REVISTA EDICIC

CLASSIFICAÇÃO DE TEXTOS: UMA ABORDAGEM COM USO DE MACHINE LEARNING

Fábio Eder Cardoso, Universidade Estadual Paulista (Unesp), Brasil, https://orcid.org/0000-0002-0309-057X

Edberto Ferneda, Universidade Estadual Paulista (Unesp), Brasil, https://orcid.org/0000-0002-8808-1217

Leonardo Castro Botega, Universidade Estadual Paulista (Unesp), Brasil, https://orcid.org/0000-0003-1495-5935

RESUMO

A classificação de textos tem sido utilizada como base para a organização do conhecimento nas mais variadas áreas, uma vez que proporciona organizar grupos de categorias para nortear recortes desses domínios. Na era da informação digital, na qual existe uma vasta quantidade de dados disseminados em ambientes de computação em nuvem, é necessário o uso de tecnologias informacionais, para auxiliar o processo de classificação desses dados. Neste contexto, a Ciência da Informação contribui no processo de produção, organização, transmissão e uso da informação, nas mais variadas áreas, dentre elas, a ciência da computação, matemática, inteligência artificial, dentre outras. Por meio da tecnologia, quando a informação é adequadamente classificada, ela pode ser disponibilizada de maneira mais eficaz para a sociedade. O objetivo geral deste artigo é abordar contextos sobre classificação de textos com uso de Machine Learning. Esta pesquisa é do tipo exploratória, de método experimental, utilizou-se a abordagem quantitativa como técnica de análise de dados. Como resultado, após utilizar o algoritmo de distância Euclidiana, estabeleceu-se uma matriz de distâncias e um agrupamento hierárquico, além de uma nuvem de palavras, retornando expressões com termos relevantes dos documentos.

Palavras-Chave: Classificação; Machine Learning; Algoritmos; Informação; Ciência da Informação.

CLASIFICACIÓN DE TEXTOS: UN ENFOQUE CON USO DE MACHINE LEARNING

RESUMEN

La clasificación de textos ha sido utilizada como base para la organización del conocimiento en las más diversas áreas, ya que permite organizar grupos de categorías para guiar el corte de estos dominios. En la era de la información digital, donde existe una gran cantidad de datos diseminados en entornos de computación en la nube, es necesario el uso de tecnologías informacionales para ayudar en el proceso de clasificación de estos datos. En este contexto, la Ciencia de la Información contribuye en el proceso de producción, organización, transmisión y uso de la información en las más variadas áreas, entre ellas, la ciencia de la computación, matemáticas, inteligencia artificial, entre otras. A través de la tecnología, cuando la información está adecuadamente clasificada, puede ser puesta a disposición de la sociedad de manera más eficaz. El objetivo principal de este artículo es abordar contextos sobre la clasificación de textos con el uso de Machine Learning. Esta investigación es de tipo exploratoria, con un método experimental, y utiliza un enfoque cuantitativo como técnica de análisis de datos. Como resultado, después de utilizar el algoritmo de distancia euclidiana, se estableció una matriz de distancias y un agrupamiento jerárquico, además de una nube de palabras, resaltando expresiones con términos relevantes de los documentos.

REVISTA EDICIC

Palabras-Clave: Clasificación; Machine Learning; Algoritmos; Información; Ciencias de la Información.

TEXT CLASSIFICATION: AN APPROACH USING MACHINE LEARNING

ABSTRACT

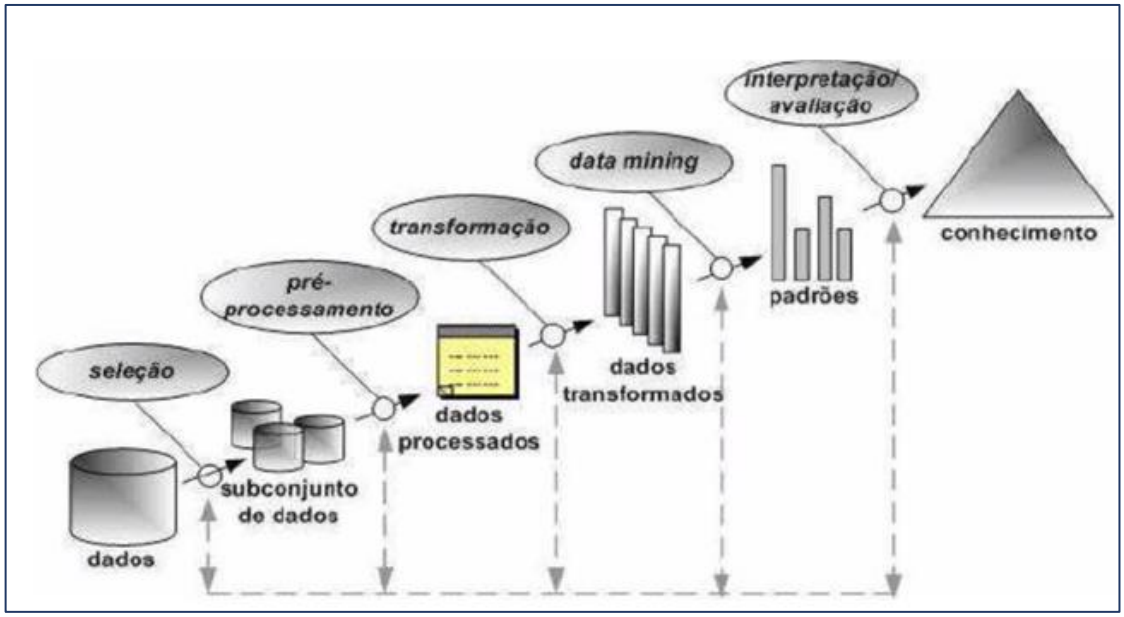
Text classification has been employed as a foundation for organizing knowledge across a wide range of fields, as it allows for the grouping of categories to guide the segmentation of these domains. In the digital information age, where there is an abundance of data spread across cloud computing environments, the use of informational technologies is essential to facilitate the classification process of this data. Within this framework, Information Science plays a pivotal role in the production, organization, transmission, and utilization of information across diverse fields, including computer science, mathematics, artificial intelligence, among others. Through technology, when information is appropriately classified, it can be made available to society more effectively. The primary aim of this article is to address contexts regarding text classification using Machine Learning. This research is exploratory, adopting an experimental method, and employs a quantitative approach as its data analysis technique. As a result, after utilizing the Euclidean distance algorithm, a distance matrix and hierarchical grouping were established, along with a word cloud, highlighting terms of significance from the documents.

Keywords: Classification; Machine Learning; Algorithms; Information; Information Science.

1 INTRODUÇÃO

A produção em massa de dados textuais em uma variedade de formatos, como mensagens eletrônicas, livros digitais, artigos e postagens nas redes sociais, é uma realidade inevitável. No entanto, processar, organizar e gerenciar manualmente esses dados textuais é uma tarefa laboriosa e frequentemente imprevisível, particularmente quando o conhecimento está embutido nos dados. Devido a isso, as instituições acadêmicas, empresas e organizações estão prestando cada vez mais atenção e usando tecnologias computadorizadas que podem organizar, gerenciar e extrair conhecimento de grandes quantidades de texto com o mínimo de intervenção manual. A busca por conhecimento nesse contexto se baseia na exploração de dados em corpora de documentos que armazenam informações, muitas vezes implícitas e desconhecidas, e as transformam em conhecimento útil e acessível que suporta a tomada de decisões importantes (Galvo e Marin, 2009). Esse processo de pesquisa envolve um número de princípios, incluindo métodos estatísticos e técnicas de inteligência artificial e mineração de textos. A Figura 1 ilustra as etapas envolvidas na descoberta de conhecimento, sendo elas: seleção, pré-processamento, transformação, mineração de dados e interpretação (Fayyad et al., 1996). Essas etapas são ainda mais importantes quando se discute classificação de textos no contexto de aprendizagem de máquina e Ciência da Informação. O processo de seleção envolve escolher textos relevantes para análise; a fase de pré-processamento envolve limpar e formatar dados textuais; a fase de transformação envolve converter textos em formatos que podem ser efetivamente manipulados por algoritmos de aprendizagem de máquina; a fase de mineração de dados envolve aplicar algoritmos de aprendizagem de máquina nos dados para identificar padrões e extrair conhecimento; e, finalmente, a fase de interpretação envolve analisar os dados.

