Processamento de Linguagem Natural para consultas de invasores na cultura da soja

Carolinne Roque e Faria[1], Cinthyan Renata Sachs C. de Barbosa[1]

[1]Programa de Pós Graduação em Ciência da Computação –
Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – 86.057-970 – Londrina – PR – Brazil

carolinne.rf@outlook.com, cinthyan@uel.br

Resumo. *As tecnologias para auxiliar na produtividade do campo estão em constantes desenvolvimentos com a expectativa de estarem em sintonia com as demandas da área. Assim, a proposta deste trabalho é apresentar ferramentas computacionais em Processamento de Linguagem Natural para auxiliar os produtores rurais no controle das principais pragas e doenças na cultura da soja para não prejudicar a produtividade e, a partir da extração de suas características que serão armazenados em um repositório, seja possível identificar os vilões que atacam a lavoura ajudando-os nas tomadas de decisões de acordo com as consultas feitas a essa Base de Dados.*

Abstract. *The technologies to assist in productivity in the field are in constant development with the expectation of being in harmony with the demands of the area. Thus, the proposal of this paper is to present computational tools in Natural Language Processing to assist farmers in the control of the main pests and diseases in soybean crops in order to avoid affecting productivity and, from the extraction of their characteristics that will be stored in a repository, it is possible to identify the villains that attack the crop, helping them in making decisions according to the queries made to this database.*

1. Introdução

No Brasil, a modernização do campo caminha em grandes passos com o intuito de aumentar a produção e melhorar as condições econômicas por meio de investimentos em inovações tecnológicas, pesquisas científicas e créditos rurais. A agricultura é uma das principais bases da economia de diversos países, principalmente no Brasil [Hiraoka e Jacopini 2018] que nesse setor tem seu potencial reconhecido globalmente. O Centro de Estudos Avançados em Economia Aplicada [CEPEA 2019] da ESALQ da USP, em

parceria com a Confederação da Agricultura e Pecuária do Brasil e com a Fundação de Estudos Agrários Luiz de Queiroz, calcularam o Produto Interno Bruto (PIB) do Agronegócio Brasileiro em 2019 e o desempenho representou significativamente 21,4% do PIB brasileiro total, o que revela uma ampliação da sua participação na economia.

A cultura da soja representa um dos principais propulsores do agronegócio brasileiro transformando a vida de milhões de cidadãos com a geração de emprego e renda, melhorias de infraestrutura e aumento do IDH *(Índice de Desenvolvimento Humano)* nas regiões produtoras [EMBRAPA 2019]. A CONAB *(Companhia Nacional de Abastecimento)* divulgou no mês de junho o 9º levantamento de grãos para a safra 2018/2019 no mercado nacional, que relata o aumento de produção de soja como vimos.

Descrição/Safra	2016/17	2017/18	2018/19 (*)
Estoque Inicial	5.405,4	7.779,8	1.390,7
Produção	114.075,3	119.281,7	114.843,3
Importação	253,7	187,0	150,0
Suprimento	119.734,4	127.248,5	116.384,0
Consumo total	43.800,0	42.600,0	44.000,0
Exportação	68.154,6	83.257,8	68.000,0
Estoque Final	7.779,8	1.390,7	3.184,0

Figura 1. Soja em grãos. Fonte: [CONAB 2019]

Devido às ciências e à produção de conhecimento, os sistemas avançados são fundamentados em práticas agrícolas. Ter uma ferramenta que auxilie o produtor sobre as principais pragas e doenças na cultura, bem como garantir a sanidade das plantas significa melhor produção e, desse modo, essa pode lidar melhor com as necessidades da lavoura. O gerenciamento de dados agrícolas depende de informações adquiridas de diversas tecnologias e sistemas que sejam capazes de auxiliar na tomada de decisão nas técnicas agrícolas. Assim, é fundamental mapear os dados (principais características de identificação de pragas e doenças na cultura da soja) e padronizá-los.

Indurkhya e Damerau (2010) salientam que o Processamento de Linguagem Natural (PLN) tem como objetivo extrair representações e significados mais completos de textos livres escritos em linguagem natural. PLN responde de forma inteligente por meio de algoritmos em uma linguagem não só escrita, mas também falada.

