Pasture degradation papers search: how can supervised and transdutive methods help on the process of classification?

Daniel Osaku[1], **Patrícia M. Santos**[2], **Brucce N. Santos**[1], **Solange O. Rezende**[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP) São Carlos – SP – Brazil

[2]Embrapa, São Carlos – SP – Brazil.

{danosaku,brucce.neves}@usp.br, patricia.santos@embrapa.br,

solange@icmc.usp.br

Abstract. *The recovery of degraded pastures has been an important topic in terms of food security. Despite the large volume of scientific papers about degraded pastures, there is a significant challenge in terms of extracting knowledge from these documents. Here two classification approaches were used, one supervised and the other transductive, aiming to improve search quality and reduce manual annotation efforts. The results showed that it is possible to separate the articles of interest with a certain level of accuracy, with SVM supervised method standing out. In other hand, the GNetMine transductive algorithm demonstrated similar performance to supervised models, using only a quarter of the labeled data. Since manual annotation of training data for supervised methods is labor-intensive and relies on expert collaboration, emphasizing the need to develop classification methods that require a smaller number of labeled data. Upon selecting the articles of interest, in future other text mining techniques can be applied to facilitate knowledge extraction and the determination of recommendations for pasture recovery in the field, contributing to the sustainable increase in food production.*

Resumo. *A recuperação de pastagens degradadas tem sido tema importante no que diz respeito à segurança alimentar. Apesar do grande volume de artigos científicos sobre "pastagens degradadas", há um grande desafio em termos de recuperação desses documentos para extração de conhecimento. Neste artigo foram exploradas duas abordagens de classificação, uma supervisionada e outra transdutiva, visando melhorar a qualidade das buscas e reduzir o esforço de anotação manual. Os resultados mostraram que é possível separar os artigos de interesse com certo nível de precisão, com destaque para o método supervisionado SVM, que apresentou o melhor desempenho. Por outro lado, o algoritmo transdutivo GNetMine, que apresentou desempenho semelhante aos modelos supervisionados utilizando apenas um quarto dos dados rotulados. Uma vez que a anotação manual de dados para treinamento dos métodos supervisionados é trabalhosa e depende da colaboração de especialista, sendo fundamental o desenvolvimento de métodos de classificação que demandem menor número de dados rotulados. A partir da seleção de artigos de interesse, futuramente outras técnicas de Mineração de Textos poderão ser aplicadas para facilitar a extração de conhecimento e a determinação de recomendações para a recuperação de*

pastagens no campo, contribuindo para o aumento da produção de alimentos de forma sustentável.

1. Introdução

A segurança alimentar da população mundial depende da preservação de recursos naturais chave, como o solo e a água. O *World Resources Institute* - WRI estima que, para atender a demanda mundial por alimentos, fibra e energia, a produção agrícola em 2050 deverá ser 50% superior à de 2012 [Searchinger et al. 2019]. Devido às restrições quanto à abertura de novas áreas, o aumento da produção agrícola brasileira deverá ocorrer principalmente com base na recuperação de áreas degradadas, incluindo áreas de pastagem. Apesar de haver consenso em relação à existência de pastagens degradadas no Brasil, há grande variação nas estimativas de área em função dos métodos utilizados.

Atualmente, há um grande volume de informações científicas sobre o processo de degradação de pastagens, suas causas e estratégias de recuperação. No entanto, o conhecimento científico precisa percorrer um caminho desde a sua produção até a sua divulgação e apropriação pelos públicos de interesse [Telles et al. 2016]. A análise de documentos relacionados ao tema pode contribuir para acelerar o processo de desenvolvimento tecnológico e transferência de tecnologia e, consequentemente, para a recuperação de pastagens no campo. Essa, no entanto, não é uma tarefa trivial. O grande volume de publicações disponível impede a execução dessa tarefa de forma manual e desestimula a busca por artigos com recomendações para a recuperação de pastagens, tornando necessária a aplicação de ferramentas para a automatização do processo.

Os artigos científicos são textos escritos em linguagem natural e o processo de Mineração de Textos pode ser aplicado a esse tipo de documento para fins de classificação, agrupamento, construção de mapas conceituais e recomendação de produtos e serviços, dentre outros [Sinoara et al. 2021].

A Mineração de Textos, também chamada de mineração de dados textuais, é um campo multidisciplinar que inclui conhecimentos de áreas como informática, estatística, linguística e ciência cognitiva, que consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos [Aranha and Passos 2006]. Também pode ser definida como um conjunto de técnicas e processos utilizados em diversas áreas como inteligência artificial, aprendizado de máquina, base de dados e estatística, para descoberta de conhecimento inovador a partir de dados textuais [Rezende 2003].

Por exemplo, [Carvalho and Tsunoda 2018] analisaram dados recuperados da *Web of Science* no contexto de Mineração de Textos para identificar padrões a partir da identificação de características dos dados como, por exemplo, autores e países com publicações, ferramentas e métodos citados neles na área, seguida pela aplicação do algoritmo Apriori [Agarwal et al. 1994] para indicar associações entre termos em cada periódico. [de Moraes and Kafure 2020] empregou técnicas e ferramentas de bibliometria e ciência de redes para realizar o levantamento bibliográfico de pesquisas em documentos recuperados da *Web of Science* para sumarização, visualização e análise das redes, que permitiram a combinação de elementos para entendimento de informações e conhecimentos da área estudada, sendo eficazes na identificação de quem são os interlocutores, o que discutem e sua produção científica.

