**UNIVERSIDADE DE SÃO PAULO**
**Centro de Ciências Matemáticas Aplicadas à Indústria**

**Robertson da Silva Pereira**

# Redes Heterogêneas para Classificação de Produtos em Notas Fiscais Eletrônicas de Compras Públicas

**São Carlos**

**2020**

**Robertson da Silva Pereira**

# Redes Heterogêneas para Classificação de Produtos em Notas Fiscais Eletrônicas de Compras Públicas

Trabalho de Conclusão do Curso em Ciências de Dados do Centro de Ciências Matemáticas Aplicada à Indústria, Universidade de São Paulo, como parte dos requisitos para a obtenção do título de MBA em Ciências de Dados.

Área de concentração: Ciência de Dados
Opção: MBA

Orientador: Prof. Dr. Alneu de Andrade Lopes

**Versão original**

**São Carlos**
**2020**

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

Pereira, Robertson da Silva
P436r Redes Heterogêneas para Classificação de Produtos em Notas Fiscais Eletrônicas de Compras Públicas / Robertson da Silva Pereira; orientador Alneu de Andrade Lopes. -- São Carlos, 2020.
34 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. CE610.4.1. 2. CE610.4.21. I. Lopes, Alneu de Andrade, orient. II. Título.

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

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.
No presente modelo consta como
folhadeaprovacao.pdf

# RESUMO

P, R. S. **Redes Heterogêneas para Classificação de Produtos em Notas Fiscais Eletrônicas de Compras Públicas**. 2020. 34p. Trabalho de Conclusão de Curso - Nome da Unidade USP, Universidade de São Paulo, São Carlos, 2020.

Uma das tarefas essenciais da auditoria governamental é monitorar como a despesa pública é executada, dessa forma, o documento fiscal é fonte de dados indispensável para o acompanhamento dos gastos públicos, entre as informações nele registradas está inclusive o possível sobrepreço de produtos. No entanto, pela quantidade de documentos fiscais emitidos diariamente a verificação manual desses documentos se torna inviável, mesmo em um contexto reduzido. Para a verificação desse grande volume de informação uma das saídas é a utilização de técnicas de ciência de dados. A ciência de dados possui diversos modelos para a representação de informação. Uma das estratégias é representar a informação no formato de redes. Tal representação, comparada com a representação em tabela atributo-valor, usual, tem como vantagens, entre outros, os seguintes fatores: 1) alguns fenômenos são naturalmente representados por redes, pois são fenômenos com diversas partes ou componentes, relacionados entre si (é o caso das compras públicas); 2) a representação em rede é simples e direta, basta definir, por exemplo, os tipos de vértices e as relações entre eles (arestas); 3) uma vez gerada a rede, sua visualização já permite perceber características importantes do sistemas modelados, por exemplo, vértices mais importantes, grupos de vértices mais densamente conectados, entre outras características. Quando vértices de uma rede possuem mais de um tipo de entidade, essas redes são definidas como redes heterogêneas. Neste trabalho foram utilizadas redes heterogêneas para classificação de produtos contidos em documentos fiscais a partir de sua descrição textual. Para tanto aplicou-se o algoritmo IMBHN, que gera um modelo indutivo baseado em rede heterogênea bipartida para classificação. Verificou-se bons resultados para um número pequeno de classes. Foi possível com a detecção de outliers, identificar superfaturamento em compras públicas.

**Palavras-chave**: Redes Heterogêneas, Notas Fiscais Eletrônicas (NF-e), Aprendizado de Máquina. Classificação de texto, IMBHN, Superfaturamento.

## ABSTRACT

P, R. S. **Heterogeneous Networks for Product Classification in Public Procurement Electronic Invoices**. 2020. 34p. Trabalho de Conclusão de Curso - Nome da Unidade USP, Universidade de São Paulo, São Carlos, 2020.

One of the essential tasks of government auditing is to monitor how public expenditure is carried out, in this way, the electronic invoices (NF-e) is an indispensable source of data for the monitoring of public expenditure, among the information recorded there is even the possible overpricing of products. However, due to the amount of electronic invoices issued daily, manual verification of these documents becomes impracticable, even in a reduced context. To verify this large volume of information, one of the ways out is the use of data science techniques. Data science has several models for representing information. One of the strategies is to represent information in the form of networks. Such representation, compared to the usual attribute-value table representation, has the following factors, among others, advantages: 1) some phenomena are naturally represented by networks, as they are phenomena with several parts or components, related to each other ( as in public procurement); 2) the network representation is simple and direct, it is enough to define, for example, the types of vertices and the relations between them (edges); 3) once the network is generated, its visualization already allows to know the important characteristics of the modeled systems, for example, more important vertices, groups of vertices more densely connected, among other characteristics. When the vertices of a network have more than one type of entity, these networks are defined as heterogeneous networks. In this work, heterogeneous networks were used to classify products contained in electronic invoices based on their textual description. For this, the IMBHN algorithm was applied, which generates an inductive model based on a heterogeneous bipartite network for classification. Good results were found for a small number of classes. It was possible with the detection of outliers, to identify overpricing in public purchases.



**Keywords**: Heterogeneous Networks, Electronic Invoices (NF-e), Machine Learning, Text Classification, IMBHN, Overpricing.

## LISTA DE FIGURAS

## LISTA DE TABELAS

SUMÁRIO

# 1 INTRODUÇÃO

## 1.1 Contextualização

Uma das tarefas essenciais da auditoria governamental é monitorar como a despesa pública é executada, dessa forma, o documento fiscal é fonte de dados indispensável para o acompanhamento dos gastos públicos, entre as informações nele registradas está inclusive o possível sobrepreço de produtos (TCU, 2009).

O Sistema de Nota Fiscal Eletrônica (NF-e) é a principal base de dados de produtos comercializados no país e até o momento, possui mais de 25 bilhões de notas fiscais eletrônicas autorizadas e 1,6 milhão de contribuintes emissores (NFE, 2020).

No entanto, pela quantidade de documentos fiscais emitidos diariamente a verificação manual desses documentos se torna inviável, mesmo em um contexto reduzido. Além disso, não existe padronização na descrição de produtos nas notas fiscais eletrônicas, existindo elevada variabilidade de grafia para os mesmos produtos (LEAL; MAHLER, 2016).

Para a verificação desse grande volume de informação uma das saídas é a utilização de técnicas de ciência de dados (SILVA, 2016).