Assim, este presente trabalho faz uso de uma grande quantidade de informações armazenadas e disponibilizadas em rede aberta e aplica técnicas de PLN em um dicionário para consultas de domínio agrícola para auxiliar o agricultor a melhorar a sua produção e evitar prejuízos. O artigo está dividido da seguinte maneira: a seção 2 descreve os materiais e métodos, na seção 3 são expostos os resultados e discussões e na seção 4 as conclusões.

2. Materiais e Métodos

O PLN desempenha um papel imprescindível na comunicação, pois suas técnicas refletem na compreensão da Linguagem Natural (LN), passando por Análise Léxico-Morfológica *(tokenização)*, Análises Sintática *(parsing)*, Semântica e Pragmática.

Para a construção deste trabalho são utilizadas algumas ferramentas auxiliares necessárias para o desenvolvimento do software em PLN para agricultura que serão descritas ao decorrer do artigo. Um chatbot será apresentado para conversar com estudantes de agronomia, agronegócio, engenheiros, produtores rurais ou pessoas que

gostariam de adquirir conhecimentos para evitar o ataque das principais pragas, insetos, nematoides, doenças e fungos na cultura da soja. A interface desse é desenvolvida em LN, o que facilita a comunicação entre o sistema e o usuário. Para permitir esse tipo de comunicação, os chatbots podem utilizar recursos de PLN [Khanna et al. 2015]. Alguns chatbots foram encontrados para uso na agricultura como, por exemplo, os trabalhos de Sawant et al. (2019) e Mostaço et al. (2018), porém nenhum deles são sobre a soja.

Para a elaboração do chatbot em PLN para consulta em Banco de Dados foi feito um sistema de interpretação e geração em linguagem natural. Alguns recursos são essenciais durante a fase de interpretação, bem como na de geração, e foram baseados em Silva et al. (2007) para este trabalho envolvendo: **Léxico:** conjunto de palavras que quando acessado por analisadores léxico, sintático e semântico reconhece os *tokens* e fornece além dos traços gramaticais, os traços semânticos de suas entradas para possibilitar a verificação semântica. **Gramática:** conjunto de regras gramaticais que definem as frases válidas em uma sentença em linguagem natural. **Modelo de Domínio:** fornece informações sobre o domínio específico da aplicação. **Modelo de Usuário:** permite reconhecer características do significado textual a partir do contexto do discurso

Após isso foi possível estabelecer os inúmeros modos de escrever, fazendo com que o sistema manuseie automaticamente a geração da tarefa e que o processo dessa transmita informações continuamente em uma base de dados em PLN.

Para este trabalho contou-se com as mais diversas fontes, tais como manuais e livros [Santos 1995] [Moreira e Aragão 2009] [Sosa-Gomez 2010] [Ávila 2017], teses/ dissertações, *sites* como da *Empresa Brasileira de Pesquisas Agropecuária* [EMBRAPA 2019], artigos e profissionais da área para selecionar as palavras para a criação do dicionário, com a finalidade de fornecer informações, permitindo que a mesma pergunta seja feita de várias maneiras, relacionada às características que se manifestam durante a produção de soja.

As palavras armazenadas em um dicionário totalizam 108 ameaças que são associadas às informações gramaticais. O registro de cada palavra foi feito de acordo com algumas categorias morfológicas, as quais são avaliadas isoladamente e separadas em classes gramaticais pelo processo de *Part of Speech* (POS) *Tagging*, isto é, para cada palavra de um texto é feita a identificação da classe gramatical a que ela pertence baseada em sua definição e no contexto, visto que uma palavra pode pertencer a mais de uma classe ou possuir mais de um significado [Moraes 2019].

Este trabalho gera uma estrutura de árvore que represente a estrutura sintática da sentença analisada e foi desenvolvido por meio da linguagem de programação Python e o analisador sintático escolhido foi o spaCy (2020) que utiliza-se de uma biblioteca de PLN, deixando-o com um desempenho superior a outras bibliotecas da área. Assim, foi possível fazer o processo de POS *Tagging* (classificar todas as sentenças do texto por categorias gramaticais e a relação de dependência entre palavras) com tal biblioteca.

