UNIVERSIDADE NOVE DE JULHO – UNINOVE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO

SABRINA DA SILVA VASCONCELLOS

INTELIGÊNCIA ARTIFICIAL COM PROCESSAMENTO DE LINGUAGEM NATURAL EM PEOPLE ANALYTICS PARA PREDIÇÃO DE CLIMA ORGANIZACIONAL

São Paulo
2025

UNIVERSIDADE NOVE DE JULHO – UNINOVE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO

SABRINA DA SILVA VASCONCELLOS

INTELIGÊNCIA ARTIFICIAL COM PROCESSAMENTO DE LINGUAGEM NATURAL EM PEOPLE ANALYTICS PARA PREDIÇÃO DE CLIMA ORGANIZACIONAL

Dissertação apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da UNINOVE como parte dos requisitos para obtenção do título de Mestre em Informática.

Orientador: Prof. Dr. Peterson Adriano Belan.
Coorientador: Prof. Dr. Wonder Alexandre Luz Alves.

São Paulo
2025

Vasconcellos, Sabrina da Silva.
Inteligência artificial com processamento de linguagem natural em people analytics para predição de clima organizacional. / Sabrina da Silva Vasconcellos. 2025.
113 f.
Dissertação (Mestrado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2025.
Orientador (a): Prof. Dr. Peterson Adriano Belan.

1. Análise de pessoas. 2. Inteligência artificial. 3. Processamento de linguagem natural. 4. Análise de sentimentos. 5. Clima organizacional.
I. Belan, Peterson Adriano. II. Título.

CDU 004

UNINOVE
Universidade Nove de Julho

ATA DE DEFESA DE DISSERTAÇÂO

Ao décimo sexto dia do mês de dezembro de dois mil e vinte e cinco, às 14H00, do programa de Pós-Graduação, desta Universidade, reuniu-se em sessão pública a Comissão Julgadora da Dissertação de Mestrado de Sabrina da Silva Vasconcellos sob o título INTELIGÊNCIA ARTIFICIAL COM PROCESSAMENTO DE LINGUAGEM NATURAL EM PEOPLE ANALYTICS PARA PREDIÇÃO DE CLIMA ORGANIZACIONAL

Integraram a comissão os professores: Prof. Dr. Peterson Adriano Belan (UNINOVE), o Prof. Dr. Fellipe Silva Martins - MACKENZIE, Prof. Dr. Edson Melo de Souza - PPGI - UNINOVE, Prof. Dr. Wonder Alexandre Luz Alves - PPGI - UNINOVE sob a presidência do primeiro, orientador da dissertação. A banca examinadora, tendo decidido aceitar a dissertação, passou à arguição pública do candidato. Encerrados os trabalhos, os examinadores deram parecer final sobre a dissertação.

	Parecer
Prof. Dr. Peterson Adriano Belan	Aprovado
Prof. Dr. Fellipe Silva Martins	Aprovado
Prof. Dr. Edson Melo de Souza	Aprovado
Prof. Dr. Wonder Alexandre Luz Alves	Aprovado

Parecer:

Em conclusão, o candidato foi considerado aprovado, no grau de Mestre em Informática e Gestão do Conhecimento. E, para constar, eu, Prof. Dr. André Felipe Henriques Librantz, diretor do Programa de Mestrado e Doutorado em Informática e Gestão do Conhecimento, lavrei a presente ata que assino juntamente com os membros da banca examinadora.

terça-feira, 16 de dezembro de 2025

gov.br Documento assinado digitalmente
PETERSON ADRIANO BELAN
Data: 13/02/2026 20:40:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Peterson Adriano Belan

gov.br Documento assinado digitalmente
FELLIPE SILVA MARTINS
Data: 19/12/2025 16:30:32-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Fellipe Silva Martins

gov.br Documento assinado digitalmente
EDSON MELO DE SOUZA
Data: 19/12/2025 16:51:33-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Edson Melo de Souza

gov.br Documento assinado digitalmente
WONDER ALEXANDRE LUZ ALVES
Data: 20/12/2025 06:02:11-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Wonder Alexandre Luz Alves

Prof. Dr. André Felipe Henriques Librantz

Dedico a Deus o autor da minha vida.

AGRADECIMENTOS

Agradeço a todos da minha família, aqueles que sempre estiveram presentes, me apoiando incondicionalmente em todos os momentos desta jornada acadêmica e pessoal.

Aos especialistas de recursos humanos que contribuíram com dedicação, generosidade e profundo comprometimento no processo de rotulagem e definição das dimensões analíticas, tornando possível a construção rigorosa desta pesquisa.

Ao meu orientador Prof. Dr. Peterson Adriano Belan, e ao meu coorientador, Prof. Dr. Wonder Alexandre Luz Alves, ambos da Universidade Nove de Julho, pela orientação competente, incentivo constante, disponibilidade e inspiração. Os senhores foram fundamentais para que eu alcançasse esta etapa e concretizasse a realização de um sonho.

Agradeço também ao Prof. Dr. Edson Melo de Souza, da Universidade Nove de Julho, cuja participação na qualificação e defesa deste trabalho foi marcada por palavras de motivação, contribuições valiosas e estímulo à melhoria contínua, enriquecendo significativamente os resultados apresentados, bem como ao Prof. Dr. Felipe Silva Martins, da Universidade Presbiteriana Mackenzie, pelas reflexões criteriosas, sugestões relevantes e contribuições acadêmicas que fortaleceram a qualidade teórica e metodológica deste estudo.

À Universidade Nove de Julho (UNINOVE), em especial ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento (PPGI), e a todos os professores que, ao longo do percurso, ampliaram meus conhecimentos e fortaleceram minha formação, possibilitando o desenvolvimento desta pesquisa.

Por fim, agradeço aos colegas que caminharam comigo durante essa trajetória, pelo apoio, pelas trocas de experiências e pela parceria ao longo do curso.

Entendo que expressar minha gratidão a cada um é também uma forma de agradecer a Deus, que me concedeu força e sabedoria para acreditar e chegar até aqui.

RESUMO

Contexto: A crescente adoção de tecnologias de inteligência artificial (IA) nas organizações tem impulsionado avanços significativos na forma como se analisa o comportamento humano no ambiente de trabalho. Embora áreas como recursos humanos (RH) coletem grandes volumes de dados em linguagem natural, como entrevistas de desligamento, *feedbacks*, pesquisas de clima e *eNPS (Employee Net Promoter Score),* esses dados ainda não são plenamente explorados em *people analytics*. **Objetivo:** Aplicar técnicas de processamento de linguagem natural (PLN) para desenvolver métodos preditivos do clima organizacional, utilizando como base textual as entrevistas de desligamento. **Método:** Para isso, foi conduzido um modelo supervisionado de análise de sentimentos, *Random Forest,* com e sem o uso de dados sintéticos. As dimensões diretas de escalas (1-5), (0-10) e os rótulos de sentimento (detrator, neutro e promotor) foram definidas manualmente por três especialistas em RH intitulados como (E1, E2 e E3) e consolidadas por um modelo de consenso intitulado como (E4), sendo um dos especialistas da empresa pesquisada, e os demais provenientes de outras organizações, ampliando a perspectiva de mercado. Também foram consideradas as dimensões indiretas do *eNPS,* definidas manualmente pelo especialista (E2), incluindo liderança, carreira, comunicação, diversidade, saúde e bem-estar, retenção, times, treinamento, inovação e engajamento, associados às perguntas e respostas das entrevistas. O desempenho do modelo foi avaliado por métricas como acurácia, precisão, *recall* e *f1-score*. Complementarmente, a pesquisa incorporou análises de sentimentos agrupadas às dimensões indiretas, estatística aplicada, visualizações e análises temporais, permitindo observar a evolução do clima organizacional ao longo do tempo. **Resultados:** O modelo *Random Forest* final, treinado exclusivamente com dados reais, apresentou o melhor desempenho, alcançando acurácia de 75%. **Conclusão:** Esse resultado demonstra o potencial da abordagem para apoiar decisões estratégicas baseadas em dados, orientar intervenções mais assertivas e contribuir para uma gestão de pessoas mais precisa e humanizada nas organizações.