## 1.2 Trabalhos relacionados

A extração de informações de Notas Fiscais Eletrônicas do Consumidor (NFC-e), modelo bem semelhante ao da Nota Fiscal Eletrônica (NF-e), foi modelada em (SANTOS, 2019). Nesse caso, utilizou-se mineração de texto e classificação de texto para agrupar produtos em classes predefinidas.

A classificação de texto, geralmente, tem como objetivo atribuir rótulos a documentos textuais ou porções de texto. Uma forma viável de realizar a classificação automática de textos é por meio de algoritmos de aprendizado de máquina, que são capazes de aprender, generalizar, ou ainda extrair padrões com base no conteúdo ou rótulos de documentos textuais (ROSSI, 2015).

A categorização de textos é a tarefa de associar um valor booleano para cada par $(d, c) \in D \times C$ onde $D$ é o domínio de documentos e $C = \{c1, ..., c_{|c|}\}$ é um conjunto de categorias predefinidas. Um valor $V$ (verdadeiro) associado ao par $(d_j, c_i)$ indica a decisão de associar o documento $d_j$ na categoria $c_i$, enquanto um valor $F$ (falso) indica a decisão de não associar $d_j$ a $c_i$. Formalmente, a tarefa é aproximar uma função alvo desconhecida $\breve{\Phi}$: $D \times C \rightarrow \{V, F\}$, que descreve como os documentos devem ser classificados por meio de uma função $\Phi$: $D \times C \rightarrow \{V, F\}$ chamada de classificador (regra, hipótese ou modelo)

tal que $\breve{\Phi}$ e $\Phi$ coincidam o máximo possível (SEBASTIANI, 2002).

Geralmente os problemas de classificação de textos utilizam o modelo espaço-vetorial para representação dos dados. Nesse modelo documentos são representados por vetores e as dimensões correspondem aos termos ou atributos dos documentos. Algoritmos baseados no modelo espaço-vetorial consideram que tanto os documentos quanto os termos ou atributos são independentes, o que pode degradar a qualidade da classificação.

Uma alternativa à representação no modelo espaço-vetorial é a representação em redes, que permite modelar relações entre entidades de uma coleção de textos. Quando vértices de uma rede possuem mais de um tipo de entidade, essas redes são definidas como redes heterogêneas (ROSSI, 2015) (FALEIROS, 2016). Um caso particular de redes heterogêneas são as redes bipartidas, que são redes heterogêneas com dois tipos de vértices e ligações apenas entre vértices de tipo diferentes.

O objetivo do trabalho é representar as informações encontradas em documentos fiscais (NFe) em redes heterogêneas. Como objetivos específicos vamos 1) utilizar o algoritmo IMBHN para classificar os produtos contidos nos documentos fiscais de acordo com os termos que aparecem na sua descrição e 2) encontrar outliers, produtos adquiridos por um valor acima da média entre seus similares.

## 1.3 Organização do Texto

O Capítulo 2 descreve os principais conceitos necessários para o entendimento do projeto, técnicas diretamente relacionadas com a proposta bem como trabalhos relacionados. O Capítulo 3 descreve como as Notas Fiscais Eletrônicas serão representadas por dois tipos de entidades em rede heterogênea bipartida, a primeira será o campo xprod da NFe que contém a descrição do produto e a segunda será formada pelos termos/palavras que compõem a descrição do produto. Essa rede será utilizada para agrupar produtos em classes predefinidas. O Capítulo 4 apresenta os resultados obtidos na classificação. O Capítulo 5 trará as conclusões acerca do modelo e também tecerá considerações sobre trabalhos futuros.

# 2 FUNDAMENTAÇÃO E TRABALHOS RELACIONADOS

Neste capítulo descrevemos os principais conceitos necessários para o entendimento do projeto, bem como as técnicas diretamente relacionadas com a proposta.

Com o intuito de conhecer o estado da arte da classificação automática de textos foi realizado levantamento nas bases de dados com acesso disponível no momento da pesquisa, Google Acadêmico, Elsevier Science Direct e IEEEXplore Digital Library com os seguintes termos de busca TERMO 1: ("text classification") AND ("machine learning"OR "deep learning"), TERMO 2 ("graph based text classification"). A consulta ao TERMO 1 na base de dados IEEEXplore Digital Library retornou 649 artigos entre 2016 e 2020. A consulta ao TERMO 2 na base de dados Elsevier Science Direct retornou 13.479 artigos entre 2016 e 2020.

Após a leitura de parte dos títulos e resumos dos trabalhos melhor relacionados ao problema proposto foram elencados alguns estudos descritos a seguir.

Em (MIKOLOV et al., 2013) Foi apresentado o modelo Word2Vec para representação de texto em espaço vetorial. Atualmente é o estado da arte para extração de atributos utilizado em deep learning.

Em (Al-Doulat; Obaidat; Lee, 2019) Foi proposta uma rede neural para classificação de artigos médicos. Foi proposta extração de características de estilo baseados em componentes sintáticos e gramaticais do texto e extração de característica de comlexidade baseada no tamanho das sentenças, número de palavras e número de sílabas das palavras.

Além das características baseadas em NLP também foram extraídas características de domínio, ou seja, termos específicos ou palavras chaves que são correlacionadas a uma classe médica específica. Como resultado obteve-se um conjunto de palavras para cada classe no banco de dados.

Para a seleção de características mais signifigativas, ou seja, com maior correlação com a classe alvo, foi utilizado o método Recursive Feature Elimination (RFE), que automaticamente busca o melhor número de características utilizando cross-validation.

A rede neural foi treinada utilizando o conjunto de características selecionadas no passo anterior. Múltiplas configurações foram testadas: 2 a 6 camadas, 20 a 100 neurônios por camada, funções de ativação no-op activation, logistic sigmoid, hyperbolic tan, and rectified linear unit. Bem como foram utilizados os otimizadores: sgd, adam, lbfgs, e diferentes taxas de aprendizado: 0.001, 0.01 e 0.1. Após identificação da melhor configuração, a rede neural resultou em 5 camadas de 100 neurônios cada.

Foi verificado que caracterísiticas baseadas em domínio obtiveram a melhor acurácia

para a predição das classes alvos. A acurária total do modelo foi de 82%, superior à acurácia de 62% alcançada pelo modelo TF/IDF padrão.

Em (Al-Doulat; Obaidat; Lee, 2019) Foi proposta uma nova forma para extração de atributos a partir de dados textuais utilizando representação a nível de caractere. O trabalho se baseia na proposta original de (ZHANG; LECUN, 2016), que representa cada caractere como um vetor de tamanho m, só que em vez de utilizar representação 1-de-m, com um alfabeto de m letras, utiliza-se a representação log(m). A proposta utiliza menos memória e é mais rápida que a original. A extração de atributos é realizada por Rede Neural Convolucional CNN, e encontrou melhores resultados que os métodos para representação Bag-of-Words e Word2Vec.