A ferramenta spaCy (2020) classifica cada palavra do texto conforme às suas classes gramaticais como artigo, conjunção, numeral, pronome, símbolos, substantivo, verbo etc., com o intuito de encontrar as características específicas sobre um determinado assunto. Ela é uma biblioteca para análise sintática e permite fazer o NER (*Named-entity recognition* - Reconhecimento de Entidades Nomeadas) que tem a função

de identificar e classificar palavras ou frases em um texto de acordo com classes definidas para o modelo [Nadeau 2007].

Neste trabalho optou-se por um Banco de Dados Não Relacional, também conhecido como Banco de Dados (BD) NoSQL (*Not Only Structured Query Language*). Esses são uma família de gerenciadores de dados que não seguem o modelo de dados relacional. O NoSQL escolhido é o tipo Grafo. Como é mostrada na Figura 2 tem-se a estruturação dos dados armazenados em nós dos Ácaros/Fitófagos e associação às informações por meio das arestas do Ácaro-Rajado, Ácaro-Branco e Ácaro-Verde.

Figura 2. Dados dos Ácaros Fitófagos no Banco de Dados NoSQL Neo4j

O objetivo é criar uma interface computacional em LN para garantir um bom desempenho das tarefas na área da agricultura, como uma ferramenta de conversação entre o sistema e o usuário que possibilita um fácil acesso às informações em um repositório da base de dados e visa aumentar esse conforme ocorre o diálogo e, então, os dados crescem exponencialmente. Assim, o BD armazena não só os dados, mas as suas relações de maneira eficiente.

Mitchell (1997) define Aprendizado de Máquina (AM) como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência. Técnicas de AM são essenciais para calcular o desempenho dos dados para o futuro e foram aplicadas a fim de detectar pragas automaticamente a partir de textos. Com o grande volume de dados propõe-se treinar os modelos de classificação para avaliar o desempenho de doenças da soja.

A fase de pré-processamento de dados pode ser considerada a mais importante para a aplicação das tarefas de AM, pois são feitas, segundo Bird, Klein and Loper (2009): - remoção de *stopwords* que são palavras como artigos ou verbos de ligação que aparecem nos textos várias vezes, mas praticamente não influenciam a classificação; - lematização que é a redução de palavras a seus radicais, removendo flexões de tempo verbal, gênero, número; - tokenização que é o processo de criação de um vetor de termos de um documento, onde cada termo ocupa um índice do vetor. Nesse contexto aplicam-se técnicas para extração e classificação de textos na identificação de pragas e doenças a partir das características da praga na planta por meio de AM, a fim de analisar em menor tempo e avaliar o grau de severidade do dano na lavoura.

Os resultados obtidos a partir da predição para representar o comportamento real da identificação é predizer o desempenho do modelo para o futuro e a principal fonte é calculada pela matriz de confusão [Olson and Delen 2008]. Ao realizar a aplicação do

modelo faz-se a medição da precisão para calcular a taxa de acerto (razão entre os valores previstos e valores reais) e erro. As predições podem ser Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN). A partir dela apresentam-se os valores previstos e reais por meio de forma tabular e baseada nessa contagem, a qual várias métricas podem ser calculadas.

Aqui foram aplicadas as principais métricas em problemas de classificação [Abonizio et al. 2019]: **(a)** acurácia que informa a taxa de acerto; **(b)** a métrica de precisão é uma equação que é medida a proporção dos valores previstos negativos com os verdadeiros negativos, ou seja, essa apresenta quão corretas estão as predições; **(c)** a revocação que é conhecida como sensibilidade, a qual calcula a proporção dos valores previstos positivos com os verdadeiros positivos e retorna a fração dos documentos que foram relevantes; **(d)** métrica F1-*score* que calcula a média da precisão e revocação.

Para validar o modelo de AM é utilizada a técnica de Árvore de Decisão, cuja fase de treinamento é feita por meio de um sistema de indução de árvores baseada na divisão recursiva [Steinberg and Colla 1995]. Após a construção da árvore pode-se classificar novos exemplos a partir dessa [Silva e Vieira 2007]. A classificação é feita percorrendo a árvore até chegar à folha que determina a classe na qual o exemplo pertence ou sua probabilidade de pertencer àquela classe.

No presente trabalho foi aplicado o aprendizado supervisionado para fornecer ao modelo os dados que já possuem significados e serão categorizados, ou seja, o algoritmo irá conferir a precisão do modelo criado utilizando o conjunto de dados para predizer a eficácia. O método de análise de dados descrito automatiza o processo de criação de modelos. Utilizando algoritmos que iterativamente aprende com os dados, o AM permite que os computadores encontrem padrões escondidos nos dados sem terem sido programados para essa finalidade.

