)

DESCRICAO	NCM	CST
MANTEIGA COM SAL ITALAC	22011000	Não informado
COIFA LADO RODA	22011000	60 - ICMS taxado por substituição
GUARDANAPO MAXIM 50 GUARDANAPOS	22011000	Não informado
FILE BOVINO FAMILIA	22011000	Não informado
HIDRAT PAIXAO 200ML ONLY YOU	22011000	Não informado

(d)

Figura 6. Respondendo a questão Q3; (a) Seleção de notas similares (agrupadas na visualização); (b) Descrição informada, valores de NCM e de CST de algumas notas selecionadas; (c) Seleção de notas similares; (d) Descrição informada, valores de NCM e de CST de algumas notas selecionadas.

Referências

Bolton, R. and Hand, D. (2001). Unsupervised profiling methods for fraud detection. *Conference on Credit Scoring and Credit Control*, 7.

Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

Cox, M. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of Data Visualization*, pages 315–347.

Di Oliveira, V., Weigang, L., and Filho, G. (2022). Eleven data-set: A labeled set of descriptions of goods captured from brazilian electronic invoices. In *18th International Conference on Web Information Systems and Technologies*, pages 257–264.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Hajek, P. and Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods. *Knowledge-Based Systems*, 128:139–152.

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). *Natural language processing: state of the art, current trends and challenges*, pages 1573–7721.

Kieckbusch, D., Filho, G., Di Oliveira, V., and Weigang, L. (2021). Scan-nf: A cnn-based system for the classification of electronic invoices through short-text product description. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies*, pages 501–508.

Kieckbusch, D., Filho, G. P. R., Oliveira, V. D., and Weigang, L. (2020). Towards intelligent processing of electronic invoices: The general framework and case study of short

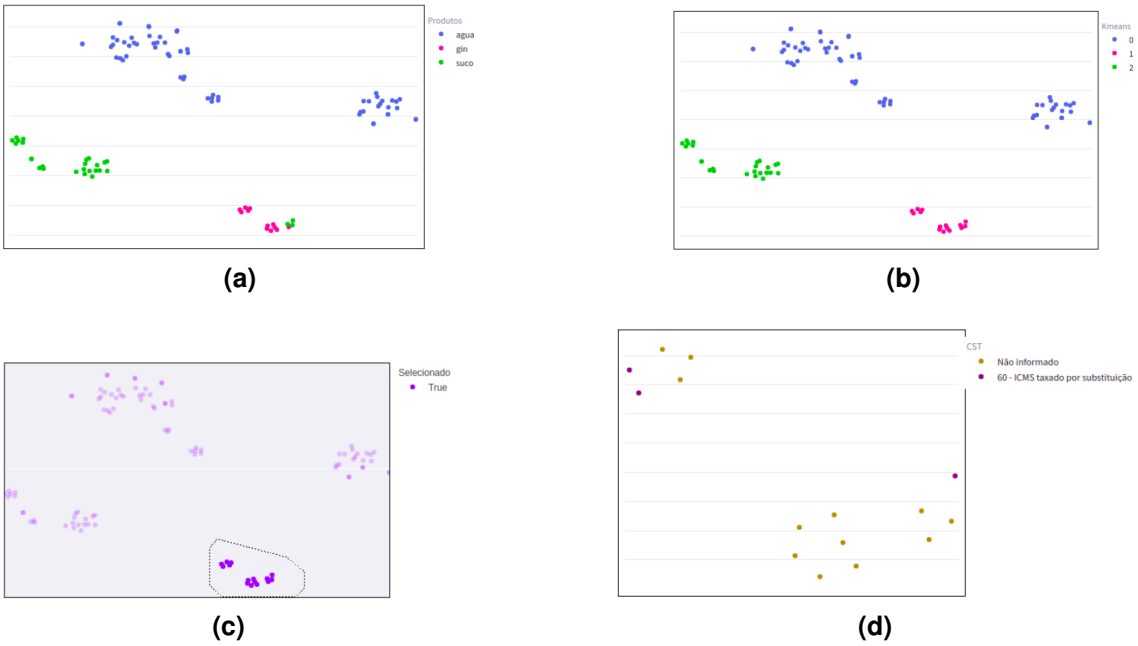


Figura 7. Respondendo a questão Q4; (a) Visualização de notas fiscais que possuem os termos "agua mineral", "gin" ou "suco" na descrição informada; (b) Visualização das notas com os pontos coloridos por agrupamento, com $K = 3$ por serem 3 categorias de produtos; (c) Seleção na visualização; (d) Visualização das notas selecionadas coloridas por CST.

text deep learning in brazil. In *International Conference on Web Information Systems and Technologies*, pages 74–92.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196. PMLR.

Marinho, M., Oliveira, V., Neto, S., Weigang, L., and Borges, V. (2022). Visual analysis of electronic invoices to identify suspicious cases of tax frauds. In *International Conference on Information Technology & Systems*, pages 185–195. Springer.

Matos, T., de Macedo, J. A. F., and Monteiro, J. M. (2015). An empirical method for discovering tax fraudsters: A real case study of brazilian fiscal evasion. In *Proceedings of the 19th International Database Engineering & Applications Symposium*, page 41–48. Association for Computing Machinery.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Resck, L. E., Ponciano, J. R., Nonato, L. G., and Poco, J. (2023). LegalVis: Exploring and inferring precedent citations in legal documents. *IEEE Transactions on Visualization and Computer Graphics*, 29.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9.

Zha, Z. (2020). Taxaa: A reliable tax auditor assistant for exploring suspicious transactions. In *Companion Proceedings of the Web Conference 2020*, pages 240–244.