Em (Gong et al., 2020) foi proposta representação gráfica de dados textuais para classificação de texto multi-label.

Conforme o autor, a representação de texto é tarefa essencial para o processamento em linguagem natural (NLP), e pode significantemente melhorar a eficiência da classificação de texto, análise de sentimentos, sistemas de pergunta e resposta etc. A maioria dos estudos focam na abordagem Bag-of-Words, na qual cada atributo corresponde a uma simples palavra. A frequência do termo e a frequência inversa do documento (TF-IDF) é um método comum de transformar texto em um representação numérica que pode ser usada para extração de atributos.

No entanto, partindo da idéia de que tanto o modelo espaço vetorial quanto os os modelos em Deep Learning (RNN e CNN) que usam geralmente o método Word2Vec, perdem quantidade significativa de informação semântica e estrutural útil para categorização do texto, a representação gráfica do documento pode preservar informação acerca da sequência, distância entre termos e sua não consecutividade.

Para capturar a informação textual de maneira mais completa, a proposta utiliza uma rede neural transformer multi-camada com mecanismo de atenção sobre a representação de palavras, de sentenças e do grafo gerado a partir do texto.

Em (Li; Karunesh; Alaniazar, 2019) foi proposto modelo para rotular dados inicialmente sem rótulo associado, considerando a grande demanda por dados rotulados utilizados por algoritmos de aprendizado supervisionado, e o gargalo que isso representa para a indústria.

Conforme o autor, a Classificação de Texto inclui a extração de características/atributos/aspectos e o modelo de classificação. A extração de atributos ajuda a converter texto livre em representações numéricas que podem ser interpretadas por algoritmos de aprendizado. Para extração de atributos existem os modelos baseados em frequência e modelos baseados em representação latente/intermediária. (embedding)

O modelo Bag-of-Word é o mais popular modelo baseado em frequência que usa

vetores com a quantidade de palavras como atributos do texto. Devido sua natureza intrínseca, modelos baseados em frequência não podem capturar informação de contexto, o que limita a capacidade da classificação de texto. Esparsidade também é outro desafio dos modelos baseados em frequência.

Na representação latente para redes neurais (word embedding), cada palavra é representada por um vetor 1-de-m, onde m é o tamanho do vocabulário. Neste caso temos uma dimensionalidade menor, bem como dados menos esparsos. Segundo o autor, o primeiro trabalho em word embedding data de 2003, uma rede neural treinada para a linguagem, predizendo a próxima palavra baseado no contexto anterior.

O modelo Word2Vec é um dos métodos de embedding mais populares, ele cria uma representação de palavras que captura significado semântico. Treinando uma grande base de dados usando rede neural simples, por exemplo, com somente duas camadas, a acurácia dos embedded vectors é significantemente aumentada. Apesar disso, a limitação do Word2Vec é sua inabilidade em distinguir palavras com múltiplos significados.

Devido às limitações descritas para a classificação de texto, somado a outras restrições como recursos computacionais, dados para treinamento e tempo de processamento o estudo propõe um framework que integra a força desses diferentes modelos para melhorar a performance da classificação.

Em (CHONG et al., 2020) foi apresentado um estudo sobre aprendizado semi-supervisionado (SSL) baseado em grafo. Algoritmos semi-supervisionados utilizam os exemplos rotulados e também exemplos não rotulados para construir um grafo capaz de propagar a informação de rótulos para os exemplos não rotulados e explora a estrutura de relacionamento existente entre os exemplos com e sem rótulos.

Em uma estrutura de grafo, cada exemplo é representado como um vértice, o grafo possui pesos que medem a similaridade entre os exemplos. A etapa chave para o SSL baseado em grafo é construir o melhor grafo que represente os dados originais, o que pode ser separado em duas etapas. Primeiro, construir um grafo de todos os exemplos com e sem rótulo; segundo, a informação de rótulos pode ser propagada para os exemplos sem rótulo através do grafo.

Entre as razões elencadas pelos autores que tornam os algoritmos SSL baseados em grafo bastante atrativos estão a convexidade, a escalabilidade, e a efetividade prática, uma vez que redes sociais, redes de comunicação e redes biológicas são naturalmente representadas por grafos.

Em (SANTOS, 2019) foi utilizado mineração de texto e aprendizado de máquina para extrair informações relevantes de Notas Fiscais Eletrônicas do Consumidor (NFC-e), processadas pela Secretaria do Estado de Tributação (SET), no estado do Rio Grande do Norte (RN). O autor cita como motivação para o trabalho: agilizar o processo de

tributação, classificar produtos em categoria/subcategoria para a partir disso, tributar os contribuintes de maneira mais correta e justa, identificar possíveis casos de sonegação de impostos, estabelecer um valor médio dos produtos vendidos no estado e assim apontar indícios de superfaturamento e fraudes em licitações.

Foi modelada uma plataforma distribuída para processamento em Big Data de Notas Fiscais Eletrônicas do Consumidor (NFC-e). O problema aqui era que SET/RN processa diariamente grande volume de dados. Em geral, são processados mais de 1000 registros de vendas por minuto. Cada registro é armazenado em formato XML, e retrata um tipo de operação de venda. Essa massa gigantesca de dados representa toda a vida fiscal dos contribuintes do RN.

Além do grande volume de dados, um dos desafios encontrados, foi a extração de informação útil a partir do campo textual da descrição do produto comercializado, por se tratar de um campo livre sem padrão de escrita.

O autor cita o atributo descrição de produtos como sendo um possível causador de inconsistência. Tendo em vista que o mesmo pode ser preenchido de acordo com o entendimento do próprio lojista, ou seja, pode descrever os mesmos produtos de forma bem diferente. Por essa razão, este atributo é sempre desconsiderado nas tarefas de extração de texto. Logo, informações sobre marca, embalagem e volume são sempre descartadas, e dessa forma, nunca aproveitadas.

Em sua análise de dados foi possível observar que há uma variedade grande de categorias de produtos, tais como: alimentos, bebidas, combustíveis, farmacêuticos, papelaria, etc. Observou-se, ainda, que cada uma delas pode ser dividida em outras subcategorias, como é o caso da categoria de bebidas, que pode ser subdividida em: refrigerantes, destilados, cervejas, sucos e água mineral.

