UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Busca de artigos sobre degradação de pastagens: como métodos supervisionados e transdutivos podem auxiliar na classificação dos textos?

Patricia Menezes Santos

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Busca de artigos sobre degradação de pastagens: como métodos supervisionados e transdutivos podem auxiliar na classificação dos textos?

USP - São Carlos

2023

Patricia Menezes Santos

Busca de artigos sobre degradação de pastagens: como métodos supervisionados e transdutivos podem auxiliar na classificação dos textos?

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Solange O. Rezende
Co-orientador: Daniel Osaku

USP - São Carlos

2023

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

M543b

Menezes Santos, Patricia
Busca de artigos sobre degradação de pastagens: como métodos supervisionados e transdutivos podem auxiliar na classificação dos textos? / Patricia Menezes Santos; orientadora Solange Oliveira Rezende; coorientador Daniel Osaku. -- São Carlos, 2023.
59 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. mineração de dados. 2. redes heterogêneas. 3. aprendizado de máquina. 4. mineração de texto. 5. degradação de pastagens. I. Oliveira Rezende, Solange , orient. II. Osaku, Daniel, coorient. III. 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

AGRADECIMENTOS

Ao Dr. Roberto Figueira Santos Filho, que me fez acreditar que seria possível fazer uma especialização na área de Inteligência Artificial e Big Data.

À Prof. Solange Oliveira Rezende pelo incentivo, orientação e apoio ao longo do desenvolvimento deste trabalho.

Ao Dr. Daniel Osaku pela coorientação e apoio ao longo do desenvolvimento deste trabalho.

A Brucce Neves dos Santos pelo apoio no desenvolvimento do trabalho.

A Marco Antônio Alvares Balsalobre, Rafael Santos Balsalobre, e Gabriel Santos Balsalobre, pela paciência e apoio ao longo de todo o curso.

RESUMO

SANTOS, P. M. **Busca de artigos sobre degradação de pastagens**: como métodos supervisionados e transdutivos podem auxiliar na classificação dos textos? 2023. 60 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

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. Nesta monografia 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 apresentou desempenho semelhante aos modelos supervisionados utilizando apenas um quarto dos dados rotulados. Tendo em vista que a anotação manual de dados para treinamento dos métodos supervisionados é trabalhosa e depende da colaboração de especialista, é 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.

Palavras-chave: mineração de dados; redes heterogêneas; aprendizado de máquina; mineração de textos; degradação de pastagens.

ABSTRACT

SANTOS, P. M. **Pasture degradation papers search**: how can supervised and transfutive methods help on the process of classification? 2023. 60 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

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.

Keywords: data mining; heterogeneous network; machine learning; text mining; pasture degradation.

SUMÁRIO

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 (WRI, 2019). Em função de sua dimensão territorial e das características edafoclimáticas favoráveis à agricultura, o Brasil ocupa posição de destaque no cenário mundial e deverá ter uma participação cada vez maior no suprimento de alimentos e outros produtos agrícolas para o mundo.

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. Por exemplo, os dados do IBGE (2019) apontam a existência de 12,1 milhões de ha de pastagens em más condições no Brasil. Já o projeto MapBiomas, com o uso de geotecnologias, estimou que, em 2021, havia aproximadamente 33 e 62 milhões de ha de pastagens em estágio de degradação severa e moderada, respectivamente, no Brasil (MAPBIOMAS, 2022).

A degradação de pastagens provoca prejuízos ambientais, sociais e econômicos que podem ser percebidos tanto na escala das fazendas quanto em escalas regionais. Na escala das fazendas, a degradação de pastagens reduz a capacidade de suporte e o desempenho dos animais, comprometendo a viabilidade do sistema com impactos econômicos e sociais negativos. Além disso, favorece a degradação do solo e a perda de biodiversidade com impactos ambientais negativos. Na escala regional, a degradação de pastagens aumenta as emissões de gases de efeito estufa por unidade de produto animal, reduz a biodiversidade e pode favorecer o desmatamento.

Atualmente, há um grande volume de informações científicas sobre o processo de degradação de pastagens, suas causas e estratégias de recuperação. Em um levantamento preliminar na Web of Science foram recuperados 6.605 registros relacionados ao tema “degradação de pastagens”, com a participação de autores brasileiros, no período de 1982 a 2022 (CLARIVATE, 2022). No entanto, o conhecimento científico precisa percorrer um caminho desde a sua produção até a sua divulgação e apropriação pelos públicos de interesse (TELLES, 2016). Ao longo desse processo, o conhecimento científico deve ser sistematizado

e transformado em conhecimento tecnológico, muitas vezes com a participação de técnicos e produtores que, ao final do ciclo, irão aplicar o novo conhecimento gerado na solução de problemas práticos que enfrentam no dia a dia.

A análise de artigos científicos pode contribuir para acelerar o processo de desenvolvimento tecnológico e transferência de tecnologia e, consequentemente, para a recuperação de pastagens degradadas no campo. Essa, no entanto, não é uma tarefa trivial. O grande volume de textos 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.

A Extração de Conhecimento de Bases de Dados (*Knowledge Discovery in Databeses* – KDD) ou Mineração de Dados (MD) é uma área multidisciplinar que busca desenvolver tecnologias de extração automática de conhecimento de bases de dados. Ela foi definida por Fayyad et al. (1996) como "...o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados". Mais especificamente, a Mineração de Textos corresponde a um conjunto de técnicas que permite extrair conhecimento a partir de dados não estruturados, seguindo uma abordagem semântica ou estatística (EBECKEN et al., 2003).

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 aplicação de técnicas de Mineração de Textos poderia melhorar o resultado das pesquisas e facilitar a recuperação de conteúdo associado às recomendações para recuperação de pastagens. No entanto, uma análise preliminar dos documentos recuperados na Web of Science sobre "degradação de pastagens" aponta alguns desafios para a aplicação das técnicas Mineração de Textos, como:

- 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.
- Extrair recomendações para a recuperação de pastagens, visto que elas não aparecem de forma explícita na maior parte dos artigos. De modo geral, os documentos científicos apontam os tratamentos que proporcionaram os melhores resultados naquelas condições específicas, sem explicitar uma recomendação prática no texto. Além disso, o mesmo conjunto de termos pode ser utilizado para

descrever os materiais e métodos dos experimentos e as práticas recomendadas para intervenção em pastagens degradadas. Isso dificulta a seleção de registros e a extração de conhecimento relacionado às recomendações de intervenção nas pastagens degradadas.

- 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 presente trabalho faz parte do Projeto "Gestão da informação e do conhecimento como suporte à gestão estratégica do Portfólio de Pastagens – Infopasto", coordenado pela Empresa Brasileira de Pesquisa Agropecuária - Embrapa (EMBRAPA, 2022). O objetivo do Projeto Infopasto é

"...mapear o conhecimento e as informações e os dados gerados sobre o domínio "pastagens", com ênfase no subtema "recuperação de pastagens" - incluindo um diagnóstico do ambiente externo, para subsidiar a gestão estratégica da informação no âmbito do Portfólio de Pastagens e como apoio ao processo de gestão da inovação na área de Pastagens no Brasil" (EMBRAPA, 2022).

Dentro desse contexto, a aplicação de técnicas de Mineração de Textos pode contribuir para o mapeamento e extração de conhecimento a partir de textos científicos relacionados ao tema "degradação de pastagens". O objetivo geral 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 capturar aspectos do contexto de uso dos termos na tarefa de classificação.

Este trabalho de conclusão de curso está organizado como segue: o Capítulo 2 apresenta o referencial teórico, incluindo informações relacionadas à degradação de pastagens e conceitos gerais sobre Extração de Conhecimento de Bases de Dados e Mineração de Textos; o Capítulo 3 apresenta a abordagem metodológica e as etapas do processo utilizado,

desde a obtenção da base de dados, pré-processamento, modelagem das redes e critérios de avaliação; o Capítulo 4 apresenta detalhes sobre a avaliação dos experimentos, resultados e discussão; o Capítulo 5 apresenta as conclusões e aponta trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Para contextualizar melhor os problemas abordados no trabalho de conclusão de curso, a revisão bibliográfica foi dividida em: degradação de pastagens no Brasil, Extração de Conhecimento, Mineração de Textos e considerações finais. No primeiro tópico são abordados aspectos relacionados ao domínio "degradação de pastagens no Brasil". Nos demais tópicos são abordados conceitos sobre Extração de Conhecimento de Bases de Dados e Mineração de Textos.

2.1 Degradação de pastagens no Brasil

A degradação de pastagens pode ser definida tanto como um processo quanto como uma condição do pasto. Telles et al. (2021), define a degradação de pastagens como:

"Processo de perda de vigor e produtividade das espécies forrageiras destinadas ao pastejo em decorrência do manejo inadequado do solo, das forrageiras e dos animais. Apresenta diferentes estágios com características distintas em cada bioma e afeta a produção e o desempenho animal, além de causar danos ao solo, sem a possibilidade de recuperação natural."

Já Dias-Filho (2011) define pastagem degradada como:

"... área com acentuada diminuição da produtividade agrícola ideal (diminuição da capacidade de suporte ideal), podendo ou não ter perdido a capacidade de manter a produtividade biológica (acumular biomassa) significativa".

Apesar das variações no conceito, há consenso entre os especialistas de que existem áreas degradadas ou em degradação no Brasil. Os dados do IBGE (2019), de fonte

declaratória, apontam a existência de 12,1 milhões de ha de pastagens em más condições no Brasil (Tabela 1). No censo agropecuário, foi considerado que pastagem em más condições:

"...corresponde à área plantada com espécies vegetais, destinada ao pastejo dos animais existentes no estabelecimento, considerada nestas condições pelo próprio produtor. Inicialmente produtiva, tal pastagem assumira esta condição devido à ausência de manutenção ou ao uso intensivo, podendo apresentar outros problemas, como erosão, plantas invasoras e cupinzeiros" (IBGE, 2019).

Tabela 1. Pastagens em más condições nas regiões brasileiras, de acordo com dados do Censo Agropecuário de 2017.

Região	Área de pastagens (milhões de ha)	
	Total	Más condições
Norte	33,2	2,3 (7%)
Nordeste	27,5	4,1 (15%)
Centro-oeste	56,5	3,0 (5%)
Sudeste	27,3	2,4 (9%)
Sul	15,0	0,3 (2%)
Brasil	159,5	12,1 (7,6%)

Fonte: IBGE (2019)

Com o uso de geotecnologias, o Projeto MapBiomas estimou que, em 2021, havia aproximadamente 33 e 62 milhões de ha de pastagens em estágio de degradação severa e moderada no Brasil, respectivamente, conforme ilustrado na Tabela 2 (MAPBIOMAS, 2022). Segundo esse levantamento, as maiores áreas de pastagens em processo de degradação estão nos biomas Amazônia e Cerrado. O levantamento da área de pastagem degradada foi feito com base no vigor das plantas, estimado a partir de índices de vegetação obtidos por imagem de satélite (MAPBIOMAS, 2022).

Tabela 2. Pastagens em grau de degradação severa, moderada ou sem degradação, de acordo com dados do Projeto Mapbiomas.

