Processamento de Linguagem Natural (PLN): Detectando Discurso de Ódio e Ofensivo nas Falas dos Deputados Federais do Brasil

Alisson Franclin Barbosa de Oliveira
UNIFESP – Universidade Federal de São Paulo
Programa de Pós-Graduação em Ciências da Computação
São Jose dos Campos, SP, Brasil
alissonf216@gmail.com

***Resumo* — Este estudo explora a aplicação de técnicas avançadas de Processamento de Linguagem Natural (PLN) na identificação de discursos de ódio e ofensivos em falas de deputados federais brasileiros. Utilizou-se uma base de dados composta por discursos parlamentares, na qual foi implementado e comparado quatro algoritmos de aprendizado de máquina - Naive Bayes, SVM (Support Vector Machines), Random Forest e Logistic Regression - com vetorização via técnica de TF-IDF (Term Frequency-Inverse Document Frequency). Todos os algoritmos passaram por um ajuste dos hiperparâmentros e Cross validação usando a técnica de GridSearchCV com 10 folds, o modelo SVM foi o modelo que melhor performou, o que resultou um f1-score equilibrado para ambas as labels de 0.97. Este resultado evidencia a eficácia na classificação precisa de discursos ofensivos, contribuindo significativamente para a moderação e análise de conteúdo em contextos políticos. Este trabalho não só demonstra a aplicabilidade do PLN na análise de discursos políticos, mas também destaca a importância de técnicas de aprendizado de máquina na identificação de linguagem ofensiva e de ódio, contribuindo para um ambiente político mais saudável e respeitoso.**

Palavras-chaves: Processamento de linguagem natural, aprendizado de máquina, Discurso de ódio, classificação, Discurso Político.

I. INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) tem emergido como uma ferramenta revolucionária na análise e interpretação de grandes volumes de texto, desempenhando um papel crucial na identificação de padrões linguísticos e discursivos em diversas esferas da comunicação humana. No cenário político brasileiro, a aplicação do PLN na análise de discursos fornece insights valiosos sobre as tendências, preocupações e intenções dos atores políticos, contribuindo significativamente para a pesquisa em ciência política e comunicação social.

Discurso de ódio, definido como qualquer forma de expressão que incite, promova ou justifique o ódio, a discriminação ou a hostilidade contra um grupo ou indivíduo com base em características como raça, religião, etnia, gênero, idade ou orientação sexual, representa um desafio crescente na sociedade contemporânea. No Brasil, a complexidade do discurso de ódio se manifesta em múltiplas dimensões sociais e políticas, tornando sua detecção e análise através do PLN uma tarefa essencial.

A detecção automática de discurso de ódio e linguagem ofensiva, especialmente em plataformas digitais e discursos públicos, tornou-se um campo de interesse crescente. Estudos recentes, como o de Raj et al. (2021), demonstram a eficácia de modelos híbridos baseados em aprendizado de máquina e PLN na detecção de cyberbullying, ressaltando a relevância de redes neurais bidirecionais e modelos de atenção para classificação de alta precisão [1]. Além disso, a pesquisa de Shakil e Alam (2022) propõe uma metodologia que combina CNN e PLN, alcançando uma precisão notável na classificação de comentários maliciosos [2]. Esses avanços são fundamentais para entender como o PLN pode ser aplicado no contexto político brasileiro.

Este estudo objetiva utilizar técnicas de PLN para identificar discursos de ódio e ofensivos nas falas de deputados federais brasileiros. Empregando algoritmos de aprendizado de máquina, como Naive Bayes, SVM, Random Forest e Logistic Regression, e utilizando a técnica de vetorização TF-IDF, este trabalho busca identificar padrões de discurso ofensivo e comparar a eficácia dessas abordagens em um contexto político específico. A escolha desses algoritmos é fundamentada na literatura existente, que destaca sua aplicabilidade e eficiência em tarefas similares de classificação de texto [1,2].

Ao integrar o rigor metodológico do PLN com a análise de discurso político, esta pesquisa visa contribuir para a criação de um ambiente político mais respeitoso e para a promoção de uma maior compreensão da dinâmica da comunicação política no Brasil.