Para a categoria de produtos bebidas, por exemplo, verificou-se que novos atributos poderiam agregar relevante valor à base de NFC-e. Assim, para conseguir aproveitar o conteúdo do campo descrição do produto, a tentativa do autor foi extrair parte da informação contida nesse campo e representá-la na forma de novos atributos, quais sejam: marca, tipo, volume e embalagem. Assim, após a mineração de texto, foi possível representar a descrição do produto da seguinte forma:

Foi especificado um conjunto de palavras que serviu de base para as comparações nessa fase de extração, no formato de *Bag of words*. Cada atributo possui assim, uma *bag of word* específica, por exemplo, a *bag of words* de possíveis marcas de bebidas foi formada pelos termos ANTARTICA","BRAHMA", "BAVARIA", "PRINCESS", "BUDWEISER", "CRYSTAL", "STER BOM", etc. Cada palavra encontrada na mineração é adicionada ao seu respectivo atributo, baseando-se na respectiva *bag of word*

Para isso, utilizou-se a biblioteca NLTK em cada interação que analisa cada palavra

Tabela 1: Exemplo de mineração de texto no campo descrição de produto

| Descrição Original | Marca | Tipo | Volume | Embalagem |
|---|---|---|---|---|
| Coca Cola LT 350ML | Coca | Cola | 350 | Lata |
| Agua Cryst S GAS 500ML | Crystal | SEM GAS | 500 | Garrafa |
| Refri Fanta Uva PET 2L | Fanta | Uva | 2000 | PET |
| Cerv Heineken Lager Beer LN 330ML | Heineken | Lager | 330 | LongNeck |
| Refri Antarctica Guarana PET 1L | Antarctica | Guarana | 1000 | PET |

Fonte: Adaptado de (SANTOS, 2019)

separada do campo descrição do produto. Nessa biblioteca foi usada a função *distance* implementando a distância de levenshtein passando as palavras a serem comparadas.

Após essa fase de mineração de texto, já com os novos atributos definidos partiu-se para a classificação dos dados. Foi Selecionado um grupo de produtos (bebidas) para predizer em que subgrupo ele se enquadrava (refrigerante, sucos, água, cerveja, destilado, diverso). Foram utilizados 3.000 registros balanceados em 500 instâncias por cada classe.

Como vimos, a maioria dos estudos utiliza Redes Neurais para a classificação textual, além de Deep Learning se utiliza também algoritmos de Machine Learning como SVM e Naive Bayes.

Foram encontrados diversos estudos sobre a extração de atributos a partir de dados textuais, ou seja, existe a preocupação em representar os dados textuais em um grupo de atributos que possam ser utilizados por um modelo de Machine Learning ou Deep Learning. Nesse contexto, o algoritmo IMBHN aparece como uma alternativa de representação para dados textuais.

## 2.1 Classificação de Texto e o Modelo Espaço Vetorial

O modelo espaço vetorial é o mais utilizado na classificação de texto. Neste modelo, cada documento é representado por um vetor, sendo que cada dimensão do vetor equivale a um atributo/característica do conjunto de dados.

Um atributo adicional deve ser incluído ao vetor para representar a classe do documento. Os documentos que possuem valor nesse atributo são denominados documentos rotulados, enquanto que, os documentos que não possuem, são denominados documentos não rotulados.

Como exemplo, na Tabela 2 temos três documentos, classificados entre as classes 1 e 2. Temos pesos que vinculam cada termo aos documentos. Uma forma simples de representar o peso de cada termo a um documento ao qual está associado é o número de ocorrências em que o termo aparece no documento.

Tabela 2: Matriz documento-termo

| | $t_1$ | $t_2$ | ... | $t_n$ | $Classe$ |
|---|---|---|---|---|---|
| $d_1$ | $P_{d_1,t_1}$ | $P_{d_1,t_2}$ | ... | $P_{d_1,t_n}$ | $c_1$ |
| $d_2$ | $P_{d_2,t_1}$ | $P_{d_2,t_2}$ | ... | $P_{d_2,t_n}$ | $c_2$ |
| $d_3$ | $P_{d_3,t_1}$ | $P_{d_3,t_2}$ | ... | $P_{d_3,t_n}$ | $c_1$ |

Fonte: adaptado de (ROSSI, 2015)

## 2.2 Redes Homogêneas e Redes Heterogêneas para Representação Textual

As duas formas mais simples de utilizar redes para representação textual são redes de documentos e redes de termos. Ambas originadas do modelo Bag of Words. No caso de rede de documentos, cada documento representa um vértice da rede, e cada aresta é obtida pela similaridade entre dois documentos relacionados (ROSSI, 2015). No caso de rede de termos, os vértices da rede são formados por termos que são conectados levando em conta sua similaridade, ordem de ocorrência no conjunto de dados ou relação semântica/sintática. (ROSSI; LOPES, 2016).

Em ambos os casos, temos que os vértices das redes são formados por um único tipo de objeto, documento ou termo, portanto estas redes são denominadas redes homogêneas. Já no caso em que se utiliza, tanto documentos quanto termos, em uma única rede para representar o conjunto de dados, estas redes são definidas como redes heterogêneas.

Por se tratar de dois tipos de vértices, com conexões somente entre objetos de tipos distintos, no caso, somente entre documento-termo, essa rede também é definida como rede bipartida (NEWMAN, 2003). Dessa forma, uma rede bipartida, também pode ser representada pelo modelo Bag of Words, uma vez que o peso das conexões entre termos e documentos pode ser quantificado pela frequência dos termos nos documentos aos quais estão relacionados.