Região	Qualidade das pastagens		
	Degradação severa	Degradação moderada	Sem degradação
	Área de pastagens (milhões de ha)		
Amazônia	8,7	22,5	24,1
Caatinga	4,4	7,6	6,8
Cerrado	12,7	18,7	16,1
Mata Atlântica	6,2	12,6	9,2
Pampa	-	-	-
Pantanal	1,3	0,9	0,4
Brasil	33,3	62,3	56,0

Fonte: MAPBIOMAS (2022)

Dias-Filho (2017) descreve duas formas de degradação de pastagens: a degradação agrícola, caracterizada pelo aumento da infestação de plantas invasoras que competem com a planta forrageia e reduzem a capacidade de suporte do pasto, e a degradação biológica, na qual a queda da produtividade do pasto está associada à deterioração do solo.

A degradação de pastagens pode ser decorrente de: falhas no estabelecimento do pasto decorrentes do preparo de solo inadequado, de sementes de baixa qualidade, da semeadura em época imprópria, de falha no manejo do primeiro pastejo, etc.; fatores abióticos, como excesso de chuva ou má drenagem do solo, déficit hídrico, baixa fertilidade do solo etc.; fatores bióticos, como infestação por plantas invasoras, ocorrência de pragas e doenças; práticas inadequadas de manejo do pasto e do pastejo, como uso do fogo, ausência de reposição de nutrientes no solo, intensidade de pastejo e períodos de ocupação e descanso inadequados (DIAS-FILHO, 2011).

A intervenção em pastagens degradadas pode envolver técnicas de recuperação direta, de renovação ou de recuperação/renovação indireta (DIAS-FILHO, 2017). As técnicas de recuperação direta são recomendadas para áreas em fase inicial de degradação e consistem, principalmente, em ajustes na carga animal e no manejo do pastejo, na correção e adubação do solo e no controle de plantas invasoras. Rocha et al. (2017) compararam os efeitos de seis estratégias para recuperação de pastagem de *Brachiaria brizantha* sobre a perda de sedimentos, água e nutrientes das pastagens, e concluíram que a reposição de nutrientes é

fundamental para evitar a erosão em áreas de pastagens degradadas. Marchi et al. (2022) observaram que a presença de plantas invasoras em áreas de reforma ou recuperação de pastagens de *Brachiaria brizantha* cv. Marandu reduz a qualidade da forrageira, e recomendaram que o controle das invasoras fosse feito até cerca de 20 dias após a intervenção.

As técnicas de renovação são recomendadas para situações em que não é possível recuperar a produtividade do pasto (por exemplo, pastagens em níveis muito avançados de degradação ou nas quais o material genético precisa ser substituído em função de problemas bióticos ou abióticos) ou quando o custo de recuperação é tão elevado que não justifica sua aplicação (por exemplo, áreas com elevada infestação por invasoras de difícil controle). A síndrome da morte do capim-marandu, uma das principais causas de degradação de pastagens na região norte do país, é causada pela baixa tolerância do capim-marandu ao excesso de umidade no solo, que favorece alterações morfológicas nas raízes e a incidência de doenças (BARBOSA, 2006; RIBEIRO et al., 2017). Manzatto et al. (2018) verificaram que a síndrome da morte do capim-marandu ocorre em solos com deficiência de drenagem e em áreas com precipitação superior a 2100 mm por ano. Nas áreas com risco de ocorrência de síndrome da morte de capim-marandu, recomenda-se a diversificação das pastagens, com plantio de gramíneas forrageiras mais tolerantes a solos com drenagem deficiente (DIAS-FILHO, 2006).

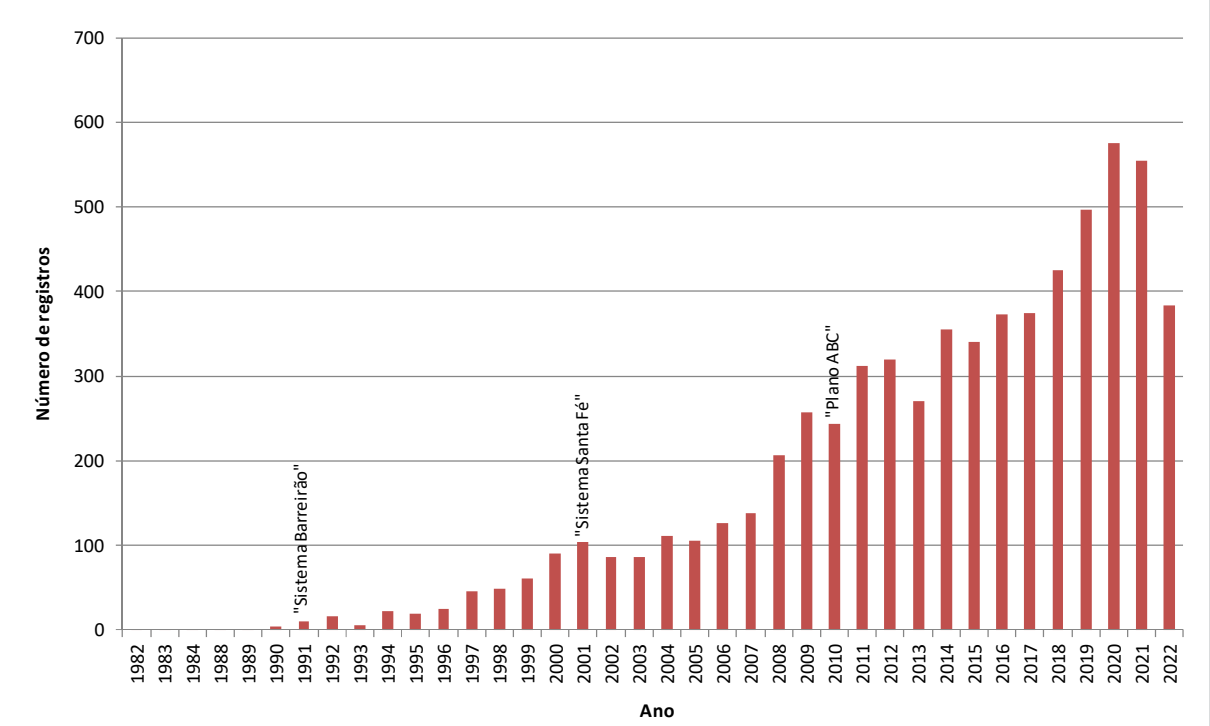
Por fim, as práticas de recuperação/renovação indireta implicam na integração da pastagem com o plantio de cultura agrícola e/ou espécies florestais (ILPF). Nos últimos anos, grande ênfase tem sido dada pela comunidade científica ao estudo dos sistemas ILPF, principalmente em função de seu potencial de sequestro de carbono (CECAGNO et al., 2018, SARTO et al., 2020). Balbino et al. (2011) define a integração lavoura-pecuária-floresta como:

"Estratégia de produção sustentável, que integra atividades agrícolas, pecuárias e florestais, realizadas na mesma área, em cultivo consorciado, em sucessão ou rotacionado, buscando efeitos sinérgicos entre os componentes do agroecossistema, contemplando a adequação ambiental, a valorização do homem e a viabilidade econômica, otimizando aumentos da produtividade com a conservação de recursos naturais."

Os técnicos e produtores, de modo geral, tem grande dificuldade de integrar o grande volume de informações disponível sobre degradação de pastagens e de identificar as melhores alternativas de intervenção para cada situação particular. Em um levantamento preliminar

feito na Web of Science foram recuperados 6.605 registros relacionados ao tema “degradação de pastagens” com a participação de autores brasileiros, no período de 1982 a 2022. A distribuição do número de registros por ano pode ser vista na Figura 1, onde também foram destacados eventos relevantes relacionados ao tema. O “Sistema Barreirão” e o “Sistema Santa Fé” foram lançados pela Embrapa em 1991 e 2002, respectivamente (KLUTHCOUSKI et al. 2003). Eles marcam o início das pesquisas com Sistemas de Integração Lavoura-Pecuária no Brasil, utilizados, dentre outras coisas, para renovação indireta de pastagens. O “Plano de Agricultura de Baixo Carbono – ABC” é um plano setorial de mitigação e de adaptação às mudanças climáticas, lançado pelo Ministério da Agricultura em 2010, sendo um de seus programas relativo à recuperação de áreas degradadas (MAPA, 2022).

Figura 1. Distribuição do número de registros por ano sobre o tema “degradação de pastagens”, recuperados em levantamento preliminar feito na Web of Science. As observações indicam os lançamentos do “Sistema Barreirão”, do “Sistema Santa Fé” e do “Plano ABC”.

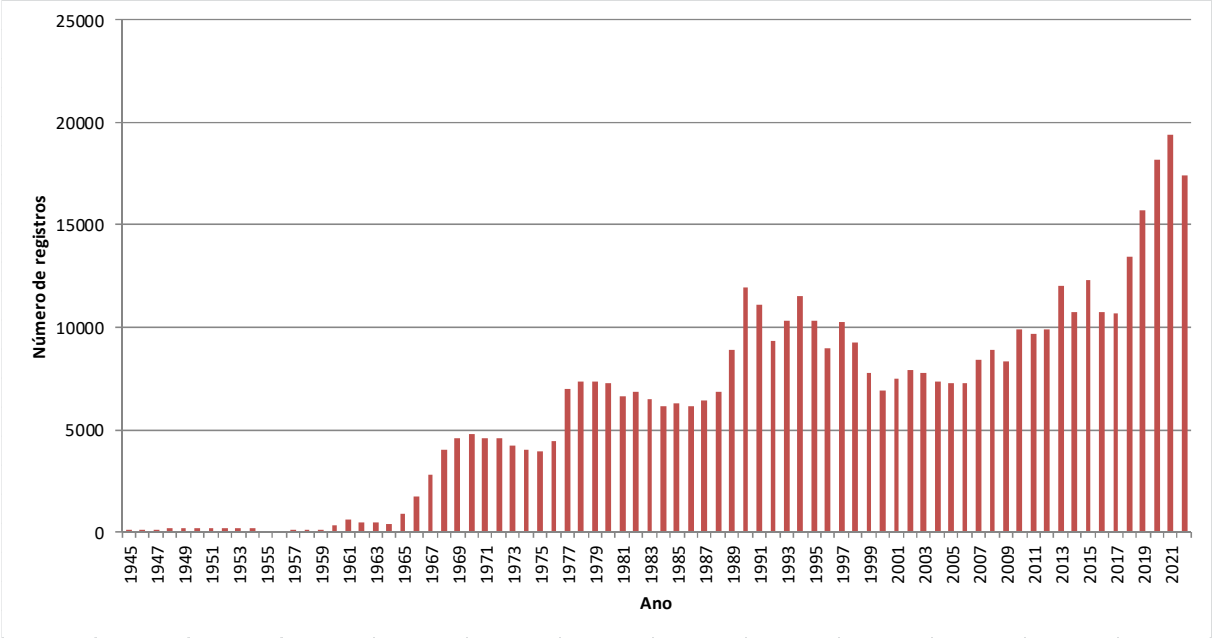


Fonte: Adaptado de CLARIVATE (2022).