Palavras-chave: Análise de pessoas, Inteligência artificial, Processamento de linguagem natural, Análise de sentimentos, Clima organizacional.

ABSTRACT

Context: The growing adoption of artificial intelligence (AI) technologies in organizations has driven significant advances in how human behavior in the workplace is analyzed. Although areas such as Human Resources (HR) collect large volumes of natural language data such as exit interviews, feedback, climate surveys, and eNPS (Employee Net Promoter Score) these data are still not fully leveraged in people analytics. **Objective:** To apply natural language processing (NLP) techniques to develop predictive methods for organizational climate, using exit interviews as the textual basis. **Method:** To this end, a supervised sentiment analysis model based on Random Forest was conducted, with and without the use of synthetic data. The direct scale dimensions (1–5), (0–10) and sentiment labels (detractor, neutral, and promoter) were manually defined by three HR specialists (E1, E2, and E3) and consolidated through a consensus model (E4). One of the specialists was from the studied organization, while the others came from different organizations, broadening the market perspective. Indirect eNPS dimensions were also considered, manually defined by specialist E2, including leadership, career, communication, diversity, health and well-being, retention, teams, training, innovation, and engagement, and associated with the questions and responses from the interviews. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. Additionally, the study incorporated sentiment analyses grouped by indirect dimensions, applied statistics, visualizations, and temporal analyses, enabling observation of organizational climate evolution over time. **Results:** The final Random Forest model, trained exclusively on real data, achieved the best performance, reaching an accuracy of 75%. **Conclusion:** This result demonstrates the potential of the proposed approach to support data-driven strategic decisions, guide more targeted interventions, and contribute to more precise and human-centered people management in organizations.

Keywords: People analytics, Artificial intelligence, Natural language processing, Sentiment analysis, Organizational climate.

LISTA DE SIGLAS

ACC - Acurácia

AIRH - Academy to innovate HR (Academia para inovação de RH)

AM - Aprendizado de máquina

ARH - Administração de recursos humanos

DHO - Desenvolvimento humano organizacional

FN - Falso negativo

FP - Falso positivo

GPTW - Great place to work (Ótimo lugar para trabalhar)

HRA - Human research analytics (Análise de recursos humanos)

IA - Inteligência Artificial

NDA - Non-Disclosure Agreement (Acordo de confidencialidade)

*PA - People analytics (*Análise de pessoas*)*

PLN - Processamento de linguagem natural

PLR - Participação nos lucros e resultados

PPR - Programa de participação nos resultados

Pr - Precision (Precisão)

Re - Recall (Revocação)

*ROI - Return on Investment (*Retorno sobre o investimento)

RH - Recursos humanos

SMOTE – Synthetic Minority Over-sampling Technique (Técnica de sobreamostragem sintética de minorias)

TA - Talent analytics (Análise de talentos)

T&D - Treinamento e desenvolvimento

TF-IDF - Term frequency–inverse document frequency (termo–inverso da frequência nos documentos)

TO - Turnover (Taxa de rotatividade dos empregados)

VN - Verdadeiro negativo

VP - Verdadeiro positivo

WA - Workforce analytics (Análise da força de trabalho)

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O RH é tradicionalmente visto como orientado para pessoas, e não para dados (Marr, 2018), contudo, esse cenário vem se transformando ao longo dos anos. Esse diagnóstico revela um dilema histórico. Embora o RH seja a área responsável por lidar com o comportamento humano nas organizações, ainda opera com baixa maturidade analítica, entendida como o grau em que uma organização é capaz de coletar, integrar, analisar e utilizar dados de forma sistemática e estratégica para apoiar a tomada de decisão, evoluindo de análises descritivas para análises preditivas e prescritivas, com impacto direto no desempenho organizacional (Gartner, 2017; West, 2020). Estudos confirmam este cenário onde a maioria das empresas permanece nos níveis descritivo e diagnóstico de maturidade analítica, com 75% em estágios iniciais e apenas 25% avançando para análises preditivas (Green, 2016; Bersin, 2024). Apesar disso, 58% das empresas planejam investir em *people analytics* nos próximos anos, reforçando a distância entre o discurso de inovação e a prática analítica aplicada ao comportamento humano (*PEOPLE ANALYTICS TRENDS; GARTNER*, 2024).

Este estudo surge como resposta a um desafio concreto e atual a subutilização de dados não estruturados no RH para análises preditivas e tomadas de decisão baseadas em evidências. Embora o clima organizacional seja reconhecido como fator central para compreender a experiência e o comportamento dos empregados (Lanzer, 2017; Daft, 2013), as organizações continuam aplicando metodologias analíticas básicas e retrospectivas, sem explorar plenamente o potencial de tecnologias como a inteligência artificial (Green, 2016; Minarelli e Oliveira, 2021). Esse cenário evidencia um paradoxo entre a digitalização crescente das organizações e a baixa adesão a análises textuais. Conforme o *GARTNER ANALYTICS ASCENDANCY MODEL* (2017), poucas empresas alcançam os níveis preditivo e prescritivo de maturidade analítica, onde se concentram os maiores ganhos estratégicos. Nesse contexto, técnicas de PLN como a análise de sentimentos, oferecem caminhos viáveis para transformar percepções organizacionais em evidências analíticas (Tunstall, Werra e Wolf, 2022).

O problema pode estar enraizado em uma cultura organizacional historicamente orientada à intuição e à subjetividade, em que clima e cultura eram considerados conceitos abstratos de difícil mensuração (Lanzer, 2017). Entretanto, a

adoção de sistemas digitais vem moldando esse cenário. Pesquisas recentes indicam que a tecnologia já é percebida como fator de melhoria de resultados por 46% dos brasileiros (*BOSCH TECH COMPASS,* 2024) e como elemento de impacto na cultura corporativa por 85% das empresas (*DELOITTE*, 2024) evidenciando uma transformação sociotécnica em curso.

A literatura recente em PLN aplicado ao RH utiliza técnicas como *Doc2Vec* e análise de grafos (Berenguer, 2024), mas ainda apresenta limitações, sobretudo pela baixa validação em ambientes reais atualizados e ausência de métricas preditivas. Boen (2022), aborda predição de saída com *Random Forest*, mas apresenta limitações com foco exclusivo em dados estruturados. Moreira (2025) emprega modelos supervisionados em dados reais de absenteísmo, porém sem análise textual ou aplicação de PLN, o que mantém aberta a lacuna sobre como captar nuances subjetivas do clima organizacional a partir de linguagem natural. Além disso, nenhum desses estudos apresenta estratégias de anonimização, etapa essencial para o tratamento ético de dados sensíveis em RH. Essa ausência reforça a lacuna entre o potencial das técnicas e sua aplicação prática na área, destacando a relevância da abordagem *TF-IDF* com *Random Forest* adotada neste estudo.

Para avançar nesta lacuna, este estudo utiliza dados reais e anonimizados de entrevistas de desligamento realizadas entre 2022 e 2025 em uma empresa do setor financeiro. Foi adotado a vetorização *TF-IDF* combinada ao modelo supervisionado *Random Forest* para análise de sentimentos, classificando os rótulos promotor, neutro e detrator com base nas dimensões diretas e indiretas do *eNPS* definidas por especialistas de RH. Apesar da utilização de escalas de dimensões diretas (1–5) e (0–10) foram empregadas apenas como mecanismo de consenso metodológico para a rotulação supervisionada dos dados textuais, não tendo como foco o cálculo do indicador *eNPS*. O desempenho foi avaliado por acurácia, precisão e *f1-score*, análises temporais e visualizações. Ao empregar dados de desligamentos reais e espontâneos, a pesquisa evita vieses comuns em pesquisas internas e reforça a segurança psicológica dos respondentes, favorecendo respostas mais genuínas (Edmondson, 2020; Clark, 2023). A metodologia contribui para a maturidade analítica do RH ao demonstrar como dados textuais podem embasar decisões baseadas em evidências e transformar percepções organizacionais em conhecimento estratégico para a gestão de pessoas.