Mais recentemente, [Limiro et al. 2022] utilizaram métodos de Mineração de Textos para realizar agrupamentos de teses e dissertações e inferência de redes de conhecimento com base na similaridade e agrupamento de tópicos de documentos científicos textuais. [de Morais 2022] utilizou técnicas de Mineração de Textos para seleção dos principais artigos relacionados ao transtorno do espectro autista em ambientes escolares e descrever os seus principais termos nas áreas de pesquisa psicologia e educação.

Dessa maneira, a aplicação de técnicas de Mineração de Textos poderiam melhorar o resultado das pesquisas e facilitar a recuperação de conteúdo associado às recomendações para recuperação de pastagens. No entanto, em uma análise preliminar de documentos sobre "degradação de pastagens", recuperados na *Web of Science*, foram identificados alguns desafios para a aplicação das técnicas Mineração de Textos, como: 1) separar textos de outras áreas relacionadas ao tema "degradação de pastagens", como "restauração de vegetação natural", "solos" e "sistemas de produção integrados", que utilizam termos semelhantes em contextos diferentes; 2) extrair recomendações para a recuperação de pastagens, visto que elas não aparecem de forma explícita na maior parte dos artigos; e 3) relacionar as recomendações de recuperação de pastagens às condições nas quais elas devem ser aplicadas (i.e. local (região, bioma, estado, município), solo, clima, sistema de produção, tipo de capim, grau de degradação e causas de degradação), visto que essas informações também não estão explícitas nos textos.

O objetivo deste trabalho é selecionar artigos científicos relacionados ao tema "pastagens degradadas" no Brasil, de forma que os textos possam posteriormente ser analisados para extração de conhecimento sobre recomendações para recuperação de pastagens em função das condições nas quais o problema se apresenta. Para isso, foram aplicadas duas abordagens de classificação, considerando tanto a necessidade de anotação manual de dados para a fase de treinamento quanto a capacidade de considerar aspectos do contexto na tarefa de classificação.

2. Materiais e métodos

Este trabalho explora diferentes abordagens para o processo de classificação de artigos científicos de interesse para o tema degradação de pastagens, que pode ser dividida em duas etapas distintas: 1) aquisição dos dados; e 2) implementação dos métodos de Mineração de Textos.

2.1. Aquisição dos dados

Nesta seção, é apresentada a base de dados para uso no presente estudo. A base de dados foi obtida por meio de busca por textos científicos publicados por autores da Empresa Brasileira de Pesquisa Agropecuária - Embrapa, indexados na *Web of Science*. A Embrapa é uma instituição de P&D vinculada ao governo federal que desenvolve projetos relacionados ao tema degradação de pastagens e possui centros de pesquisa em todo o território nacional. A escolha da instituição foi feita com base em sua capilaridade territorial e na diversidade de condições em que as pesquisas são desenvolvidas no campo. A *Web of Science* é uma base multidisciplinar que indexa os periódicos mais citados em suas respectivas áreas. A expressão de busca utilizada foi construída a partir de termos relacionados ao sub-domínio "pastagens degradadas". A primeira parte da expressão visa recuperar textos sobre o domínio "pastagens", a segunda faz o recorte do subdomínio

“pastagens degradadas”, a terceira restringe a busca aos trabalhos com endereços (AD) no Brasil e a quarta estabelece o período a partir da data de criação da Embrapa. A inclusão do campo endereço na expressão foi fundamentada na percepção de que a maior parte dos artigos relacionados aos experimentos feitos no Brasil envolve autores sediados no país.

TS=((past* OR graz* OR silvipast* OR silvopast* OR grass* OR rangeland OR gram* OR forra* OR capim) AND (abandon* OR compact* OR conserva* OR manejo OR degrada* OR desertifica* OR ecologic* OR restorat* OR restaura* OR erosa* OR erosi* OR invas* OR nativ* OR degrada* OR leaching OR lixivia* OR recov* OR recuper* OR runoff OR escoa* OR loss* OR perda* OR ecoss* OR servi* OR noxious OR nociva* OR weed* OR erva OR “water repellency” OR “repelência à água” OR daninha*))) AND AD=(Brasil OR Brazil)) AND PY=(1973-2022)

Com a aplicação da expressão de busca foram recuperados 12.886 registros. Em seguida, foram aplicados alguns filtros para reduzir o número de registros recuperados, facilitando o processo de rotulagem manual e para melhorar a qualidade da busca, aumentando a porcentagem de artigos de interesse recuperados. A definição dos filtros foi feita por meio de testes realizados na plataforma da *Web of Science*. Os filtros inseridos foram:

- “Citation Topics Meso”: “3.45 Soil Science”; “3.4 Forestry”; “3.51 Dairy and Animal Science”; “3.4 Crop Science”; “3.97 Plant Pathology”.
- Áreas de pesquisa: “Agriculture”; “Environmental Sciences Ecology”; “Plant Science”; “Forestry”; “Biodiversity Conservation”; “Water Resources”.
- Tipo de documento: “Artigo”; “Artigo de revisão”; “Artigo de conferência”; “Acesso antecipado”; “Capítulo de livro”.
- Países/Regiões: “Brazil”
- Afiliação dos autores: “Empresa Brasileira de Pesquisa Agropecuária EMBRAPA”
- Idioma: “Portuguese”; “English”

A inclusão dos filtros “Citation Topics Meso” (5226 artigos recuperados) e Áreas de pesquisa (4502 artigos recuperados) reduziu a porcentagem de artigos recuperados fora do escopo de interesse. A inclusão do filtro Tipo de documento permitiu a seleção de registros relativos a trabalhos científicos completos, dos quais será possível extrair informações contextualizadas sobre as práticas de recuperação de pastagens no futuro (4483 artigos recuperados). A inclusão do filtro Países/Regiões reduziu a porcentagem de registros relativos a experimentos feitos em outras regiões do mundo (4482 artigos recuperados). O filtro Afiliação dos autores foi incluído para restringir a busca às pesquisas realizadas por pesquisadores de uma instituição de pesquisa e desenvolvimento, reduzindo o número total de registros e facilitando o processo de anotação manual (869 artigos recuperados). O filtro Idioma foi incluído para restringir o idioma dos textos e facilitar a aplicação das técnicas de Mineração de Textos (862 artigos recuperados). Após a aplicação de todos filtros, foram recuperados os metadados de 862 registros.

