Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

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

Mayara Chew Marinho

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Vinícius Ruela Pereira Borges

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

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

Mayara Chew Marinho

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Vinícius Ruela Pereira Borges (Orientador)
CIC/UnB

Prof. Dr. Li Weigang
CIC/UnB

Prof.a Dr.a Thais Carvalho Valadares Rodrigues
EST/UnB

Prof. Dr. João Luiz Azevedo de Carvalho
Coordenador do Curso de Engenharia da Computação

Brasília, 20 de julho de 2023

# Dedicatória

Dedico esta monografia ao Luis, que foi a pessoa que mais me ajudou durante esses 5 anos de graduação. Agradeço todo o apoio que você me deu, sempre me incentivando a seguir em frente, mesmo nos momentos mais difíceis. Não posso deixar de agradecer ao meu orientador, o Prof. Dr. Vinícius Borges *(que é simplesmente o melhor professor que eu já tive)*, por me guiar ao longo dos 3 anos de pesquisa, compartilhando seus conhecimentos e me motivando a desenvolver este estudo. Sou extremamente grata pelos seus ensinamentos e por todo o tempo que dedicou a mim. Também dedico este trabalho à minha família, à empresa júnior CJR e a todos que estiveram ao meu lado, me apoiando e me incentivando ao longo dessa jornada.

Gratidão!

# Agradecimentos

O presente trabalho foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e também recebeu apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

# Resumo

A fiscalização e a detecção de fraudes fiscais têm sido um desafio significativo devido à grande quantidade de notas fiscais geradas diariamente. Nesse contexto, esta pesquisa propõe duas abordagens para auxiliar os especialistas na tarefa de auditoria, utilizando o conjunto de dados não rotulado das Notas Fiscais Eletrônicas do Consumidor do Distrito Federal. A primeira abordagem é baseada em rotulação automática por meio de similaridade de textos para a detecção de casos suspeitos de fraude, e avaliação da reprodutibilidade desses rótulos por Aprendizado de Máquina, utilizando a distância cosseno e a Edit Distance, e as formas de representação de textos Word2vec, Doc2vec, Transformer Distiluse Multilingual e BERT. A segunda abordagem é baseada em visualização interativa utilizando TF-IDF e similaridade em conjunto com as técnicas MDS, t-SNE e UMAP para a análise visual dos dados e K-Means para a definição de agrupamentos. Os melhores resultados de avaliação automática foram obtidos com rótulos criados pela Edit Distance e os de visualização foram obtidos com a combinação da distância euclideana e cosseno, t-SNE e K-Means. Foi criada uma ferramenta *web* interativa de visualização, na qual os especialistas podem explorar as notas fiscais e obter informações relevantes para a otimização do processo de detecção de inconsistências em notas fiscais.

**Palavras-chave:** visualização, detecção de fraudes, notas fiscais, similaridade, estratégias de posicionamento de pontos, análise visual

# Abstract

Fraud detection and prevention in fiscal documents have become time-consuming tasks due to the increasing number of daily issued electronic invoices that need to go through a manual audit process. In this context, this research proposes two approaches to assist specialists in the audit task, using the Consumer Electronic Invoices dataset. The first approach is based on automatic labeling by text similarity in order to identify suspected cases of fraud, and evaluation of reproducibility by Machine Learning, using cosine distance and Edit Distance, and text representation techniques such as Word2vec, Doc2vec, Transformer Distiluse Multilingual and BERT. The second approach is based on visualization using TF-IDF, similarities and techniques such as MDS, t-SNE and UMAP for the visual analysis and K-Means for clusters definition. Best results were obtained with labels created by Edit Distance and the visualization ones were obtained with the combination of Euclidean and Cosine distance, t-SNE and K-Means. An interactive visualization tool *web* was created, in which specialists can explore invoices and obtain relevant information for optimizing the process of inconsistencies detection.

**Keywords:** visualization, fraud detection, invoices, similarity, point placement strategies, visual analysis

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

**BERT** Bidirectional Encoder Representations from Transformers.

**CBOW** Continuous Bag of Words.

**CST** Código de Situação Tributária.

**DBOW** Distributed Bag of Words.

**FN** Falsos negativos.

**FP** Falsos positivos.

**ICMS** Imposto sobre Circulação de Mercadorias e Serviços.

**IDF** Inverse document frequency.

**LSTM** Long Short-Term Memory.

**MDS** Multidimensional Scaling.

**NCM** Nomenclatura Comum do Mercosul.

**NF-e** Nota Fiscal Eletrônica.

**NFC-e** Nota Fiscal Eletrônica do Consumidor.

**PCA** Principal Component Analysis.

**RNN** Redes Neurais Recorrentes.

**SMOTE** Synthetic Minority Oversampling Technique.

**t-SNE** t-Distributed Stochastic Neighbor Embedding.

**TF** Term frequency.

**TF-IDF** Term Frequency–Inverse Document Frequency.

**UMAP** Uniform Manifold Approximation and Projection.

**VN** Verdadeiros negativos.

**VP** Verdadeiros positivos.

# Capítulo 1

# Introdução

A Lei Brasileira nº 4.729/65 trata de questões relacionadas à fraude fiscal, exigindo que os comerciantes registrados declarem corretamente todas as informações referentes às notas fiscais, evitando omissões, preenchimento incorreto de documentos oficiais, a criação de documentos falsos, dentre outras práticas fraudulentas. Apesar da existência desse regulamento, a fiscalização por parte dos órgãos reguladores é um processo desafiador, uma vez que são geradas aproximadamente 5,8 milhões de notas fiscais diariamente, segundo o Ministério da Fazenda[1]. Essa grande quantidade de notas fiscais geradas diariamente torna inviável a análise manual e minuciosa realizada pelos auditores para identificar casos suspeitos de fraude.

Tendo isso em vista, uma alternativa possível seria o acréscimo de um procedimento prévio à análise individual dos casos, visando filtrar conjuntos de notas inconsistentes a serem analisadas posteriormente pelos auditores. Na literatura, existem diversos trabalhos que suportam essa linha de raciocínio por meio da criação de ferramentas visuais e aplicação de técnicas de agrupamento, que encurtam o tempo de decisão de especialistas e permitem a ampliação da quantidade de dados a serem analisados [6] [5].

Nesse contexto, a detecção de fraudes é uma tarefa que pode ser abordada por métodos não supervisionados para identificar agrupamentos por similaridade nos atributos [7], juntamente com técnicas de redução da dimensionalidade e visualização interativa. Na literatura existem estudos de caso abordando diferentes áreas, como a detecção de fraudes financeiras com o algoritmo K-Means [4], a criação de um assistente de auditoria fiscal para explorar transações suspeitas [5], a utilização de visualizações interativas para compreender propriedades de *embeddings* vetoriais [8] e um sistema web para auxiliar especialistas em direito na identificação de citações a precedentes vinculantes em documentos legais [6]. No entanto, existem lacunas a serem preenchidas no que se refere

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

à detecção de fraudes em notas fiscais eletrônicas do Distrito Federal, o que motivou a proposição dessa pesquisa.

O uso de tecnologias semelhantes às citadas anteriormente, em conjunto com a experiência prática de auditores fiscais e o conhecimento adquirido no decorrer dessa profissão podem gerar resultados efetivos na busca por inconsistências. Tendo isso em vista, os especialistas mapearam como um indício de fraude o fato de valores de Código de Situação Tributária (CST) e descrições não serem condizentes ao valor de Nomenclatura Comum do Mercosul (NCM) registrado. Portanto, a utilização de ferramentas que facilitem a identificação dessas discrepâncias pode proporcionar aos auditores fiscais um suporte para o aprimoramento de suas atividades de auditoria.

Nesse contexto, o objetivo principal desta pesquisa é desenvolver uma abordagem que permita a detecção de casos passíveis de fraude por meio da utilização de Processamento de Linguagem Natural e Visualização. Os objetivos específicos da pesquisa são citados a seguir:

- Explorar representações estruturadas para as notas fiscais;
- Propor uma medida de similaridade para ser empregada na comparação de notas fiscais;
- Criar uma rotulação de notas fiscais para viabilizar a identificação automática de notas inconsistentes;
- Investigar e desenvolver o processo de rotulação e visualização;
- Investigar técnicas para avaliar e validar a metodologia proposta;
- Desenvolver uma ferramenta que implemente a metodologia proposta.

A proposta consiste em criar uma ferramenta *web* interativa de visualização, que possibilite a análise visual das distâncias de similaridade das notas fiscais, bem como a visualização de agrupamentos e de estatísticas relevantes para auxiliar a identificação de inconsistências nos dados. Espera-se que essa abordagem visual, analítica e inteligente proporcione aos auditores uma nova perspectiva na detecção de fraudes fiscais, otimizando seus esforços e contribuindo para a eficiência do processo de fiscalização.

As principais contribuições descritas neste projeto de pesquisa são:

- Uma medida de similaridade para comparar notas fiscais eletrônicas constituídas por atributos textuais e categóricos;
- Rótulos de inconsistência criados para a definição automática de notas fiscais suspeitas de fraude;

- O uso e a avaliação de estratégias de posicionamento de pontos para visualização de notas fiscais, como a Multidimensional Scaling [9], a t-Distributed Stochastic Neighbor Embedding [10] e a Uniform Manifold Approximation and Projection [11];
- Uma abordagem baseada em visualização interativa que incorpora algoritmos de similaridade e agrupamento para identificar notas fiscais inconsistentes;
- Uma ferramenta visual interativa que aplica a metodologia proposta e pode otimizar o processo de tomada de decisão em auditorias fiscais.

# Capítulo 2

# Fundamentos

Este capítulo é dedicado a explorar os fundamentos teóricos que sustentam os experimentos realizados e as técnicas empregadas ao longo da pesquisa. A Seção 2.1 descreve os conjuntos de dados utilizados nos experimentos e suas particularidades. A Seção 2.2 detalha sobre os fundamentos de Processamento de Linguagem Natural e suas técnicas que serão empregadas na metodologia para extrair informações de textos curtos presentes nas notas fiscais, como também as medidas de dissimilaridade para dados com atributos de diferentes tipos. A Seção 2.3 detalha os princípios das tarefas de aprendizado de máquina que serão empregadas nesta pesquisa, classificação baseada em redes neurais artificiais e agrupamento de textos, além das estratégias de avaliação de desempenho de modelos de classificação. Por fim, a Seção 2.4 detalha as técnicas de visualização de textos baseadas na redução de dimensionalidade e de avaliação da qualidade de visualizações.

## 2.1 Notas Fiscais Eletrônicas do Consumidor

Segundo a Receita Federal, uma Nota Fiscal Eletrônica do Consumidor NFC-e é definida como um *"documento de existência apenas digital, emitido e armazenado eletronicamente, com o intuito de documentar as operações comerciais de venda presencial ou venda para entrega em domicílio a consumidor final (pessoa física ou jurídica) em operação interna e sem geração de crédito de ICMS ao adquirente."*[1]. A NFC-e foi implantada com um foco em consumidores finais, visando atender a demanda do varejo.

Nessa pesquisa, foi utilizado um conjunto não público de notas fiscais denominado Nota Fiscal Eletrônica do Consumidor (NFC-e). Os dados estão representados em formato tabular, em que as colunas são associadas aos atributos e linhas que descrevem as instâncias das notas fiscais. Este conjunto de dados foi criado em 2019 e contém 359.494 instâncias composta por três atributos como descritos na Tabela 2.1.

[1] `http://sped.rfb.gov.br/pagina/show/1519`, acessado em 12 de abril de 2023.

Tabela 2.1: Descrição dos atributos do conjunto de dados NFC-e.

| **Atributo** | **Tipo** | **Significado** |
|---|---|---|
| DESC | Texto | Descrição do produto feita pelo vendedor. Esse texto é livre, portanto não possui um padrão previamente definido. |
| NCM | Categórico | Nomenclatura Comum do Mercosul. Este código é padronizado para cada categoria de produto comercializado. |
| CST | Categórico | Código de Situação Tributária, utilizado como base para o cálculo do Imposto sobre Circulação de Mercadorias e Serviços (ICMS). |

No conjunto de dados empregado nessa pesquisa, as notas fiscais podem apresentar os CSTs como mostra a Tabela 2.2. Por fim, a Figura 2.1 exemplifica algumas instâncias do conjunto de dados da NFC-e, sendo que os valores ausentes de CST são mostrados como "NaN".

Tabela 2.2: Códigos de Situação Tributária.

| **CST** | **Significado** |
|---|---|
| 00 | Completamente taxado |
| 20 | Possui uma taxa base |
| 40 | Isento |
| 41 | Não taxado |
| 60 | ICMS taxado por substituição |

| | DESCRIÇÃO | NCM | CST |
|---|---|---|---|
| 0 | ESCORREDOR ARROZ N22 ALGO LUMINIO | 76151000 | 0.0 |
| 1 | ARQUIVO MORTO PRATICO AMARELO | 42021210 | NaN |
| 2 | ZUCCARDI POLIGONOS MALBEC 2016 | 22042100 | NaN |
| 3 | ALUNO ERICK RODRIGUES MAGALHAESPRODUTO GRAMEST... | 49019900 | NaN |
| 4 | -[002]-BLUSA 0125855050049 | 61061000 | 0.0 |
| 5 | PRADA VPR 62V C SVF-1O1 T 57 | 90031910 | NaN |
| 6 | MIOFLEX A 30 C D... | 30049045 | 60.0 |
| 7 | IOG C MINAS VITAMINAS 180G | 4031000 | 0.0 |

Figura 2.1: Atributos do conjunto de dados NFC-e a serem utilizados nesse projeto. (Fonte: Autoria própria).

Nesta pesquisa, será utilizada a descrição oficial concatenada do NCM de 2019, como uma forma de categoria (*ground-truth*) em relação à descrição real definida pelo comerciante no momento de registro da nota fiscal. As descrições oficiais podem ser conferidas na Figura 2.2, tanto no formato não concatenado, identificando a descrição por partes do

código NCM, quanto no formato concatenado, que junta as descrições referentes a cada parte do código NCM em uma descrição única. Esta última é denominada descrição oficial e será utilizada em experimentos posteriores desta pesquisa, pelo fato de representar a descrição de um código NCM completo, assim como é registrado no momento de geração da nota fiscal.

| | CÓDIGO | DESCRIÇÃO | DESCRIÇÃO CONCATENADA |
|---|---|---|---|
| 0 | 01 | Animais vivos. | 01: Animais vivos. |
| 1 | 01.01 | Cavalos, asininos e muares, vivos. | 01: Animais vivos. \n01.01: Cavalos, asininos ... |
| 2 | 0101.2 | - Cavalos: | 01: Animais vivos. \n01.01: Cavalos, asininos ... |
| 3 | 0101.21.00 | -- Reprodutores de raça pura | 01: Animais vivos. \n01.01: Cavalos, asininos ... |
| 4 | 0101.29.00 | -- Outros | 01: Animais vivos. \n01.01: Cavalos, asininos ... |
| 5 | 0101.30.00 | - Asininos | 01: Animais vivos. \n01.01: Cavalos, asininos ... |
| 6 | 0101.90.00 | - Outros | 01: Animais vivos. \n01.01: Cavalos, asininos ... |
| 7 | 01.02 | Animais vivos da espécie bovina. | 01: Animais vivos. \n01.02: Animais vivos da e... |

Figura 2.2: Dicionário de descrições baseadas no código NCM. A coluna da esquerda indica o código NCM, a coluna do meio indica a descrição não concatenada e a coluna da direita indica a descrição concatenada, também denominada descrição oficial. A coluna da descrição concatenada está abreviada com "..." por conta de seu tamanho. (Fonte: Autoria própria).

## 2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural é uma área de estudo que consiste em analisar a linguagem humana em seus diversos idiomas com o objetivo de permitir a interpretação e a geração automática de textos, além de outras tarefas correlacionadas, como análise de sentimentos [12] [13], tradução de idiomas [14], sumarização de textos [15], etc.

No conjunto de dados NFC-e, cada nota fiscal contém uma descrição, definida por um texto curto descrevendo informações relacionadas ao produto como, por exemplo, seu tipo, tamanho, volume, marca (ou fabricante) e peso. Essa descrição não possui um estilo de escrita padronizado e as informações contidas podem variar entre as notas fiscais, tornando desafiadoras as tarefas de reconhecimento de padrões nesses textos.

### 2.2.1 Pré-processamento

O pré-processamento é responsável pela limpeza e pelo tratamento de dados, sendo uma etapa fundamental para garantir resultados confiáveis em análises posteriores. Em textos, esse tratamento pode incluir a remoção de caracteres especiais e números, conversão para letras maiúsculas ou minúsculas, aplicação de expressões regulares, remoção de

*stopwords*, lematização, *tokenização*, dentre outros. Ressalta-se a necessidade de aplicar o pré-processamento conforme a natureza dos textos, estilo de escrita textual, idioma e também a tarefa alvo.

### Tokenização

A *tokenização* é o processo que divide um texto em suas unidades menores e sequenciais, denominadas termos (*tokens*) [16]. O critério de divisão de um texto visando a obtenção dos seus termos pode variar dependendo da tarefa alvo, sendo comumente considerados espaços em branco e símbolos de pontuação. Assim, pode-se ajustar o nível de granularidade dos *tokens* como, por exemplo, ao nível de caracteres, palavras ou frases.

A Figura 2.3 ilustra o processo de *tokenização*, em que a sequência de *tokens* foi obtida a partir da divisão da sentença de entrada de acordo com os espaços em branco. Pode-se verificar que a ordem natural dos termos foi preservada, uma vez que os termos do texto de entrada possuem uma relação de dependência. Esse aspecto é importante em um modelo de linguagem, que determina a próxima palavra considerando a semântica e o contexto obtido das palavras processadas anteriormente.

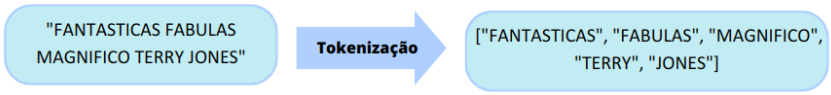



Figura 2.3: Ilustração do processo de *tokenização*. (Fonte: Autoria própria).

### Remoção de *Stopwords*

Um dos problemas que podem surgir ao lidar com textos extensos é o alto custo computacional gerado pela manipulação de matrizes multidimensionais, resultado da grande quantidade de termos que podem existir em um texto. Este problema pode ser amenizado ao remover as *stopwords*, palavras que agregam pouca informação relevante para o entendimento semântico da frase, visando assim, a redução das dimensões e consequente redução de custos computacionais. A Figura 2.4 exemplifica esse processo, com a remoção dos *stopwords* "AS" e "DO" da sentença original (à esquerda).

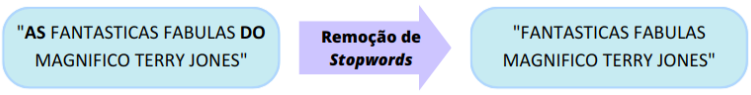



Figura 2.4: Demonstração do processo de remoção de *stopwords*. (Fonte: Autoria própria).

#### Lematização

A lematização é um processo que analisa o contexto da palavra e a converte em sua forma base anterior à inflexão, cuja base é conhecida como "lema" [17]. Isso permite a redução das palavras a formas reconhecíveis e significativas, mantendo a sua classe gramatical original. A lematização foi escolhida em vez do *stemming* porque ela reduz as palavras preservando a classe morfológica, diferente do *stemming*, que apenas remove o sufixo das palavras [16]. A Figura 2.5 exemplifica o processo de lematização.