2.2 Extração de Conhecimento

Nas últimas décadas, a humanidade tem vivenciado um aumento crescente no volume de dados coletados diariamente, o que torna seu processamento para gerar informação e conhecimento cada vez mais desafiador. Na área científica, esse processo também pode ser observado. Levantamento feito na *Web of Science* indica um aumento expressivo no número de publicações científicas indexadas na categoria "*Agronomy*" ao longo dos anos (Figura 2). Do início da década de 70 para cá, o número de artigos indexados praticamente quadruplicou, passando de cerca de 5 para quase 20 mil artigos por ano.

Figura 2. Artigos indexados pela *Web of Science*, na categoria "*Agronomy*".



Fonte: Adaptado de CLARIVATE (2022).

Além do grande número de publicações, a informação gerada e disponibilizada por meio das publicações científicas encontra-se cada vez mais segmentada em função da área de pesquisa, o que dificulta o processo de geração de conhecimento. Rezende (2003) explica que a geração de conhecimento "resulta de um processo no qual uma informação é comparada a outra e combinada em muitas ligações (hiperconexões) úteis e com significado".

A Extração de Conhecimento de Bases de Dados (*Knowledge Discovery in Databeses* – KDD) ou Mineração de Dados (MD) é uma área multidisciplinar que busca desenvolver tecnologias de extração automática de conhecimento de bases de dados. Ela foi definida por

Fayyad et al. (1996) como ”...o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados”.

Fayyad et al. (1996) descreveram os seguintes passos de um processo de Mineração de Dados: entendimento do domínio de aplicação e definição de metas; geração de base de dados; limpeza e pré-processamento dos dados; redução e projeção dos dados; escolha do método de mineração a ser aplicado; análise exploratória dos dados e seleção de modelos e hipóteses; mineração de dados; interpretação de padrões; aplicação do conhecimento gerado. Já Rezende et al. (2003) dividiram o processo de Mineração de Dados em cinco etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento (Figura 3).

Figura 3. Etapas do processo de Mineração de Dados.

Fonte: Rezende et al. (2003).

Durante a fase de identificação do problema é feito um estudo sobre o domínio ao qual a técnica de mineração de dados será aplicada, incluindo levantamento de algumas

informações prévias, e são estabelecidos os objetivos e metas do trabalho (REZENDE et al., 2003). Na fase de pré-processamento os dados são preparados para as análises posteriores por meio da aplicação de métodos de tratamento, limpeza e redução de volume de dados (REZENDE et al., 2003). Na fase de extração de padrões é feita a seleção, configuração e execução de algoritmos para extração de conhecimento a partir da base de dados pré-processada (REZENDE et al., 2003). Os métodos de extração de padrões, de modo geral, executam tarefas de predição (classificação e regressão) ou descrição (regras de associação, sumarização, agrupamento e outras) (FAYYARD et al., 1996, REZENDE et al., 2003). Por fim, no pós-processamento é verificado se os objetivos e metas definidos na primeira etapa foram alcançados e os resultados são avaliados por meio de medidas de desempenho (REZENDE et al., 2003).

2.3 Mineração de Textos

A Mineração de Textos corresponde a um conjunto de técnicas que permite extrair conhecimento a partir de dados não estruturados, seguindo uma abordagem semântica ou estatística (EBECKEN et al., 2003). É um campo multidisciplinar que agrega conhecimentos de áreas como informática, estatística, linguística e ciência cognitiva para extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos (ARANHA & 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).

Os artigos científicos são textos escritos em linguagem natural, e o processo de Mineração de Textos pode ser aplicado a esse conjunto de documentos para fins de classificação, agrupamento, construção de mapas conceituais e recomendação de produtos e serviços, dentre outros (SINOARA et al., 2021). Carvalho & Tsunoda (2018), por exemplo, analisaram dados recuperados da *Web of Science* no contexto de Mineração de Textos para identificar padrões a partir 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 & Kafure (2020) empregaram 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. Sobral et al. (2021) também realizaram análise bibliométrica sobre as aplicações da ciência de dados no âmbito das organizações hospitalares, utilizando a técnica de análise de redes sociais para explorar a literatura científica no período de 2015 a 2019 com o objetivo de encontrar termos relevantes, suas co-ocorrências e associações com a área de pesquisa, que revelaram a multidisciplinaridade existente em torno do assunto analisado.

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 Moraes (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 em psicologia e educação.

A representação estruturada de textos é necessária para a aplicação dos algoritmos de aprendizado de máquina e, consequentemente, para a Extração de Conhecimento (SINOARA et al., 2021). Desta forma, na fase de pré-processamento, além do tratamento, limpeza e redução de volume, para dados textuais também é necessário representar os documentos de forma que eles sejam processáveis pelos algoritmos para extração de padrões (SINOARA et al., 2021). O modelo booleano é um modelo binário utilizado em ferramentas de recuperação de informação (EBECKEN et al., 2003). As principais limitações do modelo booleano são: a dificuldade de seleção de termos de busca, e a impossibilidade de controlar o tamanho da saída e de ordenar os resultados de acordo com a relevância (EBECKEN et al., 2003).

No modelo espaço-vetorial é formada um matriz atributo-valor, em que cada linha representa um documento e cada coluna um atributo (ou termo) presente nos documentos (SINOARA et al., 2021). Esse modelo assume que os termos são independentes e, portanto, as relações entre eles não são levadas em consideração (SINOARA et al., 2021). Para que algoritmos baseados no modelo espaço-vetorial possam inferir relações entre termos ou documentos, é preciso que elas sejam representadas explicitamente, considerando frases ou conjuntos de palavras como termos (ROSSI, 2015). Outros problemas observados nesse modelo de representação são a elevada dimensionalidade e esparsidade (EBECKEN et al., 2003; ROSSI, 2015). Esses problemas podem ser reduzidos com a aplicação de técnicas de pré-processamento, como a padronização de caixa, remoção de palavras irrelevantes ou

ruídos, agrupamento de palavras com o mesmo significado ou seleção de palavras com auxílio de uma função sintática (ROSSI, 2015).

Os modelos de representação em redes permitem a representação de tipos de relações entre entidades dos textos, o que possibilita a captura de características das coleções de textos e seu uso para melhorar o desempenho de algoritmos na identificação de padrões (ROSSI, 2015). As redes podem ser definidas como uma tripla N = <O, R, W>, onde O corresponde ao conjunto de objetos da rede, R ao conjunto de relações entre os objetos e W o conjunto de pesos das relações entre os objetos, e ser classificadas de acordo com o sentido da relação entre os objetos (direcionadas, não direcionadas ou hipergrafos), o peso das relações entre os objetos (não ponderadas ou ponderadas), o tipo de objeto contido na rede (homogênea ou heterogênea) (ROSSI, 2015). Ji et al. (2010), por exemplo, utilizaram quatro tipos de objetos (artigo, conferência, autor e termo) e três tipos de relações (artigo x conferência, artigo x autor e artigo x termo) para modelar uma rede heterogênea para representar dados bibliográficos de uma coleção de textos.

Nas redes bipartidas, tipo específico de rede heterogênea, um determinado tipo de objeto se conecta apenas com outro tipo de objeto (ROSSI, 2015). As relações entre os objetos em uma rede podem ser explícitas, quando se referem a relações naturais ou informações explicitadas no conjunto de dados, ou implícitas, quando as relações são mineradas do conjunto de dados (ROSSI, 2015).

Apesar do bom desempenho dos algoritmos que utilizam modelo de representação em redes, Sinoara et al. (2021) alertam que as redes ainda podem ser limitadas em relação à semântica dos textos. As relações semânticas influenciam o significado do conteúdo dos textos e podem ajudar a distinguir documentos que utilizam o mesmo vocabulário para apresentar ideias diferentes. Sinoara et al. (2021) comentam que sistemas sensíveis ao contexto podem utilizar, além do comportamento e interesse do usuário, informações contextuais para fornecer recomendações mais precisas.

A classificação de textos consiste em designar rótulos aos documentos a partir de um conjunto pré-definido de classes de rótulos, de acordo com seu conteúdo (TAN et al., 2014; DENG et al., 2019). Várias técnicas de classificação têm sido aplicadas a problemas reais, como: vizinho mais próximo ("*Nearst Neighbor*" - NN), árvores de decisões ("*Decison Trees*" - DT, classificadores baseados em regras ("*ruled-based classifiers*"), redes neurais ("*neural networks*"), máquina de vetores de suporte ("*support vector machines*" - SVM) e *Naïve Bayes* (TAN et al., 2014; DENG et al., 2019).

A classificação de textos pode ser feita por técnicas supervisionadas, em que há um conjunto de treinamento com classes conhecidas, ou semisupervisionadas, quando há poucos dados rotulados para o treinamento. Além disso, a classificação automática de textos pode ser feita por algoritmos de aprendizado indutivo, em que um modelo de classificação é induzido a partir de textos rotulados e usado para classificar textos desconhecidos, ou transdutivo, quando há um número insuficiente de textos rotulados e dados não rotulados são usados para aprimorar o desempenho de classificação do algoritmo (ROSSI et al., 2016). O SVM e as redes neurais Multi-Layer Perceptron (MLP) e o *Bidirectional Enconder Representations from Transofrmers* (BERT) são exemplos de algoritmos de classificação indutiva e supervisionada.

O SVM é um dos métodos de classificação mais utilizados atualmente, em função de sua capacidade de generalização e poder de discriminação (CERVANTES et al., 2020). O SVM se baseia no princípio de minimização do risco estrutural ("*Structural Risk Minimization principal*" – SRM) para construir um hiperplano ótimo com a maior margem de separação possível entre pontos de classes distintas (TAN et al., 2014; DENG et al., 2019; CERVANTES et al., 2020). Essa estratégia minimiza os erros de classificação no conjunto de treino, ao mesmo tempo em que permite uma maior capacidade de generalização (CERVANTES et al., 2020). As principais limitações do SVM: a complexidade de algoritmo que afeta o tempo de treinamento em conjuntos de dados muito grandes, o desenvolvimento de classificadores para casos com mais de duas classes e o desempenho do algoritmo em dados desbalanceados.

O 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 MLP vêm ganhando bastante interesse nos últimos anos pela sua capacidade de generalização e pelo aumento da capacidade de processamento dos computadores.

O modelo BERT foi projetado para aprender representações bidirecionais a partir de textos não rotulados, condicionando de forma conjunta tanto o contexto à esquerda quanto à direita em todas as camadas representações de textos (DEVLIN et al., 2019). Durante a fase de pré-treinamento, o modelo é treinado em várias tarefas, o que permite um refinamento posterior para uma tarefa específica com uma quantidade muito menor de textos se comparado ao que foi usado para o pré-treinamento (DEVLIN et al., 2019). Para aprender

essas representações, o BERT utiliza o *encoder* do *Transformer*, com uma implementação quase idêntica à original (DEVLIN et al., 2019). Gozález-Carvajal & Garrido-Merchán (2021) verificaram desempenho superior do BERT em tarefas de classificação de textos efetuadas em diferentes cenários.