II. TRABALHOS RELACIONADOS

Diversos trabalhos acadêmicos têm explorado a detecção de discurso de ódio e linguagem ofensiva utilizando técnicas de Processamento de Linguagem Natural (PLN), aplicando diversos algoritmos de aprendizado de máquina. Esses estudos são fundamentais para entender as abordagens e metodologias mais eficazes nesse campo.

Um estudo [3] focou na identificação de discurso de ódio em mídias sociais, diferenciando-o de linguagem ofensiva. Intitulado "Hate Speech Detection by Classic Machine

Learning", os autores Tharwat El-Sayed, Abdallah Mustafa, Ayman El-Sayed e Mohamed Elrashidy treinaram vários modelos clássicos de aprendizado de máquina, incluindo regressão logística, árvores de decisão, random forest, naive Bayes, k-nearest neighbors e máquinas de vetores de suporte (SVM-SVC). Os algoritmos de regressão logística e SVM-SVC apresentaram bom desempenho na detecção de discurso de ódio e linguagem ofensiva, destacando a eficácia dessas técnicas em contextos de mídia social.

Outro estudo [4], "Natural Language Processing to Classify Caregiver Strategies Supporting Participation Among Children and Youth with Craniofacial Microsomia and Other Childhood-Onset Disabilities", conduzido por V. Kaelin, A. Boyd, M. Werler, Natalie Parde e M. Khetani, aplicou PLN para desenvolver um modelo preditivo que classifica estratégias de cuidadores. Neste estudo, foram utilizados recursos manualmente criados e três métodos clássicos: regressão logística, naive Bayes e máquinas de vetores de suporte (SVM). O SVM usando TF-IDF foi o modelo de melhor desempenho, demonstrando a versatilidade do PLN em diferentes contextos de aplicação.

Em "Automatic Hate Speech Detection using Ensemble Method and Natural Language Processing Techniques" [5], os autores N. V. Sahana, Prerana R, Niharika S, Rakshitha S, Bhanushree K J propuseram um modelo de conjunto para categorizar textos em ódio, ofensivo ou nenhum dos dois. O estudo considerou vários métodos de classificação, incluindo árvore de decisão, SVC, regressão logística, random forest e naive Bayes. O modelo de conjunto proposto, combinando random forest e support vector classifier, alcançou uma precisão de 90.7%, evidenciando a eficácia de abordagens de conjunto na detecção de discurso de ódio.

No trabalho "Combating Hate Speech on Q&A Forums with Machine Learning" [6], Yatendra Sahu, R. K. Gupta e S. Bharti implementaram várias estratégias de PLN e utilizaram métodos de aprendizado de máquina, incluindo regressão logística, máquinas de vetores de suporte (SVM), árvores de decisão e naive Bayes, para identificar discurso de ódio em fóruns de perguntas e respostas. O modelo de regressão logística com count vectorizer apresentou melhor desempenho, reforçando a importância de escolher o algoritmo certo para o contexto específico de aplicação.

Esses estudos fornecem uma base sólida para a compreensão das técnicas e abordagens atuais na detecção de discurso de ódio e linguagem ofensiva, utilizando PLN e algoritmos de aprendizado de máquina. Eles são fundamentais para o desenvolvimento de novas metodologias e aprimoramento de técnicas existentes no campo do PLN aplicado à análise de discurso político e social.

III. MÉTODO

Este estudo se propõe a aplicar e comparar técnicas e algoritmos sofisticadas de Processamento de Linguagem Natural (PLN) com o objetivo de identificar e analisar manifestações de discurso de ódio e linguagem ofensiva presentes nas falas dos deputados federais do Brasil. O método adotado é estruturado em uma série de etapas sequenciais e interconectadas, as quais são detalhadamente descritas a seguir:

A. Coleta e Preparação dos Dados

Foi realizada a coleta de dados através da API do portal Dados Abertos da Câmara Legislativa do Brasil. O conjunto de dados constituído abrange 1017 frases oriundas de discursos parlamentares. A seleção dos dados buscou abarcar uma ampla gama de temáticas e estilos discursivos, visando assegurar uma representatividade abrangente e diversificada.

Para que fosse possível utilizar as técnicas de aprendizado supervisionado, as frases dos discursos dos parlamentares fossem classificados manualmente, classificados como pertencentes a uma dos dois grupos discurso de ódio/ofensivo(1) ou discurso normal/neutro(0).