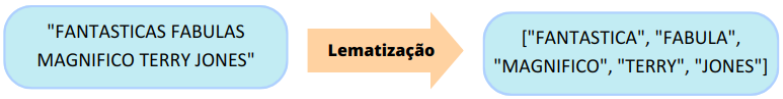



Figura 2.5: Ilustração do processo Lematização. (Fonte: Autoria própria).

### 2.2.2 Representação de Textos

Em tarefas que envolvem Processamento de Linguagem Natural, uma das etapas mais importantes após o pré-processamento é a utilização de uma representação do texto que mais se adapte ao contexto da tarefa alvo, sendo possível melhorar consideravelmente o desempenho da tarefa por meio dessa escolha [18].

Um dos métodos de representação mais simples é o One-Hot-Encoding, que consiste na criação de um vetor para cada palavra contendo o valor 1 nas posições em que ela aparecer e 0 nas outras. Embora o One-Hot-Encoding seja fácil de entender e de implementar, ele pode gerar representações de alta dimensionalidade em textos extensos, o que pode levar a problemas de eficiência computacional e gerar um espaço esparso [16]. Pela desvantagem citada, a aplicação dele é mais comum em variáveis categóricas.

### 2.2.3 Term Frequency-Inverse Document Frequency

O Term Frequency–Inverse Document Frequency (TF-IDF)) é uma forma de representação que indica a importância de uma palavra de acordo com o contexto, baseada na frequência em que ela é utilizada em um conjunto de documentos de texto [19], como exemplificam as Equações (2.1), (2.2) e (2.3).

$$TF(t) = \frac{\text{Nº de vezes que o termo t aparece no texto}}{\text{Nº total de termos no texto}} \tag{2.1}$$

$$IDF(t) = log\left(\frac{\text{Nº de textos no corpus}}{\text{Nº de textos no dataset que contem o termo t}}\right), \tag{2.2}$$

em que *texto* é o texto a ser analisado e o *dataset* é o conjunto de todos os textos da amostra de dados analisada. Por fim, a Equação (2.3) exemplifica o TF-IDF.

$$TF - IDF(t) = TF(t) \times IDF(t) \tag{2.3}$$

O valor TF-IDF de uma palavra aumenta conforme a frequência dela no texto aumenta, mas diminui conforme a frequência dela no *dataset* com todos os textos aumenta. Dessa forma, o TF-IDF consegue encontrar as palavras mais importantes em cada um dos textos e gerar um vetor esparso de alta dimensionalidade, que é fundamental para tarefas de sumarização e classificação.

## 2.3 Aprendizado de Máquina

A área de Aprendizado de Máquina abrange diversas tarefas, cada uma com suas particularidades, como classificação, agrupamento, regras de associação, regressão, dentre outras. Nesta pesquisa, a classificação será considerada na tarefa de identificação automática de inconsistências em notas fiscais, enquanto o agrupamento (*clustering*) de dados é empregado juntamente com o processo de visualização de dados.

Classificação é definida pelo processo no qual um conjunto de dados previamente rotulado é utilizado como treinamento para a construção de um modelo que classifica novas instâncias [20] cuja qualidade do resultado pode ser medido por uma métrica, que será explicada posteriormente na Seção 2.3.6. A literatura apresenta diversas técnicas para classificação de dados. Nesta pesquisa, como as notas fiscais são compostas por informações textuais, o foco será no estudo dos modelos de classificação baseados em redes neurais.

### 2.3.1 Redes Neurais Artificiais

Redes Neurais foram criadas com o objetivo de simular o processo de aprendizado biológico que ocorre no sistema nervoso de seres humanos. Assim, é possível definir componentes simplificados que remetem à ideia de neurônios, axônios, dendritos e sinapses nervosas que passam a compor uma rede neural artificial [1].

O tipo mais simples de rede neural é a Perceptron. Seja $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ um conjunto de dados, em que cada instância multidimensional $\mathbf{x_i}$ é descrita por $M$ atributos $[a_{i,1}, \ldots, a_{i,M}] = a_i$. A rede Perceptron é um modelo que recebe uma entrada $\mathbf{x_i}$ e um rótulo $y_i$ e produz uma saída $o_i$ como mostra a Equação (2.4):

$$o_i = b + \sum_{j=1}^{M} a_{i,j} w_j, \tag{2.4}$$

em que $w_j$ é um parâmetro relacionado ao peso, que pondera o valor do atributo $a_{i,j}$, enquanto $b$ é um parâmetro relacionado ao viés (*bias*), que é ajustado automaticamente pela rede neural. No caso de um problema de classificação binária, a saída $o_i$ pode ser configurada para produzir valores binários por meio de uma função de ativação. A Equação (2.5) exemplifica uma função de ativação baseada no sinal da saída $o_i$:

$$\hat{y}_i = \begin{cases} 0, & o_i < 0 \\ 1, & o_i \geq 0. \end{cases} \tag{2.5}$$

Assim, as redes neurais são apropriadas para tarefas de aprendizado supervisionado, devido à capacidade de fazer predições.

O aprendizado de uma rede neural Perceptron ocorre por meio de um processo de treinamento, em que os parâmetros $\{w_1, \ldots, w_M\}$ e $b$ são ajustados conforme os padrões reconhecidos pela rede em relação ao conjunto de dados de referência. O treinamento de uma rede neural consiste em minimizar a função objetivo descrita na Equação (2.6):

$$E = \sum_{i=1}^{N} [y_i - \hat{y}_i]^2, \tag{2.6}$$

A partir da Equação (2.6), pode-se obter a regra para atualização dos pesos, dada pela Equação (2.7):

$$w_j^{h+1} = w_j^h + \lambda a_{i,j} (\hat{y}_i - y_i), \tag{2.7}$$

em que $\lambda$ é um hiperparâmetro denominado taxa de aprendizado, que pondera o valor de atualização dos pesos, e $h$ é o instante de tempo. Pode-se notar que os pesos são alterados apenas quando a saída $\hat{y}_i$ não é correta considerando o problema de classificação binária.

O treinamento da rede Perceptron se baseia em um processo iterativo visando minimizar o valor de $E$, em que a Equação (2.7) é repetidamente calculada enquanto um erro "aceitável" não for atingido. Assim, o processo de minimização pode demandar várias iterações, sendo que uma iteração completa, que resulta na passagem por todas as instâncias de treinamento, é denominada época.

As redes neurais artificiais compostas por várias camadas (*multilayer perceptron*) foram propostas para resolver problemas mais complexos. Essas redes empregam diversos neurônios do tipo Perceptron, que possuem conexões com os neurônios da camada anterior e da camada seguinte. A Figura 2.6 ilustra uma arquitetura multicamadas, em que é possível identificar a camada de entrada, a camada intermediária e a camada de saída. A

quantidade de camadas intermediárias e a quantidade de neurônios em cada camada são hiperparâmetros que devem ser ajustados de maneira separada em relação ao treinamento.

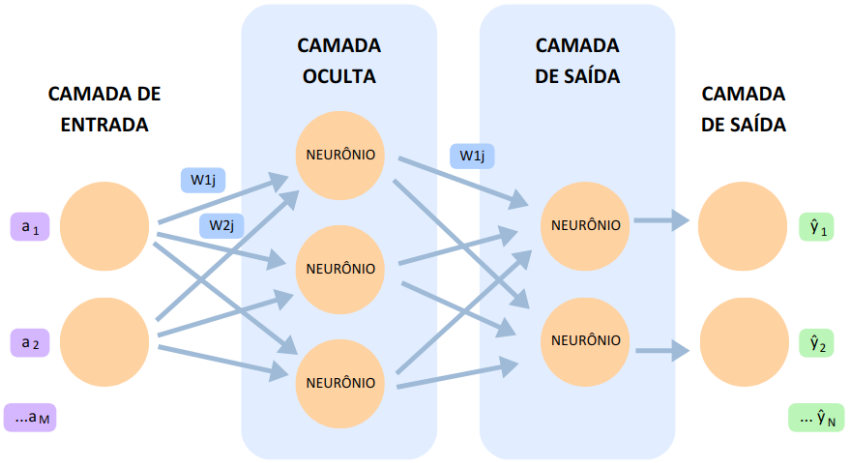



Figura 2.6: Arquitetura de uma rede neural multicamadas. (Fonte: Adaptação do livro [1]).

O treinamento de uma rede neural multicamadas é efetuado pelo algoritmo *Backpropagation*, composto por duas fases: propagação e retro-propagação. Na fase de propagação, os dados de entrada são passados pela rede, gerando-se as saídas, que são utilizadas para o cálculo de uma função de perda (*loss function*). Na fase de retropropagação, ocorre o cálculo do gradiente da função de perda, que é utilizado para ajustar os parâmetros da rede (pesos dos neurônios) desde a camada de saída até a camada de entrada. Para viabilizar os cálculos dos gradientes, é comum utilizar funções de ativação diferenciáveis nas camadas ocultas, como a função sigmóide descrita na Equação (2.8):

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2.8}$$

ou a unidade linear retificada (*rectified linear unit* - *ReLU*), como mostra a Equação (2.9):

$$relu(x) = \begin{cases} x, & x > 0 \\ 0, & \text{caso contrário,} \end{cases} \tag{2.9}$$

e, por fim, a tangente hiperbólica, exemplificada pela Equação (2.10):

$$tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}. \tag{2.10}$$

### 2.3.2 Redes Neurais Recorrentes

No que concerne ao processo de classificação de textos, modelos baseados em redes neurais multicamadas podem ser considerados. Entretanto, eles não levam em consideração a

quantidade variável de palavras nos textos de entrada, como também não consideram a ordem das palavras em um texto, o que pode ser fundamental para a preservação do contexto. Dessa forma, foram criadas Redes Neurais Recorrentes (RNN), que incorporam ambas melhorias por meio da correspondência um-a-um entre as camadas, permitindo a repetição da arquitetura a cada etapa [21]. Esse processo de correspondência entre camadas e de propagação do erro das RNNs é exemplificado pela Figura 2.7, considerando que uma instância $x_i$ seja um texto composto por $n$ termos $\{t_1, ..., t_n\}$.

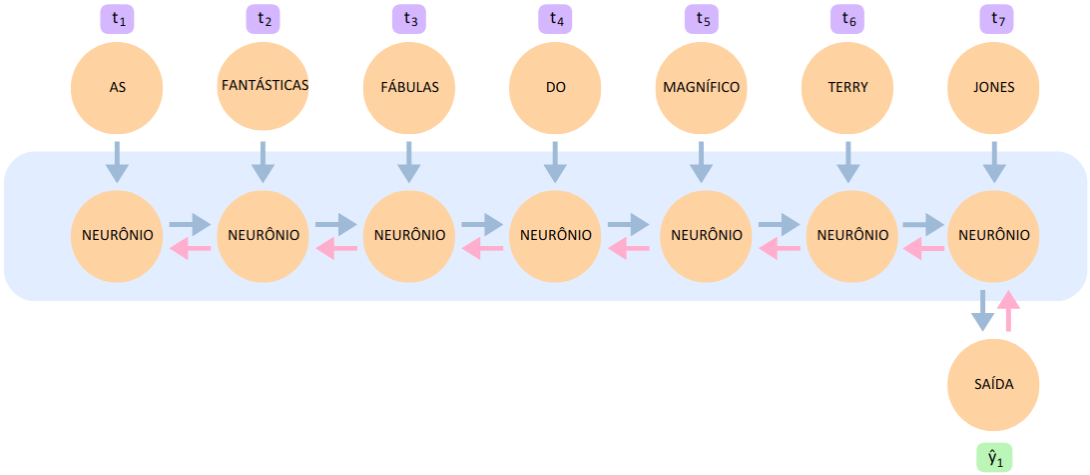



Figura 2.7: Ilustração do processo de correspondência um-a-um entre as camadas e a propagação do erro (*Backpropagation*). (Fonte: Adaptação do livro [1]).

Ressalta-se a natureza "short-term memory" destas redes neurais, devido à existência de conexões de feedback para armazenar representações de dados de entrada recentes [22]. Em termos práticos, as RNNs tradicionais favorecem a memorização de curto prazo, em seus modelos, e desfavorecem de forma proporcional ao tempo as memórias de estados anteriores.

### Long Short-Term Memory

Variações de RNNs, como a Long Short-Term Memory (LSTM), são amplamente utilizadas na literatura [23] [24] tendo em vista a interpretabilidade de textos a partir do equilíbrio entre a preservação da memória de longo prazo e a influência da memória de curto prazo e a alta capacidade de generalização [22]. Estes avanços trazem benefícios principalmente em tarefas que envolvem dados sequenciais ou de aspecto temporal. A arquitetura de uma LSTM é exemplificada pela Figura 2.8 e os principais conceitos envolvidos estão definidos abaixo.

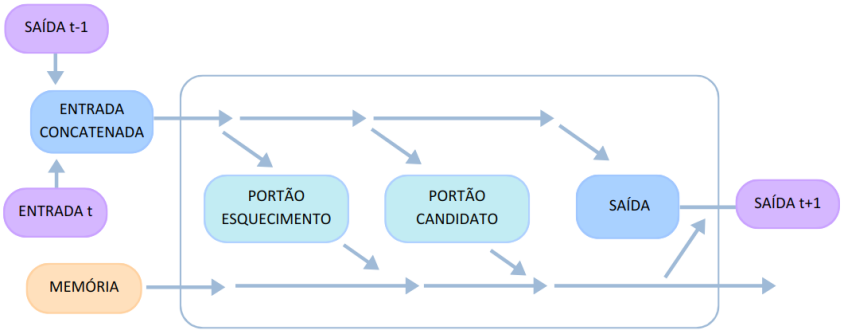



Figura 2.8: Arquitetura da camada LSTM na etapa $t$. (Fonte: Adaptação do livro [1]).

- **Portão Esquecimento (*Forget Gate*):** responsável por aplicar uma função sigmóide para selecionar as informações antigas a serem descartadas da célula de memória.
- **Portão de Entrada (*Input Gate*):** responsável por aplicar uma função sigmoidal para calcular a ativação e a função tangente hiperbólica para selecionar novas informações a serem adicionadas à célula de memória.
- **Portão de Saída (*Output Gate*):** responsável por definir a saída uma função sigmoidal, juntamente com a entrada atual e a memória atualizada.

### 2.3.3 Word Embeddings

Outras formas de representação estruturada de textos que ganharam notoriedade nos últimos anos são os *Word Embeddings*. Essas representações numéricas contínuas são geradas por meio de técnicas de transformação de dados, na maioria dos casos discretos, em espaços vetoriais [17]. Nessa representação, cada termo é estruturado como um vetor de tamanho fixo que armazena informações semânticas e de similaridade com os demais termos do conjunto de textos (*corpus*) associado [25].

A representação de palavras por meio de vetores permite o entendimento da correlação semântica entre elas [2], como exemplificam as Equações (2.11) e (2.12).

$$\vec{rei} - \vec{homem} + \vec{mulher} = \vec{rainha} \tag{2.11}$$

$$\vec{tio} - \vec{homem} + \vec{mulher} = \vec{tia} \tag{2.12}$$

A Figura 2.9 ilustra as variações em gênero pelas setas azuis e em número pelas setas vermelhas, além de gerar uma visualização dos vetores presentes nas Equações (2.11) e (2.12).

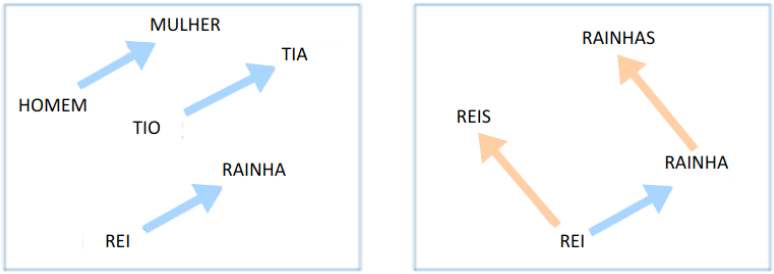



Figura 2.9: Representação vetorial de palavras com a variação de gênero e número. (Fonte: Adaptação do artigo [2]).

Outras formas de representação de palavras em um espaço vetorial também muito citadas na literatura são o Word2Vec [26] e o Doc2Vec [27], ambos utilizam arquiteturas baseadas em redes neurais, porém o Word2Vec calcula um vetor para cada palavra do texto e depois é aplicada uma operação de soma para gerar os vetores, já o Doc2Vec calcula um vetor para cada documento. Neste projeto foram utilizadas as arquiteturas Continuous Bag of Words (CBOW) [26] e Distributed Bag of Words (DBOW) [27] em cada um respectivamente, ambas são redes neurais que predizem uma palavra conforme o contexto na qual ela está inserida, sendo que a primeira possui uma abordagem continua e a segunda uma abordagem distribuída.

Em tarefas de processamento de linguagem natural, os *embeddings* são representações de tamanho fixo, que muitas vezes são obtidos a partir da saída de redes neurais profundas treinadas em grandes coleções de textos.

### Bidirectional Encoder Representations from Transformers

O *Bidirectional Encoder Representations for Transformers* (Bidirectional Encoder Representations from Transformers (BERT)) é um modelo de representação de linguagem que analisa os contextos à esquerda e à direita de uma palavra, por isso a denominação *bidirectional*. Esse é um ponto de diferenciação dele para a LSTM, visto que ela é unidirecional, logo, pode realizar predições em apenas uma direção. Por ser ter sido criado com base na arquitetura de um transformer, ele possui um mecanismo de atenção que permite acessar informações de tokens mais antigos junto com tokens atuais [17], processo denominado *random access*. Além disso, o BERT pode ser combinado com modelos de aprendizado supervisionado [28] e com um conjunto de dados de assunto específico no pré-treinamento [17] para melhorar o desempenho em tarefas específicas.

### 2.3.4 Classificação em Dados Desbalanceados

A ocorrência de desbalanceamento de dados é frequente em tarefas de classificação, na qual uma ou mais classes possuem poucas instâncias em comparação com as demais. Esse desequilíbrio pode ser prejudicial à construção de modelos, porque pode gerar um viés para as classes majoritárias, provocando um baixo desempenho na classificação das classes minoritárias.

Com o objetivo de amenizar o desbalanceamento, existem técnicas de *oversampling* e *undersampling*, que, ao aplicadas em determinadas classes, podem gerar mais instâncias semelhantes ou excluir algumas já existentes. A técnica de *undersampling* consiste em reduzir a quantidade de instâncias das classes majoritárias para diminuir a disparidade de instâncias com as classes minoritárias, enquanto o *oversampling* aplica o aumento da quantidade de instâncias das classes minoritárias por meio da duplicação ou da criação de novas instâncias.

Neste projeto, será aplicado o Synthetic Minority Oversampling Technique (SMOTE) para lidar com o desbalanceamento de dados consistentes e inconsistentes na classificação, ele será aplicado para aumentar a quantidade de instâncias com o rótulo de inconsistência e amenizar os efeitos da discrepância de quantidades das duas classes. O SMOTE é uma técnica de *oversampling* que cria novas instâncias entre duas instâncias de dados com o rótulo minoritário [29].