A principal dificuldade no desenvolvimento do sistema para identificação de pragas e doenças por meio de PLN é dialogar com o usuário e reconhecer as suas intenções a partir de uma frase e respondê-lo automaticamente. Assim, este sistema inteligente de pré-atendimento aos profissionais do campo agronômico tem o intuito de ser um canal alternativo de comunicação para facilitar o acesso às informações e auxiliar no ensino para identificar o patógeno. Para o pré-processamento da LN foram aplicados algoritmos de AM à base de dados para a construção de classificadores que pudessem predizer a causa dos sintomas das plantas da soja. A partir dos resultados é possível tomar decisões, o que permite otimizar as atividades guiadas à agricultura.

Para treinar as métricas, foram escolhidos os algoritmos de classificação Florestas aleatórias (*Random Forest* – RF) [Breiman 2001], Máquinas de Vetores Suporte (*Support Vector Machine* - SVM) [Scholkopf 2002], K-vizinhos mais próximos (*K-Nearest Neighbors* - KNN) [Weinberger, Blitzer and Saul 2006].

Os resultados do treino de classificação (Tabela 1) são compostos pela média das métricas de precisão, revocação e f1-*score* para cada praga que danifica a planta. As métricas apresentadas indicam que o modelo obteve uma performance alta, conseguindo uma média de F1-*score* de 94% e 97% de acurácia. Por meio do resultado experimental foi possível analisar o desempenho do classificador RF na identificação de 101 pragas que prejudicam a lavoura da soja e isso se mostrou robusto atingindo acurácia de 99%.

Tabela 1. Comparação entre os algoritmos

Algoritmos	Precision	Recall	F1-score	Accuracy
KNN	0.87	0.92	0.88	0.96
Random Forest	0.97	0.93	0.95	0.96
SVM	0.94	0.94	0.94	0.97

Algoritmos	Precision	Recall	F1-score	Accuracy
SVM - Kernel	1.00	0.98	0.99	0.99

Foi desenvolvida a Ferramenta CAROLINA (acrônimo para **C**onversação **A**gronômica **RO**botizada em **LI**nguagem **NA**tural) para identificação de pragas e doenças na cultura da soja e adotado o modelo em cascata para o desenvolvimento dessa que utiliza das fases de análises de requisitos, projeto, implementação, testes (validação), integração e manutenção de software. O sistema foi desenvolvido na linguagem de programação Python, com a utilização da biblioteca spaCy, para a criação de um chatbot, que elabora, por exemplo, as seguintes questões proporcionando ricas consultas, como essas: • *Os cupins são comuns em qual região?* • *Quais os prejuízos causados pela Lagarta-rosca?* • *O Ácaro-branco é o mesmo que o Ácaro-do-Bronzeamento?* • *O Piolho-de-cobra se alimenta do quê?* • *Tem uma larva atacando as folhas, flores e vagens. Que larva é?* • *O Cupim-subterrâneo ataca que região da planta da soja?* • *Minha cultura está demorando para se desenvolver. Por quê?* • *Minha planta está enfraquecendo. Por quê?* • *A minha planta teve queda das folhas durante o período de produção. Por quê?* • *As folhas estão amareladas. Por quê?* Esse sistema recebe o texto por meio de um diálogo (Figura 3) com o utilizador, analisa em um primeiro momento as palavras das sentenças isoladamente e passando pela compreensão da frase como um todo até concluir uma resposta da frase requisitada.

3. Resultados e Discussões

As etapas para o desenvolvimento do Sistema da Soja são compostas por: • definições das funcionalidades do sistema para permitir que o usuário controle as funções desse; • levantamentos das informações e técnicas necessárias para a construção das partes mais importantes do sistema para determinar a melhor maneira de realizar uma tarefa; • definições das ferramentas necessárias para a construção do sistema para o desenvolvimento de aprendizado de máquina; • construção do sistema seguindo uma metodologia de desenvolvimento de software para norteá-lo; • análises dos resultados obtidas com os testes para chegar às conclusões e verificar se o objetivo foi atingindo.