Tabela 1 – Descrição da base de dados

label	Contagem	Percentual
0	803	79%
1	214	21%

B. Análise Exploratória e Balanceamento dos Dados

Uma análise exploratória foi realizada para compreender a estrutura e a distribuição dos dados, incluindo a frequência de palavras e a distribuição das classes (ofensivo vs. não ofensivo). Identificou-se um desequilíbrio entre as classes, o que motivou a aplicação de técnicas de amostragem excessiva para equilibrar o número de instâncias em cada classe, mitigando potenciais vieses nos modelos de aprendizado de máquina.

C. Remoção de Stopwords

Muitos textos contêm palavras frequentes que, embora essenciais para a estrutura gramatical, oferecem pouco valor semântico. Essas palavras, conhecidas como 'stopwords', geralmente incluem preposições, artigos e pronomes. Elas são frequentemente removidas em processos de análise de texto, pois sua presença pode não contribuir significativamente para o entendimento do conteúdo semântico principal.

Figura 1 – Nuvem de palavras com stopwords em todo conjunto

Figura 2 – Nuvem de palavras sem stopwords em todo conjunto.

Figura 3 – Nuvem de palavras sem stopwords no conjunto da label 1.

D. Vetorização e Modelagem

Para a vetorização dos dados, adotou-se a técnica de TF-IDF (Term Frequency-Inverse Document Frequency), que converte o texto em um formato numérico, adequado para o processamento por algoritmos de aprendizado de máquina. Em seguida, implementaram-se quatro modelos distintos: Naive Bayes, SVM (Support Vector Machines), Random Forest e Logistic Regression. Cada modelo foi treinado e testado utilizando o conjunto de dados processado.

E. Avaliação dos Modelos

Para iniciar nossa avaliação, uma análise foi realizada através da matriz de confusão para cada modelo. A avaliação dos modelos de aprendizado de máquina implementados - Naive Bayes, SVM, Random Forest e Logistic Regression - foi conduzida utilizando métricas padrão de desempenho, incluindo precisão, recall e f1-score.

- Precisão:

  Esta métrica indica a proporção de identificações positivas que foram de fato corretas.

  $$Precision = \frac{Previsões\ Positivas\ Corretas}{Previsões\ Positivas}$$

- F1 Score:

  O f1-score é a média harmônica de precisão e recall, oferecendo um equilíbrio entre essas duas métricas.

  $$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- Recall:

  O recall mede a proporção de positivos reais que foram identificados corretamente.

  $$Recall = \frac{Previsões\ Positivas\ Corretas}{Exemplos\ Positivos}$$

- Acurácia:

  É o número de acertos (positivos) divido pelo número total de exemplos. Ela deve ser usada em dados com a mesma proporção de exemplos para cada classe, e quando as penalidades de acerto e erro para cada classe forem as mesmas.

  $$Acurácia = \frac{Previsões\ Corretas}{Total\ de\ Previsões}$$

- Matriz de Confusão:

  A análise da matriz de confusão permite uma compreensão mais profunda do comportamento do modelo em diferentes classes. Por exemplo, um número elevado de FN em um modelo pode indicar uma tendência à conservadorismo, evitando falsos positivos mas aumentando os falsos negativos.
  Esta análise é particularmente crucial em contextos como o da detecção de discurso de ódio, onde o equilíbrio entre minimizar falsos positivos (evitando censura indevida) e falsos negativos (não deixar passar discursos de ódio) é vital.
  A combinação dessas métricas e da análise da matriz de confusão proporciona uma avaliação abrangente e detalhada do desempenho dos modelos, permitindo identificar não apenas a eficácia geral, mas também as nuances específicas no comportamento de cada modelo.

F. Otimização de Hiperparâmetros

Todos os modelos foram submetidos a um rigoroso processo de otimização de hiperparâmetros utilizando o GridSearchCV. Além disso, empregamos a técnica de validação cruzada com 10 folds para garantir uma avaliação robusta de seu desempenho. Também ajustamos o valor do parâmetro n-gram durante o processo de vetorização na busca contínua por melhorias adicionais em seu desempenho.