Figura 1: Etapas do processo de conhecimento



Fonte: Fayyad *et al*, 1996.

Os autores He et al. (2013, p.) afirmam que a mineração de textos é uma técnica automatizada que pode "identificar, extrair, gerenciar, integrar e explorar os conhecimentos contidos em textos de forma eficaz e sistemática". Seu método envolve os seguintes passos: definição de contexto e mineração de conceitos, coleta de dados, construção de dicionário e análise de dados. Entre as várias tecnologias envolvidas, destaca-se a classificação de texto, que atribui automaticamente rótulos (identificadores de categorias predefinidas) a documentos ou segmentos de texto. Um método eficaz para classificar textos automaticamente envolve algoritmos de aprendizado de máquina que são capazes de "aprender", sintetizar ou até mesmo identificar padrões em categorias com base no conteúdo e títulos de documentos de texto.

A Classificação de Texto deu seus primeiros passos na década de 1960, mas foi somente na década de 1990 que começou a ser reconhecida como uma subárea significativa no campo dos sistemas de informação. Isso se deu devido ao crescente interesse em aplicações e ao desenvolvimento de hardware mais durável. Hoje, a classificação de texto é usada em uma ampla gama de contextos, desde indexação de documentos com base em vocabulário controlado até filtragem de documentos, a criação automática de metadados e a remoção de ambiguidades nos significados de palavras (Sebastiani, 2002).

O método mais direto para calcular a proximidade entre os termos foi usado neste trabalho, o método de distância euclidiana. No entanto, um desafio que deve ser superado é como a classificação se comportaria se outros métodos de cálculo de distância, como Manhattan, Cosseno, Pearson e Hamming, fossem usados.

Apesar de ser baseado em uma metodologia simples, este estudo tem um nível significativo de relevância. Isso indica que há necessidade de investigar a eficácia da classificação usando métodos alternativos. O uso de tecnologias como o aprendizado de máquina é uma tendência emergente que avança a ciência da informação e do conhecimento, ao mesmo tempo em que também aumenta a capacidade do sistema em processos de classificação.

2 TEORIA DA CLASSIFICAÇÃO

Ao discutir cenários relacionados à teoria da classificação, é fundamental contemplar a categoria ou categorização como instrumento principal para a análise da existência e da variabilidade dos elementos, conforme defende Barite (2000, pp. 2).

> The notion of category, from Aristotle and Kant to the present time, has been used as a basic intellectual tool for the analysis of the existence and changeableness of things.

Barite (2000, pp. 6) assevera que sobre a impossibilidade em caracterizar as categorias na teoria da classificação, sendo necessário absorver definições sobre as mais variadas áreas como, filosofia, ontologia e metafísica:

> We should first make it clear that it is not possible to characterize categories in the Theory of Classification, taking on loan the definitions provided by Philosophy, Ontology or Metaphysics.

As categorias são expressões abstratas extremamente genéricas, de modo que poderiam ser percebidas em qualquer entidade, elemento ou objeto. De acordo com o pensamento Aristotélico, há uma característica implícita que se relaciona ao caráter instrumental das categorias. Estas, são utilizadas como ferramentas para a descoberta de algumas regularidades no mundo físico, material, visto que os objetos, no mundo material, possuem certas propriedades, sendo elas uma categoria possível para analisar o mundo real. (Barite, 2000, pp. 4)

Categorias podem ser interpretadas como abstrações simplificadas que, através de classificadores, investigam os padrões dos objetos reais, tornando-os ideais para representar noções e organizar sistemas de conceitos. No contexto da teoria da classificação as categorias são relevantes enquanto instrumento de análise e organização dos objetos, fenômenos e conhecimento, incluindo como interesse, dimensões de análise aplicadas à estrutura interna do conhecimento humano e suas abstrações mais representativas, ou seja, conceitos. (Barite, 2000, pp. 5)

Barite (2000, pp. 6) define que, para os classificadores, o uso de categorias deve ser segmentado em três atividades precisas, sendo elas:

i. desenho, planejamento e estruturação de linguagens de indexação ou sistemas de conhecimento (sistemas de classificação, tesauros, taxonomias);
ii. modificação ou especificação das tabelas de classificação;
iii. a avaliação e análise de linguagens e sistemas de indexação de conceitos através de um conjunto de parâmetros capaz de estabelecer o grau de tensão recíproca entre conceitos relacionados, sua relevância e validade.

Com o advento da Web Semântica, criada por Tim Berners Lee, e das Ontologias, a ideia da necessidade de atualização no perfil e na formação dos profissionais da informação possuem um valor inestimável para o desenvolvimento de modelos conceituais consistentes. A teoria de Classificação Facetada de Ranganathan, desenvolvida na década de 1930, serve como base para a organização de conhecimentos e conteúdos informacionais. Seu trabalho é considerado como uma quebra de paradigma na construção de tabelas de classificação, visto que os modelos tradicionais dificultavam e, até mesmo, impediam a representação de temas mais novos. (Campos, Gomes e Oliveira, 2013)

A teoria da classificação facetada parte do reconhecimento das categorias nos domínios de conhecimento que se quer representar. Presume uma análise do domínio em questão, a partir da identificação do tema principal e de quais categorias este assunto englobará.

Em continuidade, inserido em cada domínio, este é pesquisado e seus termos básicos identificados e agrupados em classes, ou facetas, de acordo com características comuns. Cada faceta, pertence a uma categoria fundamental, e, entende-se por categoria, os tipos mais genéricos sob os quais se podem agrupar coisas, de natureza similar, destacando que não existe uma proposta consensual de quais categorias seriam as mais adequadas para cobrir todas as coisas existentes, sejam elas reais ou imaginárias, concretas ou abstratas. O estudo de um conjunto de tais categorias, suas características e relações, de fato, é objeto de estudo e debate entre filósofos de diferentes correntes (Loux, 2006), e, de acordo com Mazzocchi e Gnoli (2010), serviram de inspiração para Ranganathan, que propõe um conjunto de cinco categorias, referenciadas pelo acrônimo PMEST: Personalidade, Matéria, Energia, Espaço e Tempo.

3 REPRESENTAÇÃO DE TEXTO

Para se obter uma base eficiente no processamento de texto, deve-se fundamentá-la em uma boa representação, portanto, da aplicação de algoritmos de aprendizado de máquina. A qualidade dos resultados, (pós) aplicação desses algoritmos, é diretamente proporcional à qualidade da representação da coleção de texto. Um documento é uma cadeia sequencial de palavras. Portanto, cada documento, geralmente, é representado por uma série de palavras.

A coleção de todas as palavras no sistema de treinamento é denominada vocabulário ou conjunto de recursos. De acordo com Aggarwall (2014), as representações no modelo espaço-vetorial são as mais comuns na área de Machine Learning, portanto, este projeto está pautado nas dimensões do modelo espaço-vetorial e são baseadas no uso de termos. No contexto espaço-vetorial pode se utilizar um vetor do tipo binário, este podendo ser usado para representar o documento, uma vez que, ao conter uma palavra característica, será atribuído o valor 1, se a palavra não estiver no documento, será atribuído o valor 0 (Leopold e Kindermann, 2002).

Geralmente, o modelo espaço-vetorial é usado para representar repositórios de texto. Neste modelo, o documento é representado por um vetor, e o tamanho corresponde ao termo ou atributo da coleção de texto. Geralmente, a palavra “termo” é usada para indicar as dimensões geradas com base nas palavras do texto (por exemplo, palavras simples, suas sequências ou seus conjuntos) e os atributos são usados para indicar dimensões que não são baseadas em termos (por exemplo, identificadores de página “web”, a existência de determinados autores no artigo ou a localização geográfica do documento) (Rossi, 2015).