Os documento foram rotulados por um especialista de acordo com a presença (425 registros) ou ausência (437 registros) de informações sobre degradação pastagens e práticas de recuperação. A análise foi feita com base nos campos título e resumo do artigos. Para a anotação manual de cerca de vinte registros, o documento completo também

foi consultado para confirmar se o experimento havia sido realizado no Brasil e se o seu foco principal era a pastagem. Como, na maior parte dos casos, as recomendações de recuperação de pastagens não aparecem de forma explícita no texto, a classificação feita pelo especialista seguiu uma abordagem menos conservadora, considerando como de interesse também artigos nos quais as recomendações aparecem de forma implícita.

Em seguida, algumas técnicas de Mineração de Textos foram aplicadas para a identificação automática de artigos com informações sobre degradação de pastagens no Brasil e práticas de recuperação.

2.2. Métodos de Mineração de Textos

Seguindo a metodologia proposta por [Rezende 2003], a Figura 1 ilustra as etapas de mineração de dados compostas por: 1) identificação do problema; 2) pré-processamento; 3) extração de padrões; 4) pós-processamento; e 5) utilização do conhecimento.

Figura 1. Etapas do processo de mineração de dados proposto por [Rezende 2003].

Com relação à identificação do problema, os mecanismos de busca disponíveis na *Web of Science* não permitem uma seleção adequada dos artigos científicos. De acordo com a rotulagem manual realizada pelo especialista, cerca de metade dos documentos recuperados não contem informações de interesse para a tarefa de extração de conhecimento sobre degradação de pastagens.

A classificação manual dos artigos é trabalhosa e desestimula a extração de conhecimentos a partir do grande número de documentos disponível. A aplicação de técnicas de classificação automática é fundamental para a identificação dos artigos de interesse e para que o conhecimento disponível em publicações científica possa ser melhor aproveitado.

O tema "degradação de pastagens" tem relação com outras áreas de conhecimento, como "solos" e "ecologia". Os artigos relacionados a esses domínios compartilham termos semelhantes, porém aplicados em contextos diferentes, o que dificulta a

tarefa de classificação automática dos documentos. Portanto, é importante que o processo de classificação para a seleção de artigos de interesse seja capaz de incorporar aspectos de contexto à tarefa.

Dado os metadados extraídos da busca, foi realizada a limpeza dos dados utilizando a biblioteca NLTK [Bird et al. 2009] para a remoção de *stopwords* e execução da operação de *stemming*, isto é, remoção de palavras irrelevantes e diminuição de variação entre palavras com o mesmo valor semântico, respectivamente.

Posteriormente, dois modelos de representação foram utilizados: 1) modelo espaço vetorial, conhecido também como *Bag of Words* (BoW) quando esse modelo representa termos presentes em um documento [Tan et al. 2006] e 2) redes bipartidas [Rossi et al. 2016]. BoW é uma matriz de documento-termo (Figura 2a), em que cada linha representa um documento, cada coluna representa um n-grama presente na coleção de documentos e cada célula contém uma medida de frequência da palavra no respectivo documento, onde n-grama é o termo designado para descrever 1 termo ou palavra (unigrama) ou à uma sequência de termos (n-gramas).

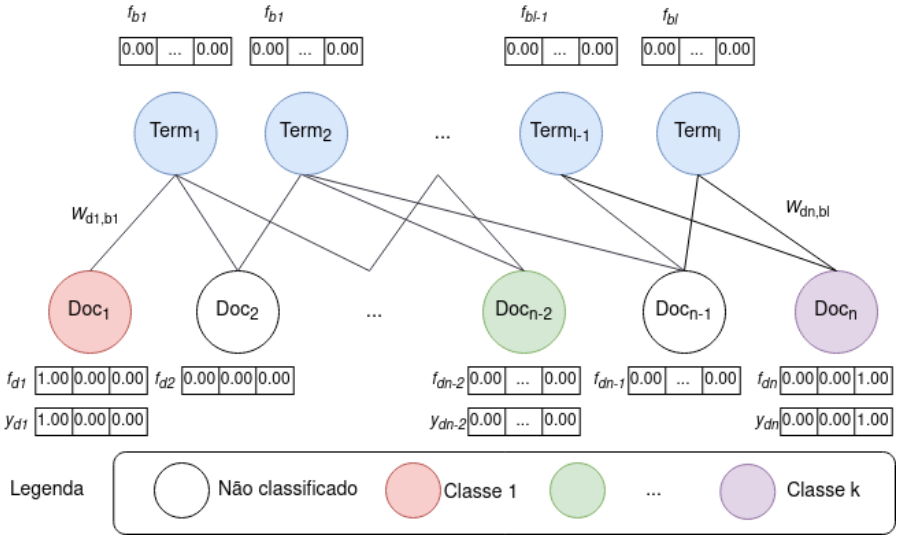
Já uma rede heterogênea bipartida representa os documentos em um grafo (Figura 2b), consistindo de 2 conjuntos distintos de vértices, cujas arestas conectam vértices de um conjunto com vértices de outro conjunto. Neste estudo, os termos extraídos do resumo e título de um artigo foram conectados ao vértice que representa esse artigo.