Este método consiste em uma busca exaustiva sobre um conjunto especificado de valores de hiperparâmetros. O processo envolve os seguintes passos:

- N-gram:

  O modelo de n-gramas é particularmente útil para prever o próximo elemento em uma série de dados. Em análises textuais mais detalhadas, ele facilita a compreensão de padrões e tendências linguísticas, contribuindo significativamente para o processamento de linguagem natural e outras aplicações relacionadas à análise de texto.

- Definição do Espaço de Parâmetros:

Inicialmente, define-se um 'grid' de hiperparâmetros potenciais para cada modelo. Este grid inclui diferentes valores para parâmetros como C (parâmetro de regularização), kernel (tipo de função do kernel), e gamma (coeficiente do kernel para 'linear','rbf', 'poly' e 'sigmoid').

- Validação Cruzada:

  O GridSearchCV emprega uma técnica de validação cruzada para avaliar cada combinação de hiperparâmetros. Geralmente, utiliza-se a validação cruzada k-fold, dividindo o conjunto de dados em k subconjuntos e realizando k iterações de treinamento e validação, cada uma usando um subconjunto diferente como conjunto de validação.

- Avaliação de Desempenho:

  Em cada iteração de validação cruzada, o modelo é avaliado com base em uma métrica de desempenho predefinida, como o f1-score. O GridSearchCV registra o desempenho de cada combinação de hiperparâmetros.

- Seleção do Melhor Conjunto de Parâmetros:

  Após a execução de todas as combinações possíveis, o GridSearchCV seleciona o conjunto de hiperparâmetros que resultou no melhor desempenho de acordo com a métrica escolhida.

- Treinamento do Modelo Final:

  Finalmente, todos os modelo são treinados usando o melhor conjunto de hiperparâmetros encontrado, desta vez utilizando todo o conjunto de treinamento.
  A otimização de hiperparâmetros via GridSearchCV é fundamental para melhorar a precisão e a eficácia do modelo, garantindo que ele esteja bem ajustado às especificidades dos dados em análise.

IV. Resultados

Este estudo buscou aplicar e comparar técnicas e algoritmos de Processamento de Linguagem Natural (PLN) na detecção de discurso de ódio e linguagem ofensiva nas falas dos Deputados Federais do Brasil. Os dados foram coletados através da API do portal de dados abertos da Câmara Legislativa, resultando em um conjunto de dados de 1017 frases classificadas manualmente com base em definições estabelecidas de discurso de ódio e linguagem ofensiva.

A. Análise de Dados e Preparação

Inicialmente, procedemos com a preparação dos dados, que envolveu a remoção de ‘stopwords’ em português, seguindo as práticas padrão em PLN para reduzir o ruído e melhorar a relevância do conteúdo textual. Além disso, enfrentamos o desafio do desequilíbrio de classes no conjunto de dados, o qual abordamos através de uma técnica de amostragem excessiva, visando equilibrar a representatividade das classes.

B. Modelagem e Avaliação

Empregamos uma série de modelos de classificação para avaliar sua eficácia na identificação de discursos de ódio. Para garantir a robustez dos resultados, todos os modelos foram testados com uma divisão padrão de dados de treino e teste, seguindo a proporção de 70% para treino e 30% para teste. Essa abordagem assegura que cada modelo seja avaliado em um conjunto de dados independente, aumentando a confiabilidade dos resultados obtidos na classificação de discursos de ódio.

- Random Forest:

Figura 4 – Matriz de Confusão – Random Forest

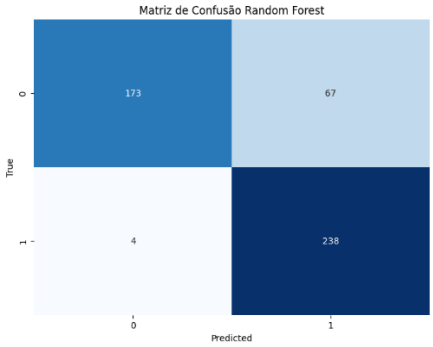

Tabela 2 – Métricas de Desempenho do Modelo

Classe	Precisão	Recall	F1-Score	Suporte
0	0.98	0.72	0.83	240
1	0.78	0.90	0.87	242
Méd/Acur.	0.88	0.85	0.85	482

- SVM (Support Vector Machine):

Figura 5 – Matriz de Confusão – SVM (Support Vector Machine)