A representação no modelo espaço-vetorial permite o uso de algoritmos tradicionais de aprendizado de máquina que processam vetores numéricos. Apenas para alguns algoritmos, tarefas ou coleções de texto específicos, o relacionamento entre as entidades (como termos ou documentos) no modelo espaço-vetorial pode melhorar o desempenho da classificação. Portanto, quando possível, esses sistemas são sempre acompanhados da aplicação de métodos de seleção de atributos. No caso de considerar apenas palavras simples como termos, a notação baseada no modelo espaço-vetorial ainda é um dos métodos mais comumente usados e eficazes na classificação automática de texto (do Prado e Ferneda, 2008; Feldman e Sanger, 2006). Na fase de análise do documento, é necessário considerar palavras, ou termos, que não são viáveis para treinar o classificador. Seja D = {d1, d2, ..., dN} uma

coleção de N documentos e T = {t1, t2, ..., tM} uma coleção de M termos que constituem uma coleção de texto. Portanto, cada um dos "N" vetores de documentos na coleção consiste em "M" dimensões. A união dos vetores de representação do documento na coleção é representada por uma matriz chamada matriz documento-termo (Tan et al., 2005) (Tabela 1).

Tabela 1: Matriz documento-termo representando N documentos e M termos

	t_1	t_2	t_3	...	t_{M-2}	t_{M-2}	t_M	Classe
d_1	w_{d_1,t_1}	w_{d_1,t_2}		...	$w_{d_1,t_{M-2}}$	$w_{d_1,t_{M-1}}$	w_{d_1,t_M}	c_{d_1}
d_2	w_{d_2,t_1}	w_{d_2,t_2}		...	$w_{d_2,t_{M-2}}$	$w_{d_2,t_{M-1}}$	w_{d_2,t_M}	c_{d_2}
d_3	w_{d_3,t_1}	w_{d_3,t_2}		...	$w_{d_3,t_{M-2}}$	$w_{d_3,t_{M-1}}$	w_{d_3,t_M}	c_{d_3}
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
d_{N-2}	w_{d_{N-2},t_1}	w_{d_{N-2},t_2}		...	$w_{d_{N-2},t_{M-2}}$	$w_{d_{N-2},t_{M-1}}$	w_{d_{N-2},t_M}	$c_{d_{N-2}}$
d_{N-1}	w_{d_{N-1},t_1}	w_{d_{N-1},t_2}		...	$w_{d_{N-1},t_{M-2}}$	$w_{d_{N-1},t_{M-1}}$	w_{d_{N-1},t_M}	$c_{d_{N-1}}$
d_N	w_{d_N,t_1}	w_{d_N,t_2}		...	$w_{d_N,t_{M-2}}$	$w_{d_N,t_{M-1}}$	w_{d_N,t_M}	c_{d_N}

Fonte: Tan *et al*. (2005).

O valor de uma célula wd_i,t_j na matriz documento-termo representa um valor ou um peso de um termo *tj* em um documento *di* . No caso da tarefa de classificação, há uma coluna adicional (última coluna) para representar a classe do documento. A classe de um documento *di* é denotada por cd_i . A coluna classe contém apenas valores nominais, que são descrições das classes dos documentos. Esses valores nominais são também denominados rótulos. O valor de cd_i é nulo caso não tenha sido definido um rótulo para um documento *di*. Os documentos que possuem valor para o atributo classe são denominados documentos rotulados, enquanto os que não possuem são denominados documentos não rotulados. Existem palavras desnecessárias, como artigos, verbos auxiliares, conjunções verbais e nominais que limitam a capacidade de análise do sistema, e devem ser filtradas e descartadas. Estas palavras são conhecidas como *stopwords* e são removidas, por meio da comparação de uma lista, durante a fase de pré-processamento. Este processo, também conhecido com seleção de recursos, é comum, pois reduz o tamanho do conjunto inicial de palavras que estão descritas na maioria dos documentos (Madsen et al. 2004).

3.1 Geração de Termos e Atributos

Um dos fatores mais importantes que implicam, diretamente, no quesito velocidade, bem como performance de classificação de textos é a geração de termos e atributos para representações baseadas no espaço-vetorial. Nesse contexto é apresentada a matriz documento-termo, denominada como bag-of-words, que tem, como características principais a alta dimensionalidade, causada pelo alto número de palavras diferentes contidas em um corpus textual e a alta esparsidade, que se origina no fato de que grande parte das palavras ocorre apenas em uma pequena parte dos documentos.

A bag-of-words é muito favorável na relação custo-benefício no que se refere ao número de termos, representação do conteúdo dos documentos para algoritmos de aprendizado de máquina e performance de classificação (Bekkerman e Allan, 2004).

No modelo espaço-vetorial a dimensionalidade e a esparsidade podem ser atenuadas por técnicas de pré-processamento, independentemente ao número de palavras

contidas nos documentos, resultando na melhoria da qualidade dos resultados dos algoritmos de aprendizado de máquina (Uysal e Gunal, 2014).

Dentre as técnicas de pré-processamento de textos destacam-se a padronização de caixas, remoção de palavras irrelevantes ou ruídos, agrupamento de palavras contendo o mesmo significado em um único atributo, ou ainda seleção de palavras de acordo com uma determinada função sintática.

Um dos procedimentos que podem ser utilizados para remoção de palavras irrelevantes é a remoção de *stopwords* que são palavras consideradas irrelevantes para os padrões aprendidos por algoritmos de aprendizado de máquina.

Normalmente são consideradas como *stopwords* preposições, pronomes, artigos e interjeições. Além dessas palavras, pode haver palavras que não são úteis apenas para um domínio de aplicação, como a palavra "introdução" em uma coleção de artigos científicos. Essas palavras consideradas irrelevantes somente em um domínio específico são denominadas stopwords de domínio. Além disso, em muitas aplicações de classificação automática de textos, palavras compostas por caracteres alfanuméricos podem ser consideradas como ruídos e podem ser removidas da coleção de textos.

3.2 Seleção de Features

O propósito do processo de seleção de Features, ou recursos, é sintetizar a dimensionalidade do conjunto de dados e limitar recursos que são considerados irrelevantes para a classificação.

Este método de conversão revela muitas vantagens, incluindo tamanho menor do conjunto de dados, requisitos de operações menores para algoritmos de classificação de texto, principalmente algoritmos com conjuntos de recursos limitados e importante atenuação de espaço para pesquisa. O objetivo é reduzir a complexidade da dimensionalidade para gerar melhor precisão de classificação. (Forman, 2003)

Outra vantagem é sua capacidade em reduzir o excesso de adaptação, ou seja, o fenômeno através do qual um classificador está sintonizado igualmente ao contingente de características dos dados de treinamento ao invés das características constitutivas das categorias e, consequentemente, para ampliar a generalização.

Métodos para seleção de subconjuntos de características para tarefa de classificação de documentos de texto utilizam uma função de avaliação que é aplicada a uma única palavra. (Soucy e Mineau, 2003)

No processo de seleção de *features*, a pontuação de palavras individuais, ou, melhores características individuais, pode ser realizada usando algumas das medidas, por exemplo, frequência do documento, frequência do termo, informação mútua e ganho de informação. O que é comum a estes métodos de pontuação de recursos é que eles concluem classificando-os por suas pontuações determinadas independentemente, e depois selecionam os melhores recursos de pontuação.

4 ALGORITMOS DE APRENDIZAGEM

Skinner (1950, p. 193) assevera que a aprendizagem é uma mudança na probabilidade de uma resposta específica, sendo que este conceito pode ser aplicado para humanos, bem como para máquinas, no instante em que um dispositivo físico artificial seja capaz de retornar respostas a entradas de dados, da mesma forma que seres vivos respondem a estímulos do meio.