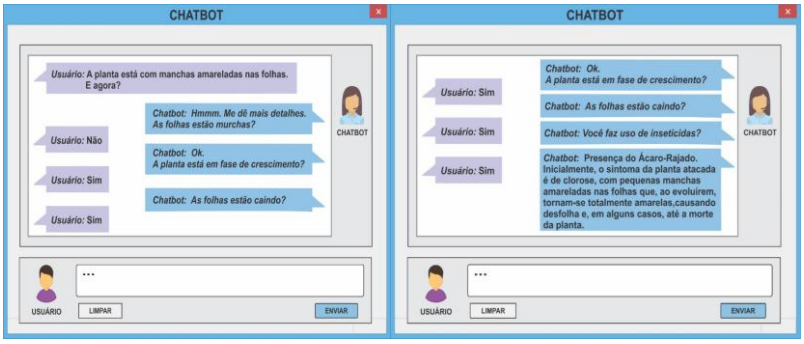

Figura 3. Interface de Consulta à Base de Dados das Pragas da Soja

A elaboração deste trabalho permitiu contribuir com a área agrícola, principalmente no que diz respeito à viabilidade de um assistente virtual, utilizando técnicas de PLN na identificação de pragas e doenças nas plantas da soja.

A partir das classificações das características das pragas e doenças da soja, um modelo foi construído por meio da biblioteca spaCy (2020), utilizando técnicas de PLN para aplicar regras gramaticais às sentenças, reconhecendo suas estruturas e extraindo seus significados, onde o pré-processamento envolveu as seguintes fases:
• **Tokenização:** subdivide a base de dados em tokens para uma pergunta como "*As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?*"
• **Lematização e Stemização**: a *lematização* é o processo de redução de palavras à sua base (raiz), enquanto a *stemização* permite checar se uma palavra é raiz de outra. São etapas importantes para ánalise de grandes volumes de textos.
• **POS *tagging*:** classifica corretamente todas as palavras de cada sentença do texto por categorias gramaticais.
• **Análise Léxico-Morfológica:** responsável por manipular o léxico que é composto por palavras que armazenam seus significados. Identificar as palavras individualmente é essencial porque ajuda a entender as frases de entrada e constrói com mais exatidão as saídas (respostas).
• **Analise Sintática:** responsável por organizar o conjunto das palavras e aplicar regras gramaticais à sentença para reconhecer sua estrutura; Posteriormente, essa ferramenta rotula todos os *tokens* a partir da análise para prever qual *tag* provavelmente se aplica nesse contexto. Para isso, conta-se com as seguintes tarefas: **Text** (texto puro), **Lemma** (reduz as palavras em seu formato base/raiz), **POS** (*tags* simples que determinam as categorias gramaticais de um *token*), **Dep** (Dependência sintática é a relação entre os *tokens* presentes dentro de uma sentença para entender o seu significado), **Shape** (classificação da palavra em maiúscula ou minúscula), **Alpha** (especificação das palavras em alfanuméricas ou não), **Stop** (indica se as palavras são consideradas *stopwords*), como é detalhada na Figura 4 para a frase: "*As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?"* O spaCy (2020) possibilita descrever a relação sintática das palavras que se conectam na formação da árvore. Isso permite percorrê-la toda, retornar uma sequência ordenada de *tokens* e verificar os atributos e domínios das palavras. Nessa fase conta-se com o Head Text (relação entre as palavras nos *tokens*), Head Pos (rotula as palavras em categorias) e Children (dependentes sintáticos do *token*).

```
for token in doc:
    print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_,
          token.shape_, token.is_alpha, token.is_stop)
```

```
As As DET <artd>|ART|F|P|@>N det Xx True True
folhas folhar NOUN <np-def>|N|F|P|@SUBJ> nsubj xxxx True False
atacadas atacar VERB <mv>|V|PCP|F|P|@ICL-N< acl xxxx True False
ficam ficar VERB <mv>|V|PR|3P|IND|@FS-STA ROOT xxxx True False
com com ADP PRP|@<ADVL case xxx True True
grandes grande ADJ ADJ|F|P|@>N amod xxxx True True
áreas área NOUN <np-idf>|N|F|P|@P< obl xxxx True False
recortadas recortar VERB <mv>|V|PCP|F|P|@ICL-N< acl xxxx True False
ou ou CCONJ <co-fcl>|<co-fmc>|<co-vfin>|KC|@CO cc xx True True
são ser VERB <cjt>|<mv>|V|PR|3P|IND|@FS-STA aux:pass xxx True True
completamente completamente ADV ADV|@ADVL> advmod xxxx True False
consumidas consumir VERB <pass>|<mv>|V|PCP|F|P|@ICL-AUX< conj xxxx True False
. . PUNCT PU|@PU punct . False False
```