Para a classificação de textos, os algoritmos SVM, MLP e BERT utilizam informações sobre os termos presentes nos documentos. No presente trabalho, os documentos apresentam termos semelhantes, o que pode dificultar a classificação dos artigos. Além disso, por serem técnicas de aprendizado supervisionado, eles demandam grande esforço de anotação manual dos dados. Por outro lado, os artigos científicos estão associados a outras informações bibliográficas como autores, palavras-chave, áreas de conhecimento, etc. A representação dos documentos na forma de redes heterogêneas permite o uso desse tipo de informação, além dos termos, o que pode melhorar o desempenho de classificação de textos com termos semelhantes.

O GNetMine é uma estrutura de regularização baseada em grafos ("*graph-based regularization framework*") desenvolvida para extrair informações de redes heterogêneas com o objetivo de contornar as principais dificuldades encontradas nesse tipo de tarefa: complexidade da estrutura da rede, ausência de características ("*features*") e ausência de rótulos (JI et al., 2010). Ela preserva a consistência das relações presentes no grafo, considerando as diferenças entre os tipos de ligações e objetos (JI et al., 2010). O GNetMine é aplicado a problemas de aprendizado transdutivo semisupervisionado em redes heterogêneas, nos quais apenas um tipo de objeto é classificado e o algoritmo tenta predizer a classe dos demais tipos de objeto da rede (JI et al., 2010). A regularização dos rótulos é feita assumindo que a classe de dois objetos conectados tende a ser similar e que a classe predita para objetos rotulados deve ser similar ao rótulo pré-estabelecido (JI el al., 2010). O GNetMine permite a troca da classe de objetos rotulados e utiliza dois parâmetros de regularização: um para atribuir peso às relações entre objetos e outro para indicar a grau de confiança atribuído aos rótulos dos objetos (JI et al., 2010; DOS SANTOS NETO et al., 2020).

2.4 Mineração de Textos na identificação de técnicas de recuperação de pastagens

A "degradação de pastagens" causa impactos econômicos, sociais e ambientais relevantes. Há um grande volume de informação técnico-científica disponível, porém técnicos e produtores têm dificuldade de acessar e processar essas informações para identificar as melhores alternativas de intervenção para cada situação particular.

O conhecimento científico precisa percorrer um caminho desde a sua produção até a sua divulgação e apropriação pelos públicos de interesse (TELLES, 2016). Ao longo desse processo, o conhecimento científico deve ser sistematizado e transformado em conhecimento tecnológico, muitas vezes com a participação de técnicos e produtores que, ao final do ciclo, irão aplicar o novo conhecimento gerado na solução de problemas práticos que enfrentam no dia a dia.

As técnicas de Mineração de Textos podem ser aplicadas para auxiliar no processo de recuperação e síntese das informações disponíveis, contribuindo para acelerar a transformação do conhecimento científico em informação tecnológica e, consequentemente, para promover a adoção de tecnologias e o sucesso da recuperação de pastagens no campo.

3 METODOLOGIA

Este trabalho explora diferentes abordagens para o processo de classificação de artigos científicos de interesse para o tema "degradação de pastagens". Esta seção está dividida em: base de dados de artigos científicos e identificação do problema; pré-processamento; e implementação dos métodos de Mineração de Textos.

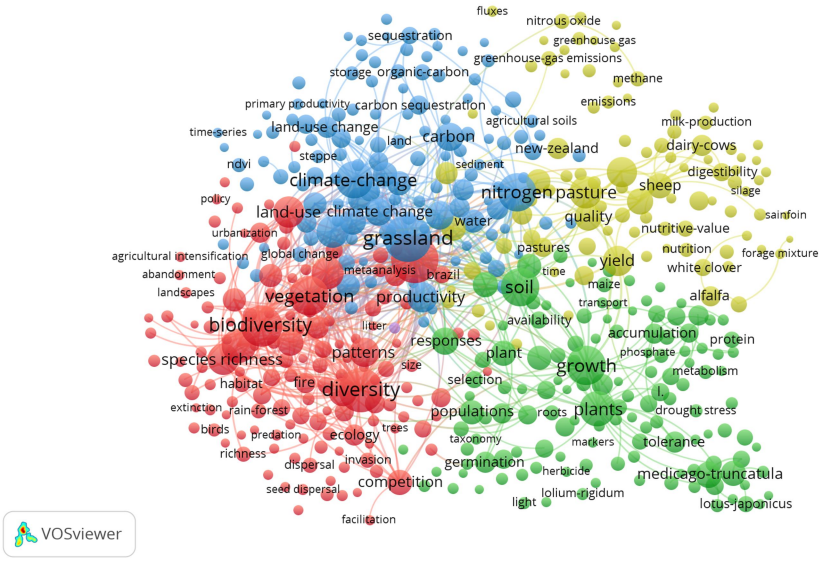
3.1 Base de dados de artigos científicos e identificação do problema

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*, no dia 3 de março de 2023. 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.

Para definir a expressão de busca, foram utilizados resultados anteriores obtidos pelo Projeto Infopasto. De forma resumida, a equipe do projeto Infopasto recuperou textos do domínio "pastagens" publicados em 2006, 2011 e 2016 e indexados na *Web of Science*, com o auxílio da expressão de busca definida por Santos et al. (2021). Em seguida, a equipe gerou

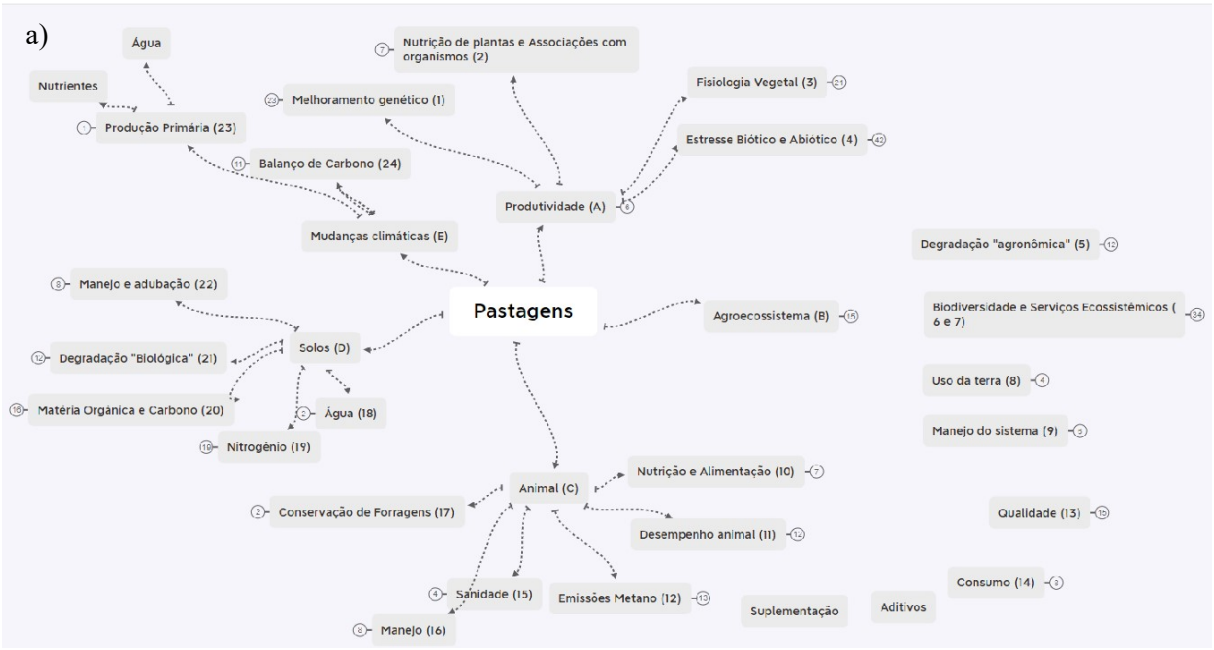
grafos do tipo "*Network Visualization*" com o auxílio do *software* VOS *Viewer* (Figura 4). A partir dos grafos, a equipe construiu um mapa mental do domínio "pastagens" e identificou termos relacionados ao tema "degradação de pastagens" (Figura 5).

Figura 4. Grafo do tipo "*Overlay Visualization*", gerado pelo *software* VOS *Viewer* a partir de documento científicos do domínio "pastagens" recuperados na *Web of Science* em 2016. Os resultados foram obtidos pela equipe do Projeto Infopasto.



Elaborado por: equipe do Projeto Infopasto.

Figura 5. Mapa mental do domínio “pastagens” construído pela equipe do Projeto Infopasto e identificação de termos relacionados ao tema “degradação de pastagens”. a) Visão geral do mapa mental; b) Detalhe do mapa mental mostrando termos relacionados à degradação “agronômica” e “biológica”.



Elaborado por: equipe do Projeto Infopasto.

A expressão de busca utilizada foi construída a partir de termos relacionados ao subdomínio “pastagens degradadas”, identificados pela equipe do Projeto Infopasto. 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)

A expressão de busca foi aplicada à *Web of Science Core Collection*. Apenas documentos publicados entre 1973 e 2022 foram recuperados, de forma a compatibilizar os resultados com outras atividades do Projeto Infopasto que recuperaram informações tecnológicas relacionadas ao tema desde a criação da Embrapa, em 1973.

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 os filtros, foram recuperados os metadados de 862 registros.

Os documentos 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 dos 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.

Uma análise preliminar dos metadados mostra que foram recuperados, entre outros, registros relacionados à restauração da vegetação natural, às práticas de plantio direto e rotação de culturas agrícolas, que fogem do escopo do trabalho. Os documentos apresentam termos semelhantes aos encontrados em textos sobre "degradação de pastagens", porém aplicados em contexto distinto (Tabela 3). As buscas na *Web of Science* são feitas por meio de palavras chaves com o auxílio de operadores boleanos, técnica que não permite a separação de artigos em função do contexto de utilização dos termos ou de aspectos semânticos.

Tabela 3. Exemplos de artigos recuperados em busca feita na *Web of Science* por termos relacionados ao tema "degradação de pastagens" que fogem ao escopo do trabalho. Os termos em negrito são semelhantes aos encontrados em artigos sobre o tema "degradação de pastagens". Descrição das referências e Resumo em inglês.