## 2.3 Classificação Baseada em Redes

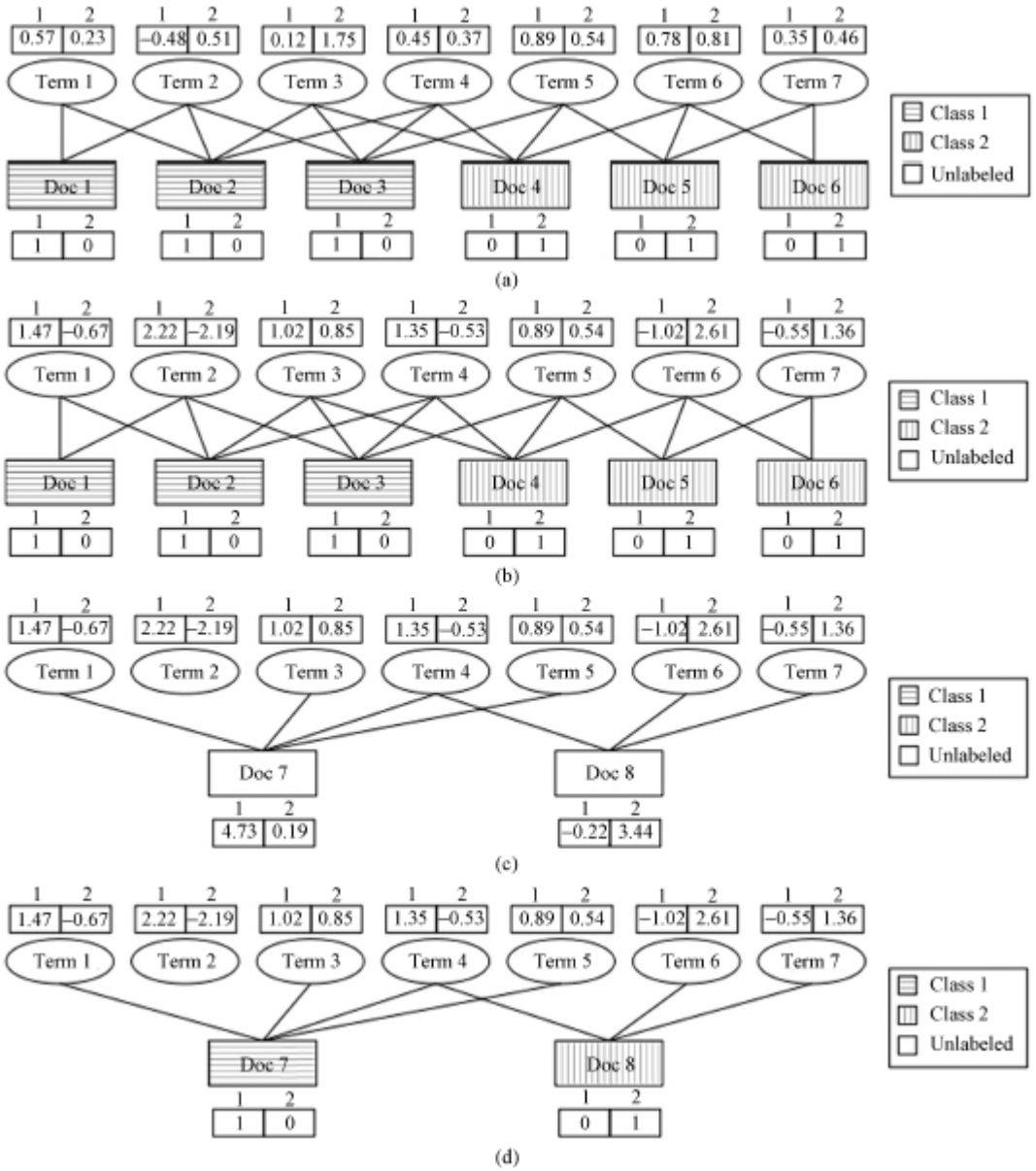
Em (Rossi et al., 2012) é proposto um algoritmo indutivo supervisionado baseado em rede heterogêneas bipartidas para classificação de texto (IMBHN - Inductive Model Based on Bipartite Heterogeneous Network). Por ser um algoritmo de aprendizado supervisionado é necessário que um grupo de documentos esteja pré-classificado para que então novos documentos possam ser classificados a partir de um modelo de classificação. O algoritmo visa induzir o peso/contribuição dos termos para cada uma das classes predefinidas e utilizar os pesos induzidos como modelo de classificação.

A Figura 1. apresenta um passo a passo do funcionamento do algoritmo IMBHN, conforme apresentado em (Rossi et al., 2012). Neste caso temos um problema de classificação binária em uma rede bipartida formada por documentos e termos. Considere que o Termo

$t_1$ apresenta valor de 1,47 para a Classe 1 e apenas -0,67 para a Classe 2, e que o Termo $t_3$ apresenta valor de 1,02 para a Classe 1 e apenas 0,85 para a Classe 2, dessa forma um novo documento $Doc_7$ que contém os dois termos $t_1$ e $t_3$ tem maior probabilidade de ser classificado na Classe 1 do que na Classe 2.

Assim, é possível propagar a informação de classes a partir dos termos de documentos já existentes e classificados. Como os termos estarão conectados tanto aos documentos classificados como aos documentos não classificados, é possível propagar os rótulos entre documentos utilizando os termos que são comuns entre eles.

Figura 1: Funcionamento do algoritmo IMBHN



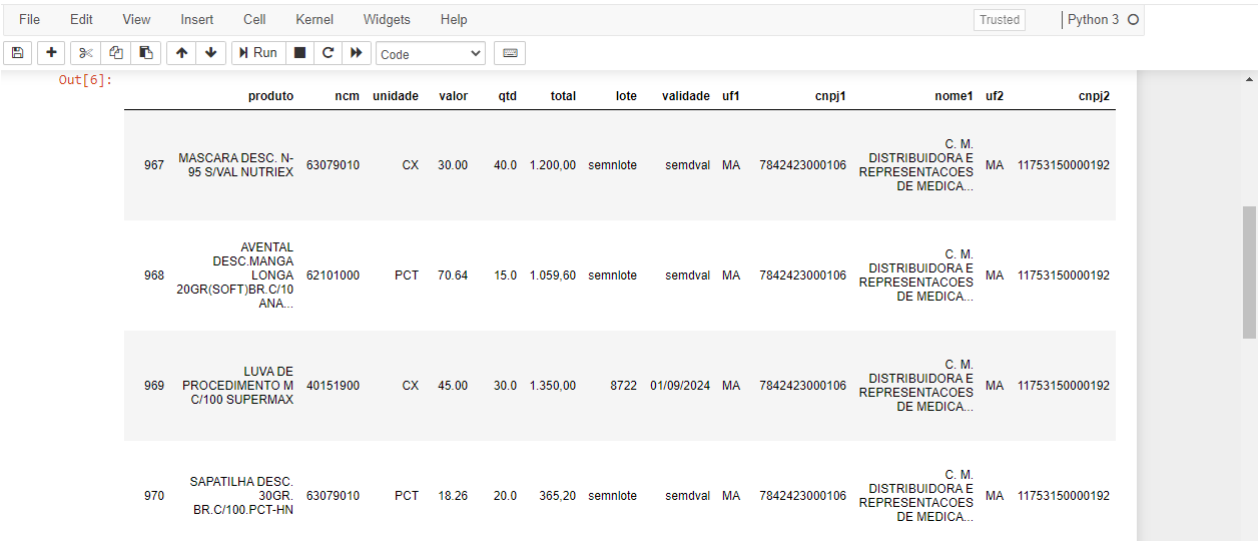

Fonte: Extraído de (Rossi et al., 2012)

# 3 PROPOSTA: USO DE REDES HETEROGÊNEAS PARA CLASSIFICAÇÃO DE PRODUTOS EM NOTAS FISCAIS ELETRÔNICAS

Para identificar produtos semelhantes, foram predefinidas classes de produtos, que serviram de rótulos para os produtos que compõem as Notas Fiscais Eletrônicas (NFe). Esses rótulos serão utilizados pela rede heterogênea para a tarefa de classificação dos produtos.