Figura 4. Parsing

• **Reconhecimento de Entidades Nomeadas:** spaCy também permite fazer o NER e tem a função de identificar e classificar palavras ou frases em um texto, de acordo com classes definidas para o modelo [Nadeau 2007]. As principais atividades feitas pelo NER são identificar os *tokens* em um texto não estruturado e classificá-los em tipos de entidades definidas de acordo com a peculiaridade do domínio [Speck e Ngomo 2014].

Ao utilizar a função displaCy por meio da ferramenta spaCy é possível destacar visualmente as entidades e os tipos para a frase: *"No Brasil há registros da ocorrência de Percevejo-castanho em várias regiões, embora os danos dessa praga tenham sido mais frequentes nos estados de Mato Grosso, Goiás e Mato Grosso do Sul"*, como é apresentado na Figura 5.

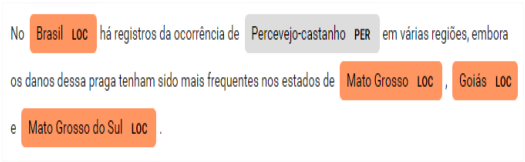

Figura 5.NER com displaCy

• ***Parsing* de dependências:** depois de classificar corretamente todas as palavras de cada sentença do texto por categorias gramaticais é feita a relação de dependência entre palavras.

4. Conclusões

A criação da Ferramenta CAROLINA teve o intuito de otimizar a análise dos dados das pragas e doenças na cultura da soja em um repositório e auxiliar o usuário na tomada de decisão por meio de uma agente conversacional, o que pode facilitar o trabalho dos profissionais da área que precisam se envolver com um amplo volume de informações na tomada de decisões. Pretende-se abranger futuramente outras culturas agrícolas na base de dados, visto que as mesmas pragas prejudicam várias plantações.

Um modelo foi criado utilizando o formato JSON *Schema* para extrair os textos do banco de dados Neo4j. Foi aplicada a função sintática em um grande volume de textos, responsável por organizar as estruturas gramaticais. Assim, foi possível a construção de um modelo por meio da ferramenta spaCy para aplicar regras gramaticais às sentenças e reconhecer as estruturas e extrair seus significados. Além disso, foi avaliado um conjunto de dados reais na identificação das doenças causadas pelas pragas na cultura da soja, a partir das extrações das características do BD. Foram escolhidos três algoritmos de classificação (RF, SVM, KNN) e dentre as validações aplicando as métricas de AM resultou em 99% de acurácia, o que é considerada uma alta taxa de acertos.

Destaca-se que para esse modelo seja realmente efetivo para auxiliar na identificação de pragas e doenças na sojicultura é necessário: analisar outras bases de dados, pois pode causar perda de eficiência treinar somente com uma base de dados estática e, ainda, para comprovar a eficácia do modelo com mais textos; melhorias quanto ao método de seleção de palavras de um documento para termos uma maior assertividade; e avaliar o desempenho de mais classificadores considerando o processamento e tempo de execução.

Ressalta-se ainda que o trabalho está diretamente relacionado aos Objetivos de Desenvolvimento Sustentável (ODS), pois a partir da construção de um sistema inteligente de pré-consulta para a identificação de pragas e doenças é possível promover a agricultura sustentável, o que evita o uso descontrolado de agroquímicos, traz oportunidade de aprendizagem sobre as pragas e doenças na planta analisada, melhora a produtividade e supera os desafios no meio agrícola, como mudanças climáticas, ausência de mão-de-obra, quantidade de terras agricultáveis disponíveis e custos.

Referências

Abonizio, H. Q., Barbosa, C. R. S. C. e Artoni, A. A. (2019) Detecção Automática dos Heterônimos de Fernando Pessoa por Aprendizado de Máquina. *XII Simpósio em Tecnologia da Informação e Linguagem Humana*, Salvador, SBC. p. 144-153

Ávila, C. J. (2017) "Pragas da soja e seu controle". https://pragas.cpao.embrapa.br/, julho.