Todas as informações disponibilizadas em forma de texto precisam ser organizadas para que os usuários realizem suas buscas da

melhor forma possível, podendo, assim, classificar, o texto, de acordo com o tema ou o tipo de cada documento. Assim, deve-se classificar os textos de acordo com seus tópicos ou com seu gênero, sendo a classificação por tópicos mais direta, sumária e a classificação por gênero mais genérica, pois, pode abranger classificações como, por exemplo, a maneira como foram escritos ou até mesmo seu público-alvo. Durante a classificação, de acordo com os gêneros, a heterogeneidade nos documentos é recorrente, uma vez que podem ser produzidos por diversas fontes. A classificação de texto possui diversos algoritmos que tentam se aproximar de uma solução ideal, nesse contexto são apresentadas algumas abordagens.

4.1 Redes Neurais

Esse modelo aproxima, por meio de simulação, as informações à forma como o cérebro humano funciona, utilizando algoritmos e estruturas de dados, realizando o processamento desses mesmos dados em paralelo onde cada um acessa a memória local individualmente. Uma série de regras que simulam os relacionamentos entre os dados é transmitido ao algoritmo, assim, este é treinado com rotinas de tratamento de hierarquia ou possíveis conexões. A eficácia do treinamento através das redes neurais está na tolerância a falhas, na capacidade de se auto-organizar e generalizar para que ocorra o processo de aprendizado (Kriesel, 2007).

Uma Rede Neural composta por diversas camadas consiste em uma grande quantidade de unidades (neurônios) agrupados de acordo com um padrão de conexão, essas unidades geralmente possuem três classes, as unidades de entrada que recebem a informação, as de saída que apresentam os resultados do processamento e as unidades encontradas entre as duas anteriores conhecidas como unidades escondidas (Kotsiantis; Zaharakis; Pintelas, 2007). Comparadas com os métodos de classificação que utilizam estatística, as redes neurais possuem tempos maiores durante o processo de treinamento do classificador.

4.2 Aprendizagem Baseada em Instâncias

Esse algoritmo utiliza instâncias específicas ao contrário dos métodos que empregam abstrações pré-compiladas durante a tarefa de classificação. Podem descrever conceitos probabilísticos, pois utilizam funções de similaridade para produzir correspondências graduais entre as instâncias. (AHA; KIBLER; ALBERT, 1991).

Pela sua simplicidade, consomem menos tempo na fase de treinamento, entretanto, demandam mais tempo no processo de classificação. (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

Dentro de um espaço bidimensional o algoritmo estima qual a classificação de uma instância com base em seus vizinhos mais próximos, no caso do algoritmo KNN.

Entretanto, pode encontrar problemas no caso de grandes bases de dados, pois o algoritmo precisa estimar qual a distância entre o elemento que deseja classificar e os demais elementos dos vetores de treinamento, os quais já possuem classe definida.

4.3 Algoritmos Genéticos

Genericamente, esses algoritmos são representados como processos de busca probabilísticos, desenvolvidos para ser aplicados em sistemas com dados abundantes, envolvendo estados que podem ser representados por “strings”. Estes métodos são intrinsecamente paralelos e utilizam um conjunto de amostras do espaço total (população de “strings”) para gerar novos

conjuntos de amostras (Goldberg; Holland, 1988).

Os modelos de classificação que utilizam esse método utilizam o paralelismo implícito. As interações entre as instâncias ocorrem através de mensagens. As condições são especificadas nos atributos das mensagens que eles recebem e as ações estão nos atributos das mensagens que eles enviam. O sistema resolve conflitos através de competição, não necessitando de algoritmos para o serviço, proporcionando uma capacidade incremental de inserção de novas regras sem atrapalhar as capacidades já estabelecidas. Algoritmos genéticos têm sido utilizados para treinar o peso das Redes Neurais e para encontrar a arquitetura das redes neurais (Goldberg; Holland, 1988).

4.4 Algoritmo de Bayes

Também conhecido como Rede Bayesiana ou algoritmos generativos, estes geram uma classificação probabilística com base em modelagens das características de palavras subjacentes em diferentes classes.

A semântica deste algoritmo é classificar o texto baseando-se na probabilidade deste estar contido nas diferentes classes, de acordo com a presença de palavras no documento.

Fundamenta-se na probabilidade de um elemento estar inserido a determinada categoria. Na aprendizagem Bayesiana o algoritmo recebe as instâncias rotuladas para o treinamento e, caso uma nova instância necessite de classificação, ele realiza o cálculo da probabilidade desta instância em receber cada rótulo, nomeando sempre a que recebe maior valor.

A fase de treinamento mais breve é a principal vantagem deste algoritmo e seu modelo, que tem a forma de um produto, pode ser convertido em uma soma por meio do uso de logaritmos trazendo vantagens computacionais (Kotsiantis; Zaharakis; Pintelas, 2007).

4.5 Árvore de Decisão

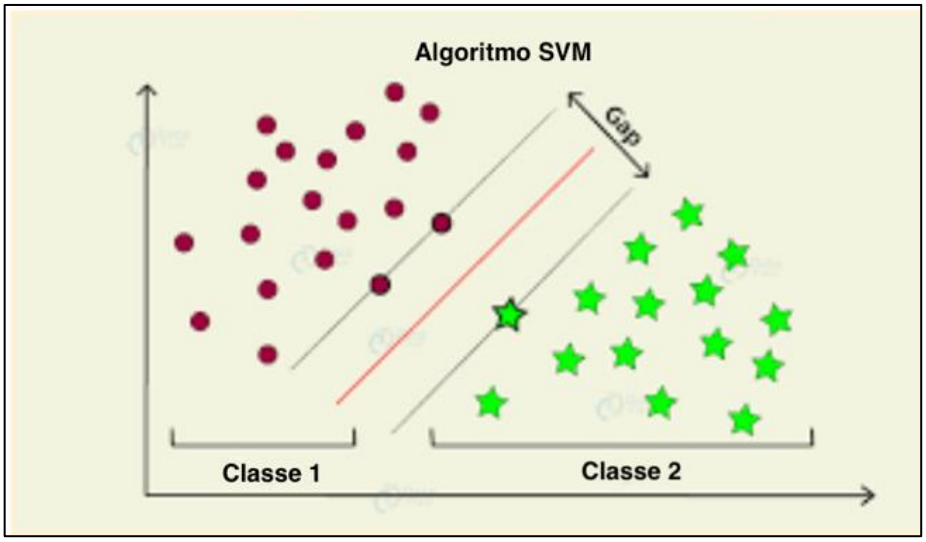
Uma árvore de decisão é produzida com base na decomposição hierárquica do espaço dos dados ou de treinamento, onde uma condição no valor dos atributos é utilizada para a divisão desse espaço, hierarquicamente, se tratando de documentos textuais essa condição geralmente é representada pela presença ou falta de um ou mais termos no texto.

A segmentação realiza-se recursivamente até que a árvore atinja uma quantidade mínima de folhas ou se chegue a uma condição estabelecida para a pureza da classe. Em determinado documento de treinamento a sequência de classes possíveis são aplicadas na estrutura da árvore criada a partir do seu topo até que se chegue á folha mais relevante correspondente à classe. (Aggarwal & Zhai, 2012).

4.6 SVM (Support Vector Machines)

É um método de aprendizagem de máquina mais recente e sua utilização se baseia na análise de dados para reconhecimento de padrões para classificação e análise de regressão. Também caracterizado como classificador linear binário não probabilístico pois, a partir de um conjunto de dados, prediz, para cada valor de entrada, qual de duas possíveis classes a entrada pertence (Figura 2).

Figura 2: Exemplo do algoritmo SVM



Fonte: https://data-flair.training/blogs/svm-support-vector-machine-tutorial/. Consultado em: 5 mar 2023.