A base de dados original é formada por notas fiscais emitidas para 198 Fundos Municipais de Saúde no estado do Maranhão no mês de maio de 2020. É formada por 68.276 produtos distribuídos em 6.382 NFe. São documentos fiscais de acesso público, disponíveis nos processos de pagamento de cada Secretaria Municipal de Saúde e disponibilizados pela SEFAZ/MA à CGU/MA, conforme Acordo de Cooperação Técnica nº 09, de 24 de agosto de 2004.

Figura 2: Base de Dados Produtos




| | produto | ncm | unidade | valor | qtd | total | lote | validade | uf1 | cnpj1 | nome1 | uf2 | cnpj2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 967 | MASCARA DESC. N-95 S/VAL NUTRIEX | 63079010 | CX | 30.00 | 40.0 | 1.200,00 | semnlote | semdval | MA | 7842423000106 | C. M. DISTRIBUIDORA E REPRESENTACOES DE MEDICA... | MA | 11753150000192 |
| 968 | AVENTAL DESC.MANGA LONGA 20GR(SOFT)BR.C/10 ANA... | 62101000 | PCT | 70.64 | 15.0 | 1.059,60 | semnlote | semdval | MA | 7842423000106 | C. M. DISTRIBUIDORA E REPRESENTACOES DE MEDICA... | MA | 11753150000192 |
| 969 | LUVA DE PROCEDIMENTO M C/100 SUPERMAX | 40151900 | CX | 45.00 | 30.0 | 1.350,00 | 8722 | 01/09/2024 | MA | 7842423000106 | C. M. DISTRIBUIDORA E REPRESENTACOES DE MEDICA... | MA | 11753150000192 |
| 970 | SAPATILHA DESC. 30GR. BR.C/100.PCT-HN | 63079010 | PCT | 18.26 | 20.0 | 365,20 | semnlote | semdval | MA | 7842423000106 | C. M. DISTRIBUIDORA E REPRESENTACOES DE MEDICA... | MA | 11753150000192 |

Fonte: autor

Já que a quantidade de produtos é muito grande, foi necessário reduzir a lista de produtos consultados. Adicionalmente, por se tratar de período de Pandemia do Covid-19, escolhemos produtos que tiveram uma grande demanda, e que por essa razão, pudessem ter sofrido com o aumento abusivo de preços. Por fim, foram escolhidos produtos com observações suficientes para gerar um modelo de classificação.

A lista de produtos foi formada pelos itens máscara, álcool, luva, soro e dipirona. Verificou-se que para cada produto existem agrupamentos com características específicas, que poderiam ser utilizadas para classificação.

O produto máscara possui duas classes principais, a classe máscara cirúrgica descartável e a classe máscara N95. Uma terceira classe foi utilizada para agrupar as máscaras que não se encaixam nos dois grupos anteriores como por exemplo máscaras de tecido e máscaras para nebulização. A classe máscara cirúrgica descartável refere-se a máscara sem elemento filtrante, de uso individual que cobre nariz e boca indicada para proteger de infecções por inalação de gotículas transmitidas à curta distância e projeção de sangue ou outros fluidos corpóreos que possam atingir as vias respiratórias e minimizar a contaminação do ambiente com secreções respiratórias geradas pelo próprio indivíduo. A classe máscara N95 refere-se a equipamento de proteção respiratória purificador de ar que possui eficiência de filtração de 95%, testada com aerossol de NaCl. É equivalente à Peça Facial Filtrante (PFF2) ou ao Equipamento de Proteção Respiratória do tipo peça semifacial com filtro P2 (ANVISA, 2009).

Esse agrupamento foi realizado para os demais produtos escolhidos. O produto Álcool possui duas classes principais, a classe álcool gel e a classe álcool líquido. O produto luva foi agrupado nas classes luva cirúrgica e luva de procedimento não cirúrgico. O soro foi agrupado nas classes soro fisiológico e soro glicosado. Para cada um desses produtos foi utilizada uma terceira classe com itens que não se encaixam nas duas classes principais. O produto dipirona foi agrupado nas classes dipirona frasco, dipirona ampola e dipirona comprimido.

Dessa forma, para cada produto há um problema de classificação ternária. A escolha de um número reduzido de classes visa aumentar a eficiência do algoritmo IMBHN, que assim como outros algoritmos de aprendizado de máquina, tem melhores resultados com um número pequeno de classes.

Além do agrupamento em classes, foi preciso ajustar a unidade de medida para cada produto, uma vez que um mesmo produto pode ser vendido em caixas ou em unidade, ou ainda em volume distinto. A tabela abaixo resume as informações sobre as classes utilizadas.