Na etapa de extração de padrões, a grande variedade de métodos aplicados torna inviável a busca exaustiva por melhores soluções, que vão desde métodos mais tradicionais como Naïve Bayes [Koch 2006] a métodos que modelam o contexto das palavras como o *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019]. Para atender o objetivo no processo de classificação serão aplicadas duas abordagens, sendo uma delas supervisionada, utilizando os métodos: i) Máquinas de Vetores de Suporte (SVM) [Cortes and Vapnik 1995]; ii) Redes *Multi-Layer Perceptron* (MLP) [Haykin 1994] e iii) BERT. A outra abordagem é transdutiva utilizando redes heterogêneas, essa abordagem se torna interessante no contexto desse trabalho devido a possibilidade de utilizar poucos dados rotulados. Porém para utilizar as redes para a tarefa de classificação é necessário utilizar um algoritmo de propagação de rótulos, nesse caso foi escolhido o GNetMine [Ji et al. 2010], que é um dos algoritmos clássicos para trabalhar com redes heterogêneas. A seguir é apresentada uma breve descrição dos métodos utilizados para a extração de padrões:

SVM: Separa as classes através da definição de hiperplanos que mapeia o espaço de características originalmente não linearmente separáveis em um espaço de mais alta dimensão, pois à medida que a dimensão é aumentada, também aumenta a probabilidade de que o problema se torne linearmente separável em relação a um espaço de mais baixa dimensão [Campbell and Ying 2022, Ma and Guo 2014, Lorena and De Carvalho 2007]. Quando os dados de treinamento são separáveis, o hiperplano ótimo no espaço de características é aquele que apresenta a máxima margem de separação [Semolini 2002]. Para maximizar a taxa de acerto usando a função de base radial (Radial Basis Function - RBF), os parâmetros C e γ precisam ser ajustados. A constante de penalização C determina o custo-beneficio entre a minimização do erro de ajuste e a maximização da margem de classificação, ao passo que γ afeta a transformação do mapeamento do espaço de dados e altera

	$term_1$	$term_2$	...	$term_{m-1}$	$term_m$	Classe
Doc1	$w_{x1,t1}$	$w_{x1,t2}$	...	$w_{x1,tm-1}$	$w_{x1,tm}$	c_{x1}
Doc2	$w_{x2,t1}$	$w_{x2,t2}$	...	$w_{x2,tm-1}$	$w_{x2,tm}$	c_{x2}
...	...	...	...	...	...	...
Doc_{n-1}	$w_{xn-1,t1}$	$w_{xn-1,t2}$	...	$w_{xn-1,tm-1}$	$w_{xn-1,ym}$	c_{n-1}
Doc_n	$w_{xn,t1}$	$w_{xn,t2}$	...	$w_{xn,tm-1}$	$w_{xn,tm}$	c_n

(a) *Bag-of-words* (BoW)



(b) Redes heterogêneas bipartidas

Figura 2. Representações de textos.

o grau de complexidade da distribuição amostral no espaço de características de mais alta dimensão [Ding 2009].

MLP: É um tipo de Rede Neural Artificial, cujas camadas são formadas por neurônios artificiais, que são inspirados no funcionamento de um neurônio, onde os sinais sinápticos de entrada são transformados e propagados para os neurônios seguintes [dos Santos Neto et al. 2020, Rodrigues 2019]. As redes MLPs consistem em uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída. As redes MLPs vem ganhando bastante interesse nos últimos anos pela sua capacidade de generalização e pelo aumento da capacidade de processamento dos computadores.

BERT: O BERT é um método que visa aprender representações de textos com o objetivo de criar um modelo de compreensão da linguagem [Devlin et al. 2019]. Ele é geralmente treinado usando conjuntos extensos de textos e permite refinamento posterior para uma tarefa específica com uma quantidade muito menor de textos se comparado ao que foi usado para o treinamento. Para aprender essas representações, o BERT utiliza exclusivamente o *encoder* do *Transformer*, com uma implementação quase idêntica à original.

GNetMine: É um algoritmo de propagação de rótulos em redes heterogêneas [Ji et al. 2010],

seu objetivo é propagar as classes considerando os diferentes tipos de relações e vértices que formam uma rede heterogênea. Além disso, o GNetMine possuí dois conjuntos de hiper-parâmetros, sendo que um deles é utilizado para atribuir um peso diferente para cada um dos tipos de relações entre vértices, em outras palavras o peso define a importância da informação de classe que transitam por relações de um determinado tipo. Enquanto o outro hiper-parâmetro é utilizado para indicar a grau de confiança (μ) atribuído aos rótulos dos vértices inicialmente rotulados [dos Santos et al. 2020, Ji et al. 2010], nesse caso se o grau de confiança for baixo isso permite que a rede altere inclusive os rótulos dos vértices inicialmente rotulados.

Para avaliar tais modelos, diferentes métricas podem ser analisadas para mensurar o desempenho dos modelos e identificar suas possíveis limitações na tarefa de classificação dos artigos. Dessa maneira, avaliar os falsos negativos produzidos pelos modelos deve ser relevante no sentido de mitigar o descarte de artigos relacionados à degradação de pastagens. Assim, os resultados serão avaliados utilizando as métricas F1-Macro, Acurácia e a Precisão. Além disso, para fazermos uma analise estatística dos modelos explorados será construído um diagrama de diferença critica (DC) utilizando Teste não paramétrico de Friedman com teste post-hoc de Nemenyi [Demšar 2006].

3. Experimentos e Análise dos Resultados

Nesta seção, são apresentados os resultados para a tarefa de Mineração de Textos em artigos científicos. Nessa avaliação, o conjunto de dados é composto por 856 artigos com resumo disponível no idioma inglês (6 artigos que não apresentavam resumo em inglês foram descartados), sendo 422 artigos da classe de interesse, enquanto 434 pertencem a outra classe. Dentre os metadados presentes nos artigos, os experimentos foram executados utilizando somente o título e o resumo do artigo.