### 2.3.5 Detalhes sobre o Ajuste do Modelo

Nas subseções abaixo serão exemplificadas as técnicas utilizadas para amostragem, ajuste de hiperparâmetros e classificação multi-classe nos experimentos. Esses detalhes são importantes para ajustar o modelo aos dados e obter resultados confiáveis.

#### Amostragem

A amostragem é uma etapa que permite selecionar uma parte representativa da população para análise, evitando a utilização de toda a população. Essa etapa é essencial quando não é possível trabalhar com todos os dados da população, economizando tempo e recursos ao longo dos processos de coleta e análise.

Nesta etapa, deve-se ter atenção para selecionar uma amostra de forma adequada para garantir a validade e a representatividade dos resultados da amostra. Existem diferentes técnicas de amostragem, como a amostragem aleatória simples, em que cada amostra da população tem a mesma probabilidade de ser selecionado, e a amostragem estratificada, em que a população é dividida em grupos homogêneos formados por determinada

característica prévia antes que seja aplicada a amostragem aleatória simples nos grupos desejados.

#### Método Holdout

O método de amostragem para treinamento do modelo de classificação utilizado nessa pesquisa é o Holdout. Ele consiste em dividir o conjunto de dados inicial em 3 partições, sendo uma para treinamento do modelo, uma para validação de hiperparâmetros e a outra para obtenção de resultados de teste. Esta divisão aumenta a confiabilidade nos resultados de desempenho do modelo, visto que os dados utilizados para treinamento não são utilizados na etapa de validação e de teste.

#### Ajuste de Hiperparâmetros

O ajuste de hiperparâmetros é crucial para que modelos de Aprendizado de Máquina entreguem o melhor resultado possível. As principais opções de ajuste são Grid Search e Random Search, o primeiro testa todas as combinações possíveis, enquanto a segunda utiliza um determinado número de combinações aleatórias de hiperparâmetros. Neste projeto será utilizada a técnica Random Search devido ao custo computacional mais baixo.

#### Classificação Multi-classe

Uma particularidade de ambos conjuntos de dados empregados nesta pesquisa é a existência de múltiplas classes. Por isso, essa tarefa se enquadra na categoria de classificação multi-classe, na qual o conjunto de dados tem mais de duas classes distintas. Nesse contexto, é necessário utilizar abordagens e algoritmos que sejam capazes de lidar com a complexidade inerente à classificação de múltiplas classes, visando a obtenção de resultados confiáveis.

### 2.3.6 Avaliação de Desempenho de Classificadores

As estratégias de avaliação de desempenho das técnicas de classificação permitem uma análise mais objetiva do desempenho de modelos de classificação, fornecendo métricas importantes para aprimorar e otimizar essas técnicas. Nas subseções abaixo serão exemplificados os procedimentos utilizados para a realização e avaliação dos experimentos.

#### Matriz de Confusão

A matriz de confusão é uma tabela construída a partir dos resultados obtidos após a classificação, que fornece uma visão geral do desempenho do modelo. Ela compara as

classes preditas pelo modelo com as classes verdadeiras, gerando as seguintes métricas: Verdadeiros positivos (VP), Verdadeiros negativos (VN), Falsos positivos (FP), Falsos negativos (FN).

- **VP:** quando a classe predita é *verdadeiro* e a real é *verdadeiro* também.
- **VN:** quando a classe predita é $falso$ e a real é $falso$ também.
- **FP:** quando a classe predita é *verdadeiro*, mas a real é $falso$.
- **FN:** quando a classe predita é $falso$, mas a real é *verdadeiro*.

A partir das métricas explicadas acima, é possível calcular também outras métricas de avaliação do modelo, como a acurácia, precisão, revocação e F1-score. Essas métricas são descritas nas seções subsequentes, considerando como base uma tarefa de classificação binária.

### Acurácia

A Acurácia mede a proporção de predições corretas em relação ao total de predições realizadas e muito utilizada para avaliar classificadores como exemplifica a Equação (2.13), porém, em conjuntos de dados desbalanceados, ela pode não refletir bem o desempenho do modelo [17]. Por exemplo, na maioria das tarefas relacionadas à detecção de fraude, poucas instâncias são fraudulentas, geralmente menos de 10%, então se o classificador definir que todos os dados do conjunto de teste são não fraudulentos temos uma alta acurácia, ao mesmo tempo que não identifica nenhuma amostra fraudulenta e consequentemente não é de muita serventia na tarefa de identificação de fraudes.

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.13}$$

### Precisão

A precisão mede a proporção de instâncias classificadas corretamente como positivas em relação ao total de instâncias classificadas como positivas, conforme a Equação (2.14), esse último caso considera tanto as classificações corretas quanto as incorretas. Esta métrica pode ser utilizada em situações nas quais precisamos minimizar os FP, ou seja, as predições positivas incorretas.

$$precisão = \frac{VP}{VP + FP} \tag{2.14}$$

#### Revocação

A revocação, também chamada de sensibilidade, mede a proporção de instâncias que foram corretamente classificadas como positivas dentre todas as instâncias positivas no conjunto de dados, conforme a Equação (2.15). Esta métrica pode ser utilizada quando precisamos minimizar os FN, ou seja, as predições negativas incorretas.

$$revocação = \frac{VP}{VP + FN} \tag{2.15}$$

#### F1-score

Calculada como a média harmônica entre precisão e revocação, a F1-score é usualmente utilizada para avaliar a qualidade da classificação de um modelo, e a sua fórmula é exemplificada pela Equação (2.16). O resultado da F1-score pode variar de 0 a 1, sendo que, quanto mais próximo de 1, melhor o desempenho do modelo.

Esta métrica é útil principalmente quando aplicada a um conjunto de dados desbalanceado, no qual possui uma distribuição não proporcional de dados em cada uma das classes. Isso ocorre porque esta métrica no modo *macro-averaged* é calculada por classe e a performance de cada uma possui a mesma importância no cálculo [17].

$$f1\_score = \frac{2 \times precisão \times revocação}{precisão + revocação} \tag{2.16}$$

No caso desta pesquisa, por lidar com dados desbalanceados em uma tarefa multiclasse, iremos utilizar a métrica F1-score para avaliar a performance da classificação. A acurácia não será utilizada tendo em vista a quantidade de dados pouco representativa de dados das classes minoritárias em relação às majoritárias, fato que pode gerar um resultado mais otimista que o real.

### 2.3.7 Agrupamento de Dados

O agrupamento de dados é uma técnica de aprendizado não supervisionado que tem como objetivo a criação de grupos para a identificação de instâncias semelhantes de acordo com a similaridade entre os seus atributos. O princípio básico consiste em agrupar instâncias semelhantes em grupos iguais e instâncias pouco semelhantes em grupos diferentes.

K-Means é um algoritmo de aprendizado não supervisionado que agrupa dados com características semelhantes em grupos [30], conforme descreve o Algoritmo 1. O processo de agrupamento utilizando o K-Means depende da definição prévia de um valor $K$, que indica o número de grupos a serem criados. A Figura 2.10 exemplifica a aplicação desse

algoritmo com $K = 3$ no espaço de alta dimensão, sendo que cada cor na visualização representa um agrupamento.

**Algorithm 1** Descrição do Algoritmo K-Means.

1: Definir um centroide aleatório para cada grupo;
2: Calcular a distância dos centróides para cada amostra;
3: Definir o grupo de cada amostra pela distância calculada, de forma que cada amostra pertença ao grupo mais próximo;
4: Calcular um novo centroide com base na média da posição de todas as amostras pertencentes ao grupo;
5: Iterar os itens 2, 3 e 4 até que o centróide não seja mais alterado no processo.

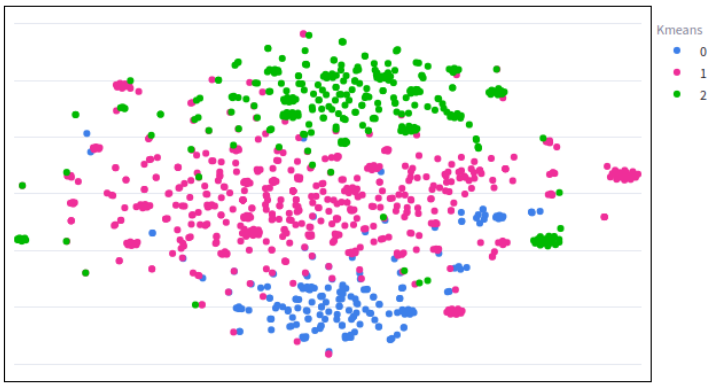



Figura 2.10: Visualização baseada em posicionamento de pontos agrupados pelo algoritmo K-Means. (Fonte: Autoria própria).

### 2.3.8 Métricas de Similaridade

O cálculo da similaridade é uma importante etapa da tarefa de agrupamento, por ser responsável por quantificar o quão semelhantes são duas ou mais instâncias de um conjunto de dados. O valor resultante desse cálculo permite a identificação de padrões, agrupamentos com dados semelhantes, *outliers* e anomalias. Existem diversas medidas de similaridade que podem ser utilizadas para comparar textos, sendo que a escolha da abordagem mais adequada depende do tipo de dado e do objetivo da classificação.

#### Edit Distance

A Edit Distance, também chamada de Distância de Levenshtein, em homenagem ao seu criador, é uma medida de similaridade entre dois textos, sendo que, quanto maior for esse valor, maior será a diferença entre os textos comparados. Supondo um texto $A$ e um texto $B$, essa distância é definida com base na quantidade mínima de operações que devem ser

feitas no texto $A$ para que se torne o texto $B$, sendo as operações de inserir, remover ou converter um caractere [31]. Este processo é realizado considerando os textos como matrizes e iterando por cada coluna e linha a fim de calcular a menor distância entre cada letra.

### Distância Cosseno

A distância cosseno é uma métrica apropriada para medir a similaridade entre palavras [32], visto que, após a redução de dimensionalidade a distância angular preserva melhor as relações entre as palavras do que a distância euclideana. Sua fórmula é exemplificada pela Equação (2.17).

$$d_{cos}(i,j) = \frac{\sqrt{(i_1 \times j_1) + (i_2 \times j_2)}}{\sqrt{(i_1^2 + i_2^2)} \times \sqrt{(j_1^2 + j_2^2)}}, \tag{2.17}$$

em que $i$ e $j$ são dois vetores que estamos comparando. A similaridade cosseno é dada pela divisão do produto interno de $i$ e $j$ pelo produto das normas de $i$ e $j$.

Como a distância é calculada com base no ângulo entre os dois vetores analisados, deve-se obter 1 caso os vetores sejam idênticos, 0 caso sejam ortogonais e -1 se forem opostos. Dessa forma, quanto mais próximo de 1 for o valor da distância cosseno, mais semelhantes os vetores de representação são entre si.

### Distância de Gower

A distância de Gower [33] pode ser usada para calcular a distância entre duas amostras cujos atributos apresentam valores de dados lógicos, numéricos, categóricos ou em formato de texto, conforme a Equação (2.18).

$$d_{gower}(i,j) = \frac{\sum_{p=1}^{m} w_p \delta_{ijp}}{\sum_{p=1}^{m} w_p}, \tag{2.18}$$

em que $d_{gower}(i,j)$ é a distância de Gower entre as instâncias $i$ e $j$, $m$ é o número de atributos, $w_p$ é o peso do atributo $p$, e $\delta_{ijp}$ é uma função que mede a diferença entre os valores do atributo $p$ para as instâncias $i$ e $j$. Para atributos nominais, $\delta_{ijp} = [x_{ip} \neq x_{jp}]$ se $x_{ip}$ é diferente de $x_{jp}$ e 0 caso contrário. Para atributos ordinais e numéricos, $\delta_{ijp} = \frac{|x_{ip} - x_{jp}|}{R_p}$, onde $R_p$ é o range do atributo $p$. A saída é um valor entre 0 e 1, em que valores próximos de zero indicam instâncias de dados semelhantes.

**Distância Customizada**

Este cálculo de distância foi proposto utilizando a média dos valores obtidos pela distância euclideana aplicada aos atributos numéricos e dos valores obtidos pela distância cosseno aplicada aos vetores TF-IDF criados a partir dos atributos numéricos. A Equação (2.19) descreve a fórmula utilizada.

$$d_{customizada}(i,j) = d_E(i,j) + d_C(i,j), \tag{2.19}$$

em que $d_{customizada}(i,j)$ é a distância customizada proposta, $d_E(i,j)$ é a distância euclideana dos atributos numéricos e $d_C(i,j)$ é a distância cosseno dos atributos textuais.

## 2.4 Visualização de Dados

Visualização é o processo de transformação dos dados em uma forma de representação visual, utilizando gráficos, posicionamento de pontos, painéis visuais, entre outras ferramentas, na qual seja possível transmitir de forma mais objetiva a informação para pessoas que não possuem necessariamente um conhecimento prévio em computação, além de possibilitar a extração de padrões visuais dos conjuntos de dados. Por meio de uma análise visual, pode-se obter um entendimento geral dos dados e possibilitar uma interpretação intuitiva que facilita a detecção de padrões e tendências.

Em contrapartida, ao lidar com visualização de dados de alta dimensão, existem diversos desafios a serem superados, sendo a "Maldição da Dimensionalidade" o mais conhecido dentre eles. Este termo introduzido por Bellman [34] se refere à geração de redundância dos dados em casos de alta dimensionalidade e à comprovada diminuição do desempenho de classificadores após um determinado número de atributos.

Conjuntos de dados textuais são suscetíveis a este problema por conta da grande quantidade de palavras e caracteres presentes. Dentre as soluções mais comuns neste caso, pode-se citar a redução das dimensões por meio da seleção de atributos ou de técnicas de projeção, como Principal Component Analysis (PCA), Multidimensional Scaling (MDS), t-Distributed Stochastic Neighbor Embedding (t-SNE) e Uniform Manifold Approximation and Projection (UMAP).

### 2.4.1 Estratégias Baseadas em Posicionamento de Pontos

A visualização baseada no posicionamento de pontos é uma forma eficaz de entender relacionamentos [35], *outliers* e comportamentos inesperados existentes em um conjunto de dados multidimensional, ou seja, com dois atributos ou mais. No caso de mais de dois atributos, devem ser aplicadas técnicas de redução de dimensionalidade com o objetivo

de criar uma projeção em duas ou três dimensões. Cada um dos pontos equivale à representação das instâncias, e a distância entre cada um dos pontos reflete a similaridade das instâncias de acordo com os atributos utilizados.

Como etapa prévia à visualização de dados multidimensionais, existe a etapa de descarte de atributos irrelevantes para a tarefa em questão e a redução das dimensões do conjunto de dados. A redução de dimensionalidade possui como objetivo principal o desenvolvimento de uma forma de representação de um conjunto de dados de alta dimensionalidade em dimensões reduzidas com a menor perda de informação possível ao longo do processo.

### 2.4.2 MDS

A técnica Multidimensional Scaling (MDS) projeta os dados de um espaço de alta dimensão para um espaço de baixa dimensão, como descreve o Algoritmo 2. Um exemplo dessa visualização pode ser encontrado na Figura 2.11. A principal vantagem desta técnica é a preservação das distâncias entre os objetos do espaço original [9], mesmo após a projeção no espaço de baixa dimensionalidade.

**Algorithm 2** MDS: Multidimensional Scaling

1: Compute a matriz de similaridades a partir das distâncias;
2: Centralize a matriz de similaridades;
3: Calcule a matriz de produtos internos centrados;
4: Calcule a decomposição em autovalores da matriz de produtos internos centrados;
5: Selecione os $k$ maiores autovalores e seus autovetores correspondentes;
6: Construa a matriz de incorporação de baixa dimensão a partir dos autovetores selecionados.

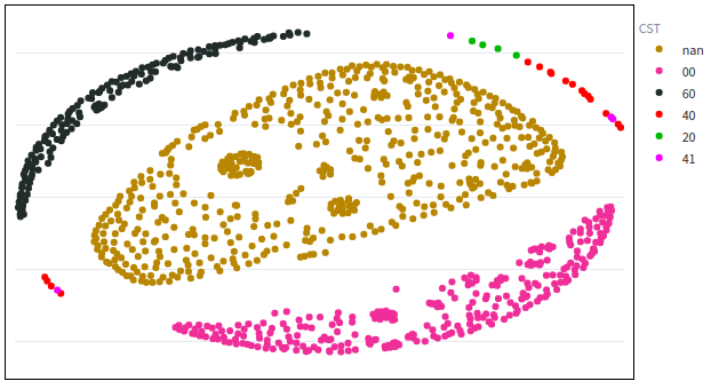



Figura 2.11: Visualização gerada com o MDS. (Fonte: Autoria própria).

### 2.4.3 t-SNE

A t-Distributed Stochastic Neighbor Embedding (t-SNE) é uma técnica de redução de dimensionalidade não linear utilizada para visualizar dados de alta dimensão em um espaço de baixa dimensão, utilizando a distribuição de probabilidade t-Student para medir a similaridade entre os pontos [10], como descreve o Algoritmo 3. Um exemplo dessa visualização pode ser encontrado na Figura 2.12. A t-SNE é mencionada na literatura como uma solução que preserva tanto estruturas locais, ou seja, a distância entre pontos próximos no espaço de alta dimensionalidade, quanto globais dos dados.

**Algorithm 3** t-SNE: t-Distributed Stochastic Neighbor Embedding

1: Calcule as similaridades de vizinhança usando a métrica de similaridade desejada;
2: Inicialize a incorporação de baixa dimensão aleatoriamente;
3: **for** *e* **in** *elementos* **do**
4: Calcule as similaridades de vizinhança condicionais na incorporação de baixa dimensão;
5: Calcule o gradiente da divergência de Kullback-Leibler entre as similaridades de vizinhança originais e as similaridades de vizinhança condicionais;
6: Atualize a incorporação de baixa dimensão usando o gradiente descendente;
7: Aplique uma redução de dimensão usando a medida de t-Student.

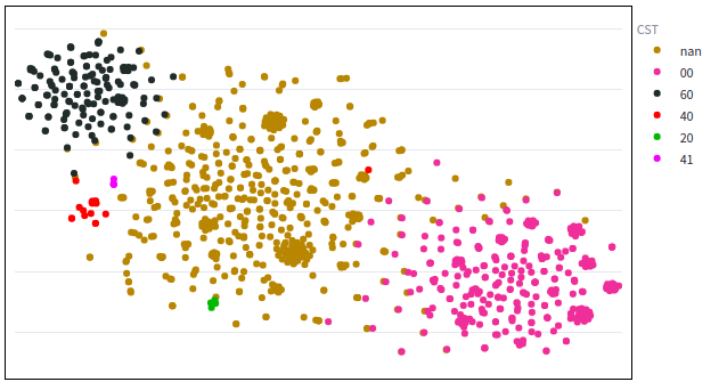



Figura 2.12: Visualização gerada com o t-SNE. (Fonte: Autoria própria).

### 2.4.4 UMAP

A Uniform Manifold Approximation and Projection (UMAP), descrita pelo Algoritmo 4, também é uma técnica de redução de dimensionalidade não linear, cuja principal vantagem é a preservação das estruturas globais dos dados [11], a custo de um tempo de processamento maior. Um exemplo dessa visualização pode ser encontrado na Figura 2.13.