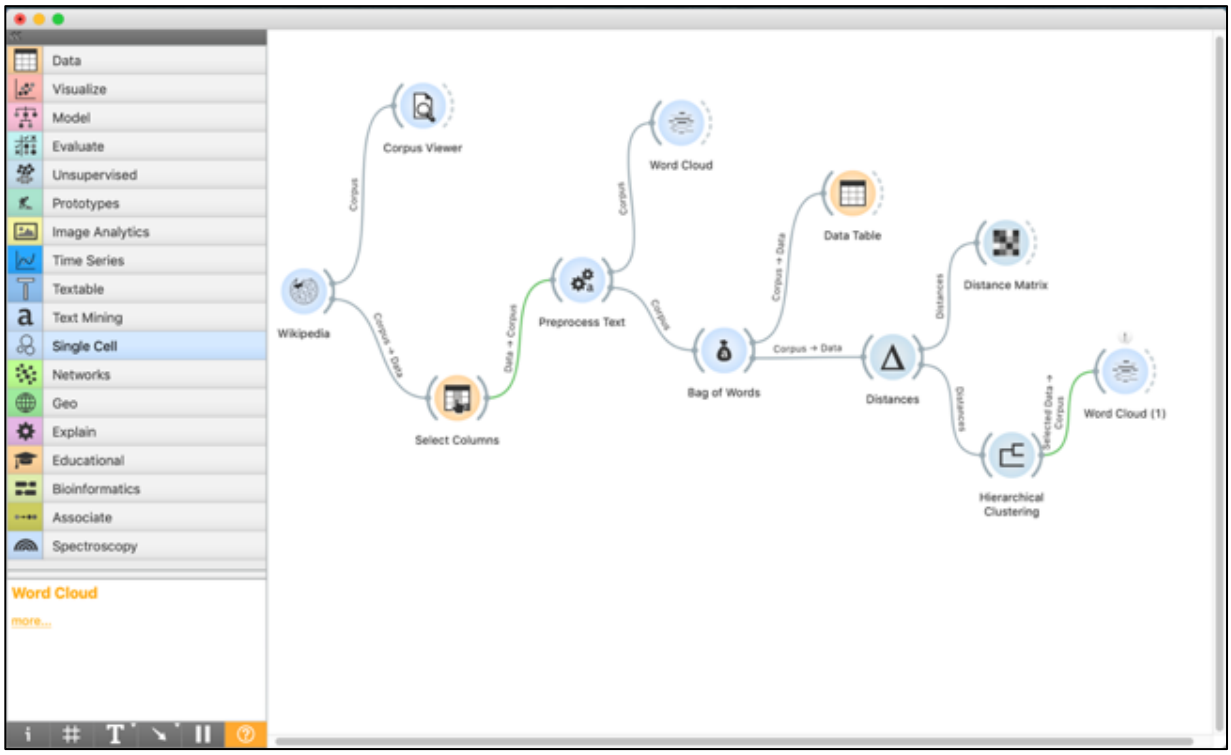
Na Figura 1 há uma linha de separação que busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes e essa linha é conhecida como hiperplano entre os dados de duas classes (class 1 / class 2). A distância entre o hiperplano e o primeiro ponto de cada classe pode ser caracterizada como margem. A SVM coloca em primeiro lugar a classificação das classes, definindo assim cada ponto pertencente a cada uma das classes, e em seguida maximiza a margem. Ou seja, ela primeiro classifica as classes corretamente e depois em função dessa restrição define a distância entre as margens. Por meio da maximização dessas margens e da segmentação das duas instâncias presentes em cada lado pode-se reduzir o erro de generalização (Kotsiantis; Zaharakis; Pintelas, 2007).

4 MATERIAIS E MÉTODOS

Quanto ao contexto de mineração de textos, há uma diversidade de ferramentas que podem ser utilizadas, dentre elas, WEKA, RapidMiner, Tanagra e Orange Canvas (Ignoatto E Webber, 2019, p. 3).

O presente trabalho utiliza a ferramenta Orange Canvas para a realização de todas as fases do processo do conhecimento (seleção, pré-processamento, transformação, mineração de dados e interpretação) uma vez que sua programação é intuitiva, a visualização permite desenhar o processo de análise de dados e possibilita a combinação de widgets para a estruturação de um framework. Essa é uma ferramenta *open source*, desenvolvida, desde 1996, no Laboratório de Bioinformática da Faculdade de Computação e Ciência da Informação da Universidade Ljubljana, na Eslovênia. Apresenta, como característica principal, uma interface onde é possível inserir widgets e, assim, elaborar trabalhos ou fluxos de análise de dados (Bennett, 2018) (Figura 3).

Figura 3: Interface de trabalho *Orange Canvas*



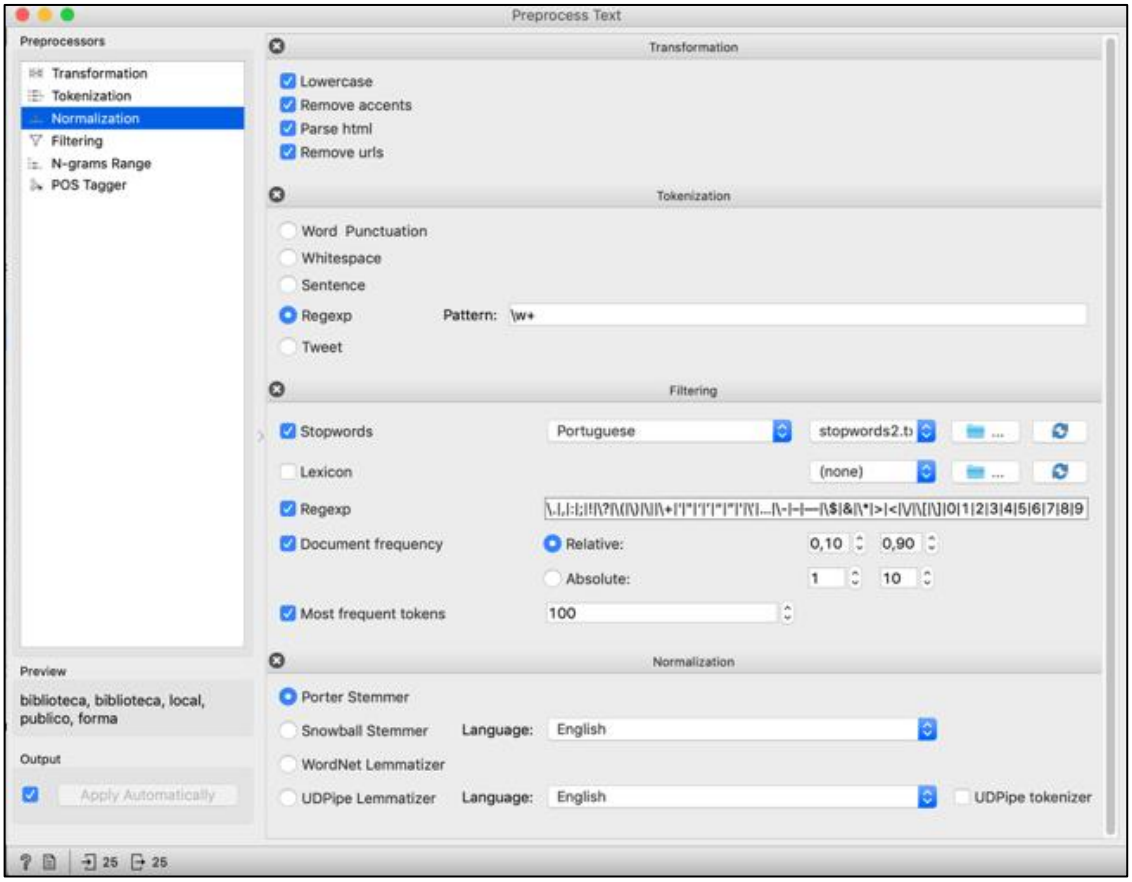
Fonte: Elaboração própria (2023).

Coleta de dados: Neste procedimento foi utilizado, como modelo, a coleta de vinte e cinco textos com o termo "Biblioteca" na plataforma "Wikipedia", visto que, na ferramenta há um *widget* próprio para realizar a recuperação textuais nessa mesma plataforma.