1.1 OBJETIVO

Este trabalho tem como objetivo identificar sentimentos por meio de técnicas de processamento de linguagem natural, analisando textos e termos emergentes em respostas não estruturadas de entrevistas de desligamento, a partir das próprias narrativas dos colaboradores, no contexto do clima organizacional.

Com base neste objetivo, foi definido, ainda os objetivos específicos a seguir:

- Validar a base de dados com entrevistas de desligamentos voluntários, ou seja, empregados que solicitaram a saída e involuntários, empregados que foram desligados pela empresa;
- Criar os rótulos manuais para o modelo supervisionado, com três especialistas de RH e com um modelo valido de concordância;
- Criar classificações manuais de dimensões diretas e indiretas de *eNPS* para perguntas e respostas garantindo o contexto;
- Realizar o pré-processamento de dados de texto;
- Treinar o modelo supervisionado para classificar as perguntas e respostas em categorias pré-definidas de sentimentos;
- Relacionar as dimensões indiretas com sentimentos;
- Criar análises exploratórias e visualizações das palavras mais frequentes para entender os principais termos mencionados nas respostas;
- Identificar padrões temporais: aplicar PNL junto com análises temporais para ver como os temas de desligamento mudam ao longo do tempo;
- Comparar e gerar Insights do modelo *Random Forest* com e sem dados sintéticos (*SMOTE)*;

1.2 DELIMITAÇÃO DO TEMA

Este estudo tem como objetivo aplicar técnicas de processamento de linguagem natural, em perguntas e respostas de entrevistas de desligamentos dos empregados, com a proposta de identificar padrões de forma automática sobre o que os empregados possuem em comum no contexto de clima organizacional. Para isso, será adotada uma abordagem de classificação supervisionada com rótulos manuais

de sentimentos, com o foco em extrair conhecimento por meio da combinação de diferentes técnicas de análise textual.

As análises incluíram: (i) análise de sentimentos por meio do modelo *Random Forest* para classificar as respostas como positivas, negativas ou neutras; (ii) construção de tabelas e gráficos para representar visualmente os termos mais frequentes nas perguntas concatenadas as respostas, apenas respostas e (iv) análise de dimensões indiretas *eNPS* com sentimentos, para mapear estratégias de sentimentos por dimensões; (v) integração com análises temporais para observar como as percepções de clima evoluem ao longo do tempo.

Contudo, o escopo do estudo se limita às respostas disponibilizadas de forma espontânea nas entrevistas de desligamento, coletadas em português do Brasil. A escolha desse conjunto de dados se justifica pela sua natureza rica em conteúdo textual, pela viabilidade de análise quantitativa com técnicas de PLN, e pela disponibilidade dos registros em ambiente institucional, facilitando a extração e o tratamento dos dados.

1.3 JUSTIFICATIVA DA PESQUISA

Apesar das possibilidades oferecidas pelas tecnologias de PLN, as áreas de RH ainda enfrentam desafios para utilizar esses dados de forma estratégica na compreensão e na predição do clima organizacional. Embora o clima seja reconhecido como fator essencial para o bem-estar e o desempenho dos empregados (Chiavenato, 2014; Daft, 2013; Lanzer, 2017; Santos, 2021), muitas organizações ainda se baseiam em métodos tradicionais e pouco exploram o potencial analítico contido em relatos, *feedbacks* e entrevistas.

A relevância desta pesquisa está em propor uma solução para esse desafio. A aplicação de técnicas de PLN aliadas à IA pode transformar grandes volumes de dados textuais provenientes de entrevistas de desligamento, pesquisas internas e feedbacks em diagnósticos preditivos sobre o clima organizacional. West (2020) destaca que *people analytics* se encontra na interseção entre estatística, ciência comportamental, tecnologia e estratégia, uma combinação ainda pouco explorada na análise do clima corporativo.

O cenário atual reforça a necessidade de uso estratégico de dados no RH. Embora 74% dos líderes reconheçam a importância da análise de dados para o futuro da gestão de pessoas, apenas 32% aplicam abordagens preditivas (*LINKEDIN GLOBAL TALENT TRENDS*, 2024). A maioria das empresas ainda opera nos níveis descritivos e diagnósticos de maturidade analítica (Green, 2016), evidenciando limitações na adoção de *people analytics* avançado. Além disso, 67% dos trabalhadores consideram o clima organizacional mais relevante que a remuneração *PWC* (2024), e 77% das empresas que utilizam ferramentas colaborativas relatam maior produtividade (*GLOBAL WORKPLACE ANALYTICS,* 2024). Assim, compreender e antecipar o clima é um diferencial competitivo.

A evolução das técnicas de PLN tem possibilitado a aplicação de métodos avançados, como análise de sentimentos, extração de tópicos e sumarização de textos em larga escala. Marr (2018) reforça que o uso de dados para prever comportamento organizacional é uma competência chave em empresas orientadas por evidências. Integrar essas tecnologias às práticas de RH pode impulsionar a maturidade analítica e promover decisões mais humanas e eficazes.

Dessa forma, esta pesquisa se mostra relevante e oportuna ao unir uma demanda organizacional real, o potencial das tecnologias emergentes e a produção científica atual. Ao integrar inteligência artificial, processamento de linguagem natural e análise preditiva, o estudo contribui tanto para o avanço teórico quanto para aplicações, promovendo práticas de gestão de pessoas mais inteligentes, humanizadas e orientadas por dados.

1.4 PROBLEMA E HIPOTESES DE RESOLUÇÃO

Apesar de seu impacto no desempenho e no bem-estar dos empregados, dados textuais provenientes de entrevistas de desligamento, feedbacks e pesquisas de clima seguem subexplorado, enquanto o uso de IA e PLN no RH permanece restrito a funções operacionais.

Nesse contexto, o problema de pesquisa consiste em investigar:

Como a IA e o PLN podem ser aplicados em *people analytics* para diagnosticar e predizer o clima organizacional de forma mais precisa e estratégica?

O estudo adota métodos computacionais de análise de sentimentos, classificação e sumarização aplicados a dados reais de uma empresa do setor financeiro, com o objetivo de superar abordagens tradicionais. Inserido no tema *Inteligência Artificial com Processamento de Linguagem Natural em People Analytics para predição de clima organizacional*, o trabalho utiliza dados textuais provenientes de entrevistas de desligamento. As variáveis independentes compreendem o tipo de desligamento (voluntário ou involuntário) e o conteúdo textual das entrevistas, enquanto as variáveis dependentes abrangem a classificação do clima organizacional (detrator, neutro ou promotor), a polaridade dos sentimentos e os padrões temáticos identificados.

Como hipótese geral, é considerado que o uso de técnicas de IA aplicadas ao PLN permite diagnosticar e predizer o clima organizacional por meio da classificação das entrevistas de desligamento em categorias de clima detrator, neutro ou promotor incorporando ainda as dimensões indiretas do *eNPS* como variáveis relevantes para a compreensão dos sentimentos expressos. Essa abordagem busca contribuir para decisões mais analíticas, estratégicas e orientadas por dados na gestão de pessoas.