**Algorithm 4** UMAP: Uniform Manifold Approximation and Projection

1: Calcule a distância entre dados e os seus vizinhos mais próximos;
2: Construa um grafo ponderado a partir das distâncias calculadas;
3: **for** *e* **in** *elementos* **do**
4: Minimize a função de custo que mede a diferença entre as distâncias no espaço multidimensional e no espaço de baixa dimensão;
5: Atualize a representação do espaço de baixa dimensão.

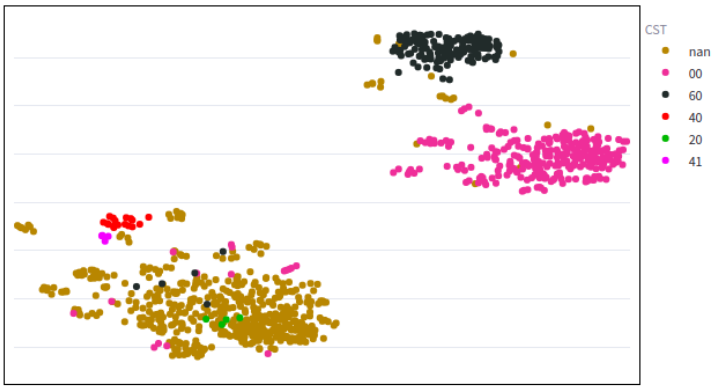



Figura 2.13: Visualização gerada com o UMAP. (Fonte: Autoria própria).

### 2.4.5 Avaliação de Qualidade de Visualizações

Existem diversas técnicas de visualização de dados multi-dimensionais, algumas delas abordadas na Seção anterior, sendo que cada uma produz resultados distintos, com isso, faz-se necessária uma análise estrutural antes da escolha pela técnica definitiva. Nesta Seção serão abordados métodos que avaliam quantitativamente a qualidade de uma projeção multidimensional.

#### Coeficiente de Silhueta

O coeficiente de silhueta mede o quão bem as instâncias de dados estão distribuídas em grupos, considerando as distâncias *intra-cluster* e *inter-cluster*. Sendo a primeira a distância média entre instâncias agrupadas em um mesmo grupo, e a segunda entre instâncias agrupadas em grupos diferentes, conforme exemplifica a Equação 2.20. Essa métrica busca minimizar a distância *intra-cluster* e maximizar a *inter-cluster*.

$$c_{silhueta}(i) = \frac{1}{N}\sum_{i=1}^{N}\frac{d_p(i) - d_g(i)}{\max\{d_g(i), d_p(i)\}}, \tag{2.20}$$

em que $c_{silhueta}(i)$ é o coeficiente de silhueta da $i$-ésima instância, $d_g(i)$ é a distância média entre a $i$-ésima instância e as outras instâncias do mesmo grupo, $d_p(i)$ é a distância média

entre a $i$-ésima instância e as instâncias do grupo mais próximo, e $max\{d_g(i), d_p(i)\}$ é o máximo entre $d_g(i)$ e $d_p(i)$.

Os resultados podem variar de $-1$ a 1, sendo que valores mais próximos de 1 indicam que as instâncias estão bem agrupadas em seus respectivos grupos e separadas dos outros grupos, enquanto valores próximos de $-1$ indicam que existem muitas instâncias agrupadas em grupos incorretos. O resultado desta métrica também pode ser utilizado para definir a quantidade de grupos de um conjunto de dados [36].

### Neighborhood Preservation

Complementar à métrica citada anteriormente, a Neighborhood Preservation analisa quantitativamente a preservação da vizinhança das instâncias após a conversão destas para o espaço multi-dimensional [35]. Quanto maior o valor, melhor a preservação da vizinhança no espaço de baixa dimensionalidade. Além disso, é esperado que esta métrica tenha uma alta variação para quantidades baixas de vizinhos, tendo em vista que predomina a localidade nessas situações, como exemplifica a Figura 2.14.

Dado um valor $k$ escolhido previamente, $\forall\ k\ \exists\ N\ k \geq 1$ e $k < número\ de\ amostras$, a Neighborhood Preservation de uma instância é obtida com base na proporção entre os $k$ vizinhos dessa instância no espaço multi-dimensional e os $k$ vizinhos dela na projeção criada. Para calcular o valor dessa métrica em um conjunto de dados, deve ser feita uma média dos valores obtidos para cada instância.

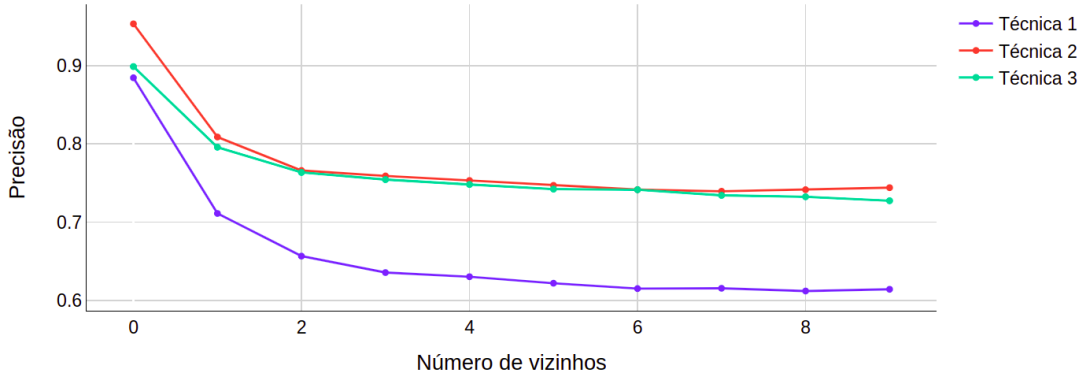



Figura 2.14: Exemplo ilustrativo de um gráfico gerado pela métrica Neighborhood Preservation. (Fonte: Autoria própria).

## 2.5 Considerações Finais

Neste capítulo foram apresentados conceitos importantes que serão utilizados na metodologia, dentre eles, as técnicas de redução de dimensionalidade t-SNE, UMAP e MDS,

além do algoritmo de agrupamento K-Means e das técnicas de classificação textos utilizando redes neurais LSTM e a F1-score como forma de avaliação de desempenho em dados desbalanceados. Com base nessa Fundamentação Teórica, pode-se notar o caráter multidisciplinar do tema, ressaltando tanto uma abordagem técnica da área de Processamento de Linguagem Natural quanto o conhecimento de auditoria de documentos fiscais.

# Capítulo 3

# Revisão de Literatura

Este capítulo é dedicado a explorar estudos de caso semelhantes ao das notas fiscais e técnicas estado-da-arte presentes na literatura nas áreas de Visualização de Dados Multidimensionais, Processamento de Linguagem Natural e Aprendizado Supervisionado em conjuntos de dados não balanceados.

A detecção de fraudes tem sido amplamente explorada na literatura, embora existam poucos documentos que abordem técnicas específicas para lidar com notas fiscais eletrônicas. Por conta do interesse em abordagens computacionais semi-automáticas, a investigação inicial foi centrada na mineração de texto, redução da dimensionalidade e técnicas de visualização.

Abordando o tema de notas fiscais eletrônicas, Diego Kieckbusch propôs um estudo de caso da auditoria fiscal [3] e uma classificação baseada em Redes Neurais Convolucionais para definir o tipo de produto, por meio da análise da descrição e do código de cadastro [37]. O processo desde a criação das notas fiscais até o processo de auditoria é exemplificado pela Figura 3.1.

Na literatura, métodos não supervisionados são amplamente utilizados em situações nas quais não se tem um conhecimento prévio do rótulo de um conjunto de dados, sendo assim, podem ser encontrados agrupamentos por semelhança nos atributos. De acordo com Richard Bolton e David Hand [7], esses aspectos citados acima podem ser utilizados para entender mais profundamente o comportamento do sistema a ser estudado e, consequentemente, auxiliar na detecção comportamentos atípicos nos dados, que, em um contexto de detecção de fraude, podem ser essenciais.

A detecção de casos suspeitos de fraude financeira foi explorado por Margarita Knyazena [4], que propôs um modelo baseado em grafos para um problema de mineração de dados de explorar e revelar grupos de domínio de usuários propensos a cometer fraudes financeiras, assim como é exemplificado na Figura 3.2. Segundo ela, a implementação de tarefas de mineração de dados são vitais para lidar com a detecção de fraude devido

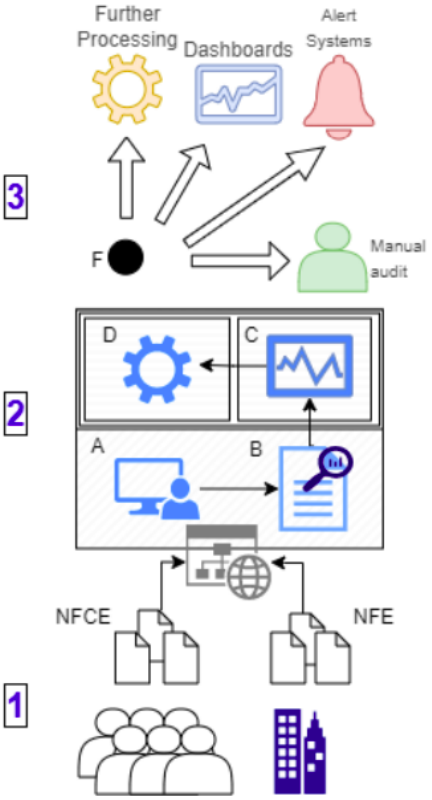



Figura 3.1: Processo de obtenção e auditoria de uma nota fiscal. (Fonte: Artigo [3]).

à grande quantidade de dados legais e financeiros não rotulados. A primeira etapa do método é a divisão dos dados em agrupamentos usando um algoritmo baseado em K-Means. A segund0,94a etapa foi baseada na criação de regras para cada agrupamento. A última etapa foi criar uma função para encontrar alguns pontos de interesse particular e conectar os escolhidos aos outros pontos, depois disso, tarefas de classificação e predição foram importantes para encontrar grupos de usuários que podem potencialmente causar fraudes.

Zhichao Zha [5] criou o TaxAA, um assistente de auditoria fiscal que ajuda os auditores fiscais a explorar transações suspeitas e obter evidências de fraude, uma tarefa de classificação semi-supervisionada. Alguns contribuintes aproveitam o processo de emissão de uma nota fiscal para evitar o pagamento de impostos e, embora existam algumas características dessa fraude, às vezes é difícil localizar evidências suficientes em um grande número de notas fiscais. A análise visual proposta possui duas abordagens, a Visão do Filtro Suspeito que usa modelos de Redes Neurais, cujo fluxograma é exemplificado na Figura 3.3, e a Visão da Descrição Detalhada que permite ao auditor explorar os dados, ambas abordagens presentes na Figura 3.4. A seção de Visão de Casos Suspeitos pode analisar informações individuais, de relacionamento e transações e detectar evidências

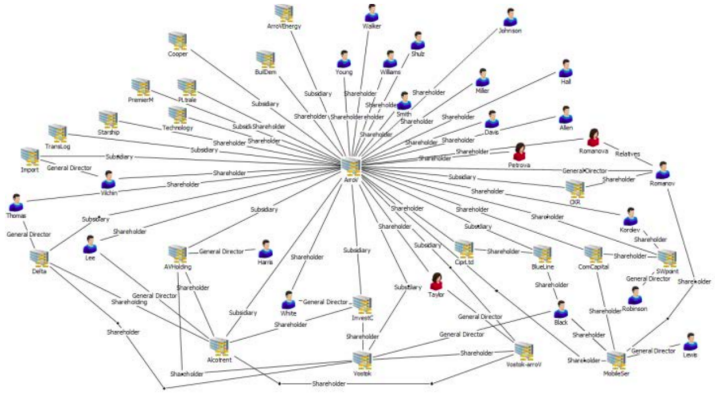


Figura 3.2: Estrutura de grafos referentes a usuários e ações. (Fonte: Artigo [4]).

reais de fraude usando a fusão da Rede de Auditoria Fiscal e da Rede Convolucional de Gráfico Hierárquico. O modelo proposto teve acurácia de 94%.

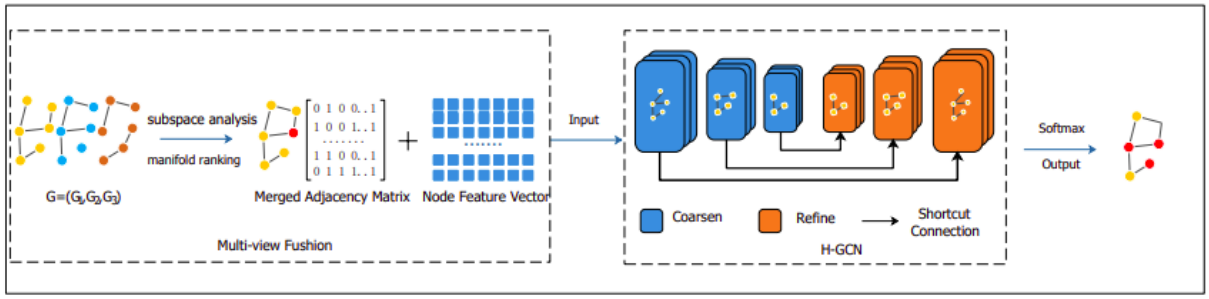



Figura 3.3: Fluxograma da metodologia proposta por Zhichao Zha para a criação do sistema TaxAA. (Fonte: Artigo [5]).

Florian Heimerl e Michael Gleicher [8] criaram visualizações interativas eficazes que auxiliam os profissionais e pesquisadores a entender e comparar melhor os espaços multi-dimensionais. Incorporações vetoriais são ferramentas importantes em tarefas de processamento de texto, no entanto, algumas propriedades e relações que essas incorporações codificam muitas vezes não são bem compreendidas. Foram criados 3 protótipos de designs para auxiliar na visualização de algumas propriedades dos *embeddings*, como vizinhos mais próximos, combinação e alinhamento de eixos.

Lucas Resck [6] propôs o LegalVis, um sistema *web* para auxiliar especialistas em direito na identificação e análise de citações a precedentes vinculantes em documentos legais, conforme a metodologia exemplificada pela Figura 3.5. O sistema utiliza técnicas de Aprendizado de Máquina e análise visual interativa para identificar citações a precedentes vinculantes, como pode ser observado na Figura 3.6. Nesse trabalho, foram propostos também requisitos do sistema e tarefas para guiar a análise pelos especialistas.

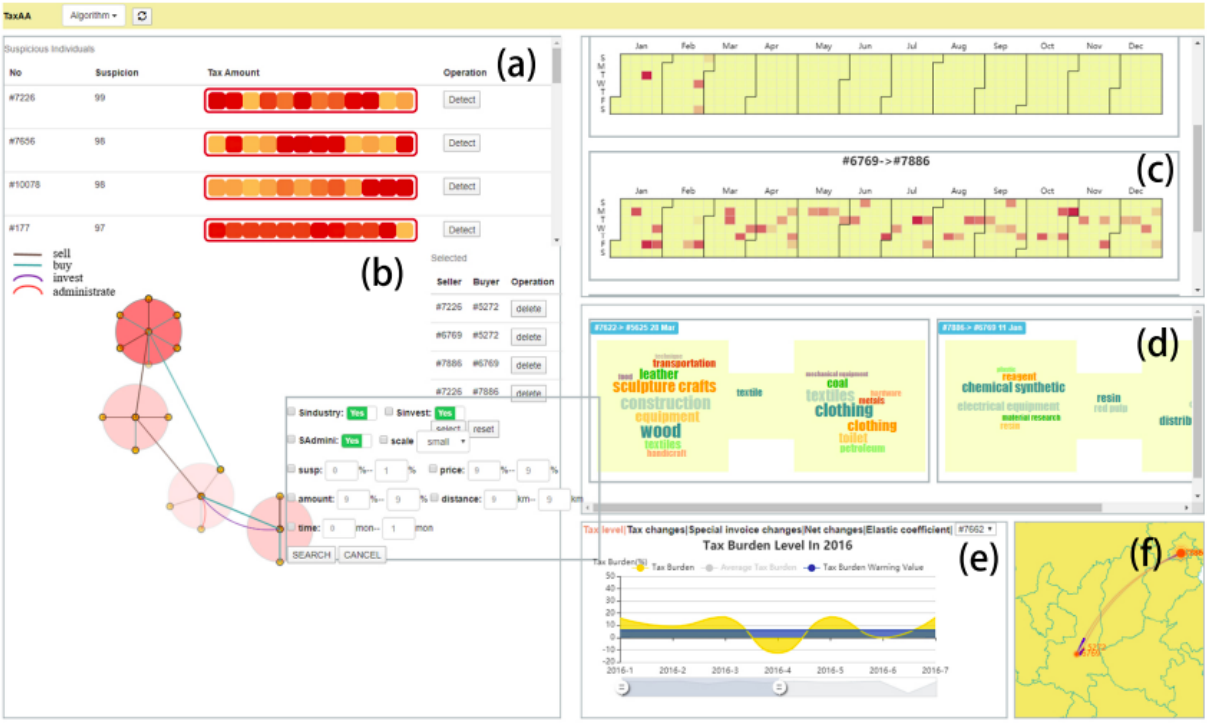



Figura 3.4: Painel principal da ferramenta TaxAA. (Fonte: Artigo [5]).

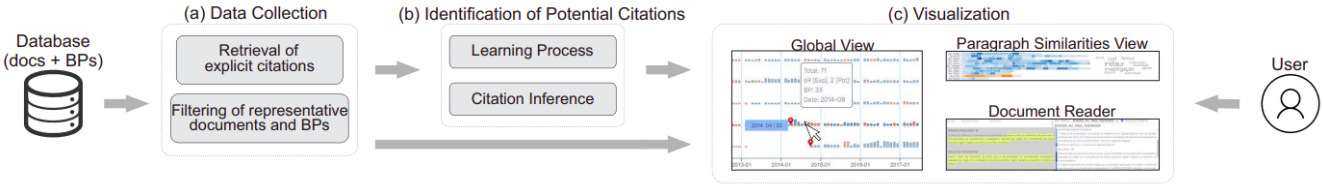



Figura 3.5: Fluxograma da metodologia proposta por Lucas Resck para a criação do LegalVis. (Fonte: Artigo [6]).

## 3.1 Considerações Finais

As NFC-e trazem alguns desafios na concepção de uma abordagem computacional para detectar casos suspeitos de fraude fiscal. Como o conjunto de dados associado não está rotulado, e as notas apresentam diferentes tipos de atributos categóricos e textuais, este cenário mostra-se apropriado para empregar visualizações interativas baseadas em estratégias de posicionamento de pontos. Uma vez que a literatura atual não consegue abordar totalmente a detecção de fraudes em notas fiscais, esta pesquisa tenta preencher esta lacuna propondo um método de visualização de notas fiscais, para que os especialistas possam identificar casos suspeitos por meio de análise visual dos *layouts* gerados com apoio de técnicas de Aprendizado de Máquina e de Similaridade de textos.


Figura 3.6: Painel principal da ferramenta LegalVis. (Fonte: Artigo [6]).

# Capítulo 4

# Identificação Automática de Inconsistências em Notas Fiscais

Este capítulo descreve os métodos e procedimentos utilizados ao longo da pesquisa para obter os resultados experimentais relacionados à identificação automática de inconsistências em notas fiscais. A Seção 4.1 descreve a metodologia utilizada nos experimentos, incluindo o pré-processamento, o cálculo de distância e a definição de rótulos de inconsistência em notas fiscais. A Seção 4.2 descreve os experimentos de geração de um rótulo de inconsistência para auxiliar o auditor fiscal na detecção de fraudes e uma avaliação automática da qualidade deles, com base em informações oficiais sobre o NCM e em distintas técnicas de representação textual, visto que este conjunto de dados não é rotulado previamente.