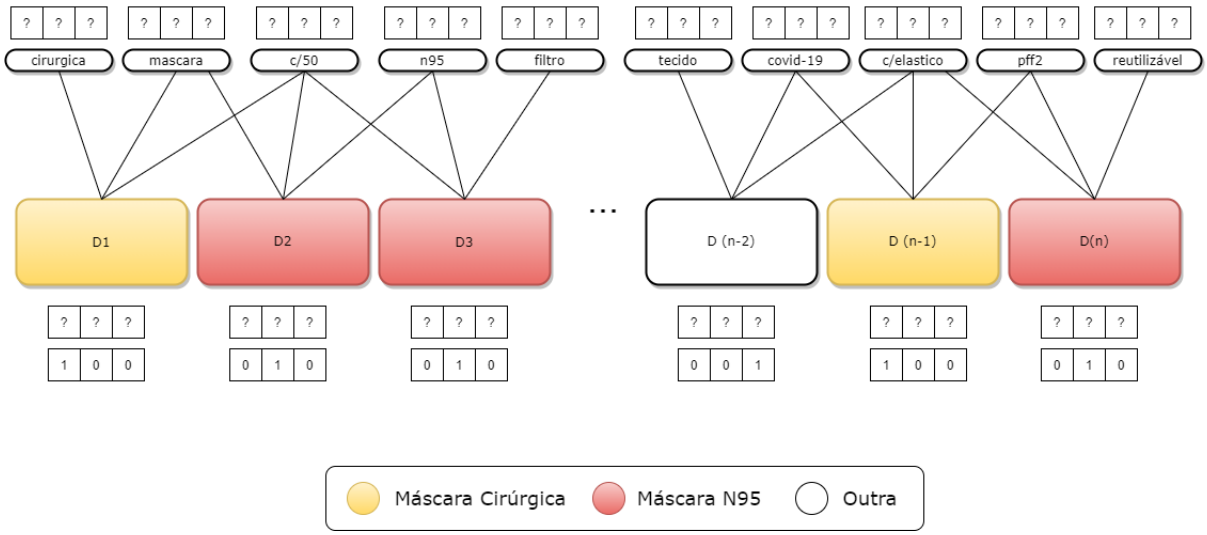
Figura 3: Produtos e sua classificação

| # | produto | embalagem | valor-ajustado | ncm | unidade | classe |
|---|---|---|---|---|---|---|
| 1 | MASCARA CIRURGICA CX/ COM 50 | 50 | 3,21 | 63079010 | CX | mascara-cirurgica |
| 2 | MASCARA N95 | 1 | 34,13 | 90200010 | UND | mascara-n95 |
| 3 | ALCOOL ETILICO HIDRATADO 1000ML | 12 | 9,370833333 | 0 | CX | alcool-liquido |
| 4 | ALCOOL GEL 500ML | 6 | 16,65 | 0 | CX | alcool-gel |
| 5 | LUVA DE PROCEDIMENTO M C/100 SUPERMAX | 100 | 0,45 | 40151900 | CX | luva-procedimento |
| 6 | LUVA CIRURGICA ESTERIL N 7,5 | 1 | 3,5 | 40151100 | UND | luva-cirurgica |
| 7 | SORO FISIOLOGICO 0,9% 500ML | 1 | 2,2 | 30049099 | BOL | soro-fisiologico |
| 8 | SORO GLICOSADO 5% 500ML SF C/24 | 1 | 7,42 | 30049099 | FR | soro-glicosado |
| 9 | DIPIRONA SODICA 500MG COMP. | 1 | 0,23 | 30049069 | CPR | dipirona-comprimido |
| 10 | DIPIRONA INJ.DE 1G.CX-C/100 AMP.DE 2ML.SANTIDOR | 1 | 2,1 | 30039099 | AMP | dipirona-injetavel |

Fonte: autor

Após a escolha dos rótulos para cada produto, passamos para a representação em rede bipartida, com o intuito de agrupar produtos nas classes predefinidas. A imagem a seguir apresenta a representação da rede bipartida para termos e produtos extraídos de Notas Fiscais Eletrônicas para o produto Máscara.

Figura 4: Rede Bipartida de Termos/Descrição de produto e Classes para o item máscara composto pelas Classes Máscara Cirúrgica e Máscara N95




Fonte: Autor, baseado em (ROSSI, 2015).

Como apresentado na figura, em um primeiro momento, não conhecemos o peso que cada termo possui relacionado a cada classe (cirúrgica, n95, outra). Após a construção do modelo, os pesos serão conhecidos e, a partir deles, cada produto será associado a uma classe específica.

Foram realizadas duas etapas de pré-processamento sobre os termos de produtos, ambas com a utilização da biblioteca nltk da linguagem python. Na primeira etapa, foi realizado o stemming dos termos, com a utilização do pacote stem e em seguida foi realizada a retirada de stop words, com a utilização do pacote corpus.

Foi utilizado o algoritmo IMBHN para a classificação dos produtos. O Algoritmo IMBHN é um algoritmo de aprendizado supervisionado, portanto, a base de dados foi dividida entre observações de treinamento e observações de teste. Os produtos são classificados previamente utilizando os rótulos predefinidos, as observações de treinamento são utilizadas para treinar o modelo. É realizada a classificação dos dados de testes utilizando o modelo gerado em treinamento e em seguida é verificada a acurácia do modelo, ou seja, é verificado qual o percentual de acerto é alcançado pelo modelo gerado. O Algoritmo IMBHN está disponível no github [1] e foi ajustado para utilizar a base de dados de produtos

[1] https://github.com/edilsonacjr/IMBHN

como uma Bag-of-Words, onde cada linha representa um produto e cada coluna representa um termo associado ao produto.

# 4 RESULTADO

Foram utilizados cinco banco de dados distintos para a representação dos produtos de Nota Fiscal Eletrônica NFe. Verificou-se que com o aumento do número de classes o algoritmo perde em acurácia. Assim, cinco modelos foram gerados para a classificação, cada modelo utiliza três classes por vez. Foram selecionadas as primeiras 250 ocorrências para cada produto, o que resultou em bases de dados desbalanceadas. Cada base possui três classes distintas para cada produto.

Foi utilizado o método K-fold Cross Validation para partição dos dados entre treinamento e teste. Neste método os dados são particionados em K partes, cada uma contendo 1/K observações: K-1 partes são usadas, em cada rodada, para treinamento e a partição restante é utilizada para validação (ANGUITA et al., 2009). Foi predefinido o valor de K=10. Utilizou-se a função StratifiedKFold do pacote SkLearn da linguagem python. Essa função garante que a proporção entre o número de observações pertencentes a cada classe seja mantida para cada partição k em cada rodada. A tabela abaixo apresenta os testes realizados e os resultados encontrados.

Tabela 3: Resultados da Classificação com IMBHN

| Produto | Teste | Acurácia | Classe 1 | Classe 2 | Classe 3 | Termos |
|---|---|---|---|---|---|---|
| Álcool | 10 Fold CV | 0.704 | 140 | 82 | 28 | 234 |
| Dipirona | 10 Fold CV | 0.644 | 114 | 80 | 56 | 180 |
| Luva | 10 Fold CV | 0.764 | 114 | 82 | 54 | 224 |
| Máscara | 10 Fold CV | 0.82 | 136 | 69 | 45 | 259 |
| Soro | 10 Fold CV | 0.784 | 86 | 81 | 83 | 215 |

Fonte: Autor

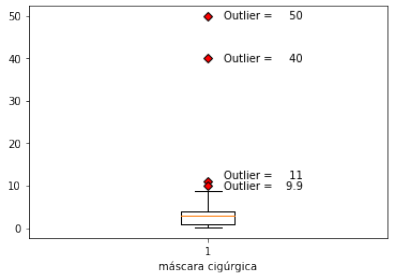
Verificou-se que o algoritmo IMBHN se comporta melhor com o balanceamento do número de observações por cada classe, assim faz-se necessário um ajuste na seleção das observações para iniciar a execução do algoritmo com o mesmo número de observações para cada classe.