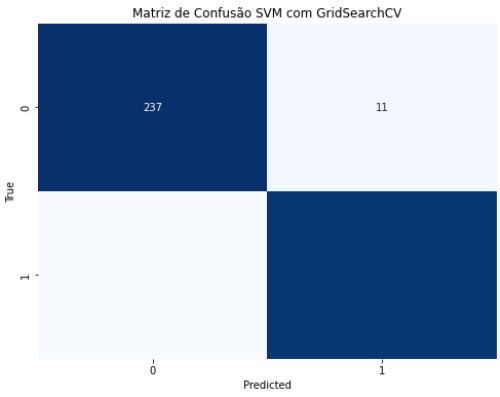

Tabela 3 – Métricas de Desempenho do Modelo

Classe	Precisão	Recall	F1-Score	Suporte
0	0.98	0.96	0.97	248
1	0.95	0.98	0.97	234
Méd/Acur.	0.97	0.97	0.97	482

- Regressão Logística:

Figura 6 – Matriz de Confusão – Regressão Logística

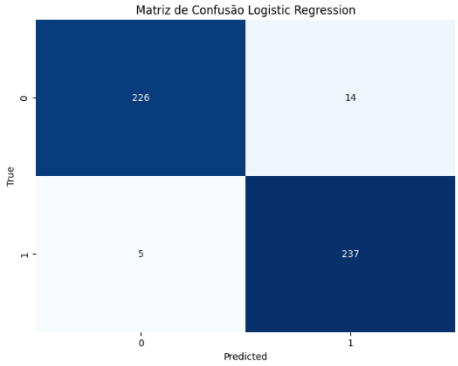

Tabela 4 – Métricas de Desempenho do Modelo

Classe	Precisão	Recall	F1-Score	Suporte
0	0.98	0.94	0.96	240
1	0.94	0.98	0.96	242
Méd/Acur.	0.96	0.96	0.96	482

- Naive Bayes:

Figura 7 – Matriz de Confusão – Naive Bayes

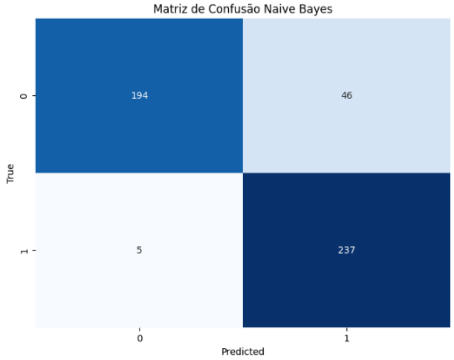

Tabela 5 – Métricas de Desempenho do Modelo

Classe	Precisão	Recall	F1-Score	Suporte
0	0.97	0.81	0.88	240
1	0.84	0.98	0.90	242
Méd/Acur.	0.91	0.89	0.89	482

Todos os modelos demonstraram um desempenho notável na classificação das falas, evidenciando a eficácia das técnicas de PLN na identificação de discursos de ódio. No entanto, o modelo SVM destacou-se particularmente em termos de precisão, recall e F1-Score.

Tabela 6 – Parâmetros otimizados: SVM

Parâmetro	Valor
C	10
gamma	scale
kernel	linear

Tabela 7 – Parâmetros otimizados: Naive Bayes

Parâmetro	Valor
alpha	100

Tabela 8 – Parâmetros otimizados: Random Forest

Parâmetro	Valor
max_depth	Nome
n_estimators	100

Tabela 8 – Parâmetros otimizados: Regressão logistica

Parâmetro	Valor
C	10

Descrição dos Parâmetros:

C: Parâmetro de regularização. Quanto maior o valor de C, menor a regularização. Um valor de 10 indica uma regularização moderada.

gamma: Define a influência de um único exemplo de treinamento. O valor 'scale' é calculado automaticamente com base no número de características do conjunto de dados.

kernel: Especifica o tipo de kernel a ser usado no algoritmo. 'linear' indica um kernel linear.

n_estimators: número de arvores na floresta.

alpha: suavização para evitar probabilidades nulas ou zero.