3.1. Tratamento dos Dados

Com exceção do BERT, os outros algoritmos utilizados não conseguem lidar com os textos diretamente. Esses algoritmos precisam de uma representação estruturada, por isso é preciso realizar um pré-processamento nesses textos e transforma-los em uma representação no modelo espaço vetorial, nesse caso conhecido como *Bag-of-Words* (BoW). Nessa representação, os documentos são representados pelas linhas, enquanto as palavras/termos formam as colunas e o valor de cada célula indica a presença do termo no documento ou sua frequência. Nesse pré-processamento foram realizados os seguintes passos:

- O texto foi convertido para caixa baixa (minúsculo);
- Foi feita a remoção das *stop words*, juntamente das *stop words* de domínio, como também foi removido as pontuações e *tokens* numéricos, uma vez que esses não contribuem para o processo de classificação;
- Foi feita a radicalização (*Stemming*) das palavras, utilizando o algoritmo de Porter [Porter 1980];

Para construir a BoW, dentre as abordagens existentes foi escolhido por meio de experimentos preliminares a frequência que os termos ocorrem nos textos, mais especificamente foi utilizado o *CountVectorizer* presente na biblioteca do *Scikit-Learn*, a

ideia geral é que os valores de cada par (documento, termo) na BoW seja composta pela frequência que esse termo ocorre no texto. Além de palavras/termos individuais também foi gerados n-gramas, que são sequência de n palavras que ocorrem nos textos, nesse experimento foi utilizado $n = 4$, ou seja, foram geradas colunas contendo de 1 até 4 palavras, porém a representação acaba por se tornar ainda mais esparsa. Para lidar com a esparsidade foi necessário filtrar quais seriam os termos ou n-gramas que seriam utilizados pelos modelos, nesse caso, todas as colunas que tinham uma frequência de documento menor do que 2 foram removidas. Além disso, também foram removidas as colunas que tinha uma frequência maior que 90%, por fim foram selecionadas as 700 colunas com a maior frequência no conjunto de dados.

3.2. Avaliação

Para permitir a avaliação comparativa, foi utilizada uma validação cruzada com 5 *folds*, nesse cenário a BoW foi construída utilizando somente os documentos dos 4 *folds* referentes ao conjunto de treinamento. Lembrando que o BERT faz uso diretamente dos textos sem a necessidade dos pré-processamentos realizados na seção 3.1.

No caso da rede heterogênea que foi utilizada pelo algoritmo GNetMine, ela é uma rede bipartida composta por dois conjuntos de vértices, um deles sendo os documentos e o outro sendo composto pelos termos selecionados na BoW, por fim as arestas que ligam os documentos e os termos é a frequência que o termo ocorre naquele documento. Por ser uma rede bipartida, não existe conexão entre termos nem entre documentos.

Além disso, a rede é criada com a mesma estrutura da BoW que os modelos avaliados foram treinados, isso implica que a rede também é recriada toda vez que os *folds* de treinamento mudam. Sendo a única diferença o fato de que os pesos das arestas tiveram que ser normalizados para ficarem entre 0 e 1, uma vez que esse é um dos requisitos para a utilização dos algoritmos de propagação de rótulos. Para realizar a normalização, para cada documento foi somado a frequência de todos os termos que ele possuí e por fim cada termo é dividido por esse valor. A seguir é descrito os hiper-parâmetros utilizados em cada um dos modelos:

- SVM: foi utilizado o *kernel* RBF e função de decisão *one-vs-rest*. Os parâmetros C e γ foram otimizados usando busca em grade com os possíveis valores C = [0.1, 1, 10, 100] e γ = [1, 0.1, 0.01, 0.001]. A implementação utilizada foi da biblioteca *Scikit-Learn*.
- MLP: foi treinada uma rede com 2 camadas *fully connected* de 128 e 64 neurônios com normalização do *batch*, a função de ativação escolhida foi a *ReLu*, sendo que a camada de decisão utiliza a função de ativação sigmóide, essa rede foi treinada por 50 épocas com tamanho do *batch* de 8, esses hiper-parâmetros foram escolhidos empiricamente. A implementação utilizada foi da biblioteca Keras [Chollet et al. 2015].
- BERT: foi utilizada a versão pré-treinada "bert-base-uncased" e o refinamento dos pesos foi feito com 10 épocas, escolhida empiricamente. A implementação utilizada foi da biblioteca *HuggingFace*[1].
- GNetMine: o único hiper-parâmetro utilizado foi $\mu = 1$, pois como a rede só possui um tipo de relação, no caso entre documentos e termos, então não é necessário

[1]https://huggingface.co/bert-base-uncased

definir a importância de que cada tipo de relação. A implementação utilizada foi da biblioteca GraphTLP[2].

Como foram comparados dois tipos de abordagem (supervisionada e transdutiva), foi necessária uma análise cautelosa para que a avaliação fosse feita de forma justa entre os modelos. Então, para a avaliação do GNetMine, foi selecionado um processo um pouco diferente do realizado para os algoritmos supervisionados, pois o objetivo é que o GNetMine utilize menos dados rotulados e obtenha resultados competitivos. Sendo assim existem três formas que poderiam ser utilizadas para avaliar os resultados:

- Uma primeira abordagem que poderia ser utilizada para avaliar os modelos, seria treinar todos eles com poucos dados rotulados, porém essa abordagem seria injusta com os modelos supervisionados que dependem de mais dados rotulados.
- Outra abordagem seria treinar o GNetMine com uma pequena parte dos dados e realizar a avaliação com o restante dos dados, nesse caso se o GNetMine fosse treinado com 1 *fold* e avaliado com os outros 4, estaríamos penalizando mais o GNetMine do que os demais algoritmos.
- Por fim, poderia ser utilizada a mesma estrutura de treino e teste dos modelos supervisionados, porém no caso do GNetMine, seria utilizado somente 1 dos *folds* do treinamento como rótulos a serem propagados pela rede.