Referência e título	Resumo
Bustamante et al. (2019). Ecological **restoration** as a strategy for mitigating and adapting to climate change: lessons and challenges from Brazil.	Climate change is a global phenomenon that affects biophysical systems and human well-being. The Paris Agreement of the United Nations Framework Convention on Climate Change entered into force in 2016 with the objective of strengthening the global response to climate change by keeping global temperature rise this century well below 2 degrees C above pre-industrial levels and to pursue efforts to limit the temperature increase even further to 1.5 degrees C. The agreement requires all Parties to submit their nationally determined contributions (NDCs) and to strengthen these efforts in the years ahead. Reducing carbon emissions from deforestation and forest **degradation** is an important strategy for mitigating climate change, particularly in developing countries with large forests. Extensive tropical forest loss and **degradation** have increased awareness at the international level of the need to undertake large-scale ecological **restoration**, highlighting the need to identify cases in which **restoration** strategies can contribute to mitigation and adaptation. Here we consider Brazil as a case study to evaluate the benefits and challenges of implementing large-scale **restoration** programs in developing countries. The Brazilian NDC included the target of **restoring** and reforesting 12 million hectares of forests for multiple uses by 2030. **Restoration** of native vegetation is one of the foundations of sustainable rural development in Brazil and should consider multiple purposes, from biodiversity and ecosystem services conservation to social and economic development. However, ecological **restoration** still presents substantial challenges for tropical and mega-diverse countries, including the need to develop plans that are technically and financially feasible, as well as public policies and monitoring instruments that can assess effectiveness. The planning, execution, and monitoring of **restoration** efforts strongly depend on the context and the diagnosis of the area with respect to reference ecosystems (e.g., forests, savannas, **grasslands**, wetlands). In addition,

	poor integration of climate change policies at the national and subnational levels and with other sectorial policies constrains the large-scale implementation of **restoration** programs. The case of Brazil shows that slowing deforestation is possible; however, this analysis highlights the need for increased national commitment and international support for actions that require large-scale transformations of the forest sector regarding ecosystem **restoration** efforts. Scaling up the ambitions and actions of the Paris Agreement implies the need for a global framework that recognizes landscape **restoration** as a cost-effective nature-based solution and that supports countries in addressing their remaining needs, challenges, and barriers.
Bamberg et al. (2009). Bulk density of an Alfisol under cultivation systems in a long-term experiment evaluated with Gamma Ray computed tomography.	The sustainability of irrigated rice (Oryza sativa L.) in lowland soils is based on the use of crop rotation and succession, which are essential for the control of red and black rice. The effects on the soil properties deserve studies, particularly on **soil compaction**. The objective of this study was to identify **compacted** layers in an Albaqualf under different cultivation and tillage systems, by evaluating the soil bulk density (Ds) with Gamma Ray Computed Tomography (TC). The analysis was carried out in a long-term experiment, from 1985 to 2004, at an experimental station of Embrapa Clima Temperado, Capao do Leao, RS, Brazil, in a random block design with seven treatments, with four replications (T1-one year rice with conventional tillage followed by two years fallow; T2-continuous rice under conventional tillage; T4-rice and soybean (Glycine Max L.) rotation under conventional tillage; T5-rice, soybean and corn (Zea maize L.) rotation under conventional tillage; T6-rice under no-tillage in the summer in succession to **rye-grass (Lolium multiflorum L.)** in the winter; T7-rice under no-tillage and soybean under conventional tillage rotation; T8-control: uncultivated soil). The Gamma Ray Computed Tomography method did not identify compacted soil layers under no-tillage rice in succession to **rye-grass**; two fallow years in the irrigated rice production system did not prevent the formation of a compacted layer at the soil surface; and in the rice, soybean and corn rotation under conventional tillage two compacted layers were identified (0.0 to 1.5 cm and 11 to 14 cm), indicating that they may restrict the agricultural production in this cultivation system on

	Albaqualf soils.
Campoe et al. (2014). Atlantic forest tree species responses to silvicultural practices in a **degraded pasture** restoration plantation: From leaf physiology to survival and initial growth.	Deforestation has led to ecosystem **degradation** in many tropical regions. Re-establishment of native tree species on **degraded** land presents challenges due to environmental stressors such as water and nutrient limitations, particularly from weed competition. Ecophysiological studies can help assess responses of native tree species to silvicultural practices and improve our understanding of processes that influence their establishment and growth. Silvicultural treatments borrowed from commercial tree plantations such as greater nutrient applications and complete **weed** control can improve best silvicultural practices in forest restoration. Two contrasting silvicultural treatments, traditional based on common management practices for reforestation of native trees and intensive based on commercial plantation silviculture, were evaluated based on tree mortality, biomass, photosynthesis, chlorophyll content, soluble proteins, and nutritional status of 20 native Brazilian species, 2.5 years after planting. Intensive silviculture increased tree survival by 20%, showed higher aboveground biomass from 13% to 7-fold and increased photosynthesis of similar to 20% from 15.8 mu mol m(-2) s(-1) to 18.7 mu mol m(-2) s(-1), compared to traditional silviculture. Total soluble proteins were 14% higher with 6.7 mu g cm(-2) in intensive silviculture compared to 5.9 mu g cm(-2) under traditional silviculture. Eighty percent of trees showed greater N content, with a 13% higher average than under traditional silviculture (2.60 g m(-2) versus 2.92 g m(-2)). Average values of chlorophyll A, B, and total were similar to 8% higher under intensive silviculture, but not significantly different between treatments. Overall, intensive silviculture provided a positive impact on the **restoration** plantation. During the initial years of plantation establishment, intensive silviculture methods were effective in leading to significant increases in growth and survival.

A partir da análise preliminar da base de dados textuais, o primeiro desafio identificado foi a seleção de um método para classificação de artigos de interesse, que depois possam ser analisados para extração de conhecimento sobre recomendações para recuperação de pastagens em função das condições nas quais o problema se apresenta.

3.2 Pré-processamento dos artigos

A limpeza dos dados dos campos título e resumo da matriz de metadados foi realizada 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, onde os vetores de atributos são gerados pelo método conhecido como *Bag of Words* (BoW) (Tan et al., 2006) e 2) redes bipartidas (Rossi et al., 2016). BoW é uma matriz de documento-termo, 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 uma sequência de termos (n-gramas).

Já uma rede heterogênea bipartida representa os documentos em um grafo, consistindo em 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.

3.3 Análise de dados e extração de conhecimento

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). Desta forma, para atender ao objetivo no processo de classificação serão aplicadas duas abordagens, sendo uma delas supervisionada, utilizando os métodos: i) Máquinas de Vetores de Suporte (SVM) (CORTES & VAPNIK, 1995); ii) Redes *Multi-Layer Perceptron* (MLP) (HAYKIN, 1994) e iii) BERT. A outra abordagem é transdutiva, utilizando redes heterogêneas, o que permite explorar outras características dos artigos que não são usadas pelos métodos da primeira abordagem. Para isso foi utilizado um algoritmo de propagação de rótulos, 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 & YING, 2022; MA & GUO, 2014; LORENA & 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 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 *encoder* do *Trasformer*, 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 possui 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 NETO et al., 2020; JI et al., 2010), permitindo que a rede altere os rótulos dos vértices rotulados se o grau de confiança for baixo.

3.4 Análise de resultados

Diferentes métricas podem ser analisadas para mensurar o desempenho dos modelos e identificar suas possíveis limitações na tarefa de classificação dos artigos e, no caso da seleção de artigos sobre "degradação de pastagens", avaliar os falsos negativos produzidos pelos modelos é relevante no sentido de mitigar o descarte de artigos de interesse. Assim, os resultados serão avaliados utilizando as métricas F1-Macro, Acurácia e a Precisão. Além disso, para análise estatística dos modelos explorados será construído um diagrama de diferença critica (DC) utilizando teste não paramétrico de Friedman com teste *post-hoc* de Nemenyi (DEMSAR, 2006).

4 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.

4.1 Tratamento dos dados

Com exceção do BERT, que não exige pré-processamento, 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 transformá-los em uma representação no modelo espaço vetorial, nesse caso a *Bag of Words*. Nesse pré-processamento foram realizados os seguintes passos:

- O texto foi convertido para caixa baixa (minúsculo);
- Foi feita a remoção das *stopwords*, juntamente das *stopwords* 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).

A frequência de ocorrência dos termos nos textos foi escolhida como abordagem para construção da BoW, após alguns experimentos preliminares. 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 foram gerados *n*-gramas, que são sequência de *n* palavras que ocorrem nos textos. Neste experimento foi utilizado n=4, ou seja, foram geradas colunas contendo de 1 até 4 palavras. Para reduzir a esparsidade foi necessário filtrar 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% e foram selecionadas as 700 colunas com a maior frequência no conjunto de dados.

4.2 Avaliação

Para permitir a avaliação comparativa, foi utilizada uma validação cruzada com 5 *folds*, sendo que a BoW foi construída utilizando somente os documentos dos 4 *folds* referentes ao conjunto de treinamento.

No caso do algoritmo GNetMine é construída uma rede bipartida composta por dois conjuntos de vértices: os documentos e os termos selecionados na BoW. As arestas que ligam os documentos e os termos da rede correspondem às frequências que os termos ocorrem naquele documento. Por ser uma rede bipartida, não existe conexão entre termos nem entre documentos.

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. A única diferença corresponde ao fato de os pesos das arestas serem 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 é

somada a frequência de todos os termos que ele possui e por fim cada termo é dividido por esse valor. A seguir são descritos os hiper-parâmetros utilizados em cada um dos modelos:

- SVM[1]: 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].
- MLP[2]: foi treinada uma rede com 2 camadas ocultas com tamanhos respectivos 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 sigmoide. A rede foi treinada por 50 épocas com tamanho do *batch* de 8. Os hiper-parâmetros foram escolhidos empiricamente.
- BERT[3]: foi utilizada a versão pré-treinada "BERT-base-uncased" e o refinamento dos pesos foi feito com 10 épocas, escolhida empiricamente.
- GNetMine[4]: o único hiper-parâmetro utilizado foi μ = 1, pois como a rede só possui um tipo de relação, no caso entre documentos e termos, então não é necessário definir a importância de que cada tipo de relação. A implementação utilizada foi da biblioteca GraphTLP.

Como foram comparados dois tipos de abordagem (supervisionada e transdutiva), foi necessário identificar uma alternativa que permitisse a avaliação dos modelos de forma justa. Para comparar as duas abordagens de classificação (supervisionada e transdutiva) havia três possibilidades:

- Treinar todos eles com poucos dados rotulados, o que tende a prejudicar os modelos supervisionados que dependem de mais dados rotulados.
- 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, o GNetMine seria penalizado em relação aos demais algoritmos.

[1] Foi utilizada a biblioteca *Scikit-Learn*.
[2] Foi utilizada a biblioteca *Keras* (Chollet et al., 2015).
[3] Foi utilizada a biblioteca *HuggingFace* (https://huggingface.co/bert-base-uncased).
[4] Foi utilizada a biblioteca *GraphTLP* (https://github.com/BrucceNeves/GraphTLP).

- Utilizar a mesma estrutura de treino e teste dos modelos supervisionados, porém no caso do GNetMine, utilizar somente 1 dos *folds* do treinamento como rótulos a serem propagados pela rede.

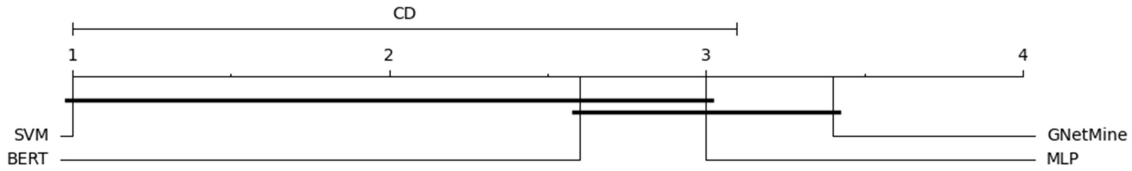
Dentre as formas de avaliação descritas, a que parece permitir uma melhor comparação entre as duas abordagens é a última. Para a avaliação do GNetMine, foi realizado um processo um pouco diferente daquele utilizado para os algoritmos supervisionados, visto que dependendo do *fold* selecionado para treinamento é possível obter diferentes resultados. Assim, 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*.