## 4.1 Metodologia

Nesta seção serão apresentadas as técnicas empregados para coletar, analisar e interpretar os dados para a identificação automática de inconsistências em notas fiscais eletrônicas. O fluxograma da Figura 4.1 ilustra as etapas que compõem o método proposto, desde o pré-processamento do conjunto de dados até a definição dos rótulos.

### 4.1.1 Coleta e amostragem

Foram realizados experimentos com o conjunto de dados da NFC-e, porém considerando 10.000 instâncias devido à alta complexidade computacional dos modelos. Além disso, foi utilizado também um dicionário com todos os NCMs e a descrição oficial de cada um para fins de comparação com a descrição do produto definida pelo comerciante no momento de registro da nota.

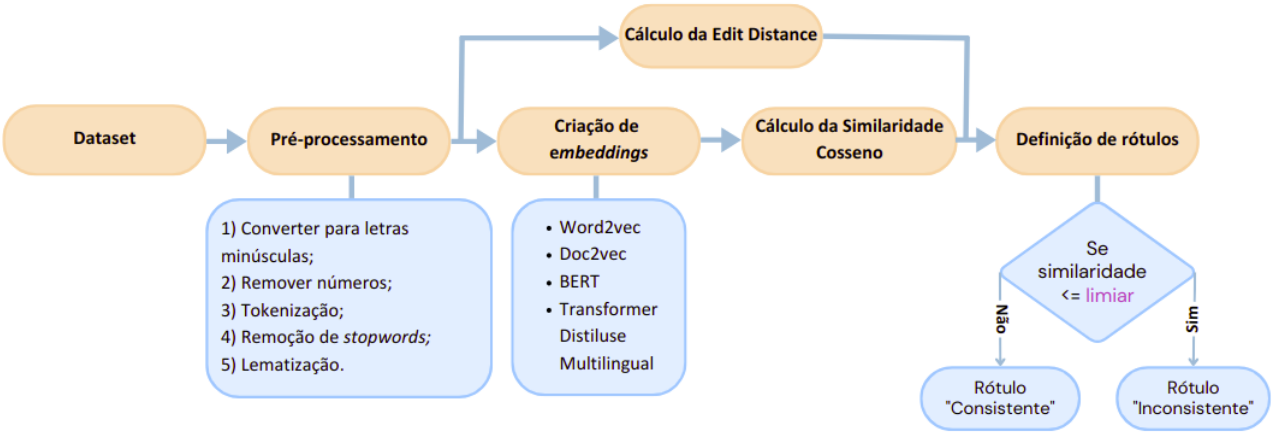



Figura 4.1: Fluxograma da metodologia de criação de rótulos de consistência. (Fonte: Autoria própria).

### 4.1.2 Pré-processamento

Nesta etapa, foi aplicado um pré-processamento nos dados selecionados com a finalidade de prepará-los para as técnicas de visualização e de aprendizado de máquina. A partir da descrição do produto e da descrição oficial, vistas anteriormente nas Figuras 2.1 e 2.2, foi feita uma conversão de letras para minúsculo, remoção de números, *tokenização*, remoção de *stopwords* e lematização. A Tabelas 4.1 e 4.2 exemplificam esse pré-processamento.

### 4.1.3 Cálculo de Distâncias

Após o pré-processamento, foi feito um cálculo de distância entre a descrição do produto registrada no momento da criação da nota fiscal e a descrição oficial relacionada ao NCM do produto em questão no dicionário. O objetivo é definir o quão similares são os dois textos. A comparação de similaridade pode ser útil para detectar valores de descrição do produto pouco compatíveis com o esperado, o que pode caracterizar uma inconsistência.

Nas subseções a seguir, são explicados os métodos para se comparar os textos das notas fiscais, utilizando o algoritmo da Edit Distance e a geração de *embeddings* pelas técnicas Word2Vec, Transformer Distiluse Multilingual e BERT, com a distância calculada pela similaridade cosseno.

#### Cálculo da Edit Distance

Partindo da premissa de que o valor de NCM definido no momento de registro da nota fiscal é o correto, foi aplicada a Edit Distance entre a descrição do produto e a descrição oficial do respectivo NCM para o cálculo da similaridade entre estes dois textos. Essa

| Descrição | Descrição Oficial |
|---|---|
| JOGO DE PALHETA DIANTEIRA CORSA MONTANA | 85: Máquinas, aparelhos e materiais elétricos, e suas partes; aparelhos de gravação ou de reprodução de som, aparelhos de gravação ou de reprodução de imagens e de som em televisão, e suas partes e acessórios. 85.44: Fios, cabos (incluindo os cabos coaxiais) e outros condutores, isolados para usos elétricos (incluindo os envernizados ou oxidados anodicamente), mesmo com peças de conexão; cabos de fibras ópticas, constituídos por fibras embainhadas individualmente, mesmo com condutores elétricos ou munidos de peças de conexão. 8544.30.00: - Jogos de fios para velas de ignição e outros jogos de fios do tipo utilizado em quaisquer veículos |
| FERRO A VAPOR ANT BR/DOURADO 220V GCSTBS5907 OSTER | 85: Máquinas, aparelhos e materiais elétricos, e suas partes; aparelhos de gravação ou de reprodução de som, aparelhos de gravação ou de reprodução de imagens e de som em televisão, e suas partes e acessórios. 85.16: Aquecedores elétricos de água, incluindo os de imersão; aparelhos elétricos para aquecimento de ambientes, do solo ou para usos semelhantes; aparelhos eletrotérmicos para arranjos do cabelo (por exemplo, secadores de cabelo, frisadores, aquecedores de ferros de frisar) ou para secar as mãos; ferros elétricos de passar; outros aparelhos eletrotérmicos de uso doméstico; resistências de aquecimento, exceto as da posição 85.45. 8516.40.00: - Ferros elétricos de passar |
| CASTANHA DO PARA KG 0,200KG X 79,00 | 08: Fruta; cascas de citros (citrinos) e de melões. 08.01: Cocos, castanha-do-brasil (castanha-do-pará) e castanha-de-caju, frescos ou secos, mesmo com casca ou pelados. 0801.3: - Castanha-de-caju: 0801.31.00: – Com casca |
| CONVERSOR MIGTEC DE VIDEO COMPONENTE (YPBPR) P/ HD | 85: Máquinas, aparelhos e materiais elétricos, e suas partes; aparelhos de gravação ou de reprodução de som, aparelhos de gravação ou de reprodução de imagens e de som em televisão, e suas partes e acessórios. 85.43: Máquinas e aparelhos elétricos com função própria, não especificados nem compreendidos noutras posições do presente Capítulo. 8543.70: - Outras máquinas e aparelhos 8543.70.3: Máquinas e aparelhos auxiliares para vídeo 8543.70.39: Outros |

Tabela 4.1: Tabela comparativa dos atributos antes do pré-processamento.

| Descrição Pré-processada | Descrição Oficial Pré-processada |
|---|---|
| ['jogo', 'palheta', 'dianteira'] | ['máquina', 'aparelho', 'material', 'elétrico', 'parte', 'aparelho', 'gravação', 'reprodução', 'som', 'aparelho', 'gravação', 'reprodução', 'imagem', 'som', 'televisão', 'parte', 'acessório', 'fio', 'cabo', 'cabo', 'condutor', 'uso', 'elétrico', 'peça', 'conexão', 'cabo', 'fibra', 'fibra', 'condutor', 'elétrico', 'munido', 'peça', 'conexão', 'jogo', 'fio', 'vela', 'ignição', 'jogo', 'fio', 'tipo', 'veículo'] |
| ['ferro', 'ant', 'br', 'gcstbs'] | ['máquina', 'aparelho', 'material', 'elétrico', 'parte', 'aparelho', 'gravação', 'reprodução', 'som', 'aparelho', 'gravação', 'reprodução', 'imagem', 'som', 'televisão', 'parte', 'acessório', 'aquecedor', 'elétrico', 'água', 'imersão', 'aparelho', 'elétrico', 'aquecimento', 'solo', 'uso', 'aparelho', 'eletrotérmico', 'arranjo', 'cabelo', 'exemplo', 'secador', 'cabelo', 'frisadore', 'aquecedor', 'ferro', 'mão', 'ferro', 'elétrico', 'aparelho', 'eletrotérmico', 'uso', 'resistência', 'aquecimento', 'posição', 'ferro', 'elétrico'] |
| ['castanha', 'kg'] | ['casca', 'citro', 'citrino', 'melão', 'coco', 'castanha', 'castanha', 'castanha', 'caju', 'fresco', 'seco', 'casca', 'castanha', 'caju', 'casca'] |
| ['conversor', 'migtec', 'componente', 'hd'] | ['máquina', 'aparelho', 'material', 'elétrico', 'parte', 'aparelho', 'gravação', 'reprodução', 'som', 'aparelho', 'gravação', 'reprodução', 'imagem', 'som', 'televisão', 'parte', 'acessório', 'máquina', 'aparelho', 'elétrico', 'função', 'posição', 'capítulo', 'máquina', 'aparelho', 'máquina', 'aparelho', 'auxiliar', 'vídeo'] |

Tabela 4.2: Tabela comparativa dos atributos após o pré-processamento.

etapa é uma das opções apresentadas na metodologia, em contraposição à extração de características em conjunto com o cálculo da distância cosseno.

#### Extração de Características

Foram utilizadas distintas formas de representação de texto para fins de comparação, como Word2vec calculado pela soma e com a arquitetura Continuous Bag of Words (CBOW), Doc2vec, Transformer Distiluse Multilingual e BERT. Estes últimos foram aplicados nos experimentos por meio da biblioteca Hugging Face, sendo considerados os modelos "distiluse-base-multilingual-cased-v1" e "bert-base-multilingual-cased", respectivamente.

#### Cálculo da Distância Cosseno

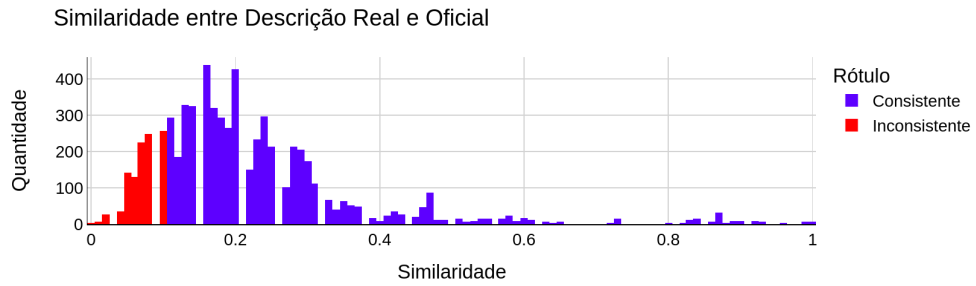
Após a extração de características, foi aplicada a distância cosseno para as representações descritas anteriormente, com o objetivo de obter uma medida do quão similares são a descrição da nota fiscal e a descrição oficial de cada produto analisado.

### 4.1.4 Definição de Rótulos

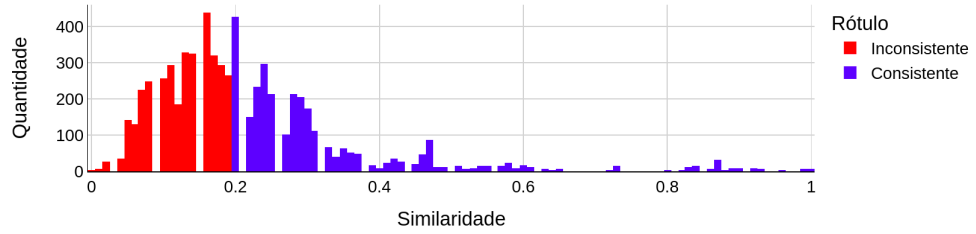
Com base na medida de similaridade calculada anteriormente, é definido um limiar para a criação de rótulos de consistência. Sendo assim, quanto menos similares as duas descrições são, mais inconsistente é a nota fiscal em relação à descrição oficial, visto que a descrição e o NCM registrados serão incompatíveis nesse caso. Com isso, o limiar irá definir o valor de similaridade mínima para uma nota ser considerada consistente, assim, notas que possuem similaridade abaixo desse limiar serão rotuladas como inconsistentes. Para fins de definição e comparação de limiar, a distância cosseno, que possui valores entre $[-1, 1]$, foi convertida para a escala $[0, 1]$, a mesma escala da Edit Distance, em um processo de normalização de distâncias entre pares de instâncias, de forma semelhante à implementação da t-SNE [10].

Segundo especialistas em auditoria fiscal, cerca de 25% das notas fiscais são inconsistentes, porém essa é uma estimativa baseada em experiência e pode variar dependendo do conjunto de notas fiscais analisado. Para os experimentos dessa pesquisa, foram analisados limiares de similaridade equivalentes a 0,1, 0,2 e 0,3 para a definição dos rótulos de consistência, como exemplifica a Figura 4.2.

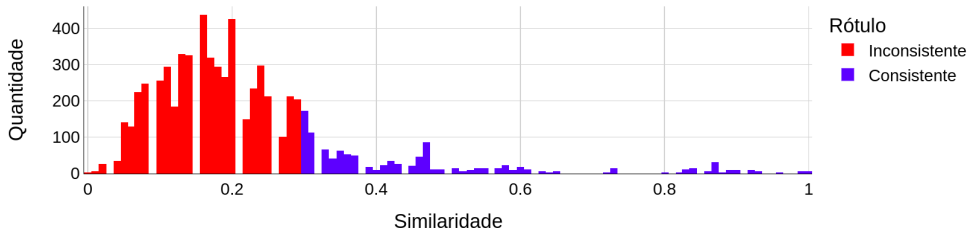
Sendo assim, o referido limiar deve ter o seu valor ajustado por parte de um especialista, aumentando o valor caso o objetivo seja rotular como consistentes apenas as notas que tenham uma similaridade maior, e diminuindo o valor caso o intuito seja ter mais instâncias rotuladas como consistentes. Isso se deve pelo fato do limiar definir o valor




(a) Limiar = 0,1, com 83,5% de rótulos consistentes.




(b) Limiar = 0,2, com 45,62% de rótulos consistentes.




(c) Limiar = 0,3, com 17,19% de rótulos consistentes.

Figura 4.2: Gráfico de quantidade de notas rotuladas por similaridades geradas pela Edit Distance. (Fonte: Autoria própria).

de similaridade mínimo a partir do qual serão definidos as notas fiscais com o rótulo de consistente.

## 4.2 Experimentos

Foram realizados experimentos com base na metodologia apresentada anteriormente a fim de avaliar automaticamente a qualidade dos rótulos de consistência, visto que não existem dados rotulados previamente pelo auditor fiscal neste conjunto de dados. Nessa comparação, foi realizada a tarefa supervisionada de classificação, utilizando o modelo LSTM com um texto único composto pela descrição do produto, o NCM e o CST como entradas, e o rótulo de inconsistência como a classe.

Com o objetivo de obter resultados quantitativos para validar a metodologia, foi aplicado o método Holdout no conjunto de dados, gerando um subconjunto de treinamento com 70% dos dados, um de validação com 10% dos dados e um de teste com 20% dos dados. Por fim, foi aplicado o SMOTE para lidar com os dados de treino desbalanceados, gerando novas instâncias inconsistentes com base nas que já existiam até que a quantidade de cada classe se igualasse.

### 4.2.1 LSTM

A arquitetura utilizada nos experimentos é composta por uma camada LSTM, uma Dropout e uma Totalmente Conexa com função de ativação sigmóide ou softmax. Ambas funções foram testadas em conjunto com os hiperparâmetros, como é explicado na próxima subseção. A função de perda utilizada foi a Binary Crossentropy e a otimização foi a Adam. Além disso, também foi utilizada a técnica de Early Stopping para o modelo parar o treinamento a cada 5 resultados iguais obtidos pela função de perda, visto que o treinamento após o fim da diminuição da função de perda aumenta as chances de que ocorra um sobreajuste dos dados de treinamento [38]. A LSTM foi escolhida com base na sua capacidade de lidar com textos de distintos tamanhos, que reflete a realidade das descrições dos produtos.

### 4.2.2 Otimização de Hiperparâmetros

A otimização de hiperparâmetros é uma etapa importante para melhorar o desempenho de redes neurais profundas, de acordo com a literatura [39]. Sendo assim, foi feita uma otimização com o conjunto de dados de validação a partir do espaço de busca definido a seguir:

- **Unidades da LSTM:** amostragem linear, desde o valor 8 até o valor 96 com um incremento de 8;
- **Função de ativação da LSTM:** funções retificadora, tangente hiperbólica e sigmóide;
- **Taxa da camada de Dropout:** amostragem linear, desde o valor $10^{-1}$ até o valor $5 \times 10^{-1}$ com um incremento de $10^{-1}$;
- **Função de ativação da camada Densa:** funções sigmóide e softmax;
- **Taxa de aprendizado:** amostragem logarítmica, desde o valor $1 \times 10^{-6}$ até o valor $5 \times 10^{-4}$;
- **Tamanho do Batch:** 8, 16, 24 e 32;
- **Épocas:** desde o valor 4 até o valor 64 com um incremento de 4;

Devido ao alto custo computacional e ao tempo necessário para rodar cada treinamento, foram realizadas apenas 5 tentativas randômicas de definição de hiperparâmetros para todos os experimentos, com 1 execução para cada tentativa. Para isso, foi utilizado o modo de busca Random Search [40], com a vantagem de encontrar a combinação que mais se adeque aos dados em menos tempo e a desvantagem de não ser a melhor combinação possível [39]. Apesar da quantidade baixa de tentativas, foi possível obter valores de F1-score satisfatórios, sendo que a F1-score foi a métrica utilizada para definir os melhores hiperparâmetros.

### 4.2.3 Resultados experimentais

Experimentos foram realizados no conjunto de dados da NFC-e, com o objetivo de aplicar e avaliar quantitativamente os resultados provenientes da metodologia proposta. Para isso, foram realizados experimentos tanto com todos os produtos do conjunto de dados, quanto com um produto específico, simulando o procedimento real de um especialista. Vale ressaltar que os resultados obtidos na etapa de avaliação são utilizados apenas com a finalidade de avaliar a reprodução dos rótulos, e que, nesse caso, uma tarefa de classificação que compara as distintas formas de representação propostas poderia ser aplicada apenas se fosse possível obter dados rotulados pelo auditor fiscal.

#### NFC-e

Os resultados obtidos com todos os dados e com os dados de "cigarro" da NFC-e estão exemplificados nas Figuras 4.3 e 4.4, com exceção do Word2Vec com distância cosseno para

o limiar de similaridade 0,1 do subconjunto com todos os dados, visto que a quantidade de instâncias inconsistentes era insuficiente para seguir com a análise. A escolha do produto "cigarro" foi baseada na opinião de um especialista que indicou este produto como um grande alvo de sonegação de imposto nas notas fiscais.