Dentre as formas de avaliação descritas, a que parece permitir uma melhor comparação entre as duas abordagem é a última. Porém dependendo do *fold* que é selecionado para treinamento é possível obter diferentes resultados, a fim de mitigar isso, dado os 4 *folds* de treinamento, cada um deles é utilizado individualmente no processo de propagação de rótulos e avaliado no *fold* de teste, por fim é feita uma média dos resultados das 4 execuções para esse *fold* de teste, esse processo se repete para todos os *folds*.

Os resultados obtidos estão ilustrados na Tabela 1, todos os modelos foram capazes de refinar o resultado da pesquisa, separando melhor os artigos de interesse. O modelo que apresentou o melhor desempenho na tarefa de classificação de artigos científicos relacionados ao tema "degradação de pastagens" foi o SVM, com maiores valores de F1-Macro, Acurácia e Precisão. Além disso, na Figura 3 está ilustrado um diagrama de diferença critica (DC) utilizando Teste não paramétrico de Friedman com teste post-hoc de Nemenyi [Demšar 2006], o diagrama foi construído com base nos rankings médios da Precisão, em cada um dos *folds*, quando não existe diferença estatística entre os modelos avaliados eles são conectados por uma linha horizontal.

Tabela 1. Desempenho dos modelos SVM, MLP, GNetMine e BERT na tarefa de classificação de artigos científicos juntamente do desvio padrão.

Modelos	F1-Macro	Acurácia	Precisão
SVM	**0.7733 ± 0.0367**	**0.7744 ± 0.0358**	**0.7861 ± 0.0465**
MLP	0.7660 ± 0.0614	0.7663 ± 0.0610	0.7549 ± 0.0635
BERT	0.7678 ± 0.0261	0.7687 ± 0.0254	0.7573 ± 0.0364
GNetMine	0.7220 ± 0.0464	0.7249 ± 0.0456	0.7512 ± 0.0649

[2]https://github.com/BrucceNeves/GraphTLP

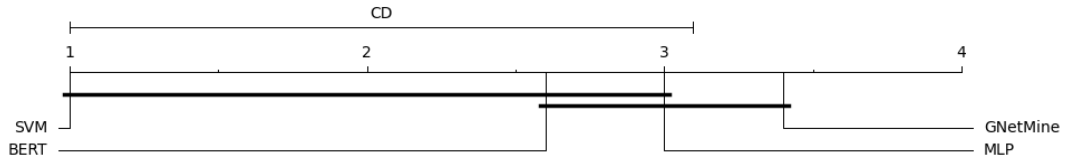


Figura 3. Diagrama de diferença crítica (DC) usando o teste não paramétrico de Friedman com teste post-hoc de Nemenyi entre cada um dos modelos.

Como é possível observar na Tabela 1, o BERT e o MLP tiveram desempenhos similares nas métricas avaliadas, sendo confirmado pelo diagrama DC, mostrando que ambos não possuem diferença estatística entre eles. Isso se dá pela similaridade do conteúdo dos artigos, fazendo com que ambas as classes fiquem bem próximas, conforme ilustrado na Figura 4. O domínio "pastagens" e o sub-domínio "degradação de pastagens" tem forte interface com domínios como: "solos" (balanço de carbono, emissões de gases de efeito estufa, dinâmica de matéria orgânica e qualidade química, física e biológica), "ecologia" (restauração de vegetação nativa e biodiversidade) e outros, sendo que vários termos são compartilhados entre eles, contribuindo para a similaridade do conteúdo dos artigos. Por exemplo, na área de solos muitos experimentos são realizado com o objetivo de avaliar o sequestro de carbono na forma de matéria orgânica em diferentes condições, inclusive em áreas de pastagens. Esses experimentos, que tem como principal objeto de estudo o solo, nem sempre trazem informações úteis para a identificação de práticas de recuperação de pastagens, mas utilizam termos muitos semelhantes a outros artigos nos quais o impacto de práticas de recuperação de pastagens sobre a dinâmica de matéria orgânica e sequestro de carbono no solo são analisados. Nesse sentido, modelos de linguagem neural como o BERT tem um desempenho pior uma vez que ambas as classes estão em um contexto próximo, apesar disso tanto o BERT quanto o MLP tiveram desempenho próximo ao SVM, mostrando que não existe uma diferença estatística entre esses métodos.

Por outro lado, o GNetMine utilizando somente 1 *fold* para treinamento (equivalente a 20% dos dados rotulados) obteve resultados bem próximos aos demais modelos avaliados que utilizaram 4 *folds* (equivalente a 80% dos dados rotulados). Sendo que olhando a métrica de Precisão, que nesse trabalho é a mais relevante, a diferença entre o GNetMine, BERT e MLP são minímas, mesmo tendo uma grande diferença entre a quantidade de dados rotulados. Isso comprova a robustez desse método, que pode ser considerado como uma opção para rotulação de artigos, uma vez que utilizou $\frac{1}{4}$ dos dados rotulados se comparado com os outros métodos avaliados e mesmo assim teve um desempenho comparativo muito bom.