Legenda das Tabelas: 2,3,4,5,7:

Classe: Identificador da classe (0 ou 1).
Precisão: Proporção de identificações positivas corretas (quanto maior, melhor).
Recall: Proporção de casos positivos reais identificados corretamente (quanto maior, melhor).
F1-Score: Média harmônica de precisão e recall (quanto maior, melhor).
Suporte: Número de ocorrências reais de cada classe no conjunto de dados.

V. CONCLUSÃO

Este estudo demonstrou a viabilidade e eficácia do uso de técnicas de Processamento de Linguagem Natural (PLN) para identificar discursos de ódio e linguagem ofensiva no contexto político brasileiro. Comparando quatro algoritmos de aprendizado de máquina - Naive Bayes, SVM, Random Forest e Logistic Regression - com vetorização via TF-IDF, destacamos o SVM como o modelo mais eficiente. A otimização de hiperparâmetros, realizada através do GridSearchCV, reforçou a eficácia deste modelo, evidenciada por um f1-score equilibrado de 0.97 para categorias de discurso ofensivo e não ofensivo.

A relevância do PLN na moderação de conteúdo político é destacada por estes resultados, não apenas contribuindo para a filtragem e análise de discursos ofensivos, mas também promovendo um ambiente político mais respeitoso e saudável. Este trabalho também ressalta a importância da combinação entre técnicas de aprendizado de máquina e PLN na identificação de linguagem prejudicial, abrindo caminho para futuras pesquisas e aplicações práticas neste campo. A capacidade de identificar e classificar discursos de ódio e ofensivos de maneira precisa e eficiente pode contribuir para a devida responsabilização dos autores, potencializando a manutenção de uma esfera pública equilibrada e respeitosa, especialmente no contexto da política brasileira.

Uma importante limitação do estudo está relacionada a classificação individual das frases dos deputados federais, o que pode ser influenciado por pontos de vista pessoais do autor do estudo, visto que o conteúdo ofensivo não foi validado por outras pessoas pertencentes aos grupos ofendidos. A limitação do hardware disponível também impediu a exploração de técnicas computacionais mais avançadas, como LDA e métodos baseados em redes neurais.

Alternativas tecnológicas, como aprendizado semi-supervisionado e Active Learning, também são recomendadas. Além disso, a utilização do método de Transfer Learning, aplicado para adaptar estudos sobre detecção de discurso de ódio em inglês para o contexto da língua portuguesa, representa uma direção promissora para futuras investigações.

REFERÊNCIAS

[1] Raj, C., Agarwal, A., Bharathy, G., Narayan, B., & Prasad, M. (2021). Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. Electronics, 10(22), 2810.

[2] Shakil, M. H., & Alam, M. G. R. (2022). Hate Speech Classification Implementing NLP and CNN with Machine Learning Algorithm Through Interpretable Explainable AI. IEEE.

[3] EL-SAYED, Tharwat; MUSTAFA, Abdallah; EL-SAYED, Ayman; ELRASHIDY, Mohamed. Hate Speech Detection by Classic Machine Learning. In: 2023 3rd International Conference on Electronic Engineering (ICEEM), Menouf, Egypt, 07-08 October 2023. IEEE, 2023. DOI: 10.1109/ICEEM58740.2023.10319569.

[4] KAELIN, V.; BOYD, A.; WERLER, M.; PARDE, Natalie; KHETANI, M. Natural Language Processing to Classify Caregiver Strategies Supporting Participation Among Children and Youth with Craniofacial Microsomia and Other Childhood-Onset Disabilities. Journal of Pediatric Rehabilitation Medicine, [S.l.], v. 16, n. 3, p. 1-14, 18 Sep. 2023. DOI: 10.1007/s41666-023-00149-y.

[5] AHANA, N. V.; R, Prerana; S, Niharika; S, Rakshitha; K J, Bhanushree. Automatic Hate Speech Detection using Ensemble Method and Natural Language Processing Techniques. In: 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), [S.l.], 01-02 September 2023. IEEE, 2023. DOI: 10.1109/NMITCON58196.2023.10276372.

[6] SAHU, Yatendra; GUPTA, R. K.; BHARTI, S. Combating Hate Speech on Q&A Forums with Machine Learning. In: 2023 World Conference on Communication & Computing (WCONF), RAIPUR, India, 14-16 July 2023. IEEE, 2023. DOI: 10.1109/WCONF58270.2023.10235146.