Todos os modelos foram capazes de refinar o resultado da pesquisa, separando melhor os artigos de interesse (Tabela 4). 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 7 está ilustrado um diagrama de diferença critica (DC) utilizando Teste não paramétrico de Friedman com teste *post-hoc* de (DEMSAR, 2006). O diagrama foi construído com base nos rankings médios de Precisão, em cada um dos *folds*, sendo que não existe diferença estatística entre os modelos avaliados quando eles são conectados por uma linha horizontal.

Tabela 4. 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

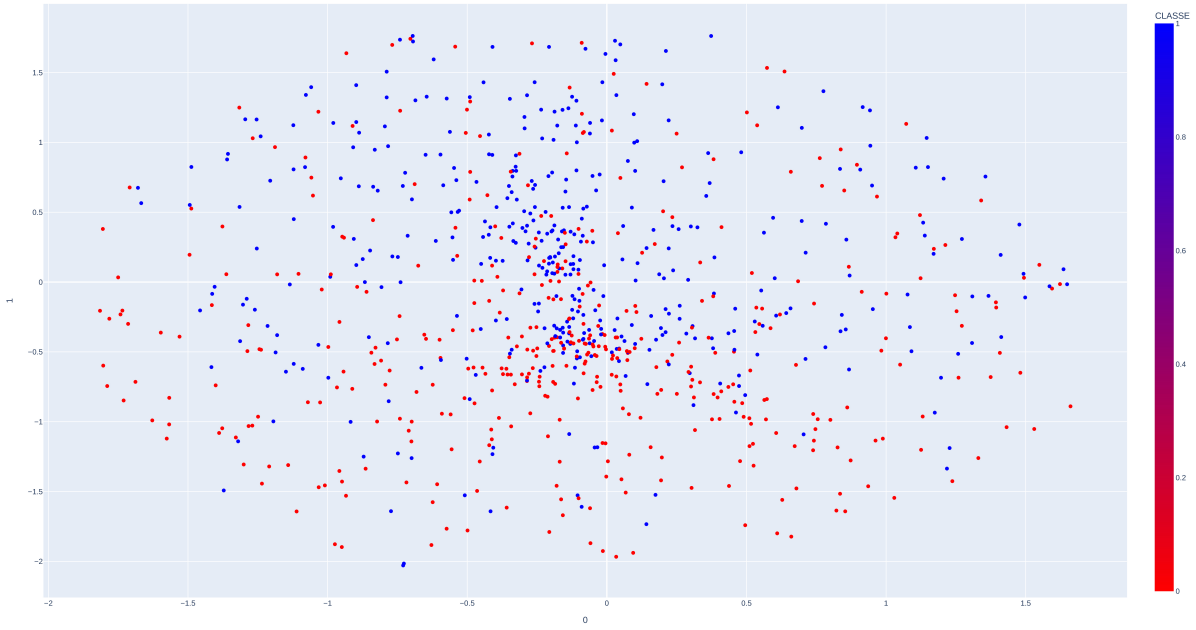
Figura 7. Diagrama de diferença crítica (DC) usando teste não paramétrico de Friedman com teste *post-hoc* de Nemenyi entre cada um dos modelos, construído com base nos rankings médios de precisão.



Como é possível observar na Tabela 4, o BERT e o MLP tiveram desempenhos similares em todas as 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 8, onde foi apresentada a distribuição dos artigos em um espaço 2D, obtida *t-Distributed Stochastic Neihbor Embedding* (t-SNE) (VAN DER MAATEN & HILTON, 2008).

O domínio “pastagens” e o subdomínio “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 realizados 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.

Figura 8. Distribuição dos artigos em um espaço 2D utilizando o *t-Distributed Stochastic Neihbor Embedding* (t-SNE).



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). Além disso, as diferenças na métrica de Precisão, que nesse trabalho é a mais relevante, são mínimas entre o GNetMine, BERT e MLP, mesmo tendo uma grande diferença entre a quantidade de dados rotulados. Isso comprova a robustez do GNetMine, que pode ser considerado como uma opção para rotulação de artigos, uma vez que utilizou 1/4 dos dados rotulados se comparado com os outros métodos avaliados e mesmo assim teve um desempenho comparativo muito bom.

4.3 Oportunidades futuras

Para investigar as oportunidades futuras de aplicação dos resultados deste trabalho, os artigos de interesse foram agrupados com o auxílio de uma rede K-NN, considerando os três vizinhos mais próximos e a métrica de cosseno. Na Tabela 5 podem-se observar os 20 primeiros grupos e seus descritores.

Tabela 5. Grupos de artigos relacionados ao tema “degradação de pastagens”, número de artigos por grupo e descritores.

Grupo	Número de artigos	Descritores
7	28	*soil*, *system*, *crop*, *pastur*, *use*, *manag*, *physic*, *integr*, *properti*, *differ*, *year*, *studi*, *eros*, *evalu*
6	21	*n*, *fertil*, *pastur*, *nitrogen*, *rate*, *kg*, *grass*, *forag*, *ha*, *plant*, *urea*, *system*, *product*, *use*
19	15	*system*, *soil*, *c*, *carbon*, *integr*, *organ*, *stock*, *agricultur*, *som*, *increas*, *organ matter*, *nativ*, *matter*, *evalu*
4	13	*soil*, *c*, *fraction*, *organ*, *n*, *matter*, *organ matter*, *cm*, *pastur*, *area*, *year*, *soil organ*, *stock*, *depth*
34	13	*system*, *tree*, *product*, *plant*, *integr*, *speci*, *increas*, *integr system*, *year*, *biomass*, *pastur*, *area*, *growth*, *studi*
8	12	*system*, *product*, *use*, *cattl*, *pastur*, *livestock*, *brazil*, *increas*, *shade*, *manag*, *intensif*, *soil*, *product system*, *improv*
23	12	*emiss*, *n2o*, *system*, *n2o emiss*, *soil*, *n*, *integr*, *ch4*, *flux*, *kg*, *silvopastor*, *season*, *silvopastor system*, *cattl*
16	11	*land*, *use*, *forest*, *pastur*, *increas*, *land use*, *agricultur*, *amazon*, *product*, *system*, *soil*, *ecosystem*, *flow*, *sustain*
10	11	*speci*, *pastur*, *tree*, *nativ*, *forest*, *rich*, *plant*, *forag*, *system*, *establish*, *use*, *nt*, *densiti*, *field*
12	11	*n*, *ha*, *kg*, *kg ha*, *soil*, *pastur*, *fertil*, *effect*, *dose*, *yield*, *forag*, *evalu*, *k*, *applic*
17	11	*pastur*, *soil*, *degrad*, *forest*, *system*, *studi*, *year*, *show*, *use*, *cattl*, *manag*, *area*, *grass*, *region*
27	9	*c*, *pastur*, *soil*, *ha*, *stock*, *mg*, *manag*, *forest*, *year*, *soil c*, *c stock*, *studi*, *mg ha*, *carbon*
1	9	*soc*, *soil*, *stock*, *carbon*, *system*, *pastur*, *use*, *soc stock*, *manag*, *studi*, *organ*, *organ carbon*, *chang*, *soil organ*
32	9	*soil*, *organ*, *system*, *carbon*, *matter*, *organ matter*, *soil organ*, *manag*, *organ carbon*, *manag system*, *fraction*, *differ*, *stock*, *evalu*
0	8	*soil*, *water*, *loss*, *crop*, *soil loss*, *use*, *system*, *conserv*, *soil water*, *runoff*, *eros*, *cover*, *total*, *concentr*
29	8	*soil*, *organ*, *carbon*, *organ carbon*, *pastur*, *forest*, *fraction*, *area*, *studi*, *organ*

		matter, *matter*, *differ*, *use*, *physic*
9	8	*area*, *pastur*, *nativ*, *cerrado*, *microaggreg*, *studi*, *veget*, *nativ veget*, *soil*, *sampl*, *depth*, *grass*, *ferralsol*, *show*
5	8	*soil*, *organ*, *matter*, *organ matter*, *chemic*, *crop*, *soil organ*, *soil organ matter*, *nutrient*, *increas*, *addit*, *acid*, *product*, *reduc*
3	8	*system*, *soil*, *pastur*, *forest*, *agroforestri*, *agroforestri system*, *crop*, *qualiti*, *evalu*, *use*, *differ*, *one*, *result*, *indic*
18	7	*soil*, *system*, *physic*, *qualiti*, *studi*, *area*, *agroforestri*, *integr*, *graze*, *total*, *agroforestri system*, *attribut*, *manag*, *physic quality*

A análise dos descritores dos grupos sugere que os artigos selecionados podem estar relacionados ao diagnóstico e caracterização do processo de degradação (por exemplo, o grupo 16), aos seus impactos (por exemplo, grupos 12 e 13) ou às estratégias de recuperação (por exemplo, grupo 3, 7 e 21), sendo que nem todos devem conter informações relevantes para a definição de recomendações para recuperação de pastagens, mesmo que de forma indireta. A análise de trechos extraídos documentos completos, relacionados aos objetivos, aos tratamentos testados e às conclusões, poderia ajudar a refinar a seleção de artigos com potencial para oferecer informações relevantes para a definição de recomendações para recuperação de pastagens. Para isso, seria necessário recuperar os artigos completos na internet e selecionar os trechos relacionados a estas partes dos documentos para aplicação das técnicas de classificação.

Após a seleção de artigos de interesse, outras técnicas de Inteligência Artificial poderiam ser utilizadas para automatizar a Extração de Conhecimento a partir dos textos científicos. A escolha das estratégias de intervenção para recuperação de pastagens deve ser baseada em aspectos como: local (região, bioma, estado, município), solo, clima, sistema de produção, tipo de capim, grau de degradação e causas de degradação. Apesar do contexto ao qual determinado conhecimento deve ser aplicado nem sempre ficar explícito nas publicações científicas, informações contidas no corpo do texto, principalmente na seção “material e métodos” podem contribuir para essa definição. Na descrição dos experimentos, normalmente são colocadas informações referentes ao local em que ele foi realizado, ao tipo de solo, clima e capim. Em algumas situações, também são inseridas informações sobre as causas e o grau de degradação da área. Na Tabela 6, foram transcritos alguns trechos de artigos que trazem informações sobre o contexto no qual os resultados foram gerados.

Tabela 6. Trechos de artigos recuperados em busca feita na *Web of Science*, por termos relacionados ao tema "degradação de pastagens" no Brasil, que contêm indicações sobre o contexto no qual os resultados foram gerados. Os trechos com informações de interesse foram realçados em negrito. Descrição das referências em inglês. Trechos extraídos em inglês ou português, conforme o artigo original.