4. Conclusão

Atualmente, existe uma vasta quantidade de informações científicas disponíveis sobre o processo de degradação de pastagens. Analisar documentos relacionados a esse tema pode acelerar o desenvolvimento tecnológico, a transferência de tecnologia e, consequentemente, a recuperação das pastagens no campo. No entanto, essa não é uma tarefa simples. O grande volume de publicações disponíveis impede a execução manual desse processo e desencoraja a busca por artigos que contenham recomendações para a recuperação de pastagens, tornando necessário o uso de ferramentas para automatizar o processo.

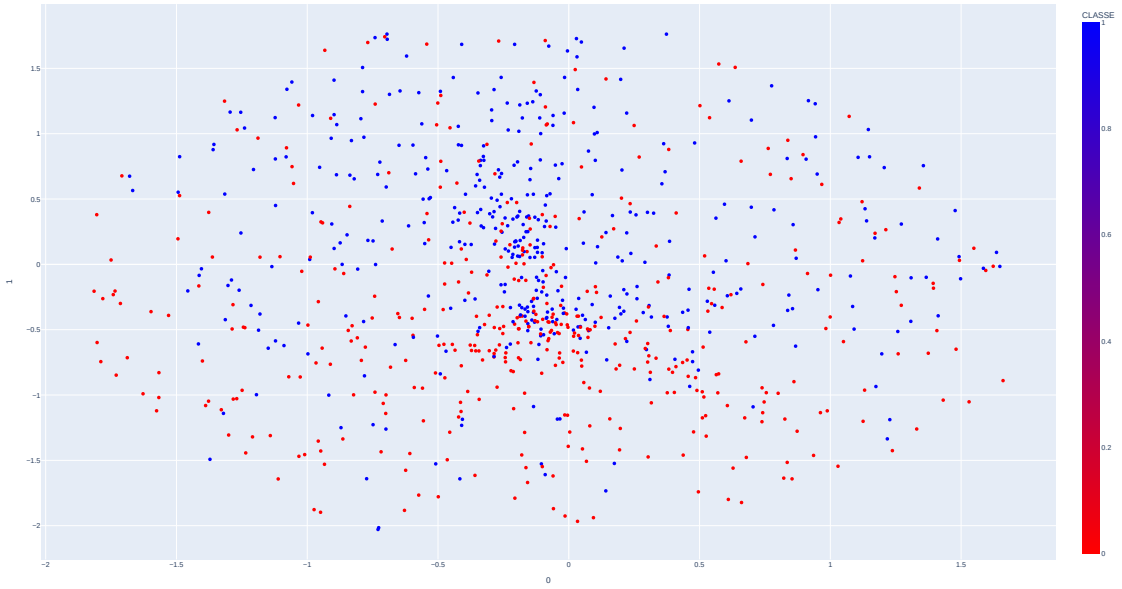


Figura 4. Distribuição dos artigos em um espaço 2D utilizando o *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [Van der Maaten and Hinton 2008].

A automatização do processo de extração de conhecimento sobre degradação de pastagens a partir de artigos científicos, no entanto, apresenta algumas dificuldades. O tema degradação de pastagens tem forte interface com outras áreas de conhecimento, como "solos" e "ecologia". Os artigos relacionados a essas áreas compartilham termos semelhantes, mas que são aplicados em contextos diferentes. A semelhança dos documentos dificulta a separação dos artigos com informações relevantes sobre degradação de pastagens a partir de buscas em bases de dados de publicações científicas. Além disso, a anotação manual dos documento é trabalhosa e demanda ajuda de especialista. Dessa forma há dois desafios, o primeiro é selecionar um modelo que utilize poucos dados rotulados. O segundo desafio está relacionado à dificuldade que esse modelo terá em separar os artigos de interesse dos demais que possam ser recuperados.

Com isso, esse artigo explorou duas abordagens, uma supervisionada que conta com modelos robustos, incluindo o BERT que tem se destacado em diversas tarefas de processamento de linguagem natural. A segunda foi uma abordagem transdutiva que visa a utilização de poucos dados rotulados, sendo esse um ponto importante para esse trabalho, uma vez que a quantidade de artigos publicados, ou seja, não rotulados, tende a ser muito maior do que os artigos rotulados, sendo que o processo de rotulação é custoso em diversos sentidos.

Com relação ao segundo desafio, os resultados mostraram que é possível separar os artigos de interesse dos demais artigos com um certo nível de precisão, com destaque para o método SVM, que apesar de não ter diferença estatística em relação ao BERT e o MLP, conseguiu melhores médias nas métricas utilizadas. Em relação ao primeiro desafio, foi observado que a rede proposta utilizando o algoritmo GNetMine para propagação de rótulos, alcançou resultados promissores em relação aos métodos supervisionados, mesmo em um cenário em que utilizou somente $\frac{1}{4}$ dos dados rotulados que foi utilizado

pelos modelos supervisionados, com resultados próximos aos obtidos por esses modelos.

A aplicação de técnicas de Mineração de Textos, portanto, permite um refinamento dos resultados das buscas realizadas em bases de dados científicos, diminuindo o esforço do usuário na seleção de artigos relevantes no domínio alvo, uma vez que boa parte dos artigos de interesse foram classificados corretamente com um baixo número de falsos negativos, por todas as abordagens.

Para trabalhos futuros, o objetivo é melhorar a precisão e a acurácia dos modelos supervisionados. Para isso, pode ser utilizada uma abordagem híbrida que faça uso do aprendizado transdutivo, em um cenário com poucos dados rotulados, a fim de aumentar o conjunto de treinamento com menor esforço para que esse possa ser utilizado por algoritmos supervisionados.

Outras possibilidades estão relacionadas a exploração de abordagens como "*one class*", que parece ser bem promissora para esse tema, uma vez que reduz o esforço do usuário na rotulação, já que ele passa a rotular somente dados da classe de interesse. Por fim, também é possível explorar a inclusão de outros tipos de vértices e relações na rede, afim de melhorar a precisão.