No processo de coleta são recuperados todos os termos dispostos no texto, como, título, conteúdo, tags, url (*Uniform Resource Locator*), números de página, dentre outros. Assim, é necessária a aplicação de um filtro para a escolha dos campos que serão recuperados no texto. O *widget "Select Columns*" realiza este procedimento que, neste caso, a definição de quais termos serão classificados é estritamente manual.

A seguir, deve-se realizar o pré-processamento do texto para que se elimine palavras sem relevância para a mineração, termos como artigos, pontos, números, e palavras sem sentido devem ser excluídas. O *widget "Preprocess Text"* realiza esta tarefa, filtrando as *stopwords* e normalizando os dados (Figura 4).

Figura 4: Uso da *widget* "*Preprocess Text*"



Fonte: Elaboração própria (2023).

Após a fase de pré-processamento, é imperativo que se junte as palavras mais relevantes, e, neste caso, é utilizado a *widget "Bag of Words"*. Nesse modelo, um texto é representado por um conjunto ou agrupamento (*Bag*) de suas palavras, desconsiderando a gramática e, consecutivamente, a ordem das palavras, entretanto, mantendo a multiplicidade. O modelo "*Bag of Words*" é comumente usado em métodos de classificação de documentos onde a frequência de cada palavra é empregada como uma característica para treinar um classificador.

Consecutivamente, deve-se utilizar o cálculo de distância entre os termos, e, neste caso foi utilizada a distância Euclidiana, conforme fórmula a seguir é selecionada na *widget* "*Distances*":

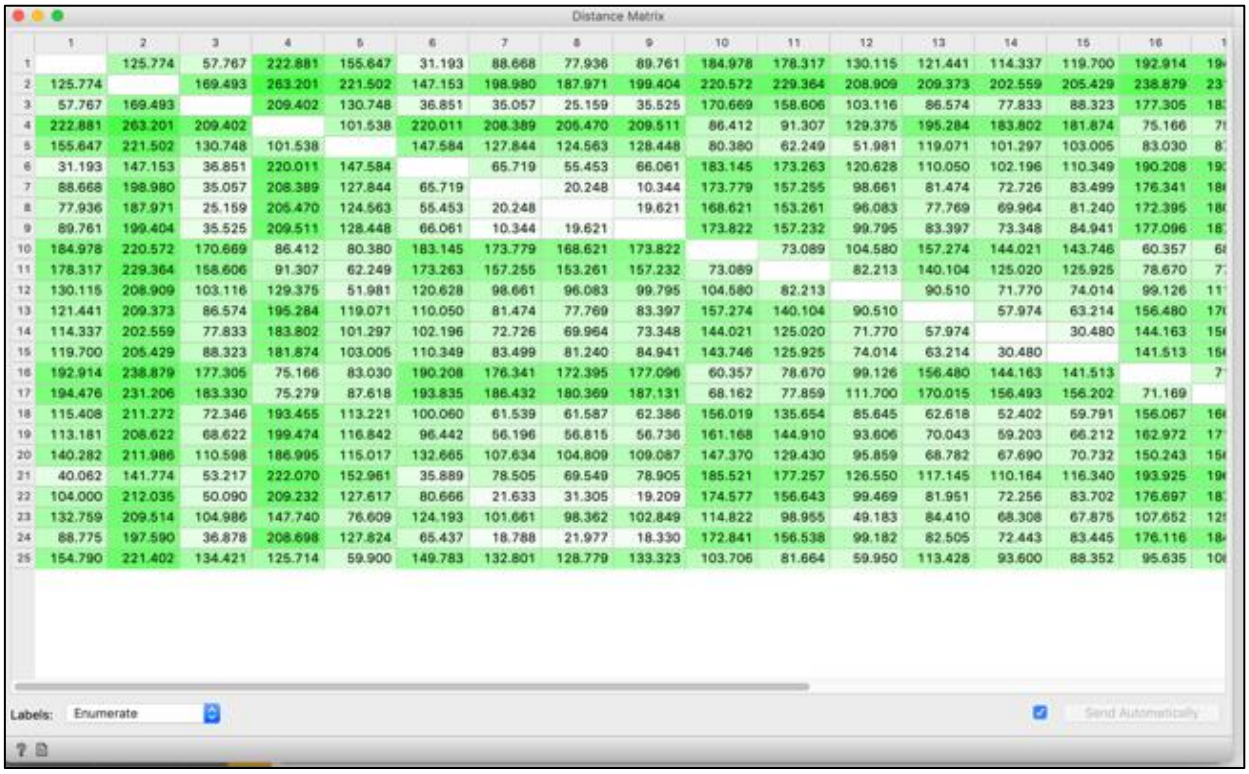
$$DE(x,y) = \sqrt{\sum_{i}^{p}.(X_i - Y_i)^2}$$

Onde, Xi é o termo ou palavra contida no primeiro texto e Yi é o termo ou palavra contido no segundo texto; o valor "i" indica o início da quantidade de palavras e o valor "p" indica o número máximo de palavras.

5 RESULTADOS

O resultado do cálculo desse algoritmo, que utiliza a distância euclidiana para determinar a matriz de distâncias entre os pontos, está explícito na widget “Distance Matrix” (Figura 5).