Outra questão que foi verificada diz respeito à quantidade de atributos utilizados na classificação. Como pode-se observar na tabela de resultados, o pior desempenho ocorreu na base de dados do produto "Dipirona", que além de possuir dados desbalanceados também possui o menor número de atributos gerados, apenas 180, dessa forma também é necessário um ajuste para extrair termos suficientes para o uso do algoritmo.

Após classificar os produtos em grupos de semelhantes, foi possível verificar Outliers, produtos adquiridos por um valor acima da média entre seus similares.

A Figura abaixo apresenta o Boxplot para os valores de venda do produto Máscara pertencente à classe máscara cirúrgica. A média observada foi de R$ 3,42, com valor para o primeiro quartil de R$ 0,79 e para o terceiro quartil de R$ 4,00. Como verificado, existem pontos com maior distanciamento da média observada.

Figura 5: Outliers para o produto Máscara Cirúrgica




Fonte: Autor

Entre as observações merece destaque o valor de R$ 9,90 referente ao produto com descrição "MASCARA TRIPLA C/50", contido na Nota Fiscal nº 106 de 11/05/2020. Após Auditoria realizada verificou-se de fato o superfaturamento na venda, o que culminou em ação de controle contra os responsáveis pela aquisição do produto [1].

A Auditoria confirmou o superfaturamento apontado pelo estudo, com a vantagem deste último utilizar detecção automática de Outliers, processo que pode ser realizado a partir da data de emissão das Notas Fiscais contidas na base de dados.

[1] https://www.gov.br/cgu/pt-br/assuntos/noticias/2020/06/cgu-e-pf-combatem-fraudes-em-compras-de-mascaras-cirurgicas-em-sao-luis-ma

# 5 CONCLUSÃO

Verificou-se a existência de diversas técnicas para representação de texto por meio de seus atributos. Entre os modelos de representação os dois principais são o modelo espaço-vetorial e a representação em redes.

O modelo espaço-vetorial divide-se entre modelos baseados em frequência (por exemplo: Bag of Words, TF-IDF) e em representação latente (por exemplo: Word2Vec).

Entre os modelos baseados em redes podemos citar o IMBHN, os Grafos Hierárquicos (Gong et al., 2020), e diversos algoritmos semi-supervisionados baseados em grafos (CHONG et al., 2020).

Entre as limitações práticas encontradas na classificação de texto pode-se citar a grande quantidade de informação não rotulada (Li; Karunesh; Alaniazar, 2019). Diversos estudos já avançam no sentido de aproveitar informação não rotulada com o objetivo de classificação de texto.

O algoritmo IMBHN possui relevância tanto por ser uma alternativa viável para representação de dados textuais bem como pela possibilidade de ser utilizado em aprendizado semi-supervisionado.

A utilização do algoritmo IMBHN para a classificação de produtos de Notas Fiscais Eletrônicas (NF-e) retornou bons resultados que podem ser melhorados com o balanceamento do número de observações entre classes e também com a extração de um número suficiente de atributos.

Foi possível a partir do uso do algoritmo IMBHN sobre a base de dados de Notas Fiscais Eletrônicas (NF-e) classificar produtos e comparar o valor de aquisição entre similares, o que permitiu a detecção de Outliers na aquisição de produtos. O estudo apontou superfaturamento que foi confirmado em Auditoria realizada sobre compras públicas.

# REFERÊNCIAS


Al-Doulat, A.; Obaidat, I.; Lee, M. Unstructured medical text classification using linguistic analysis: A supervised deep learning approach. In: **2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)**. [S.l.: s.n.], 2019. p. 1–7.

ANGUITA, D. et al. **K–Fold Cross Validation for Error Rate Estimate in Support Vector Machines**. [S.l.: s.n.], 2009.

ANVISA. **BRASIL. Cartilha de proteção respiratória contra agentes biológicos para trabalhadores**. [S.l.: s.n.], 2009.

CHONG, Y. et al. Graph-based semi-supervised learning: A review. **Neurocomputing**, v. 408, p. 216 – 230, 2020. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231220304938>.

FALEIROS, T. d. P. **Optimizing the class information divergence for Transductive classification of texts using propagation in bipartite graphs**. [S.l.: s.n.], 2016.

Gong, J. et al. Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification. **IEEE Access**, v. 8, p. 30885–30896, 2020.

LEAL, J. P.; MAHLER, P. R. **Atenção Com O Preço Justo De Mercado – Economizando Nas Compras Públicas Com O Uso Da Nota Fiscal Eletrônica**. [S.l.: s.n.], 2016.

Li, X.; Karunesh, A.; Alaniazar, S. Mixed-model text classification framework considering the practical constraints. In: **2019 Second International Conference on Artificial Intelligence for Industries (AI4I)**. [S.l.: s.n.], 2019. p. 67–70.

MIKOLOV, T. et al. **Efficient Estimation of Word Representations in Vector Space**. 2013.

NEWMAN, M. E. J. **The structure and function of complex networks**. [S.l.: s.n.], 2003.

NFE. **Portal da Nota Fiscal Eletrônica**. [s.n.], 2020. Disponível em: <https://www.nfe.fazenda.gov.br/portal/infoEstatisticas.aspx>. Acesso em: 01 jul. 2020.

ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. 2015. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2015.

Rossi, R. G. et al. Inductive model generation for text categorization using a bipartite heterogeneous network. In: **2012 IEEE 12th International Conference on Data Mining**. [S.l.: s.n.], 2012. p. 1086–1091.

ROSSI, S. O. R. R. G.; LOPES, A. de A. **Term network approach for transductive classification**. [S.l.: s.n.], 2016.

SANTOS, D. S. dos. **Uma Plataforma Distribuída de Mineração de Dados para Big Data: um Estudo de Caso Aplicado à Secretaria de Tributação do Rio Grande do Norte**. 2019. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte - UFRN, Natal-RN, 2019.

SEBASTIANI, F. **Machine Learning in Automated Text Categorization**. [S.l.: s.n.], 2002.

SILVA, L. A. D. **Uso de técnicas de inteligência artificial para subsidiar ações de controle. Revista do TCU**. [s.n.], 2016. Disponível em: <https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/1385>. Acesso em: 04 abr. 2020.

TCU. **Análise de documentos fiscais relacionados a fraudes na administração pública**. [S.l.: s.n.], 2009.

ZHANG, X.; LECUN, Y. **Text Understanding from Scratch**. 2016.