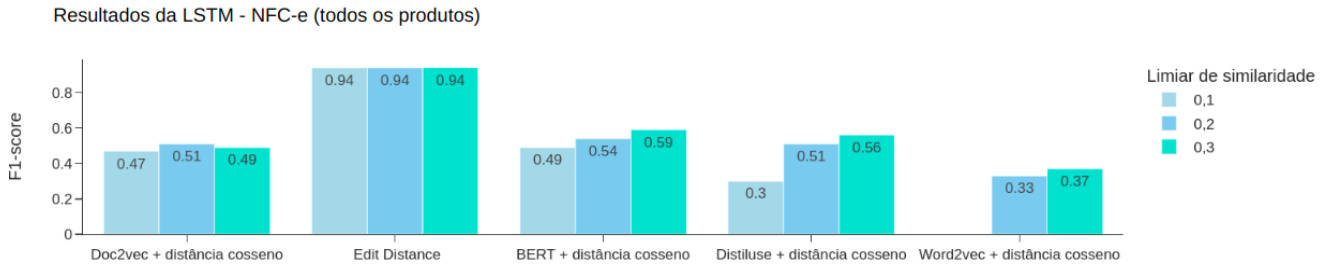



Figura 4.3: Resultados da rede LSTM como forma de avaliar automaticamente os rótulos de inconsistência criados em uma amostra sem seleção de produtos, nos dados da NFC-e. (Fonte: Autoria própria).

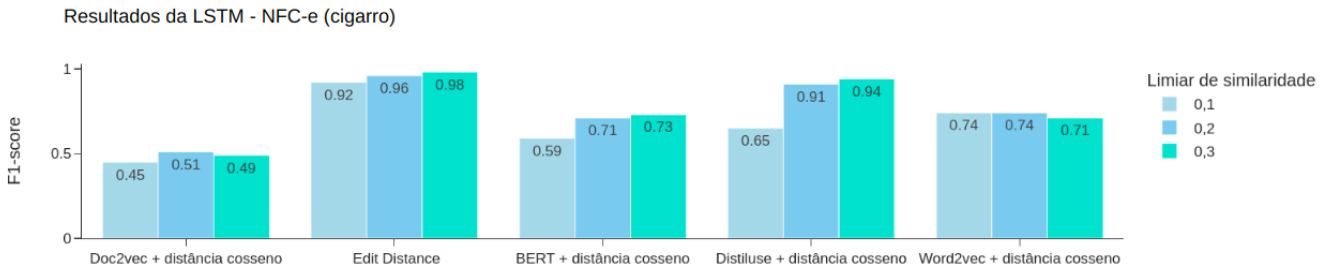



Figura 4.4: Resultados da rede LSTM como forma de avaliar automaticamente os rótulos de inconsistência criados em uma amostra de notas fiscais com "cigarro" na descrição, nos dados da NFC-e. (Fonte: Autoria própria).

Ao analisar a métrica F1-score, nota-se que a Edit Distance produziu rótulos que apresentaram um bom desempenho preditivo na etapa de avaliação automática, tanto ao utilizar todos os produtos, quanto utilizando um produto específico. Assim, resultando no valor de 94% de F1-score para os limiares analisados, enquanto técnicas como BERT, Transformer Distiluse Multilingual, Word2Vec e Doc2Vec com distância cosseno obtiveram resultados abaixo de 60% na análise de todos os produtos da Figura 4.3. Este resultado implica que a rede LSTM conseguiu reconhecer os padrões nas notas fiscais com base nos rótulos de inconsistência criados pela Edit Distance, mas não foi possível reproduzir os rótulos gerados por representações que empregam a distância cosseno.

A Edit Distance é uma medida de similaridade menos complexa e pode ter se adaptado melhor aos padrões textuais das notas fiscais devido a sua simplicidade e ao tamanho

dos textos de entrada, que possuem em média 4,5 palavras e um número máximo de 13 palavras por descrição, no caso de todos os produtos. Em relação à utilização das representações de texto com a distância cosseno, pode-se observar valores de F1-score mais baixos. Esse fato pode ser explicado pela rede LSTM ter tido dificuldade em reproduzir estes rótulos a partir das palavras analisadas.

Algo interessante de se observar é o desempenho das representações de texto no caso do produto "cigarro", que foi consideravelmente maior que no caso geral apresentado anteriormente, como exemplifica a Figura 4.4. Dessa forma, quanto mais específicas as descrições, melhor o desempenho da rede LSTM e das representações de texto para a geração de rótulos, tendo em vista que os textos curtos enviados como entrada são mais parecidos entre si, por se tratarem do mesmo produto.

### 4.2.4 Considerações Finais

Neste capítulo, foram apresentados os métodos e os procedimentos utilizados na pesquisa, visando a identificação automática de inconsistências em notas fiscais. A metodologia utilizada nos experimentos foi descrita detalhadamente, abrangendo o pré-processamento dos dados, cálculo de distâncias e definição de rótulos de consistência e validação automática deles. Foram realizados experimentos com a LSTM, utilizando hiperparâmetros otimizados e a métrica F1-score, para avaliar a confiabilidade dos rótulos criados através do cálculo de similaridade entre as descrições, usando a Edit Distance. Tendo em vista que os rótulos criados nesse capítulo fornecem informações de inconsistência sem que haja uma participação direta do especialista, também foi desenvolvida uma abordagem de visualização complementar a esta, com o objetivo de fornecer mais controle ao especialista no processo de detecção de inconsistências em notas fiscais.

# Capítulo 5

# Visualização Exploratória de Notas Fiscais

Este capítulo apresenta os métodos e procedimentos utilizados ao longo da pesquisa para obter os resultados experimentais relacionados à visualização exploratória de notas fiscais. A Seção 5.1 descreve a metodologia utilizada nos experimentos, incluindo o pré-processamento, a estratégia baseada em posicionamento de pontos e o agrupamento de pontos por meio do algoritmo K-Means. A Seção 5.2 descreve os experimentos de visualização para comparar as distintas técnicas de representação de dados. A Seção 5.3 exemplifica a implementação da metodologia através de um sistema *web*, utilizando como base requisitos do sistema e tarefas para guiarem as análises.

## 5.1 Metodologia

Nesta seção serão apresentados os procedimentos e técnicas empregados para processar, visualizar e interpretar os dados por meio de análises exploratórias, visando encontrar inconsistências entre os dados. O fluxograma da Figura 5.1 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 criada uma ferramenta *web* que implementa a metodologia proposta.

### 5.1.1 Pré-processamento

O pré-processamento da etapa de visualização se deu de forma distinta da etapa rotulação, visto que as métricas de Coeficiente de Silhueta e Neighborhood Preservation não apresentaram resultados satisfatórios considerando os atributos originais da NFC-e. Abaixo estão descritos os atributos criados a partir dos dados originais na etapa de pré-processamento:

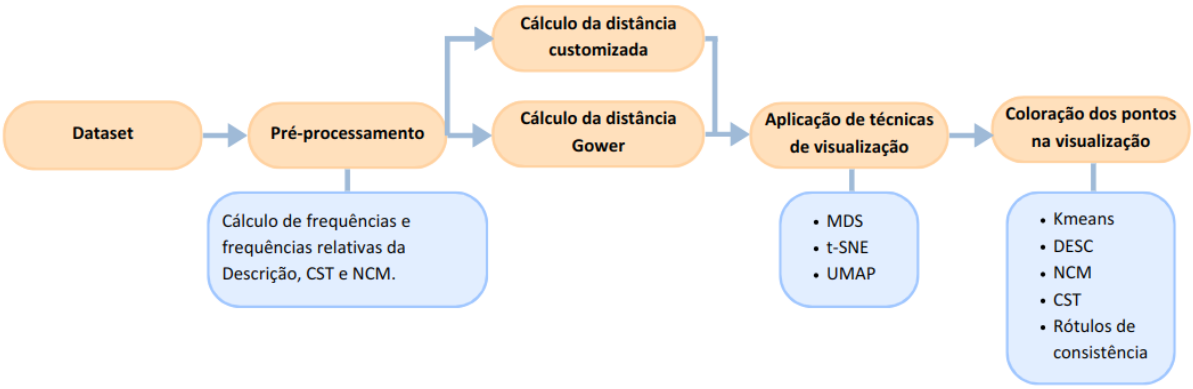



Figura 5.1: Fluxograma da metodologia aplicada nos experimentos de visualização. (Fonte: Autoria própria).

- **DESC_1W:** Seleção da primeira palavra da descrição, antes do primeiro espaço em branco.
- **FREQ_1W:** Cálculo da quantidade de ocorrências da DESC_1W.
- **DESC_2W:** Seleção da segunda palavra da descrição, após o primeiro espaço em branco. Este campo pode conter uma ou mais palavras, dependendo da descrição.
- **FREQ_2W:** Cálculo da quantidade de ocorrências da DESC_2W em relação à DESC_1W.
- **FQREL_2W:** Cálculo da frequência relativa da DESC_2W em relação à DESC_1W.
- **FREQ_NCM:** Cálculo da quantidade de ocorrências de um NCM em relação à DESC_1W e DESC_2W, conjuntamente.
- **FQREL_NCM:** Cálculo da frequência relativa de um NCM em relação à DESC_1W e DESC_2W, conjuntamente.
- **FREQ_CST:** Cálculo da quantidade de ocorrências de um CST em relação à DESC_1W, DESC_2W e NCM, conjuntamente.
- **FQREL_CST:** Cálculo da frequência relativa de um CST em relação à DESC_1W, DESC_2W e NCM, conjuntamente.

Foi aplicada a técnica One-Hot-Encoding nos atributos categóricos, como CST e NCM, transformando cada categoria em uma nova coluna de valores binários. Em atributos textuais, foi aplicado o TF-IDF para a extração de vetores de características. Todos os atributos acima foram considerados para o processo de visualização, que será exemplificado nas subseções a seguir.

### 5.1.2 Cálculo de Distâncias

Após o pré-processamento, foram aplicados dois cálculos de distâncias para serem comparados nos experimentos: a distância de Gower e a distância customizada, explicados anteriormente na subseção 2.3.8. A distância customizada foi proposta como uma alternativa à distância de Gower para lidar com o conjunto de dados pós-processado, que contém dados textuais, numéricos e categóricos.

### 5.1.3 Estratégia Baseada em Posicionamento de Pontos

Dada a grande quantidade de palavras distintas no *corpus*, a representação estruturada proposta possui alta dimensionalidade. Por isso, foram aplicadas três técnicas estado-da-arte para a visualização das notas fiscais: MDS, t-SNE e UMAP.

No contexto deste estudo, cada instância de dados associada a uma NFC-e é representada por um ponto bidimensional no espaço visual. O objetivo das estratégias de posicionamento de pontos é mapear dados multidimensionais em uma dimensão menor, preservando ao máximo as relações entre eles. O posicionamento dos pontos na visualização pode revelar padrões globais e locais, visto que instâncias semelhantes tendem a aparecer mais próximas, enquanto as dissimilares aparecem distanciadas.

Após a geração da representação gráfica bidimensional das notas fiscais, pode-se pensar em distintas formas de colorir os pontos na visualização, visto que o conjunto de dados NFC-e não possui rótulos originais. A estratégia de coloração dos pontos auxilia a identificação de padrões específicos relacionados aos atributos das notas fiscais. Sendo assim, as cores dos pontos no espaço visual podem ser definidas utilizando as técnicas a seguir:

- Conforme a cor dos atributos CST, NCM e descrição;
- De acordo com os agrupamentos obtidos pelo algoritmo K-Means;
- A partir dos rótulos de consistência definidos por similaridade com a descrição oficial.

### 5.1.4 Agrupamento

Nas visualizações geradas, os pontos podem ser coloridos por meio do agrupamento via K-Means. Este algoritmo foi escolhido com base na aplicação dele em um contexto semelhante, pela Margarita Knyazena [4]. O objetivo da criação de agrupamentos diferenciados por cor nas visualizações é fornecer ao especialista grupos semelhantes que podem estar indicando valores consistentes ou não, sendo necessária uma análise aprofundada nesse segundo caso.

## 5.2 Experimentos

Como esse processo de identificação de notas fiscais inconsistentes é manual e subjetivo, foram exploradas estratégias de validação das visualizações, visto que a qualidade delas afeta o desempenho do especialista na identificação de casos suspeitos de fraude por meio da análise visual. Assim, foram realizados experimentos com o objetivo de validar o processo de visualização de notas fiscais, e avaliar a qualidade das visualizações com base no posicionamento de pontos.

Os experimentos foram realizados em uma amostra de notas fiscais $\mathcal{S}$ relacionadas a produtos que podem apresentar indícios de fraude entre eles, segundo os especialistas. Esta amostra $\mathcal{S}$ descreve produtos relacionados a "gin", "vodka" e "cigarro" e contém 5.000 instâncias.

A Figura 5.2 apresenta os resultados da métrica Neighborhood Preservation, que quantifica a preservação de vizinhança após a redução do espaço dimensional. Observa-se um valor de preservação mais alto e estável para o t-SNE com a distância customizada, seguido pelo t-SNE com a distância Gower, o MDS com a distância customizada e o UMAP com a distância customizada.

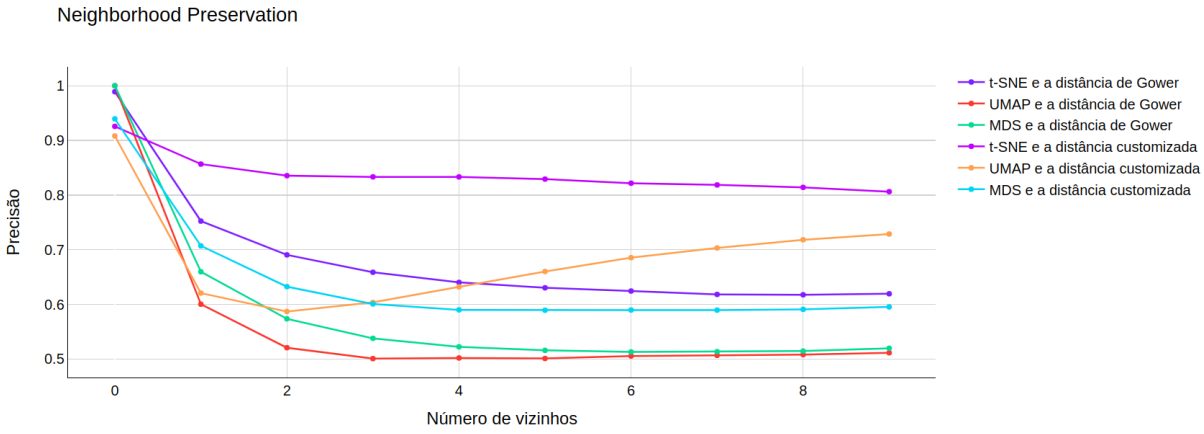
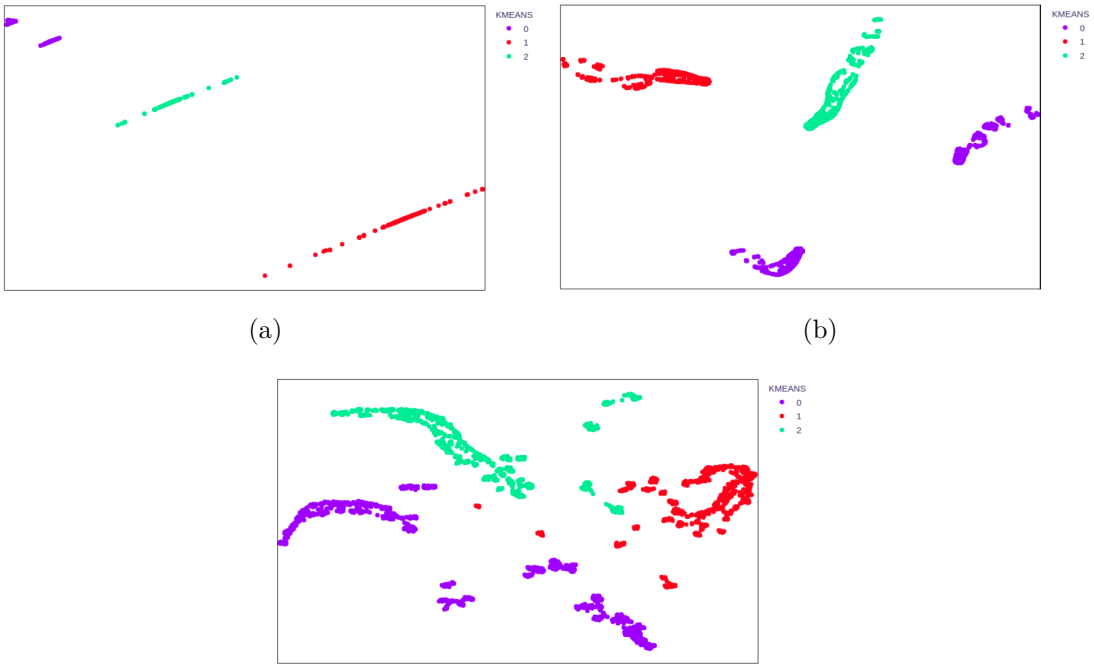



Figura 5.2: Resultados da métrica Neighborhood Preservation. (Fonte: Autoria própria).

A Figura 5.3 ilustra as visualizações obtidas pelas técnicas de visualização usando a distância customizada. Os pontos foram coloridos a partir dos rótulos obtidos pelo K-Means para $K = 3$, em que este valor foi escolhido como palpite inicial pelo especialista devido ao número de produtos considerados em $\mathcal{S}$: gin, vodka e cigarro. Pode-se observar que t-SNE (Figura 5.3(b)) e MDS (Figura 5.3(a)) produziram *layouts* nos quais os grupos apresentam mais separabilidade quando comparados a UMAP (Figura 5.3(c)).

Foram realizados experimentos com o objetivo de avaliar visualmente e quantitativamente a qualidade da coloração dos pontos utilizando o algoritmo K-Means, a DESC_1W




(a) (b)

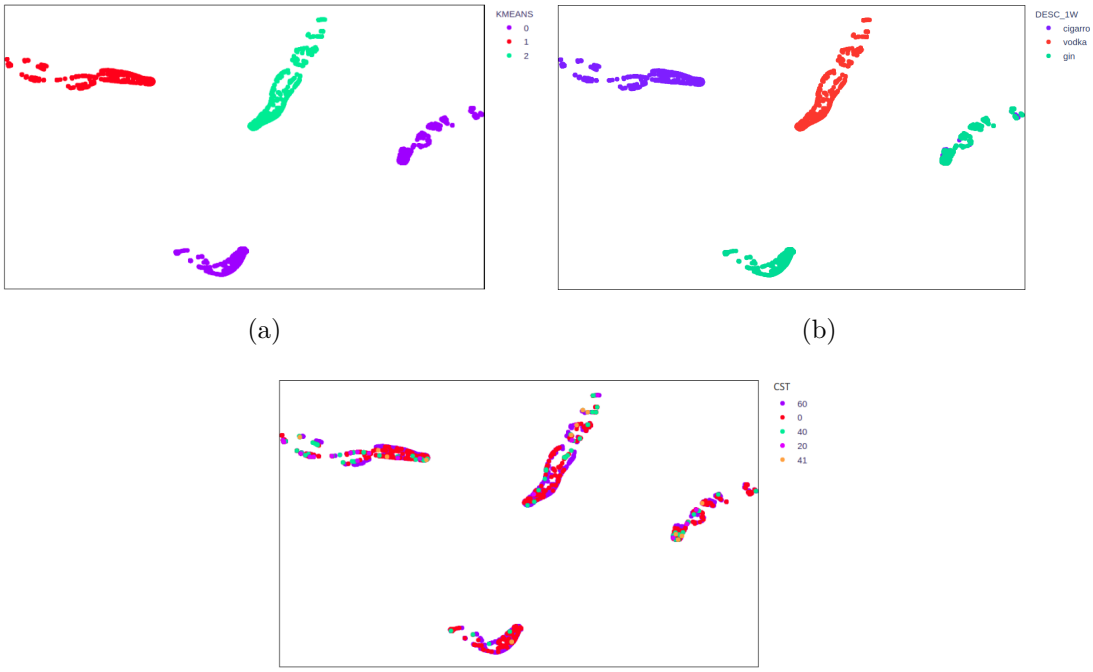
(c)

Figura 5.3: Visualizações obtidas utilizando a distância customizada. Os pontos foram coloridos com os rótulos do agrupamento K-Means, com $K = 3$; (a) MDS; (b) t-SNE; (c) UMAP. (Fonte: Autoria própria).

e o CST, como pode ser observado nas visualizações da Figura 5.4 e na métrica Coeficiente de Silhueta da Tabela 5.1. Devido à grande quantidade de NCMs existentes, optou-se por não adicioná-la como possibilidade de coloração dos pontos nessa análise.