Figura 5: Matriz de Distância

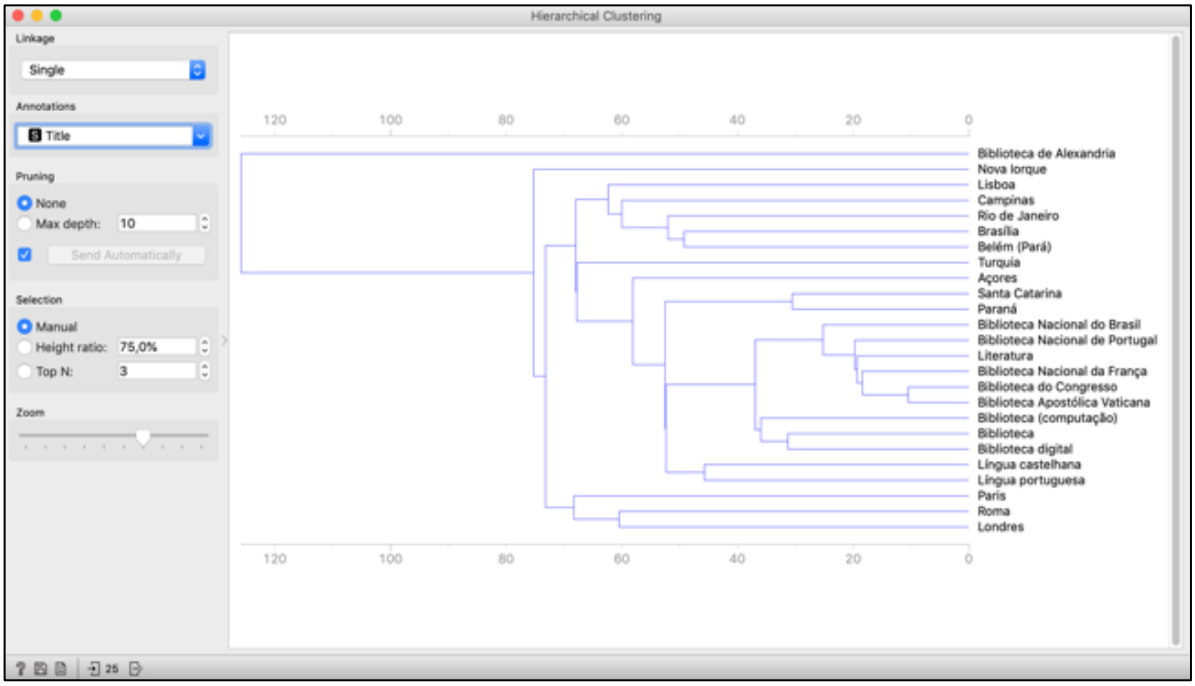


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1
1		125.774	57.767	222.881	155.647	31.193	88.668	77.936	89.761	184.978	178.317	130.115	121.441	114.337	119.700	192.914	19
2	125.774		169.493	263.201	221.502	147.153	198.980	187.971	199.404	220.572	229.364	208.909	209.373	202.559	205.429	238.879	23
3	57.767	169.493		209.402	130.748	36.851	35.057	25.159	35.525	170.669	158.606	103.116	86.574	77.833	88.323	177.305	18
4	222.881	263.201	209.402		101.538	220.011	208.389	205.470	209.511	86.412	91.307	129.375	195.284	183.802	181.874	75.166	7
5	155.647	221.502	130.748	101.538		147.584	127.844	124.563	128.448	80.380	62.249	51.981	119.071	101.297	103.005	83.030	8
6	31.193	147.153	36.851	220.011	147.584		65.719	55.453	66.061	183.145	173.263	120.628	110.050	102.196	110.349	190.208	19
7	88.668	198.980	35.057	208.389	127.844	65.719		20.248	10.344	173.779	157.255	98.661	81.474	72.726	83.499	176.341	18
8	77.936	187.971	25.159	205.470	124.563	55.453	20.248		19.621	168.621	153.261	96.083	77.769	69.964	81.240	172.395	18
9	89.761	199.404	35.525	209.511	128.448	66.061	10.344	19.621		173.822	157.232	99.795	83.397	73.348	84.941	177.096	18
10	184.978	220.572	170.669	86.412	80.380	183.145	173.779	168.621	173.822		73.089	104.580	157.274	144.021	143.746	60.357	6
11	178.317	229.364	158.606	91.307	62.249	173.263	157.255	153.261	157.232	73.089		82.213	140.104	125.020	125.925	78.670	7
12	130.115	208.909	103.116	129.375	51.981	120.628	98.661	96.083	99.795	104.580	82.213		90.510	71.770	74.014	99.126	11
13	121.441	209.373	86.574	195.284	119.071	110.050	81.474	77.769	83.397	157.274	140.104	90.510		57.974	63.214	156.480	17
14	114.337	202.559	77.833	183.802	101.297	102.196	72.726	69.964	73.348	144.021	125.020	71.770	57.974		30.480	144.163	15
15	119.700	205.429	88.323	181.874	103.005	110.349	83.499	81.240	84.941	143.746	125.925	74.014	63.214	30.480		141.513	15
16	192.914	238.879	177.305	75.166	83.030	190.208	176.341	172.395	177.096	60.357	78.670	99.126	156.480	144.163	141.513		7
17	194.476	231.206	183.330	75.279	87.618	193.835	186.432	180.369	187.131	68.162	77.859	111.700	170.015	156.493	156.202	71.169	
18	115.408	211.272	72.346	193.455	113.221	100.060	61.539	61.587	62.386	156.019	135.654	85.645	62.618	52.402	59.791	156.067	16
19	113.181	208.622	68.622	199.474	116.842	96.442	56.196	56.815	56.736	161.168	144.910	93.606	70.043	59.203	66.212	162.972	17
20	140.282	211.986	110.598	186.995	115.017	132.665	107.634	104.809	109.087	147.370	129.430	95.859	68.782	67.690	70.732	150.243	15
21	40.062	141.774	53.217	222.070	152.961	35.889	78.505	69.549	78.905	185.521	177.257	126.550	117.145	110.164	116.340	193.925	19
22	104.000	212.035	50.090	209.232	127.617	80.666	21.633	31.305	19.209	174.577	156.643	99.469	81.951	72.256	83.702	176.697	18
23	132.759	209.514	104.986	147.740	76.609	124.193	101.661	98.362	102.849	114.822	98.955	49.183	84.410	68.308	67.875	107.652	12
24	88.775	197.590	36.878	208.698	127.824	65.437	18.788	21.977	18.330	172.841	156.538	99.182	82.505	72.443	83.445	176.116	18
25	154.790	221.402	134.421	125.714	59.900	149.783	132.801	128.779	133.323	103.706	81.664	59.950	113.428	93.600	88.352	95.635	10

Fonte: Elaboração própria (2023).

Para que haja uma melhor interpretação das distâncias Euclidianas entre os termos foi utilizado, nesse experimento, um Cluster Hierárquico onde retorna a relação entre os títulos dos textos. A figura 6 ilustra o modelo da widget “Hierarchical Clustering”.

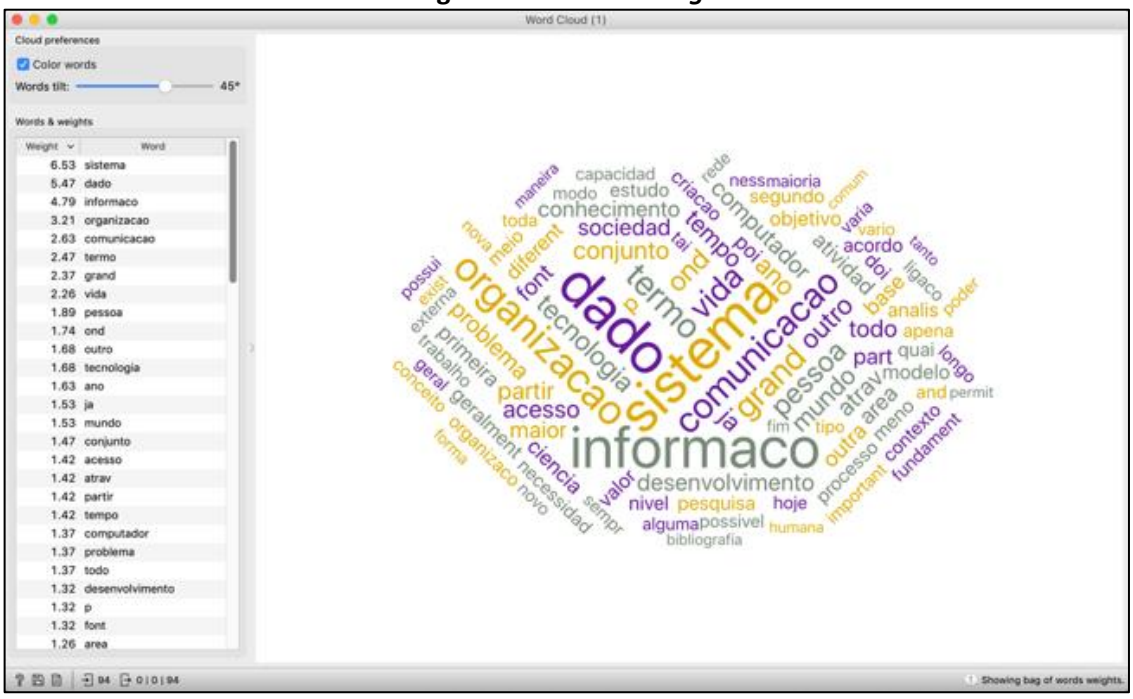
Figura 6: Hierarchical Clustering



Fonte: Elaboração própria (2023).

Outro resultado relevante foi a apresentação em nuvem de *tags* por meio da *widget "World Cloud",* que, neste cenário apresentou o peso do termo "biblioteca" em 92.50. A figura 7 ilustra o resultado.

Figura 7: Nuvem de *Tags*



Fonte: Elaboração própria (2023).

Durante o processo de classificação, foi utilizada uma base da Internet para a coleta de dados. Neste exemplo foi pesquisado o termo “Biblioteca” e foram recuperados vinte e cinco documentos que continham essa expressão. Após o uso do algoritmo de comparação por meio da aplicação do método de distância Euclidiana, observa-se que vários textos possuem aproximação.

Os resultados na matriz de distância ilustram a proximidade das palavras, que são representadas por valores numéricos, impossibilitando uma melhor interpretação. Entretanto, quando se utiliza o modelo de agrupamento por clusterização, a visualização da proximidade entre as palavras é mais compreensível.