Bird, S, Klein, E. and Loper, E (2009) *Natural Language Processing with Python.* 1st edition. Cambridge: O'Reilly Media Inc.

Breiman, L. (2001) Random forests. *Machine Learning*, Califórnia, Springer, v.45, n.1, Out. p. 5-32.

CEPEA (2019). "PIB-AGRO/CEPEA: PIB do agronegócio encerra 2019 com alta de 3,81%" https://www.cepea.esalq.usp.br/br/releases/pib-agro-cepea-pib-do-agronegocio-encerra-2019-com-alta-de-3-81.aspx, Dezembro.

EMBRAPA (2019). "Pragas da Soja". https://pragas.cpao.embrapa.br/ , Julho.

Hiraoka, E. e Jacopini, V. (2018) O papel da tecnologia na evolução da agricultura. *Sociedade Nacional de Agricultura*. http://www.sna.agr.br/o-papel-da-tecnologia-na-evolucao-da-agricultura, março.

Indurkhya, N. and Damerau, F. J. (2010) In: *Handbook of Natural Language Processing.* 2nd edition. Chapman & Hall/CRC.

Khanna, A. et al. (2015) A Study of Today's A.I. through Chatbots and Rediscovery of Machine Intelligence. In *International Journal of u-and e-Service, Science and Technology*, v.8, pages 277–284.

Mitchell, T. M. (1997) *Machine learning.* 37ª ed. Burr Ridge, IL: McGraw Hill.

Moraes, M. P. (2019) *Mineração de Dados Aplicada à Identificação de Notícias Falsas.* Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro, Rio de Janeiro. Trabalho de Conclusão de Curso.

Moreira, H. J. C. e Aragão, F. D. (2009) *Manual de Pragas da Soja.* Campinas: FMC Agricultural Products.

Mostaço, G. M., Campos, L. B., Souza, I. R. C. and Cugnasca, C. E. (2018). AgronoBot: a smart answering Chatbot applied to agricultural sensor networks. *14th International Conference on Precision Agriculture*, p.1-13.

Nadeau, D. (2007) *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision.* Institute for Computer Science de Ottawa-Carleton. Ottawa, Canada. Master Thesis.

Neo4J (2018) "Neo4j Basics". https://neo4j.com/ product/#basics, Julho.

Olson, D. L. and Delen, D. (2008) *Advanced data mining techniques.* Berlim: Springer Science & Business Media.

Santos, O. S. (1995) *A Cultura da Soja, 1. Rio Grande do Sul-Santa Catarina-Paraná.* 2ª edição. São Paulo: Globo.

Sawant, D., Jaiswal, A., Singh, J. and Shaw, P. (2019) AgriBot – An Intelligent interactive interface to assist farmers in agricultural activies. In *IEEE Bombay Section Signature Conference*, pages 1-6.

Silva, B. C. D., Montilha, G., Rino, L. H. M., Specia, L., Nunes, M. G. V., Oliveira Junior, O. N., Martins, R. T. e Pardo, T. A. S. (2007) *Introdução ao Processamento das Línguas Naturais e suas Aplicações.* Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional da Universidade de São Paulo. São Carlos.

Silva, C. F. e Vieira, R. (2007) Categorização de Textos da Língua Portuguesa com Árvores de Decisão, SVM e Informações Linguísticas. *V Workshop em Tecnologia da Informação e da Linguagem Humana*, Rio de Janeiro, SBC. p.1650-1658

Sosa-Gomez, D. R. et al. (2010) *Manual de identificação de insetos e outros invertebrados da cultura da soja*. Londrina, PR.: Embrapa Soja. Documentos 269.

Spacy (2020) https://spacy.io/ julho.

Speck, R. and Ngomo, A. C. N. (2014) Ensemble learning for named entity recognition. In *International semantic web conference*, v.8796, n.1, Out., pages 519–534. Switzerland, Springer.

Steinberg, D. and Colla, P. (1995) *CART: Tree-Strutured NonParametric Data Analysis.* San Diego, CA: Salford Systems.

Weinberger, K. Q., Blitzer, J. and Saul, L. K. (2006) Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems,* pages 1473-1480. MIT Press.