As hipóteses específicas são:

(i) existe associação significativa entre as perguntas e respostas das entrevistas de desligamento e a classificação do clima organizacional (detrator, neutro ou promotor);

(ii) técnicas de PLN são capazes de classificar corretamente o clima organizacional percebido nas entrevistas, com desempenho superior ao acaso;

(iii) modelos de aprendizado supervisionado como *Random Forest* apresentam desempenho satisfatório na predição da classificação de clima organizacional.

Dessa forma, as hipóteses deste estudo avaliam se técnicas de inteligência artificial aplicadas ao processamento de linguagem natural são capazes de predizer o clima organizacional a partir de perguntas e respostas de entrevistas de desligamento. Ao examinar os padrões textuais e as percepções presentes nesses relatos, este estudo procura produzir evidências empíricas que fortaleçam o uso estratégico de dados não estruturados no RH para compreender o clima organizacional.

1.5 ESTRUTURA DA PESQUISA

A presente pesquisa está organizada em seis seções principais. A primeira seção apresenta o objetivo da pesquisa e sua relevância, além de abordar a delimitação, a justificativa do estudo, o problema e a estrutura proposta. A segunda seção discute 30 trabalhos correlatos publicados entre 2016 e 2025 e as lacunas identificadas. A terceira seção traz a fundamentação teórica, reunindo os conceitos essenciais para o entendimento da investigação. A quarta seção contempla a caracterização da pesquisa, bem como os materiais e métodos utilizados na condução dos experimentos. Na sequência, a quinta seção apresenta os resultados obtidos a partir dos experimentos realizados com a abordagem desenvolvida. Por fim, a sexta seção compreende o encerramento do trabalho, com as conclusões e considerações finais.

2 REVISÃO DA LITERATURA E LACUNAS

Este estudo adotou uma revisão integrativa da literatura para mapear e analisar pesquisas sobre processamento de linguagem natural e análise de sentimentos no contexto de *people analytics* e clima organizacional. Foram analisados 30 trabalhos publicados entre 2016 e 2025, nos idiomas inglês, espanhol, italiano, português do Brasil e de Portugal. As buscas contemplaram artigos científicos indexados em bases internacionais, como *Scopus* e *Web of Science (SCIE, SSCI* e *ESCI),* além de bases complementares como *IEEE Xplore, PubMed/PMC* e *SciELO*. Também foram considerados estudos identificados via *Google* Acadêmico e *ResearchGate*, bem como literatura acadêmica qualificada proveniente de repositórios institucionais universitários, utilizada como suporte metodológico e contextual. Os descritores empregados incluíram *"People Analytics", "Sentiment Analysis", "Natural Language Processing"* e *"Organizational Climate",* combinados por operadores booleanos. Como critérios de inclusão, foi considerado artigos revisados por pares, estudos empíricos e revisões teóricas alinhados ao contexto organizacional, sendo excluídos trabalhos duplicados, estudos fora do escopo da gestão de pessoas e publicações sem acesso ao texto completo. A revisão mapeia contribuições teóricas, metodológicas e técnicas, organizadas por eixos temáticos, oferecendo uma visão estruturada e cronológica do estado da arte e apontando direções para a integração entre tecnologia, linguagem natural e a gestão de pessoas orientada por dados.

A literatura recente apresenta avanços, mas também desafios no uso de IA e PLN aplicados ao contexto de *people analytics* e clima organizacional. Estudos iniciais, como Ribeiro et al. (2016), apontam limitações técnicas nos métodos de análise de sentimentos, incluindo vieses na detecção de emoções negativas e falta de padronização. Zangaro (2020) complementa esse debate ao oferecer uma crítica sociológica ao uso de sentimentos como mecanismo de gestão subjetiva. Propostas como o modelo *aspect-based* de Dina e Juniata (2020), aplicado a dados públicos do *Glassdoor*, ilustram o potencial analítico, mas ainda revelam distância entre experimentação e aplicação prática em ambientes internos sensíveis.

A relevância das fontes textuais internas também é destacada por diversos autores. Colladon e Scettri (2019) identificam relação entre positividade linguística e

desempenho organizacional. Polzer (2022) demonstra o valor de *e-mails* e mensagens para o estudo da cultura e da socialização. Entretanto, Gutiérrez et al. (2022) observam que métricas de sentimentos continuam subutilizadas no RH. Embora Ishikawa et al. (2021) e Gaye et al. (2021) comprovem a viabilidade técnica do PLN em larga escala, esses trabalhos ainda se concentram em redes sociais e avaliações externas, distantes do contexto corporativo confidencial que caracteriza os dados internos de clima.

Outro eixo importante da literatura envolve governança e confiabilidade na adoção de IA em RH. Hearne (2023), Bar-Gil et al. (2024) e Henriques (2024) destacam a importância da transparência, da confiança institucional e da aceitação tecnológica para a consolidação dessas soluções. Mayfield (2024) propõe uma abordagem híbrida de IA para análise de sentimentos, enquanto Ribeiro et al. (2016) reforça que limitações metodológicas ainda comprometem a robustez interpretativa dessas análises.

No campo conceitual, trabalhos como Rigamonti et al. (2023), Benin (2023) e Stevenson (2024) apontam confusões teóricas e desafios metodológicos que dificultam a consolidação do PLN no RH. Em paralelo, surgem iniciativas que ampliam o horizonte de aplicações, como a escuta ativa mediada por *chatbots* (Souza et al., 2022) e o uso de gêmeos digitais humanos para monitoramento contínuo de sentimentos (Martin, 2024), embora ainda sem validação empírica em dados organizacionais internos.

Do ponto de vista técnico, há evidências do desempenho de modelos supervisionados em diferentes contextos. Matos et al. (2023) aplicam *Naive Bayes* a interações via *WhatsApp.* Lopes (2024) utiliza métodos *ensemble* com *TF-IDF* e *BoW.* Jim et al. (2024) revisam mais de 160 modelos, de *CNN* e *RNN* a *BERT* e *GPT,* indicando a superioridade de abordagens baseadas em *deep learning*. Loures et al. (2024) validam *BERT* e *GPT* em bases públicas e destacam resultados promissores, enquanto estudos como Sajid et al. (2022) e Prestes (2024) aproximam-se do ambiente organizacional ao analisar avaliações espontâneas online, ainda que com escopo limitado.

A literatura também evidencia o papel estratégico dos sentimentos como preditores de clima, engajamento e desligamento. Colladon e Scettri (2019) relacionam padrões linguísticos ao valor de mercado. Sun et al. (2021) utilizam

Doc2Vec para estimar satisfação. Saxena et al. (2023) identificam correlação entre sentimentos positivos e produtividade. Boen (2022) emprega *Random Forest* para prever *turnover*. Moreira (2025) utiliza *Random Forest* para prever absenteísmo. Pedrosa (2022) reforça a importância da escuta estratégica. Berenguer (2024) propõe uma plataforma *SaaS* de análise de sentimentos em *e-mails* corporativos, sugerindo caminhos promissores para o monitoramento contínuo do clima organizacional.

Esse panorama torna claras as principais lacunas metodológicas o uso restrito de dados internos, ausência de anonimização, dependência de bases públicas e validação insuficiente. O interesse pelo clima organizacional ganhou força após a *COVID*-19, quando bem-estar e clima passaram a ser prioridade estratégica (*MCKINSEY*, 2021). No entanto, a *GALLUP* (2019) já apontava baixo engajamento no Brasil, indicando que os desafios são anteriores à pandemia. Apesar disso, apenas 36% das empresas utilizam análise de sentimentos conforme *PEOPLE ANALYTICS TRENDS*, (2024) e somente 28% levam os resultados aos gestores, o que limita a adoção de ações efetivas. Grande parte das pesquisas em RH segue abordagem qualitativa ou teórica, com pouca aplicação prática. Além disso, poucos estudos incluem a classe neutra, incorporada explicitamente neste trabalho para captar nuances do clima organizacional. Trabalhos como Leidner e Stevenson (2024), Pedrosa (2022), Peeters e Voorde (2020) e Rigamonti et al. (2022) reforçam a necessidade de ampliar o uso de PLN em todo o RH e elevar a maturidade analítica. Estudos como de Guo et al. (2021) e Prestes (2024) utilizam bases externas, mas enfrentam desbalanceamento, ambiguidade textual e ausência de modelos intermediários, como o *BERTimbau*, além da falta de análises temporais.