Nas visualizações, pode-se notar uma coloração mais uniforme e agrupada utilizando o algoritmo K-Means e o atributo DESC_1W, enquanto na visualização colorida pelo atributo CST não foi possível identificar um padrão entre as cores e os grupos de pontos na visualização. Esse fato indica a existência de diversos tipos de taxa tributária para um mesmo agrupamento ou um mesmo produto, o que deve ser analisado com uma perspectiva de identificação de notas fiscais inconsistentes.

Como indica a Tabela 5.1, o resultado da métrica Coeficiente de Silhueta foi satisfatório para o espaço de alta dimensão utilizando a distância customizada e rótulos do K-Means ou do atributo DESC_1W, obtendo valores aproximados de 0,90 e 0,83 respectivamente. No geral, obter resultados melhores no espaço de alta dimensão era esperado, visto que existe uma perda de informação ao longo do processo de redução de dimensionalidade. Além disso, foi comprovado que, para este subconjunto de dados, a distância customizada se adaptou melhor do que a distância de Gower.




(a) (b)

(c)

Figura 5.4: Estratégias para colorir os pontos em visualizações criadas com distância customizada e t-SNE; (a) Algoritmo K-Means; (b) DESC_1W; (c) CST. (Fonte: Autoria própria).

| Espaço | Rótulo \ Distância | ***Customizada*** | ***Gower*** |
|---|---|---|---|
| Bidimensional (visual) | Rótulos do K-Means | 0.4078 | 0.2926 |
| | DESC_1W | 0.3886 | 0.6223 |
| | CST | -0.1295 | -0.1264 |
| Alta dimensionalidade (original) | Rótulos do K-Means | 0.8963 | 0.2809 |
| | DESC_1W | 0.8353 | 0.2803 |
| | CST | -0.1066 | 0.1575 |

Tabela 5.1: Coeficientes de Silhueta obtidos a partir das medidas de dissimilaridade. O t-SNE foi aplicado em espaços de baixa dimensionalidade para gerar o espaço visual.

## 5.3 Visão Geral do Sistema

Como uma forma de implementação da metodologia citada anteriormente, foi criado um sistema *web* visando auxiliar os especialistas na detecção de subconjuntos de dados inconsistentes, para que estes possam ser investigados de maneira aprofundada. O sistema está exemplificado pela Figura 5.5. Nas subseções a seguir são explicados os requisitos mapeados para que este sistema cumpra o seu objetivo e as respectivas tarefas para a validação dos requisitos.


Figura 5.5: Painel principal da ferramenta construída com base na metodologia proposta. (Fonte: Autoria própria).

### 5.3.1 Requisitos do Sistema

São citadas abaixo as questões consideradas para a construção do sistema. Para a definição delas, foi necessário consultar um especialista em análise de inconsistências em notas fiscais para aumentar o entendimento acerca de como uma ferramenta poderia auxiliá-lo nesse processo. As questões fornecidas são citadas a seguir:

- Q1. Quais notas fiscais possuem valores de NCM e CST incompatíveis?
- Q2. Quanto as descrições das notas fiscais são distintas das descrições oficiais provenientes do NCM?
- Q3. Em um tipo de produto específico, existem notas fiscais com valores de descrição ou CST distintos?
- Q4. Em agrupamento por similaridade, existem notas fiscais com valores de descrição, NCM ou CST distintos da maioria?

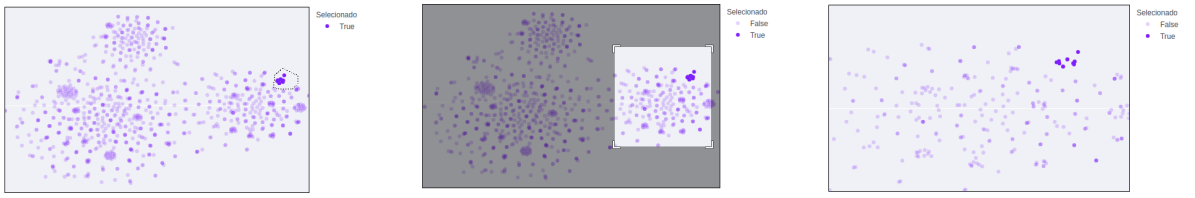
### 5.3.2 Tarefas

A partir das questões citadas anteriormente, foram definidas tarefas práticas para simular o processo real de busca de inconsistências. Ao final de cada tarefa será citada entre

parênteses a questão relacionada. Logo em seguida, as tarefas são simuladas utilizando a ferramenta *web* proposta. As tarefas definidas são citadas a seguir:

- **T1 - Visão geral dos dados conforme a similaridade entre cada um:** O sistema deve disponibilizar uma visualização de aspecto global contendo todas as instâncias em questão, cuja distância entre elas seja definida por quão similares elas são entre si (Q1,Q2,Q3).
- **T2 - Filtro e seleção de instâncias:** O sistema deve suportar o filtro e a seleção de instâncias pela descrição, NCM e CST (Q3).
- **T3 - Visualização de agrupamentos:** O sistema deve identificar e agrupar instâncias similares (Q4).
- **T4 - Ênfase de Notas Fiscais com descrições inconsistentes:** O sistema deve identificar Notas Fiscais que possuem uma descrição com uma alta dissimilaridade em relação às descrições oficiais provenientes do NCM (Q2).
- **T5 - Identificação de produtos com o mesmo NCM e distintos valores de CST:** O sistema deve permitir a análise do NCM e do CST de um conjunto de instâncias (Q1).

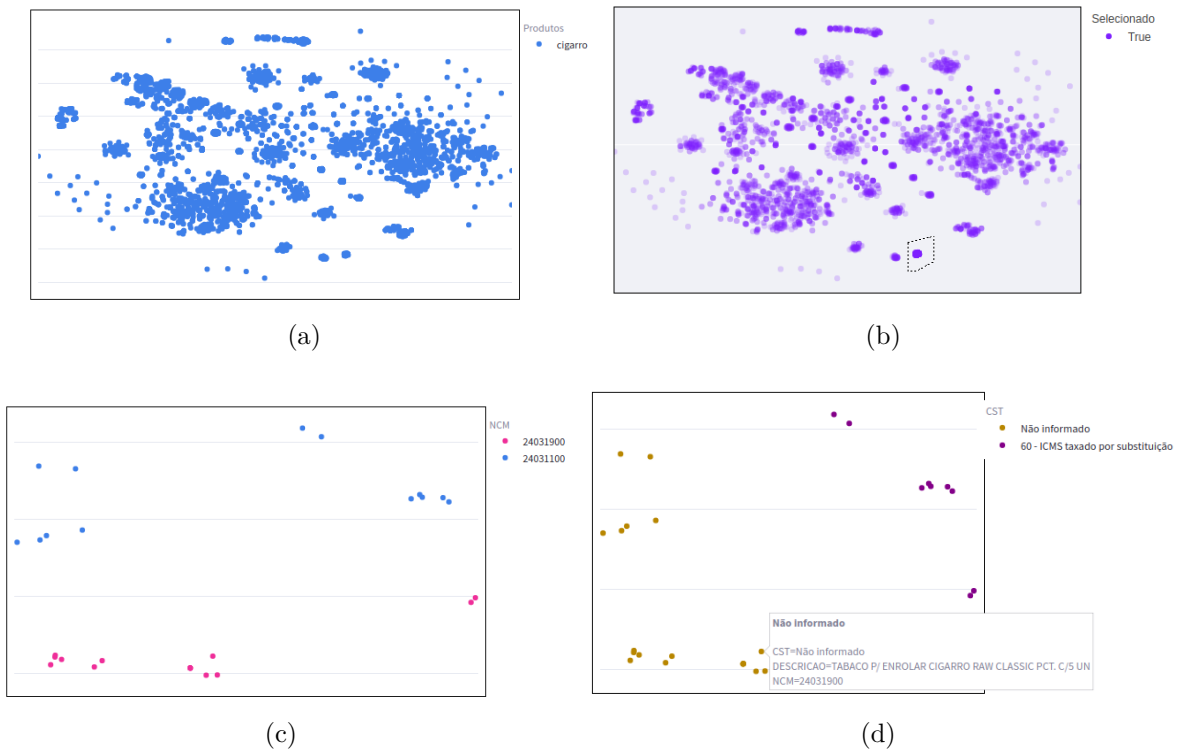
A Figura 5.6 demonstra o processo de seleção de pontos associados às notas fiscais na visualização para análises mais específicas e de aplicação de zoom para facilitar a análise de cada instância individualmente.




(a) (b) (c)

Figura 5.6: Visualização da ferramenta com suporte a filtro e seleção de instâncias; (a) Seleção de um subconjunto de dados na visualização; (b) Aplicação de zoom na área em que o subconjunto foi selecionado; (c) Subconjunto selecionado com zoom. (Fonte: Autoria própria).

A questão Q1 pode ser exemplificada pela Figura 5.7, que inicia o processo de busca por inconsistência através da seleção de um conjunto de pontos cujas descrições DESC_1W 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) (b)

(c) (d)

Figura 5.7: Processo de busca de notas fiscais com o mesmo NCM, mas distintos valores de CST, conforme descreve a questão Q1; (a) Visualização t-SNE gerada com produtos que continham a palavra "cigarro" na descrição; (b) Seleção de um grupo menor de pontos; (c) Visualização dos pontos coloridos pelo NCM; (d) Visualização dos pontos coloridos por CST. (Fonte: Autoria própria).

A questão Q2 pode ser exemplificada pela Figura 5.8, que utiliza notas fiscais de diversos produtos. Após a visualização e seleção de pontos cujo CST seja do tipo "Não informado", foi aplicada a coloração por rótulo de consistência da similaridade Edit Distance com a descrição oficial, definido no capítulo anterior. Neste caso, o especialista em auditoria pode ajustar o valor de similaridade a partir do qual a nota será considerada inconsistente, gerando assim, um conjunto de notas a serem analisadas de forma mais aprofundada.

A questão Q3 pode ser exemplificada pela Figura 5.9, que utiliza notas fiscais de uma categoria específica de produtos que possuem o $NCM$ = "22011000" para analisar os valores de descrição, NCM e CST. Como os primeiros 2 números do código NCM é "22", essas notas idealmente possuem a classificação geral de "Bebidas, líquidos alcoólicos e vinagres", e, por terem os 2 próximos números "01", elas possuem a classificação específica de "Águas, incluindo as águas minerais, naturais ou artificiais, e as águas gaseificadas, não adicionadas de açúcar ou de outros edulcorantes nem aromatizadas; gelo e neve", conforme as normas de tributação do Ministério da Fazenda.

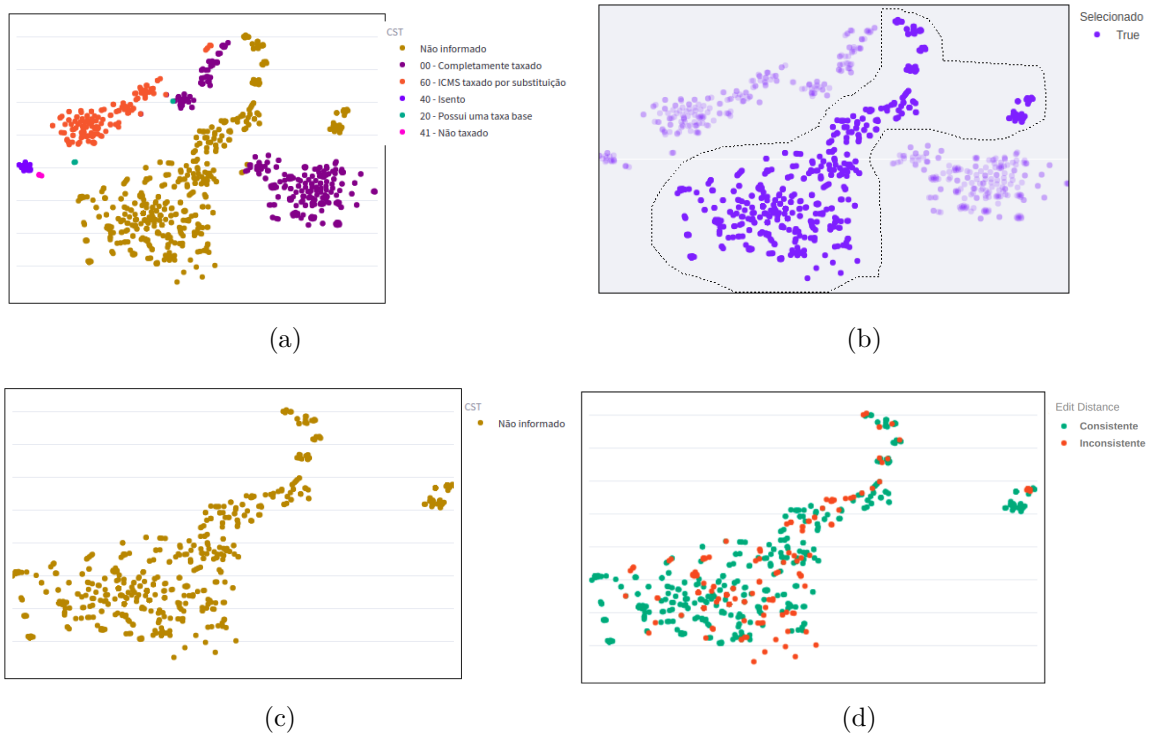



Figura 5.8: Processo de busca de notas fiscais com descrições distintas em relação às descrições oficiais, conforme descreve a questão Q2; (a) Visualização t-SNE de um subconjunto de notas fiscais; (b) Seleção de pontos que representam notas fiscais com valor CST não informado; (c) Resultado da seleção de pontos da imagem anterior; (d) Visualização dos pontos coloridos pelo rótulo de consistência. (Fonte: Autoria própria).

De forma contrária ao esperado, pode-se notar que existem conjuntos de pontos que não possuem descrições equivalentes a águas, como, por exemplo, “guardanapo naxim 50 guardanapos” e “coifa lado roda”, além do CST também variar. Notou-se também a existência de um produto com a descrição “gt gin tonica” e CST “Não informado”, que deveria se enquadrar em um NCM com o código inicial “2208” de classificação específica de “Álcool etílico não desnaturado, com um teor alcoólico, em volume, inferior a 80% vol.; aguardentes, licores e outras bebidas espirituosas”, ao invés de “2201”. Todos estes casos citados devem ser investigados, por se tratarem de notas fiscais com valores de descrição e CST distintos, como prevê a questão Q3.

Por fim, o processo de resolução da questão Q4 pode ser exemplificada pela Figura 5.10, que faz uma comparação entre pontos gerados por notas que possuem a descrição com “água mineral”, “gin” e “suco”, e a coloração pelo algoritmo K-Means. Por meio dessa comparação, percebe-se que foram agrupados pontos de distintos produtos em um mesmo agrupamento e que alguns deles possuem CST “Não informado”, assim, estes pontos devem ser analisados por serem a resposta da questão Q4.

A formação de agrupamentos prevista na questão Q4 também pode ser explorada por

meio da execução do K-Means com distintos valores de K, visando encontrar subgrupos de produtos nas visualizações. Estes subgrupos pertencentes a um mesmo produto indicam um conjunto de notas fiscais semelhantes entre si que foram agrupadas por possuírem características mais distintas das outras notas do mesmo produto. Com isso em vista, essas situações devem ser analisadas como possíveis inconsistências, assim como exemplifica a Figura 5.11.

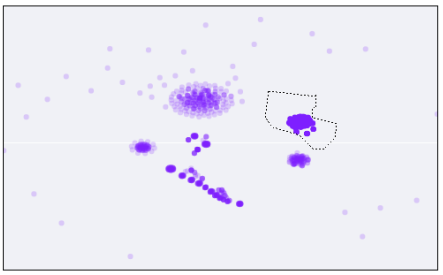
### 5.3.3 Considerações Finais

Neste capítulo foram apresentados a metodologia e os resultados experimentais, enfatizando uma abordagem visual interativa por meio de um sistema *web*, apoiado por técnicas que detectam automaticamente dados inconsistentes por meio de agrupamentos e similaridade. Foram realizados experimentos visuais e quantitativos para avaliar as visualizações, além da criação de requisitos necessários para que os especialistas em auditoria fiscal utilizem o sistema desenvolvido.