Referência e título	Trechos
Assis et al. (2018). Identification of stylo lines with potential to compose mixed pastures with higher productivity.	The study was performed in two phases **in the Brazilian state of Acre, located in the Amazon region.** The first phase evaluated the agronomic characteristics of the lines of ***S. guianensis*** under a system of clippings in the **Experimental Field of Embrapa Acre, in the municipality of Rio Branco, in the state of Acre (10°01′34″S, 67°42′13″W, Datum WGS 84) and 160 m of altitude.** The second phase evaluated the most promising lines under grazing by Nelore beef cattle at the **Fazenda Manjerona, in the municipality of Senador Guiomard, Acre (09°37′57″S, 67°17′17″W, Datum WGS 84) at 170 m of altitude**. **The climate in the region is classified according to Köppen (1900) as Am (hot and humid), with average maximum and minimum temperatures of 31.6°C and 21.3°C, respectively, average annual precipitation of 2,015 mm (between 1997 and 2015), relative humidity of approximately 85%, a rainy period between October and May and a dry period between June and September (Figure 1)**. The agronomic evaluation was performed on a **Red Latosol and the evaluation under grazing on a Yellow Latosol**… A previous essay of **resistance to anthracnose** was performed for these lines, and all were highly resistant…. After selection cuttings, the five selected lines were seeded in **consortium with *Brachiaria brizantha* cultivar Xaraés**….
Costa et al. (2008). Nitrogen doses and sources in marandu pasture. I - Changes in soil chemical properties.	O experimento foi conduzido de julho de 2003 a março de 2006 na **Fazenda Modelo do curso de Zootecnia da Universidade Estadual de Goiás, em São Luís de Montes Belos-GO, a 579 m de altitude, 16 ° 31 ' 30 "de latitude sul e 50 22 ' 20 " de longitude oeste**… A pastagem já se encontrava **estabelecida há mais de 10 anos, com**

	estádio moderado de degradação, apresentando pouca cobertura do solo, com baixa produção de forragem, devido à exploração intensiva de animais e à falta de reposição de nutrientes no solo... O solo foi classificado como **Latossolo Vermelho distrófico (Embrapa, 2006) de textura argilosa**… **A adubação nitrogenada, em cada ano, foi parcelada em três épocas após cada corte de avaliação da forrageira: a primeira aplicação foi realizada em dezembro, a segunda em janeiro e a terceira em fevereiro, todas com intervalo de 30 dias.** **Durante os três anos do experimento, foram realizados três cortes por ano do capim-marandu, 30 dias após a aplicação do N**...

Técnicas de reconhecimento de entidades nomeadas poderiam ser aplicadas para a identificação do contexto no qual os experimentos foram realizados. Além das bibliotecas existentes para a identificação de informações sobre local, tabelas já disponíveis com informações sobre classes de solo, clima e tipo de capim poderiam utilizadas na tarefa. Para exemplificar, na Tabela 7 são apresentadas as classes de solos existentes no Brasil, segundo a classificação do Sistema Brasileiro de Classificação de Solos (SiBCS), e sua equivalência em relação a sistemas utilizados anteriormente.

Tabela 7. Classes de solo existentes no Brasil, segundo a classificação do Sistema Brasileiro de Classificação de Solos (SiBCS), e sua equivalência em relação a sistemas utilizados anteriormente.

SiBCS	WRB/FAO	*Soil Taxonomy*	Classificação anteriormente utilizada pela Embrap Solos
Argissolos	Acrisols	Ultisols	Rubrozéns, Podzólicos Bruno-Acinzentados Distróficos ou Álicos, Podzólicos Vermelho-Amarelos Distróficos ou Álicos Ta, Podzólicos vermelho-Amarelos Tb, pequena parte de Terra Roxa Estruturada, de Terra Roxa Estruturada Similar, de Terra Bruna Estruturada e de Terra Bruna Estruturada Similar com gradiente textural necessário para B textural, em qualquer caso Eutróficas, Distróficas ou Álicas, e mais recentemente Podzólicos Vermelho-Escuros Tb com B textural e Podzólicos Amarelos.
	Lixisols	Oxisols (Kandic)	
	Alisols		
Cambissolos	Cambisols	Inceptisols	Cambissolos Eutróficos, Distróficos e Álicos Ta e Tb, exceto os Cambissolos Eutróficos com horizonte A chernozêmico e com argila de atividade alta.
Chernossolos	Chernozems	---	Rendzinas, Brunizéns, Brunizéns Avermelhados e Brunizéns Hidromórficos.
	Kastanozems	Molisols (apenas os Ta)	
	Phaeozems	---	
Espodossolos	Podzols	Spodosols	Podzol, inclusive Podzol Hidromórfico.
Gleissolos	Gleysols	Entisols (Aquents), Alfisols (Aqualfs) e Inceptisols (Aquepts)	Glei Pouco Húmicos, Glei Húmicos, parte dos Hidromórficos Cinzentos (sem mudança textural abrupta), Glei Tiomórficos e Solonchaks com horizonte glei.

Gleissolos Sálicos	Solonchaks	Aridisols e Entisols	
Latossolos	Ferralsols	Oxisols	Latossolos, excetuadas algumas modalidades anteriormente identificadas como Latossolos Plínticos.
Luvissolos	Luvisols	Alfisols, Aridisols (Argids)	Brunos Não Cálcicos, Podzólicos Vermelho-Amarelos Eutróficos Ta, Podzólicos Bruno-Acinzentados Eutróficos e Podzólicos Vermelho-Escuros Eutróficos Ta.
Neossolos	---	Entisols	Litossolos, Solos Litólicos, Regossolos, Solos Aluviais e Areias Quartzosas (Distróficas, Marinhas e Hidromórficas).
Neossolos Flúvicos	Fluvisols	Entisols (Fluvents)	
Neossolos Litólicos	Leptsols	Entisols (Lithic... Orthents; Lithic... Psamments)	
Neossolos quartzarênicos	Arenosols	Entisols (Quartzipsamments)	
Neossolos Regolíticos	Regosols	Entisols (Psamments e Orthents)	
Nitossolos	Nitisols	Ultisols, Oxisols (Kandic), Alfisols	Terra Roxa Estruturada, Terra Roxa Estruturada Similar, Terra Bruna Estruturada, Terra Bruna Estruturada Similar, alguns Podzólicos Vermelho-Escuros Tb e alguns Podzólicos Vermelho-Amarelos Tb.
	Lixisols		
	Alisols		
Organossolos	Histosols	Histosols	

Planossolos	Planosols	Alfisols	Planossolos, Solonetz Solodizados e Planossolos Hidromórficos Cinzentos com mudança textural abrupta.
Planossolos Nátricos	Solonetz	Alfisols (Natrustalfs e Natrudalfs)	
Planossolos Háplicos	Planosols	Ultisols (Albaquults e Plintaquults) e Alfisols (Albaqualfs e Plintaqualfs)	
Plintossolos	Plinthosols	Alfisols (Plintaqualfs), Ultisols (Plintaquults) e subgrupos Plinthic de várias classes de Oxisols, Ultisols, Alfisols, Entisols e Inceptsols	Lateritas Hidromórficas, parte dos Podzólicos Plínticos, parte dos solos Glei Húmicos e dos Glei Pouco Húmicos Plínticos e alguns dos possíveis Latossolos Plínticos.
Vertissolos	Vertisols	Vertisols	Vertissolos, inclusive os hidromórficos.
Não classificaados no Brasil	Cryosols	Gelisols	
	Anthrosols	---	
	Andosols	Andisols	
	Umbrisols	---	
	Gypsisols	Vários subgrupos de	
		Aridisols	
	Durisols	Vários grandes	

		grupos Dura de Alfisols, Andisols, Aridisols, Inceptisols, etc.	
	Calcisols	Vários subgrupos de Vertisols, Molisols, Inceptisols, Alfisols, etc.	
	Albeluvisols	Alfisols (Glossaqualfs, Glossocryalfs, Glossudalfs, etc.)	

Fonte: Santos et al. (2018).

5 CONSIDERAÇÕES FINAIS

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.

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 documentos 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 1/4 dos dados rotulados que foi utilizado pelos modelos supervisionados, ficando 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 foi classificada 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, a fim 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 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 aos 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.

Os resultados deste trabalho foram aceitos para publicação no 20º Encontro Nacional de Inteligência Artificial e Computacional, a ser realizado em Belo Horizonte, MG. A referência bibliográfica do trabalho é:

OSAKU, D.; SANTOS, P.M.; SANTOS, B.N.; REZENDE, S.O. Pasture degradation papers search: how can supervised and transductive methods help on the process of classification? In Encontro Nacional de Inteligência Artificial e Computacional, 2023. Belo Horizonte, Brasil. (no prelo)

REFERÊNCIAS

AGARWAL, R., SRIKANT, R., et al. Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference, volume 487, page 499. 1994.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v.5, n.2, 2006.

ASSIS, G. M. L. de; BEBER, P. M.; MIQUELONI, D.P.; SIMEAO, R.M. Identification of stylo lines with potential to compose mixed pastures with higher productivity. **Grass and Forage Science**, v.73, n.4, p.897-906, 2018.

BALBINO, L. C.; BARCELLOS, A. de O.; STONE, L. F. (ed.). **Marco referencial.** Integração lavoura-pecuária-floresta. Brasília, DF: Embrapa, 2011. 130 p. Disponível em : http://ainfo.cnptia.embrapa.br/digital/bitstream/item/103901/1/balbino-01.pdf. Acesso em: 2 março de 2023.

BAMBERG, Y.A.L.; PAULETTO, E.A.; SILVA, T.R. Bulk density of an alfisol under cultivation systems in a long term experiment evaluated with gamma ray computed tomography . **Revista Brasileira de Ciência do Solo**, v.33, n.5, p.1079-1086, 2009.

BARBOSA, R.A. (ed) **Morte de pastos de braquiárias**. Campo Grande, MS: Embrapa Gado de Corte, 2006.

BUSTAMANTE, M.M.C.; SILVA, J.S.; NOBRE, C.E. Ecological restoration as a strategy for mitigating and adaptating to climate change: lessons and challenges from Brazil. **Mitigation and Adaptation of Global Change**, v.24, n.7, p.1249-1279, 2019.

CAMPOE, O.C.; IANNELLI, C.; STAPE, J.L.; COOK, R.L.; MENDES, J.C.T.; VIVIAN, R. Atlantic forest tree species responses to silvicultural practices in degraded pasture restoration plantation: from leaf physiology to survival and initial growth. **Forest Ecology and Management**, v.313, p.233-242, 2014.

CARVALHO, M. B.; TSUNODA, D. F. Data analysis on articles retrieved from web of science (wos). **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v.23, n.esp., p.112-125, 2018.

CECAGNO, D.; GOMES, M.V.; COSTA, S.E.V.G.D.; MARTINS, A.P.; DENARDIN, L.G.D.; BAYER, C.; ANGHINONI, I.; CARVALHO, P.C.F. Soil organic carbono in an integrated crop-livestock system under different grazing intensities. **Revista Brasileira de Ciências Agrárias**, v.13, n.2, 2018.

CERVANTES, J.; GARCIA-LAMONT, F.; RODRIGUES-MAZAHUA, L.; LOPEZ, A. A comprehensive survey on support vector machine classification: applications, challenges and trends. **Neurocomputing**, v.408, p.189-215, 2020.