Para concluir o processo de observação, utilizou-se a widget “Word Cloud”, popularmente caracterizada como “Nuvem de Tags” que permitiu uma análise mais detalhada e retornou alguns “pesos” pertinentes à pesquisa. No cenário apresentado, às cinco expressões mais recuperadas foram: “sistema” com peso 6,53; “dado” com peso 5,47; “informação” com peso “4,79”; “organização” com peso 3,21; e, “comunicação” com peso 2,63.

Com base nas análises é possível concluir que o experimento apresentado processou todas as fases adequadamente e apresentou resultados satisfatórios, assim, outros resultados podem ser investigados a partir da inserção de outros valores de entrada.

6 CONSIDERAÇÕES FINAIS

A pesquisa sobre classificação é atual e abrangente e desempenha um papel significativo para a organização do conhecimento. Entretanto, há que se considerar toda a estrutura para a implementação de tecnologia neste processo com a finalidade de obter resultados mais rápidos e precisos. Assim como Ranganathan rompeu paradigmas na construção de tabelas de classificação, o uso de aprendizagem de máquina e inteligência artificial estão se comportando da mesma maneira, visto que a capilaridade de recursos é elevada.

Entretanto, surgem algumas questões que podem gerar novas propostas de trabalho, sendo algumas delas: Quais modelos de características são viáveis para um bom desempenho dos classificadores? Alterar o modelo do espaço-vetorial baseado em conceitos terá maior relevância para a categorização do texto? Até qual dimensão do corpus a distância Euclidiana suporta analisar?

Outros problemas a serem discutidos na classificação de textos estão orientados aos termos relacionados à polissemia, onde uma palavra possui vários significados e sinonímia, onde palavras distintas podem ter o mesmo ou semelhante significado.

Ferramentas como a Orange Canvas permite aplicar diversos modos de classificação, sendo ela uma aplicação que abarca diversos padrões de análise, como, por exemplo, análises de sentimentos, predição e classificação de textos, análises de mapas epidemiológicos durante a pandemia Covid-19, clusterização de textos, dentre outras.

Gostaria de expressar meus agradecimentos ao Programa de Pós-Graduação em Ciências da Informação da UNESP Marília e ao Departamento de Computação da Fundação Educacional do Município de Assis. A ajuda financeira fornecida por ambas as instituições foi fundamental para a apresentação do artigo científico.

7 REFERÊNCIAS

Aggarwal, C. C., Zhao, Y., e Yu, P. S. (2014). On the use of side information for mining

text data. IEEE Transactions on Knowledge and Data En-gineering, 26(6):1415–1429.

Aha, David W; KIBLER, Dennis; ALBERT, Marc K. Instance-based learning algorithms. Machine learning 6.1, p. 37-66, 1991.

Barite, M.G., The Notion of "Category": Its Implica-tions in Subject Analysis and in the Construction and Evaluation of Indexing Languages. School of Library Science University of the Republic of Uruguay. 2000.

Bekkerman, R. e Allan, J. (2004). Using bigrams in text categorization. Relatório Técnico IR-408, Center of Intelligent Information Retrieval, UMass Amherst.

Bennett, J., Orange Data Mining, in https://www.predictiveanalyticstoday.com/Orange-data-mining/. 2018. Acesso em 03 de maio de 2023.

Breve, F. A., Zhao, L., Quiles, M. G., Pedrycz, W., e Liu, J. (2012). Particle competition and coope-ration in networks for semi-supervised lear-ning. IEEE Transactions on Knowledge and Da-ta Engineering, 24(9):1686–1698.

Burke, W. W., & Nourmair, D. A. (2001). The role of personality assessment in organization development. In J. Waclawski & A. H. Church (Eds.), Organization development: A data-driven approach to organizational change (pp. 55-77). Jossey-Bass.

Campos, M.L.A.; Gomes, H.E.; Oliveira, L.L. As Categorias de Ranganathan na organização dos conteúdos de um portal científico. Data-GramaZero, Rio de Janeiro, v. 14, n.3, jun. 2013.

DiFonzo, N., & Bordia, P. (2007). Rumor psychology: Social and organizational approaches. American Psychological Association.

Prado, H. A. do, E. Ferneda, E., editors (2008). Emerging Technologies of Text Mining: Techni-ques and Applications. Information Science Re-ference.

Fayyad, U.M., G.Piatetsky–Shapiro, P.Smyth. Kno-wledge Discovery and Data Mining: Towards a Unifying Framework. Proceeding of the Second International Conference on Knowledge Disco-very and Data Mining (KDD-96), Portland, Ore-gon, august, 1996.

Forman, G., An Experimental Study of Feature Se-lection Metrics for Text Categorization. Journal of Machine Learning Research, 3 2003, pp. 1289-1305

Galvão, N. D.; Marin, H. F. Técnica de mineração de dados: uma revisão da literatura. Acta Paulista de Enfermagem, São Paulo, v.22, n.5, p. 686-690, 2009.

Goldberg, David E; HOLLAND, John H. Genetic algo-rithms and machine learning. Machine lear-ning 3.2, p. 95-99, 1988.

He, W., Zha, S. & Li, L. social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Infor-mation Management, 33(3), 464-472. 2013.

Ignoatto M. L., Webber C. G., "Inteligência Competi-tiva nas Mídias Sociais: Um Estudo de Caso na Moda". Revista SCIENTIA CUM INDUSTRIA, V. 7, N. 2, PP. 156 — 164, 2019

Ikonomakis, M; kotsiantis, Sotiris; Tampakas, V. Text classification using machine learning tech-niques. WSEAS Transactions on Computers 4.8, p. 966-974, 2005.

King, M. L., Jr. (2010). Stride toward freedom: The Montgomery story. Beacon Press.

Kotsiantis, Sotiris B; zaharakis, I; pintelas, Pa-nayiotis. Supervised machine learning: A re-view of classification techniques. p. 3-24, 2007

Kriegel, David. A brief introduction to neural networks. 2007.

Leopold, Edda & Kindermann, Jörg, "Categorização de Texto com Máquinas V etoriais de Apoio". Como representar textos no espaço de entra-da", Machine Learning 46, 2002, pp. 423 - 444.

Madsen R. E., Sigurdsson S., Hansen L. K. e Lansen J., "Pruning the Vocabulary for Better Context Recognition", 7th International Conference on Pattern Recognition, 2004

Mazzochi, F. Gnoli, C. S.R. Ranganathan´s PMEST Categories: Analyzing their Philosophical Back-ground Cognitive Function. Information Studies, v.16, p. 133-147, 2010.

Posluszny, D., Spencer, S., & Baum, A. (2007). Post-traumatic stress disorder. In S. Ayers, A. Baum, C. McManus, & et al. (Eds.), Cambridge handbook of psychology, health and medicine (2nd ed.). Cambridge University Press.

Rossi, Rafael G., Classificação automática de textos por meio de aprendizado de máquina baseado em redes. Tese – Programa de Pós-graduação em Ciências de Computação e Matemática Computacional. ICMC/USP. São Carlos. 2015

Sebastiani, F. (2002). Machine learning in automa-ted text categorization. ACM Computing Sur-veys, 34(1):1–47.

Skinner, Burrhus F. Are theories of learning neces-sary? Psychological review 57.4, p. 193, 1950.

Somers, C. L., Day, A. G., Niewiadomski, J., Sutter, C., Baroni, B. A., & Hong, J. S. (2018). Under-standing how school climate affects overall mood in residential care: Perspectives of ado-lescent girls in foster care and juvenile justice systems. Juvenile & Family Court Journal, 69(4), 43-58. https://doi.org/10.1111/jfcj.12120.

Soucy P. e Mineau G., "Feature Selection Strategies for Text Categorization", AI 2003, LNAI 2671, 2003, pp. 505-509.

Uysal, A. K. e Gunal, S. (2014). The impact of pre-processing on text classification. Information Processing & Management, 50(1):104–112.