(a)



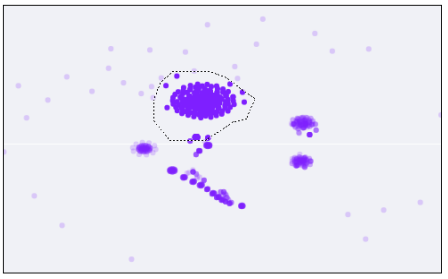

(b)




(c)

| | DESCRICAO | NCM | CST |
|---|---|---|---|
| 10 | AGUA MINERAL CRYSTAL C GAS 500ML | 22011000 | 60 - ICMS taxado por substituição |
| 11 | AGUA MINERAL (IFOOD) | 22011000 | 60 - ICMS taxado por substituição |
| 18 | AGUA MINERAL INDAIA C/GAS 1.5L | 22011000 | 60 - ICMS taxado por substituição |
| 58 | AGUA MINERAL LIA COPO 200MLDE:1,63 | 22011000 | 60 - ICMS taxado por substituição |
| 58 | AGUA MINERAL LIA COPO 200MLDE:1,63 | 22011000 | 60 - ICMS taxado por substituição |
| 71 | AGUA MINERAL LIA COPO 200MLDE:1,63 | 22011000 | 60 - ICMS taxado por substituição |

(d)

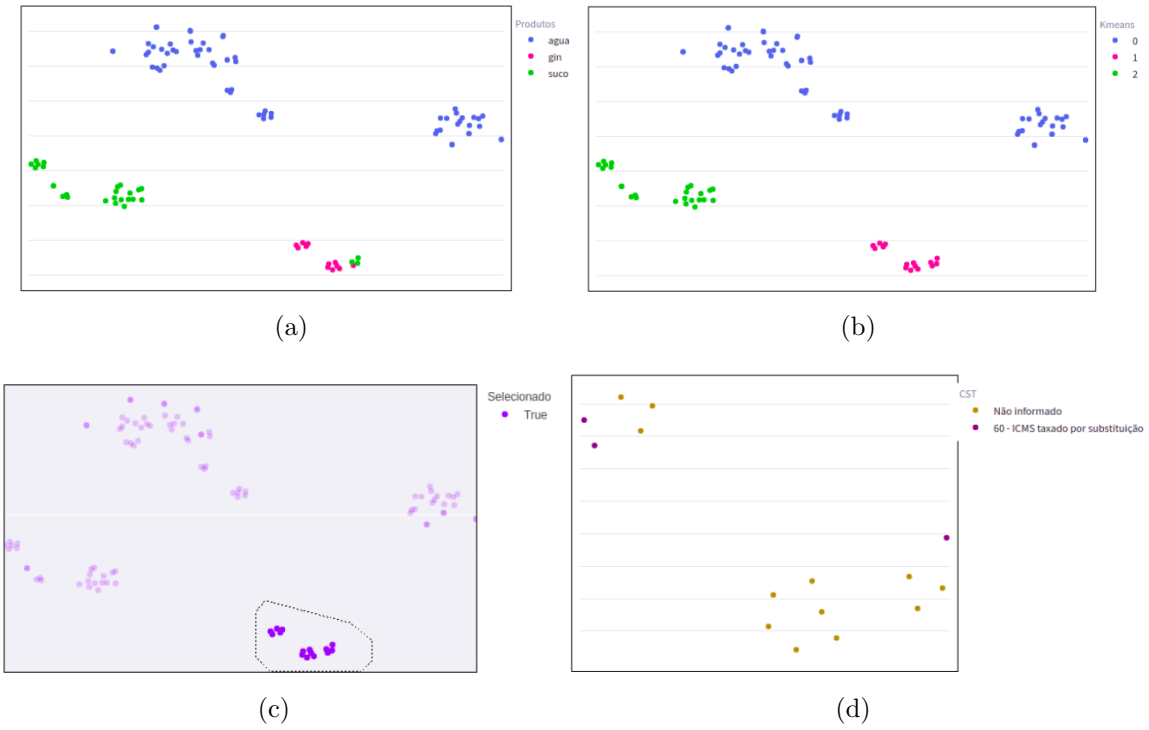



(e)

| | DESCRICAO | NCM | CST |
|---|---|---|---|
| 98 | MANTEIGA COM SAL ITALAC | 22011000 | Não informado |
| 99 | COIFA LADO RODA | 22011000 | 60 - ICMS taxado por substituição |
| 101 | GUARDANAPO MAXIM 50 GUARDANAPOS | 22011000 | Não informado |
| 106 | FILE BOVINO FAMILIA | 22011000 | Não informado |
| 108 | HIDRAT PAIXAO 200ML ONLY YOU | 22011000 | Não informado |
| 122 | IFOOD - ENERGETICO RED BULL SUGAR I | 22011000 | Não informado |

(f)

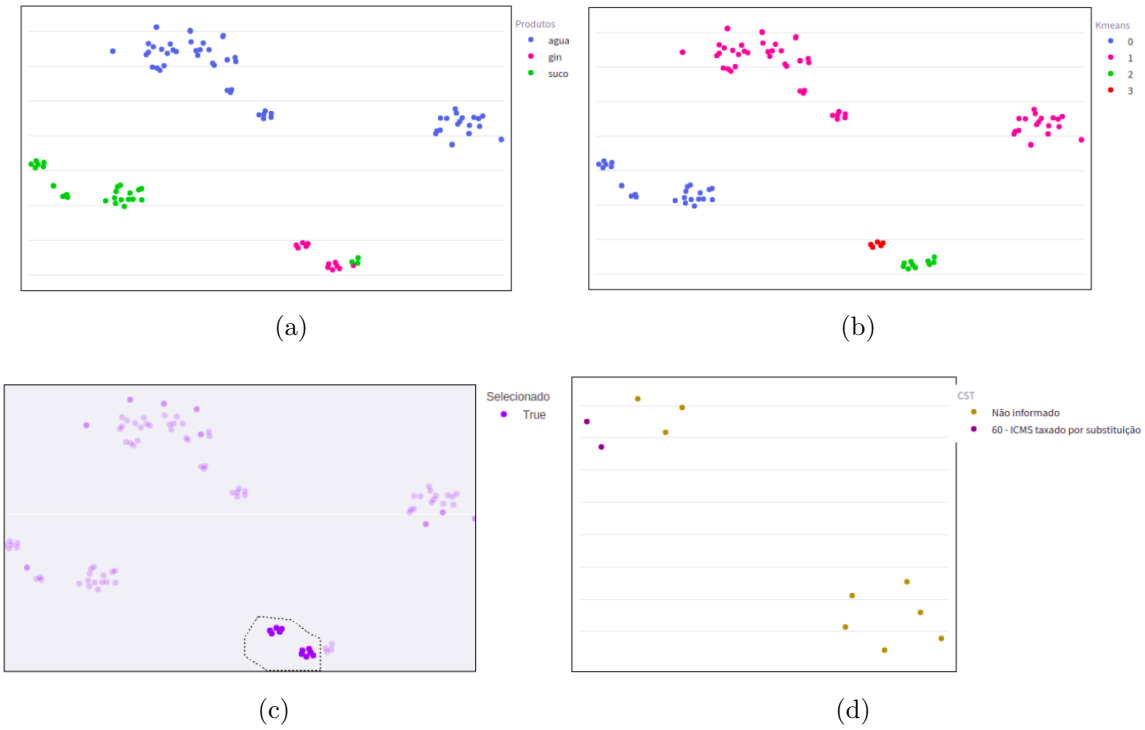
Figura 5.9: Processo de busca de notas fiscais de um produto específico que possuem valores de descrição e NCM distintos, conforme descreve a questão Q3; (a) Visualização t-SNE de pontos que representam notas fiscais com NCM igual a “22011000”; (b) Visualização de pontos coloridos por descrições que contenham a palavra “gin”; (c) Seleção de um conjunto de pontos similares conforme a estratégia de posicionamento de pontos; (d) Descrição, NCM e CST de algumas das notas selecionadas; (e) Nova seleção de conjunto de pontos similares conforme a estratégia de posicionamento de pontos; (f) Descrição, NCM e CST de algumas das notas selecionadas. (Fonte: Autoria própria).




(a) (b)

(c) (d)

Figura 5.10: Processo de busca de notas fiscais de um mesmo agrupamento que possuem valores de descrição, CST ou NCM distintos, conforme descreve a questão Q4; (a) Visualização t-SNE de pontos representando notas fiscais que continham as palavras “agua mineral”, “gin” ou “suco” na descrição; (b) Visualização de pontos coloridos por agrupamentos gerados pelo K-Means; (c) Seleção de pontos na visualização; (d) Visualização de pontos coloridos por CST. (Fonte: Autoria própria).




(a) (b)

(c) (d)

Figura 5.11: Processo de busca de notas fiscais de um mesmo produto que fazem parte de agrupamentos distintos; (a) Visualização t-SNE de pontos representando notas fiscais que continham as palavras “agua mineral”, “gin” ou “suco” na descrição; (b) Visualização de pontos coloridos por agrupamentos gerados pelo K-Means, com $K = 4$; (c) Seleção de pontos na visualização; (d) Visualização de pontos coloridos por CST. (Fonte: Autoria própria).

# Capítulo 6

# Conclusão

Esta pesquisa apresentou uma abordagem inovadora para a fiscalização e detecção de casos de inconsistências fiscais, enfrentando o desafio de lidar com a grande quantidade de notas fiscais geradas diariamente. O objetivo foi auxiliar os especialistas na análise das NFC-es e, para esse propósito, duas abordagens foram propostas: uma baseada em rotulação por similaridade de descrições e outra baseada em visualização interativa de notas fiscais.

Por meio de técnicas de Processamento de Linguagem Natural combinadas com medidas de similaridade, foi definido um rótulo de inconsistência para as notas fiscais, auxiliando o especialista na fase de descobrimento de casos para análise. Foram exploradas também diversas representações de texto para reproduzir os rótulos criados, como Word2vec, Doc2vec, Transformer Distiluse Multilingual e BERT, em conjunto com medidas de similaridade cosseno e a Edit Distance.

O desenvolvimento de um método para identificação de descrições de produtos inconsistentes, por meio da comparação com as descrições oficiais, e a criação de agrupamentos pode auxiliar os auditores fiscais no quesito de eficiência. As visualizações geradas considerando os produtos gin, vodka e cigarro obtiveram um Coeficiente de Silhueta acima de 0,83 no espaço de alta dimensionalidade e uma preservação de vizinhança (Neighborhood Preservation) acima de 0,80 para t-SNE com distância customizada, demonstrando a qualidade do posicionamento de pontos e da preservação das distâncias mesmo após a redução de dimensionalidade.

Para aplicação em produção da metodologia, foi criada uma ferramenta *web* interativa de visualização, permitindo que os analistas fiscais analisem visualmente a similaridade entre as notas fiscais, identifiquem agrupamentos e acessem estatísticas relevantes para a tomada de decisão. A interatividade da ferramenta em conjunto com o modelo preditivo fornece um maior controle ao especialista no processo de detecção de inconsistências em notas fiscais.

## 6.1 Trabalhos Futuros

A partir dos resultados alcançados nesta pesquisa, é importante destacar que há oportunidades para aprimorar ainda mais a performance da ferramenta de análise de fraudes fiscais, que atualmente fornece resultados em um tempo razoável para 1.000 notas por vez. Nesse contexto, podem ser utilizados bancos integrados à aplicação para diminuir o tempo de obtenção de dados.

Além disso, é essencial explorar novos conjuntos de dados relacionados à fraude fiscal, como, por exemplo, os conjuntos de dados abertos Nota Fiscal Eletrônica (NF-e), disponibilizado recentemente pelo Portal de Transparência do Governo, e o ELEVEN [41], que possui dados de produtos rotulados. Estes novos conjuntos de dados possuem mais atributos e, consequentemente, mais possibilidades de atuação nessa área de pesquisa. A utilização desses novos atributos pode potencialmente aprimorar a capacidade da ferramenta de detectar padrões de fraude mais complexos e sutis.

Por fim, os próximos passos também devem envolver a distinção entre inconsistências explícitas, nas quais existe a intenção de pagar menos imposto, e implícitas, nas quais é inserido um erro não intencional, além de uma avaliação qualitativa da execução das tarefas por meio de especialistas, assim como foi feito na ferramenta LegalVis, proposta por Lucas Resck [6]. Dessa forma, a ferramenta pode ser melhorada para que os especialistas a utilizem como uma forma de otimização de esforços na busca por inconsistências em notas fiscais eletrônicas.

## 6.2 Publicações obtidas

Em 2022 foi publicado o artigo *Visual Analysis of Electronic Invoices to Identify Suspicious Cases of Tax Frauds* na conferência internacional ICITS, *International Conference on Information Technology & Systems*. Ele contém os avanços iniciais da pesquisa apresentada por meio dessa monografia e utiliza um subconjunto de dados pré-processados da NFC-e. Esse artigo apresenta estratégias de posicionamento de pontos para representar as notas fiscais visualmente, permitindo a identificação de notas semelhantes em termos de descrição do produto, mas com diferentes categorias tributárias. Resultados experimentais e um estudo de caso demonstraram a eficácia do método proposto na detecção de notas suspeitas de fraude, enfatizando a importância da análise visual e interativa para investigações mais aprofundadas pelos especialistas.

# Referências


[1] Hobson Lane, Cole Howard e Hannes Max Hapke: *Natural Language Processing.* Hanning, 2019. ix, 9, 11, 12, 13

[2] Mikolov, Tomas, Wen tau Yih e Geoffrey Zweig: *Linguistic regularities in continuous space word representations.* Em *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2013. ix, 13, 14

[3] Diego Santos Kieckbusch, Geraldo Pereira Rocha Filho, Vinicius Di Oliveira Li Weigang: *Towards intelligent processing of electronic invoices: The general framework and case study of short text deep learning in brazil.* Em *International Conference on Web Information Systems and Technologies*, páginas 74–92. Springer International Publishing, 2020. ix, 27, 28

[4] Knyazeva, Margarita, Alexander Tselykh, Alexey Tselykh e Elena Popkova: *A graph-based data mining approach to preventing financial fraud: a case study.* Em *8th International Conference on Security of Information and Networks*, páginas 109–113, 2015. ix, 1, 27, 29, 43

[5] Zha, Zhichao: *Taxaa: A reliable tax auditor assistant for exploring suspicious transactions.* Em *Companion Proceedings of the Web Conference 2020*, páginas 240–244, 2020. ix, x, 1, 28, 29, 30

[6] Resck, Lucas E., Jean R. Ponciano, Luis Gustavo Nonato e Jorge Poco: *LegalVis: Exploring and inferring precedent citations in legal documents.* IEEE Transactions on Visualization and Computer Graphics, 29, 2023. x, 1, 29, 30, 31, 56

[7] Bolton, Richard e David Hand: *Unsupervised profiling methods for fraud detection.* Conference on Credit Scoring and Credit Control, 7, setembro 2001. 1, 27

[8] Heimerl, Florian e Michael Gleicher: *Visual exploration of word vector embeddings.* IEEE Visualization Poster Proceedings, 2017. `http://graphics.cs.wisc.edu/Papers/2017/HG17`. 1, 29

[9] Cox, Michael AA e Trevor F Cox: *Multidimensional scaling.* Em *Handbook of data visualization*, páginas 315–347. Springer, 2008. 3, 22

[10] Maaten, Laurens Van der e Geoffrey Hinton: *Visualizing data using t-sne.* Journal of machine learning research, 9(11), 2008. 3, 23, 35

[11] McInnes, Leland, John Healy e James Melville: *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv preprint arXiv:1802.03426, 2018. 3, 23

[12] Liu, Ruijun, Yuqian Shi, Changjiang Ji e Ming Jia: *A survey of sentiment analysis based on transfer learning*. IEEE Access, 7:85401–85412, 2019. 6

[13] Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev e Dimitar Trajanov: *Evaluation of sentiment analysis in finance: From lexicons to transformers*. IEEE Access, 8:131662–131682, 2020. 6

[14] Maruf, Sameen, Fahimeh Saleh e Gholamreza Haffari: *A survey on document-level neural machine translation: Methods and evaluation*. ACM Comput. Surv., 54(2), mar 2021, ISSN 0360-0300. `https://doi-org.ez54.periodicos.capes.gov.br/10.1145/3441691`. 6

[15] Yadav, Divakar, Rishabh Katna, Arun Kumar Yadav e Jorge Morato: *Feature based automatic text summarization methods: A comprehensive state-of-the-art survey*. IEEE Access, 10:133981–134003, 2022. 6

[16] Lane, H., H. Hapke e C. Howard: *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications, 2019, ISBN 9781617294631. `https://books.google.com.br/books?id=UyHgswEACAAJ`. 7, 8

[17] Hagiwara, Masato: *Real-World Natural Language Processing*. Hanning, 2021. 8, 13, 14, 17, 18

[18] Babić, Karlo, Sanda Martinčić-Ipšić e Ana Meštrović: *Survey of neural text representation models*. Information, 11(11), 2020, ISSN 2078-2489. `https://www.mdpi.com/2078-2489/11/11/511`. 8

[19] Luhn, H. P.: *A statistical approach to mechanized encoding and searching of literary information*. IBM Journal of Research and Development, 1(4):309–317, 1957. 8

[20] Han J., Pei J., Kamber M.: *Data Mining: Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*. Elsevier, 2011. 9

[21] Aggarwal, Charu C.: *Neural Networks and Deep Learning*. Springer, 2018, ISBN 978-3-319-94462-3. 12

[22] Hochreiter, Sepp e Jürgen Schmidhuber: *Long short-term memory*. Neural Comput., 9(8), 1997, ISSN 0899-7667. 12

[23] Greff, Klaus, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink e Jürgen Schmidhuber: *LSTM: A search space odyssey*. CoRR, 2015. 12

[24] Murthy, Dr, Shanmukha Allu, Bhargavi Andhavarapu e Mounika Bagadi: *Text based sentiment analysis using lstm*. International Journal of Engineering Research and, V9, 2020. 12

[25] Almeida, Felipe e Geraldo Xexéo: *Word embeddings: A survey*, 2023. 13

[26] Mikolov, Tomas, Kai Chen, Greg Corrado e Jeffrey Dean: *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013. 14

[27] Le, Quoc e Tomas Mikolov: *Distributed representations of sentences and documents*. Em *International Conference on Machine Learning*, páginas 1188–1196. PMLR, 2014. 14

[28] Devlin, Jacob, Ming Wei Chang, Kenton Lee e Kristina Toutanova: *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018. 14

[29] Bowyer, Kevin W., Nitesh V. Chawla, Lawrence O. Hall e W. Philip Kegelmeyer: *SMOTE: synthetic minority over-sampling technique*. CoRR, abs/1106.1813, 2011. `http://arxiv.org/abs/1106.1813`. 15

[30] Hartigan, J. A. e M. A. Wong: *Algorithm AS 136: A K-Means clustering algorithm*. Applied Statistics, 28:100–108, 1979, ISSN 00359254. 18

[31] Levenshtein, Vladimir Iosifovich: *Binary codes capable of correcting deletions, insertions and reversals*. Soviet Physics Doklady, 10(8), 1966. 20

[32] Singhal, Amit: *Modern information retrieval: A brief overview*. IEEE Data Eng. Bull., 24:35–43, 2001. 20

[33] Gower, John C: *A general coefficient of similarity and some of its properties*. Biometrics, páginas 857–871, 1971. 20

[34] BELLMAN, RICHARD: *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961, ISBN 9780691079011. `http://www.jstor.org/stable/j.ctt183ph6v`, acesso em 2023-07-13. 21

[35] Paulovich, Fernando V e Rosane Minghim: *Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections*. IEEE Transactions on Visualization and Computer Graphics, 14(6):1229–1236, 2008. 21, 25

[36] Rousseeuw, Peter J: *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. J. of Computational and Applied Mathematics, 20:53–65, 1987. 25

[37] Diego Kieckbusch, Geraldo Filho, Li Di Oliveira Vinícius e Weigang: *Scan-nf: A cnn-based system for the classification of electronic invoices through short-text product description*. Em *Proceedings of the 17th International Conference on Web Information Systems and Technologies (WEBIST 2021)*, páginas 501–508, 2021. 27

[38] Prechelt, Lutz: *Early Stopping - But When?*, páginas 55–69. Springer Berlin Heidelberg, 1998, ISBN 978-3-540-49430-0. `https://doi.org/10.1007/3-540-49430-8_3`. 37

[39] Bakhashwain, Norah e Alaa El. Sagheer: *Online tuning of hyperparameters in deep lstm for time series applications*. International Journal of Intelligent Engineering and Systems, 2021. 37, 38

[40] Bergstra, James e Yoshua Bengio: *Random search for hyper-parameter optimization*. J. Mach. Learn. Res., 13:281–305, 2012. 38

[41] Filho, Vinícius Di Oliveira; Li Weigang; Geraldo: *Eleven data-set: A labeled set of descriptions of goods captured from brazilian electronic invoices*. Em *18th International Conference on Web Information Systems and Technologies*, páginas 257–264, 2022. 56