Embora a literatura evidencie avanços no uso de IA e PLN em *people analytics*, persistem limitações relevantes. A maioria dos estudos utiliza bases externas e públicas, não envolve dados internos confidenciais e, por isso, não apresentam técnicas de anonimização transparentes. Além disso, há baixa validação em ambientes organizacionais reais, o que reduz a utilidade prática dos modelos.

Diante dessas lacunas, o presente estudo aplica *Random Forest* a dados reais e anonimizados de entrevistas de desligamento, oferecendo evidências empíricas para a predição do clima organizacional a partir de linguagem natural.

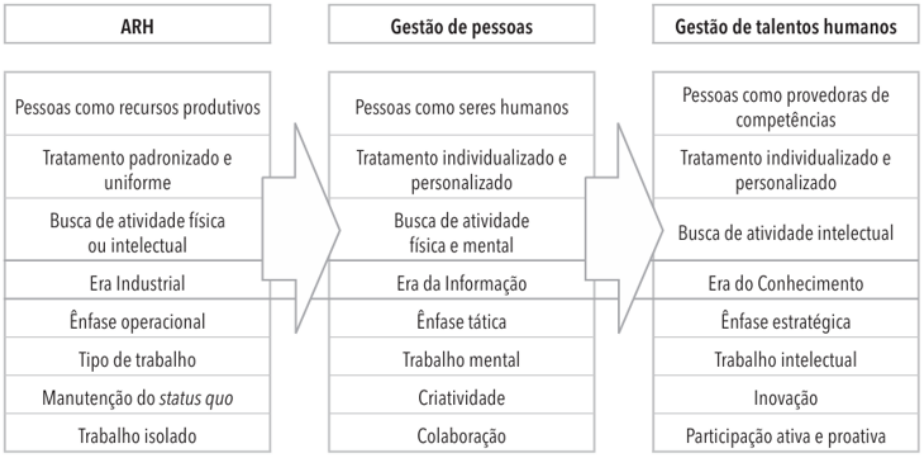
3 FUNDAMENTAÇÃO TEÓRICA

Nesta seção será apresentada a fundamentação teórica relacionada ao desenvolvimento desta dissertação, contemplando os seguintes tópicos: Recursos Humanos (seção 3.1); Clima organizacional (seção 3.1.1); *People Analytics* (seção 3.1.2); Inteligência Artificial (seção 3.2); Aprendizado de Máquina (seção 3.2.1); *Random Forest* (seção 3.2.2); *TF-IDF* (seção 3.2.3); Processamento de Linguagem Natural (seção 3.2.4); Análise de sentimentos (seção 3.2.5); Tipos de dados (seção 3.3); Métricas (seção 3.4) e Principais autores (seção 3.5).

3.1 RECURSOS HUMANOS (RH)

A área de RH é uma das que mais têm passado por transformações, inclusive na própria nomenclatura como Administração de Recursos Humanos (ARH), Gestão de Pessoas (GP) e Gestão do Talento Humano. Segundo Chiavenato (2014), cada termo reflete papéis e significados distintos, conforme apresentado na Figura 1.

Figura 1 – Os saltos gradativos da área de RH



Fonte: Chiavenato (2014).

A gestão de recursos humanos pode ser dividida em cinco subsistemas interligados, que visam a gestão completa do ciclo de vida do empregado dentro da organização, (Chiavenato, 2014).

Os cinco subsistemas de RH são:

1. Provisão: relacionado à busca por profissionais para preencher vagas, incluindo recrutamento e seleção.
2. Aplicação: envolve a utilização dos recursos humanos, como o processo de *onboarding*, orientação inicial e aplicação prática do conhecimento adquirido.
3. Manutenção: abrange a gestão de remuneração, benefícios, ambiente de trabalho e relações de trabalho, visando garantir a satisfação e motivação dos colaboradores.
4. Desenvolvimento: foco no desenvolvimento e aprimoramento das habilidades e competências dos colaboradores, incluindo treinamento, educação corporativa e planejamento de carreira.
5. Monitoramento: acompanhamento do desempenho dos colaboradores, avaliação de desempenho, *feedback* e gestão de conflitos, visando garantir o cumprimento das metas e objetivos da organização.

Esses subsistemas são interativos e interdependentes, ou seja, a gestão de cada um deles é fundamental para o sucesso da gestão de RH como um todo. "*RH é a área responsável por atrair, reter, desenvolver e recompensar pessoas, atuando como elo entre os objetivos organizacionais e as necessidades humanas*." Chiavenato (2004), sendo assim, "*A função de RH busca alcançar resultados organizacionais por meio das pessoas, garantindo um ambiente de trabalho favorável ao desempenho*." Maximiano (2000), isso reforça a importância das pessoas como fator crítico de sucesso organizacional. Para Robbins e Coulter (2011), "*A gestão de pessoas é a base da eficácia organizacional. Administrar pessoas envolve mais do que processos; envolve compreender motivações, culturas e comportamentos*."

3.1.1 CLIMA ORGANIZACIONAL

Clima organizacional é uma medida coletiva de como os empregados se sentem, como percebem interpretam e reagem ao ambiente organizacional construído pela estrutura, pela cultura e pelos sistemas de gestão, Daft (2013), corroborando com Chiavenato (2014), o clima organizacional é "*o ambiente interno existente entre os membros da organização e que está intimamente relacionado com o grau de*

motivação de seus participantes". Esse clima pode ser favorável ou desfavorável e impacta o comportamento das pessoas no ambiente de trabalho.

Para Luz (2003), o clima organizacional é o conjunto de propriedades mensuráveis do ambiente de trabalho, percebidas direta ou indiretamente pelos indivíduos, que influenciam a motivação e o comportamento organizacional. Pode ser avaliado por meio de pesquisas, entrevistas e indicadores relacionados à produtividade, rotatividade, absenteísmo e engajamento. Medeiros (2002) complementa ao afirmar que o clima reflete a qualidade da relação entre a organização e seus empregados, sendo um importante indicador da saúde organizacional. Climas positivos tendem a favorecer cooperação, criatividade e retenção, enquanto climas negativos estão associados a conflitos, desmotivação e aumento do *turnover*. Apesar de sua relevância, Santos (2021) destaca a dificuldade de operacionalizar o conceito de clima organizacional de forma válida e confiável, especialmente diante da complexidade da comunicação, da cultura e das relações humanas no trabalho.

Para superar essa dificuldade, diversas dimensões têm sido propostas como categorias de análise. Litwin e Stringer (1968), sugerem dimensões indiretas como estrutura, responsabilidade, recompensa, risco, apoio, conflito, padrões, identidade e pressão. Essas dimensões representam variáveis comportamentais e organizacionais que influenciam satisfação, motivação, comprometimento e identificação fatores que, segundo Reichheld (2003), determinam se um colaborador se torna promotor ou detrator no *eNPS*.

Chiavenato (2005; 2014), as dimensões são variáveis distribuídas entre dois eixos, o clima organizacional, composto por fatores perceptíveis do ambiente de trabalho nas dimensões de liderança, comunicação interna, recompensas e reconhecimento, engajamento e satisfação, desenvolvimento e carreira, trabalho em equipe e cooperação, treinamento e desenvolvimento (T&D), qualidade de vida no trabalho (QVT)), e cultura organizacional, formada por valores, crenças e normas compartilhadas que orientam o comportamentos relacionados à equidade e justiça organizacional e inovação e aprendizagem. Essa estrutura converge com a abordagem de Daft (2013), ao compreender o clima como resultado das percepções e reações dos empregados ao ambiente organizacional moldado pela estrutura, pela

cultura e pelos sistemas de gestão. As dez principais dimensões indiretas estão sintetizadas no Quadro 1.