CHOLLET, F. et al. **Keras**. https://github.com/fchollet/keras. 2015. Acessado em 20 de junho de 2023.

CLARIVATE. **Web of Science**. 2022. Disponível em: https://www.webofscience.com/wos/. Acesso em: 2 de março de 2023.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, V.20, n.3, p.273‑297, 1995.

COSTA, K.A.P.; FAQUIM, V.; OLIVEIRA, I.P.; RODRIGUES, C.; SEVERIANO, E.C. Doses e fontes de nitrogênio em pastagens de capim-marandu. I- Alterações nas características químicas do solo. **Revista Brasileira de Ciência do Solo**, v.32, n.4, p.1591-1599, 2008.

DE MORAES, L. L.; KAFURE, I. 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**, v.18, n.e020016‑e020016, 2020.

DE MORAES, M. V. B.. Comparação bibliográfica sobre ensino de matemática para pessoas com transtorno autista utilizando técnica de mineracção de texto. **REMAT: Revista Eletrônica da Matemática**, v.8, n.1, p.e2002, 2022.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v.7, p.1‑30, 2006.

DENG, X; LI, Y.; WENG, J. ZHANG, J. Feature selection for text classification: a review. **Multimed Tools Applied**, v.78, p.3797-3816, 2019.

DEVLIN, J., CHANG, M.W., LEE, K., and TOUTANOVA, K. 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. 2019.

DIAS-FILHO, M.B. **Degradação de pastagens**: o que é e como evitar. Brasília, DF: Embra-pa, 2017.

DIAS-FILHO, M.B. **Degradação de pastagens: processos, causas e estratégias de recuperação**. 4ed. Belém, PA: Ed. Do Autor, 2011. 215p.

DIAS-FILHO, M.B. Opções forrageiras para áreas sujeitas ao encharcamento ou alagamento temporário. Belém: Embrapa Amazônia Oriental, 2006. 34p. (Embrapa Amazônia Oriental. Documentos, 239).

DING, S. 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. 2009.

DOS SANTOS NETO, B.N.; ROSSI, R.G.; REZENDE, S.O. MARCACINI, R.M. A two stage regularization framework for heterogeneous event networks. **Patter Recognition Letters**, v.138, p.490-496, 2020.

EBECKEN, N.F.F.; LOPES, M.C.S.; COSTA, M.C.A. Mineração de textos. In: Rezende, S.O.. **Sistemas Inteligentes**: fundamentos e aplicações. Barueri, SPS: Manole, 2003. p.337-370.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. **Sistema Gerencial Ideare**. Brasília, Embrapa, 2022. Disponível em: https://sistemas.sede.embrapa.br/ideare. Acesso em: 15 de fevereiro de 2022.

FAYAAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in databases. **AI Magazine**, v.17, n.2, p.37-53,1996.

GONZÁLEZ-CARVAJAL, S.; GARRIDO-MERCHÁN, E.C. Comparing BERT Against traditional machine learning text classification. Journal of Computational and Cognitive Engineering, 2021.

HAYKIN, S. **Neural networks: a comprehensive foundation**. Prentice Hall PTR. 1994.
Instituto Brasileiro de Geografia e Estatística – IBGE. **Censo Agropecuário**: resultados definitivos 2017. Rio de Janeiro: IBGE, 2019. Disponível em: IBGE | Biblioteca | Detalhes | Censo agropecuário : resultados definitivos 2017. Acesso em: 15 de janeiro de 2023.

JI, M.; SUN, Y.; DANILEVSKY, M.; HAN, J.; GAO, J. Graph regularized transductive classification on heterogeneous information networks. In: Balcázar, J.L. (Eds). ECML PKDD 2010, Part 1, ÇNAI, 6321. P. 570-586. 2010.

KLUTHCOUSKI, J.; STONE, L.F.; AIDAR, H. (eds) **Integração Lavoura-Pecuária**. Santo Antônio de Goiás: Embrapa Arroz e Feijão, 2003.

KOCH, K.R. Bayesian inference with geodetic applications, volume 31. Springer, Germany. 2006.

LIMIRO, R. M.; DA SILVA, N. R.; CORDEIRO, D. F. Mineração de textos para agrupamento de teses e dissertações por meio de análise de similaridade. **Revista Brasileira de Biblioteconomia e Documentação**, v.18, p.1‑20, 2022.

LORENA, A. C.; DE CARVALHO, A. C. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v.14, n.2, p.43‑67, 2007.

MA, Y.; GUO, G. Support vector machines applications, volume 649. Springer. 2014.

MANZATTO, C.V.; PEREIRA, S.E.M.; PEDREIRA, B.C. Zoneamento do risco de ocorrência da síndrome da morte do capim-marandu no Estado do Mato Grosso, 2018. 29 p. -- (Boletim de Pesquisa e Desenvolvimento/Embrapa Meio Ambiente, 74).

MAPA. 2022. Plano ABC – Agricultura de Baixo Carbono. Diponível em: https://www.gov.br/agricultura/pt-br/assuntos/sustentabilidade/plano-abc/plano-abc-agricultura-de-baixa-emissao-de-carbono. Acesso em: 13 de fevereiro de 2023.

MAPBIOMAS. **Projeto MapBiomas** – Coleção [7] da Série Anual de Mapas de Uso e Cobertura da Terra do Brasil. Disponível em: https://plataforma.brasil.mapbiomas.org/pastagem?activeBaseMap=6&layersOpacity=100&a

ctiveModule=quality_of_pasture_data&activeModuleContent=quality_of_pasture_data%3Aquality_of_pasture_data_main&activeYear=2000%2C2021&mapPosition=-15.072124%2C-51.416016%2C4&timelineLimitsRange=2000%2C2021&activeLayers=estados&baseParams[territoryType]=1&baseParams[territories]=1%3BBrasil%3B1%3BPa%C3%ADs%3B-33.751177993999946%3B-73.9904499689999%3B5.271841077000019%3B-28.847639913999956&baseParams[activeClassTreeOptionValue]=quality_of_pasture_main&baseParams[activeClassTreeNodeIds]=79%2C80%2C81&baseParams[activeSubmodule]=quality_of_pasture_data_main&baseParams[activeClassesLevelsListItems]=1%2C7%2C8%2C9%2C10%2C2%2C11%2C12%2C13%2C14%2C15%2C16%2C3%2C17%2C18%2C27%2C37%2C38%2C39%2C40%2C41%2C28%2C42%2C43%2C44%2C19%2C20%2C4%2C21%2C22%2C23%2C24%2C5%2C25%2C26%2C6. Acesso em: 08 de fevereiro de 2022

MARCHI, S.R.de; MARQUES, R.F.; ARAÚJO, P.P.S.; SILVA, I.T.D.; MARTINS, D. Weed interference in Marandu palisadegrass pastures under renewal or maintenance conditions. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.26, n.3, p.166-172, 2022.

PORTER, M. F. An algorithm for suffix stripping. **Program**, v.14, n.3, p.130‑137, 1980.

REZENDE, S.O. Introdução. In: Rezende, S.O. **Sistemas Inteligentes: fundamentos e aplicações**. Barueri, SPS: Manole, 2003. P.3-11.

REZENDE, S.O.; PUGLIESI, J.B.; MELANDA, E.A.; PAULA, M.F.de. Mineração de Dados. In: Rezende, S.O. **Sistemas Inteligentes**: fundamentos e aplicações. Barueri, SP: Manole, 2003. P.307-335.

RIBEIRO, N.G.; SILVA, I.V.da; ARAÚJO, C.F.de; FAGUNDES, O.D.; GERVAZIO, W. **Iheringia Serie Botanica**, v.72, p.127-132, 2017.

ROCHA, P.R. da; AANDRADE, F.V.; MENDONÇA, E.D.; DONAGEMMA, G.K.; FERNANDES, R.B.A.; BHATTHARAI, R.; KALITA, P.K. Soil, water, and nutriente losses from management alternatives for degraded pastures in Brazillian Atlantic Rainforest biome. **Science of the Total Environment**, v. 583, p. 53-63, 2017.

RODRIGUES, W. G. Predição de diâmetros e cálculo de volume de clones de eucalipto: uma abordagem com redes multi layer perceptron e long-short term memory. Dissertação (Mestrado). Universidade Federal de Goiás. 2019

ROSSI, R. G.; DE ANDRADE LOPES, A.; REZENDE, S. O. Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts**. Information Processing & Management**, V.52, n.2, p.217‑257, 2016.

ROSSI, R.G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. 2015. Tese (Doutorado em Ciência da Computação e Matemática Computacional). Instituto de Cinências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2015.

SANTOS, H. G. dos; JACOMINE, P. K. T.; ANJOS, L. H. C. dos; OLIVEIRA, V. A. de; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A. de; ARAUJO FILHO, J. C.

de; OLIVEIRA, J. B. de; CUNHA, T. J. F. **Sistema Brasileiro de Classificação de Solos**. 5 ed., rev. e ampl. Brasília, DF: Embrapa, 2018. 356p.

SANTOS, P.M. et al. Grass and forage research indexed by the Web of Science from 2005 to 2015. São Carlos: Embrapa Pecuária Sudeste, 2021. 25p. (Embrapa Pecuária Sudeste. Documentos, 132).

SARTO, M.V.M.; BORGES, W.L.B.; SARTO, J.R.W.; RICE, W.C.; ROSOLEN, C. Deep soil carbon stock, origin, and root interaction in a tropical integrated crop–livestock system. **Agroforest Systems**, v.94, p.1865–1877, 2020.

SEMOLINI, R. **Support Vector Machines, Inferência Transdutiva e o Problema de Classificação**. Tese (Doutorado). Universidade Estadual de Campinas. 2002.

SINOARA, R.A.; MARCACINI, R.M.; REZENDE, S.O. Mineração de textos e semântica: desafios, abordagens e aplicações. **Revista Brasileira de Informação da FSMA**, n.27, p.41-53, 2021.

SOBRAL, N. V.; LIMA, G. L. d. Q.; SOBRAL, A. S. P. d. M. (2021). Produção científica sobre hospitais no contexto da ciência de dados: um estudo a partir da web of science. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**. V.26, n. esp. P.01-16, 2021.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Harlow, Essex: Pearson Education Limited, 2014. 732p.

TELLES, M.A. **Da produção do conhecimento científico à transferência de informações: análise da circulação de saberes no âmbito de duas redes de pesquisa agropecuária**. 2016. Tese (Doutorado em Ciência da Informação). Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

TELLES, M.A. et al. **Glossário ILPF**: Integração Lavoura-Pecuária-Floresta. Colombo: Embrapa Florestas, 2021.85p. (Embrapa Florestas. Documentos, 350)

VAN DER MAATEN, L.; HINTON, G.. Visualizing data using t-sne. **Journal of machine learning research**, v.9, n.86, p.2579-2605. 2008.

World Resources Institute, WRI. **Creating a Sustainable Food Future.** A menu of solutions to feed nearly 10 billion people by 2050. 2019. Disponível em: WRR_Food_Full_Report_0.pdf (wri.org). Acesso em: 27 de fevereiro de 2022.