Após a seleção de artigos com informações relevantes sobre degradação de pastagens, outras técnicas de inteligência artificial podem ser aplicadas para a identificação do contexto em que os experimentos descritos nas publicações científicas foram realizados. A aplicação de práticas agropecuárias para a recuperação de pastagens depende de fatores relacionados ao bioma, tipo de solo, clima, tipo de capim, sistema de produção e outros. Com o uso de de técnicas envolvendo entidades nomeadas, é possível extrair essas informações do material e métodos dos artigos científicos. A partir destas informações, os artigos podem ser agrupados em função das condições nas quais os experimentos foram realizados e, em seguida sumarizados. A partir da análise agregada dos resultados de cada grupo de artigos, os especialistas podem identificar as práticas agropecuárias mais adequadas para a recuperação de pastagens em diferentes condições de campo, com maior segurança. Pode ser feita ainda uma adaptação de linguagem, de forma a tornar a informação disponível nas publicações mais acessível a profissionais da área de ciências agrárias que não tenham formação científica e que atuem diretamente no campo, prestando assistência técnica e consultoria ao produtores rurais.

A automação do processo de extração de conhecimento a partir de publicações científicas pode contribuir para a recuperação de pastagens degradadas no Brasil e, consequentemente, para aumentar a produção de alimentos de forma sustentável. Além disso, esse processo contribui para aumentar o potencial de impacto das pesquisas realizadas por instituições de ciência e tecnologia no país.

5. Agradecimentos

Este trabalho foi executado no Centro de Inteligência Artificial (C4AI-USP) com apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation. Os autores também agradecem à Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/25010-5), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq processo 309575/2021-4), à Embrapa e ao ICMC USP São Carlos pelo apoio financeiro e técnico, que contribuíram para que este trabalho tenha sido realizado.

Referências

Agarwal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, volume 487, page 499.

Aranha, C. and Passos, E. (2006). A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*, 5(2).

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Campbell, C. and Ying, Y. (2022). *Learning with support vector machines*. Springer Nature.

Carvalho, M. B. and Tsunoda, D. F. (2018). Data analysis on articles retrieved from web of science (wos). *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação; Edição Especial-v. 23, n. esp. 1 (2018); 112-125*, 24(2):125–112.

Chollet, F. et al. (2015). Keras. https://github.com/fchollet/keras. Acessado em 20-06-2023.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

de Moraes, L. L. and Kafure, I. (2020). Bibliometria e ciência de dados: um exemplo de busca e análise de dados da web of science (wos). *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, 18:e020016–e020016.

de Morais, M. V. B. (2022). Comparação bibliográfica sobre ensino de matemática para pessoas com transtorno autista utilizando técnica de mineração de texto. *REMAT: Revista Eletrônica da Matemática*, 8(1):e2002–e2002.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ding, S. (2009). Feature selection based f-score and aco algorithm in support vector machine. In *Second International Symposium on Knowledge Acquisition and Modeling*, volume 1, pages 19–23.

dos Santos, B. N., Rossi, R. G., Rezende, S. O., and Marcacini, R. M. (2020). A two-stage regularization framework for heterogeneous event networks. *Pattern Recognition Letters*, 138:490–496.

dos Santos Neto, L. A., Maniesi, V., Querino, C. A. S., da Silva, M. J. G., and Brown, V. R. (2020). Modelagem hidroclimatologica utilizando redes neurais multi layer perceptron em bacia hidrográfica no sudoeste da amazônia. *Revista Brasileira de Climatologia*, 26.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD, September 20-24, 2010, Proceedings, Part I 21*, pages 570–586, Barcelona, Spain. Springer.

Koch, K.-R. (2006). *Bayesian inference with geodetic applications*, volume 31. Springer, Germany.

Limiro, R. M., Da Silva, N. R., and Cordeiro, D. F. (2022). Mineração de textos para agrupamento de teses e dissertações por meio de análise de similaridade. *Revista Brasileira de Biblioteconomia e Documentação*, 18:1–20.

Lorena, A. C. and De Carvalho, A. C. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67.

Ma, Y. and Guo, G. (2014). *Support vector machines applications*, volume 649. Springer.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

Rodrigues, W. G. (2019). Predição de diâmetros e cálculo de volume de clones de eucalipto: uma abordagem com redes multi layer perceptron e long-short term memory. Master's thesis, Universidade Federal de Goiás.

Rossi, R. G., de Andrade Lopes, A., and Rezende, S. O. (2016). Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. *Information Processing & Management*, 52(2):217–257.

Searchinger, T., Waite, R., Hanson, C., Ranganathan, J., Dumas, P., Matthews, E., and Klirs, C. (2019). Creating a sustainable food future: A menu of solutions to feed nearly 10 billion people by 2050. final report. `https://files.wri.org/d8/s3fs-public/wrr-food-full-report.pdf`. Accessado em 10 de Julho de 2023.

Semolini, R. (2002). *Support Vector Machines, Inferência Transdutiva e o Problema de Classificação*. PhD thesis, Universidade Estadual de Campinas.

Sinoara, R. A., Marcacini, R. M., and Rezende, S. O. (2021). Mineraçao de textos e semântica: desafios, abordagens e aplicaçoes. *Revista de Sistemas de Informaçao da FSMA*, 27(2021):41–53.

Tan, P., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Always learning. Pearson Addison Wesley.

Telles, M. A. et al. (2016). *Da produção do conhecimento científico à transferência de informações: análise da circulação de saberes no âmbito de duas redes de pesquisa agropecuária*. PhD thesis, Instituto Brasileiro de Informação em Ciência e Tecnologia/Universidade Federal do Rio de Janeiro.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).