Quadro 1 – Resumo das 10 principais dimensões indiretas

Dimensões (variáveis)		Dimensões (resumo)	Palavras-chave (indicadores)	Eixos
1	Comunicação interna	Comunicação	Informação, transparência, feedback	Clima Organizacional
2	Desenvolvimento e carreira	Carreira	Promoção, capacitação, crescimento	
3	Engajamento e satisfação	Engajamento	Motivação, pertencimento, compromisso	
4	Liderança	Liderança	Gestor, orientação, apoio	
5	Qualidade de vida no trabalho (QVT)	Saúde e bem-estar	Equilíbrio, carga de trabalho, ergonomia,	
6	Recompensas e reconhecimento	Retenção	Salário, bônus, benefícios, mérito	
7	Trabalho em equipe e cooperação	Time	Colaboração, integração, parceria	
8	Treinamento e desenvolvimento (T&D)	Treinamento	Cursos, aprendizagem, habilidades	
9	Equidade e justiça organizacional	Diversidade	Inclusão, ética, imparcialidade	Cultura Organizacional
10	Inovação e aprendizagem organizacional	Inovação	Melhoria contínua, experimentação	

Fonte: Elaborado pela autora (2025)

O quadro sintetiza dez dimensões indiretas que ajudam a interpretar o clima organizacional a partir de temas recorrentes no comportamento humano nas organizações. Cada dimensão é apresentada com um nome resumido e palavras-chave que funcionam como indicadores semânticos, agrupadas em dois grandes eixos diretamente ligados clima organizacional e cultura organizacional. Esses indicadores refletem aspectos resumidos como comunicação, carreira, engajamento, liderança, saúde e bem-estar, reconhecimento, trabalho em equipe, treinamento, diversidade e inovação. Daft (2013), Chiavenato (2005; 2014).

Nesse contexto, o *Employee Net Promoter Score (eNPS)* é conhecido como uma métrica utilizada para mensurar clima organizacional e engajamento. Derivado do Net Promoter Score proposto por Reichheld (2003) para medir lealdade de clientes, o *eNPS* foi adaptado ao contexto interno das organizações a partir de 2006, com difusão em práticas empresariais e estudos organizacionais. Tradicionalmente, o

eNPS utiliza uma escala de 0 a 10 para classificar colaboradores como promotores (notas 9 e 10), neutros (notas 7 e 8) ou detratores (notas de 0 a 6), sendo calculado pela diferença percentual entre promotores e detratores. Entretanto, em pesquisas de clima organizacional, é comum a adoção de escalas do tipo Likert, de 1 a 5 pontos, por sua maior adequação metodológica e familiaridade aos respondentes (Likert, 1932). Essa adaptação mantém a lógica conceitual do *eNPS*, classificando como promotores as respostas 4 e 5, neutros a resposta 3 e detratores as respostas 1 e 2, preservando a simplicidade e a interpretabilidade da métrica. Estudos empíricos indicam que essa adaptação é válida quando há coerência metodológica e alinhamento com os objetivos da pesquisa (Silva & Oliveira, 2021; West, 2020).

Além disso, escalas mais curtas reduzem a carga cognitiva, diminuem a fadiga dos respondentes e tendem a gerar respostas mais consistentes e estáveis, favorecendo análises e comparações internas (Preston & Colman, 2000; Dawes, 2008). Assim, o uso do *eNPS* adaptado à escala Likert é configurado como um instrumento para a mensuração do clima organizacional, especialmente quando integrado a análises mais amplas baseadas em dimensões indiretas e indicadores semânticos.

3.1.2 PEOPLE ANALYTICS (PA)

Até meados dos anos 2000, o processo de análise de dados no RH costumava ser caro e demorado, porém, com o avanço da tecnologia, os softwares tornaram-se mais acessíveis e precisos para a área, como explica Waber (2022). *People analytics* é um fenômeno simultaneamente antigo e novo. Guenole, Ferrar e Feinzig (2017), aborda a história, evolução e práticas de *people analytics*, além de contribuir com uma linha do tempo conceitual e prática que contextualiza o avanço de *people analytics* nas organizações.

People analytics integra a gestão de pessoas à ciência de dados, unindo os fundamentos clássicos de recursos humanos, como os propostos por Chiavenato (2014), às tecnologias modernas, como a inteligência artificial (Mitchell, 1997; Russell; Norvig, 2010) e o aprendizado de máquina (Haykin, 1999). A cultura orientada por dados é essencial para uma atuação estratégica da área de recursos humanos, e a análise de sentimentos vem ganhando destaque como ferramenta para captar

percepções e prever o clima organizacional. Assim, a metodologia de *people analytics* representa a síntese entre conhecimento organizacional, tecnologia e métodos analíticos, promovendo uma gestão mais inteligente, proativa e alinhada às transformações digitais nas organizações. A evolução do campo de *people analytics*, também conhecido como *Workforce Analytics* (WA), *Talent Analytics* (TA) e *HR Analytics* (HRA), surgiu no avanço da medição e análise de pessoas nas últimas décadas. Segundo Guenole, Ferrar e Feinzig (2017), a transformação da área de recursos humanos tradicional para uma função estratégica e baseada em dados ocorre à medida que as organizações passam a integrar análises quantitativas à gestão de pessoas. *People analytics* deixa de ser descritivo e passa a incorporar análises preditivas e prescritivas para apoiar a tomada de decisão e gerar cada vez mais valor à gestão de pessoas e consequentemente aos negócios.

A linha do tempo apresentada no Quadro 2 mostra a estrutura dessa transição, destacando os marcos na história de *people analytics* e como as empresas adotaram essa abordagem para melhorar o desempenho organizacional.

Quadro 2 – Linha do tempo da metodologia de *people analytics*

1970 - 1980	RH é operacional e visto como centro de custo.
1984	Marco inicial da quantificação do RH.
1990 - 2000	Surgem *benchmarks* e métricas de *ROI* em análise de pessoas, com foco descritivo.
2000 - 2010	Avanço da análise descritiva com *dashboards* e *KPIs,* a introdução de conceitos de *people data* em contextos de comportamento organizacional com sensores e dados sociométricos.
2010 - 2015	Abordagem preditiva e prescritiva com tomada de decisão baseada em evidências. Popularização de métodos supervisionados para prever comportamento humano no trabalho.
2015 - 2018	Consolidação de práticas e modelos de maturidade. Aplicações consolidadas de técnicas de PLN em RH. Modelos como Random Forest com *TF-IDF* para detectar temas em comentários abertos de pesquisas internas.
2018 - Presente	Além das discussões de ética e governança, este período marca a transição entre técnicas como *Random Forest* com *TF-IDF* para sentimento e clima e abordagens avançadas com IA generativa e *LLMs* para análise de dados não estruturados.

Fontes: Fitz-enz (1984), Breiman (2001), Waber (2013), West (2016), Green (2018) e (*Sundmark, 2018*).

O Quadro 2 apresenta a evolução de *people analytics*, destacando a transição do RH operacional (1970–1980) para uma atuação estratégica baseada em dados. Em 1984, começa a quantificação do RH, evoluindo com o uso de *benchmarks, KPIs* e sensores organizacionais. Entre 2010 e 2015, surge a aplicação de análises preditivas e prescritivas, em 2017, é consolidado os modelos de maturidade, e, desde 2018, cresce a atenção à ética, governança e ao uso de IA generativa e *LLMs* para explorar dados não estruturados relacionados ao clima e comportamento organizacional.

Ao longo dos anos, tanto a nomenclatura quanto a prática organizacional adotaram diferentes termos conforme o foco analítico e abordagem:

- *HR Analytics:* ênfase na eficiência dos processos tradicionais de RH (recrutamento, folha, benefícios).
- *Workforce Analytics:* abordagem voltada à força de trabalho como ativo estratégico, com foco em produtividade e custos.
- *Talent Analytics:* foco no ciclo de vida do empregado atração, retenção, desempenho e desenvolvimento
- *People Analytics:* termo mais recente e abrangente, integrando dados de sentimentos, relacionais, culturais e de performance.

Essas variações refletem a evolução da área, marcada pelo aumento da maturidade técnica e pela ampliação das fontes de dados. Novas informações, como sensores (Waber, 2013), redes de relacionamento *Organizational Network Analysis* (*ONA*), linguagem natural (Green, 2015; West, 2017) e análises de sentimentos, expandiram o papel do RH, possibilitando uma atuação mais digital, estratégica e orientada à predição.

A consolidação conceitual aparece em obras como Guenole, Ferrar e Feinzig (2017) e é aprofundada por autores como Waber (2013), que introduz o uso de sensores para compreender comportamentos sociais no trabalho, e Green (2016), que discute maturidade analítica, ética algorítmica e adoção organizacional de *people analytics*.

Segundo Waber (2013), "*quando usamos dados para entender os comportamentos dentro do ambiente de trabalho que tornam as pessoas eficientes, felizes, criativas, especialistas, líderes, seguidores, estamos usando people analytics*".

O autor também define a abordagem como um método científico aplicado à gestão de pessoas. Isson e Harriott (2016) complementam como um conjunto de etapas analíticas de definição do problema, coleta, análise, predição e decisão, integrando ciência de dados e estratégia organizacional.

De acordo com os autores *people analytics* se sustenta em três eixos centrais, dados, pessoas e negócio, sendo assim, a metodologia não se limita a uma única ferramenta específica. *People analytics* é uma metodologia de análise de pessoas através de dados para o negócio. Para West (2020), *people analytics* representa a interseção entre estatística, ciência comportamental, tecnologia e estratégia de pessoas, formando uma estrutura metodológica ampla e integrada, representada na Figura 2.

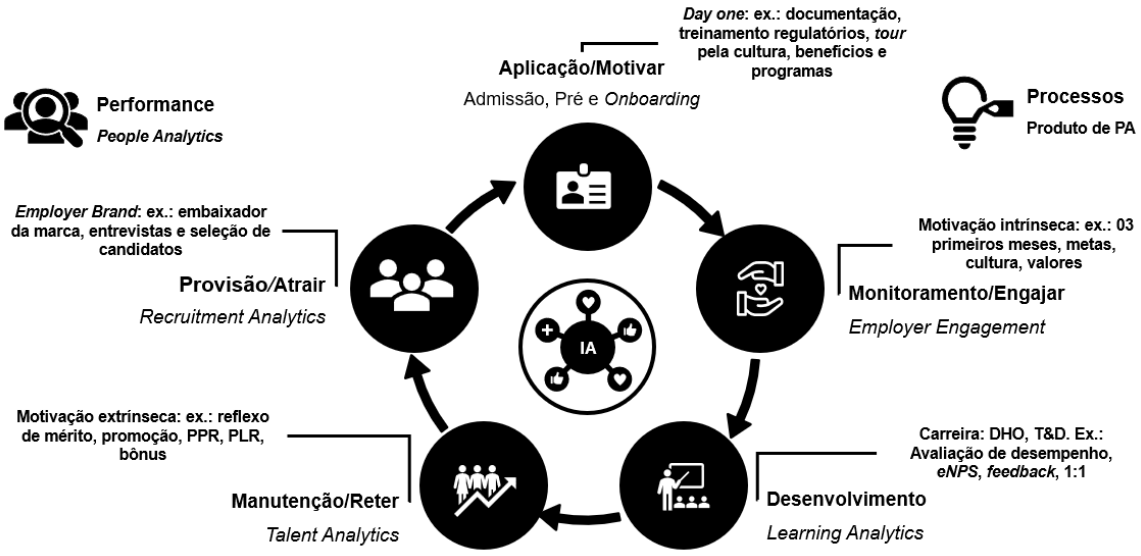
Figura 2 – Estrutura do *People Analytics*

Fonte: WEST (2020).

Apesar de *people analytics* está diretamente ligada a área de humanas em RH, a estrutura de West (2020) reforça a descrição de Green (2016) onde *people analytics* consiste em analisar dados. Em vez de ou apenas confiar na intuição, *people analytics* ajuda as organizações a confiar em dados, ou seja, traz o rigor científico da área de exatas em sua atuação. *People analytics* é uma área diretamente ligada ao RH, porém com uma dinâmica multidisciplinar entre áreas de exatas e humanas.

Baseado no ciclo de vida do empregado, através dos 05 subsistemas, o conceito de motivação intrínseca: satisfação interna, realização, crescimento, autonomia, propósito e motivação extrínseca: recompensas externas, remuneração, benefícios, promoções, reconhecimento de Chiavenato (2014) e a representação multidisciplinar de West (2020), a Figura 3 “Jornada de dados no RH *(HR Data Journey)*, inteligência artificial e processamento de linguagem natural”, apresenta onde a IA pode atuar, ou seja, no centro deste ciclo em cada subsistema do RH.

Figura 3 – Jornada de dados no RH (*HR Data Journey*), IA e PLN



Fonte: Elaborado pela autora (2025).

A Figura 3 apresenta o ciclo de *people analytics* aplicado às etapas da jornada do empregado, integrando IA aos processos de provisão, aplicação, monitoramento, desenvolvimento, manutenção e performance. Cada etapa é associada a um tipo de *analytics (Recruitment, Employer, Learning, Talent* e *People Analytics)* e exemplificada por práticas como *onboarding,* avaliações*, feedback*, treinamentos e ações de retenção. A Figura 3 também relaciona fatores de motivação intrínseca e extrínseca que influenciam engajamento e a retenção. Ao operar de forma paralela aos subsistemas de RH, *people analytics* contribui para reduzir vieses nas análises táticas. Os processos são apresentados como propostas de produtos analíticos de RH para ampliar a aplicação estratégica dos dados de pessoas.

3.2 INTELIGÊNCIA ARTIFICIAL (IA)

Inteligência Artificial (IA) é o ramo da ciência da computação que desenvolve sistemas capazes de realizar tarefas que normalmente exigiriam inteligência humana, como aprender, raciocinar, perceber o ambiente e tomar decisões. Para Russell e Norvig (2010), IA é "*o estudo de agentes que percebem o ambiente e tomam ações que maximizam suas chances de alcançar seus objetivos*". A IA é categorizada por capacidade funcional, incluindo IA reativa (ex: *Deep Blue*), IA com memória limitada (ex: carros autônomos), IA da teoria da mente em desenvolvimento e IA autoconsciente ainda teórica.

Já Bostrom (2014) propõe três níveis de IA, a fraca, geral e superinteligência, conforme sua capacidade de generalização e consciência:

1. IA Fraca (ou Estreita): realiza tarefas específicas com eficiência, sem consciência ou compreensão real;
2. IA Geral (ou Forte): teria capacidade cognitiva equivalente à humana, com adaptação a qualquer tarefa intelectual;
3. Superinteligência: IA que excederia a inteligência humana em praticamente todos os domínios.

Para Muller et al. (2020), a IA teve vários falsos começos e interrupções ao longo dos anos, em parte pela dificuldade das pessoas em compreender seu propósito e potencial. Mitchell (1997) define aprendizado de máquina, um dos pilares da IA, como o estudo de algoritmos que permitem que sistemas melhorem seu desempenho com base na experiência, definição essencial para compreender os modelos atuais aplicados à linguagem natural e à análise preditiva.

Embora a IA seja considerada um campo fascinante e promissor, Mueller et al. (2020) destacam que ela ainda não foi definida de maneira precisa. Russell e Norvig (2000) reforçam que existem diferentes abordagens para conceituá-la, considerando tanto a racionalidade quanto a semelhança com o pensamento humano. A IA representa um campo amplo, que abrange desde aprendizado de máquina até percepção e tomada de decisão em ambientes complexos.

Um marco inicial foi o modelo de neurônios artificiais de McCulloch e Pitts (1943), que demonstrou que redes neurais poderiam computar qualquer função lógica. Hebb (1949) introduziu a regra de aprendizado hebbiana, que inspira métodos de

atualização em redes neurais. Haykin (1999) sistematiza esses fundamentos e apresenta redes como o *Perceptron*, redes multicamadas (*MLP*), algoritmos de retropropagação (*backpropagation*) e a importância da função de ativação e da generalização, consolidando o entendimento teórico e prático de redes neurais artificiais componentes centrais do aprendizado profundo (*deep learning*).

Apesar dos avanços, para West (2020), o estado atual da Inteligência Artificial ainda exige participação humana significativa em *people analytics*, considerando que essas ferramentas são úteis apenas quando tarefas bem definidas permitem que algoritmos atuem de forma autônoma com eficácia.

3.2.1 APRENDIZADO DE MÁQUINA (AM)

O termo *Machine Learning* (AM - Aprendizado de Máquina) foi introduzido por Samuel (1959), pioneiro da IA na IBM, ao criar um programa que aprendia a jogar damas. Apesar disso, foi apenas a partir dos anos 2000, com o avanço da capacidade computacional e da disponibilidade de dados, que o campo ganhou força e aplicações práticas em diferentes áreas. O aprendizado de máquina é a ciência e a arte de programar computadores para aprender com dados, sendo definido por Mitchell (1997) como o processo em que um sistema melhora seu desempenho em uma tarefa a partir da experiência. Géron (2019) reforça que o objetivo central é desenvolver algoritmos capazes de identificar padrões e tomar decisões sem programação explícita, enquanto autores como Russell e Norvig (2010), Bishop (2006) e Grus (2016) destacam suas bases estatísticas, matemáticas e a conexão com agentes inteligentes.

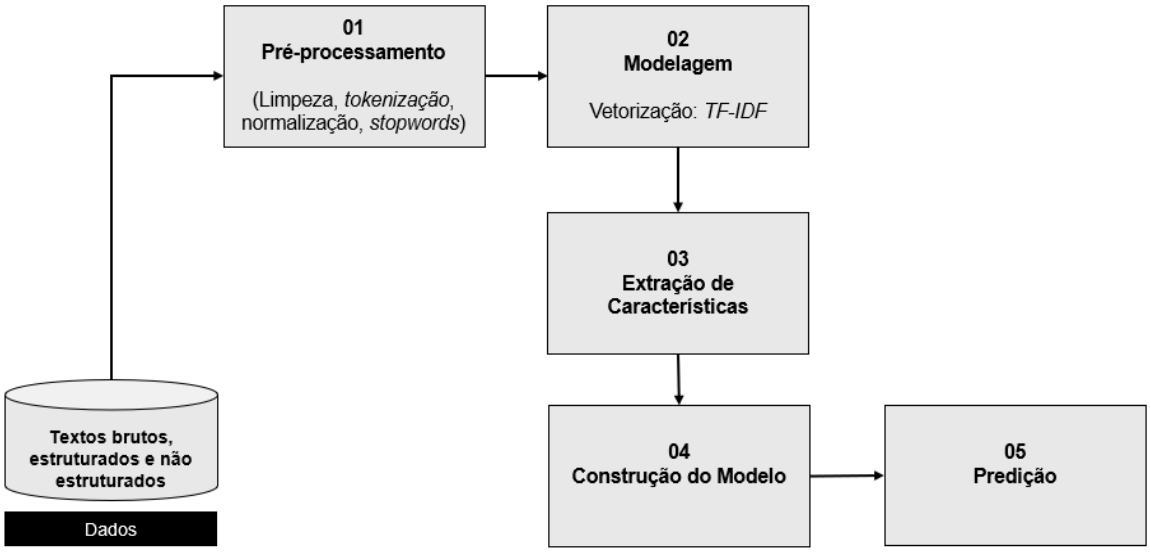
O aprendizado de máquina é geralmente classificado em três categorias o supervisionado, em que os dados possuem rótulos e o modelo aprende com exemplos conhecidos como regressão, classificação, *Random Forest*, redes neurais, não supervisionado, no qual o sistema identifica padrões ocultos em dados não rotulados (ex.: *clustering*, *PCA, LDA, BERTopic*), e por reforço, em que o algoritmo aprende por meio de recompensas em interação com o ambiente (ex.: robótica e jogos). Outras variações incluem o aprendizado semi-supervisionado e auto-supervisionado, aplicados em contextos de escassez de dados rotulados ou em modelos de linguagem (Géron, 2019).

Na prática, modelos supervisionados são utilizados em PLN. Um exemplo é o uso de *BERTopic*, que inicialmente pode atuar como técnica não supervisionada de descoberta de tópicos em entrevistas de desligamento e, posteriormente, ser integrado a modelos supervisionados como o *BERT*, compondo *pipelines* híbridos preditivos. Esses modelos de redes neurais exigem cuidados a capacidade de generalização, definição clara do problema e prevenção de *overfitting* (Haykin, 2001; Mitchell, 1997). Da mesma forma, técnicas supervisionadas, como *Random Forest* combinada ao vetor de características *TF-IDF*, são aplicadas para análise, em que textos rotulados, são utilizados para treinar o modelo a identificar automaticamente novas respostas.

Antes de tratar de AM é necessário compreender o conceito de modelo, entendido como a representação matemática ou probabilística da relação entre variáveis. O AM consiste justamente na criação e uso desses modelos aprendidos a partir de dados, também chamados de modelos preditivos, Grus (2016).

A Figura 4 ilustra um *pipeline* de PLN com *Random Forest* e *TF-IDF*.

Figura 4 – *Pipeline* de modelagem com PLN



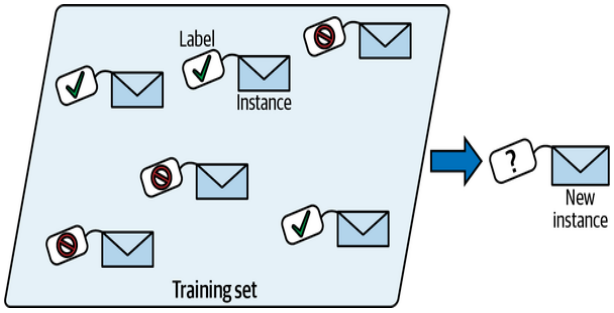
Fonte: Adaptado de Géron (2019).

O *pipeline* inicia com textos brutos, que passam pelo pré-processamento, incluindo limpeza, normalização, *tokenização* e remoção de *stopwords*. Em seguida,

na etapa de modelagem, é aplicado a vetorização *TF-IDF*, que transforma o texto em matrizes numéricas. A fase de extração de características organiza essas representações para alimentar a construção do modelo de classificação. Após o treinamento, o sistema realiza a etapa de predição, permitindo classificar automaticamente novos textos de acordo com os rótulos definidos.

O aprendizado supervisionado utiliza dados rotulados, ou seja, exemplos de entrada já associados às saídas corretas. O algoritmo aprende a mapear essas instâncias para seus respectivos rótulos, conforme apresentado na Figura 5.

Figura 5 – Um conjunto de treinamento rotulado para aprendizado supervisionado

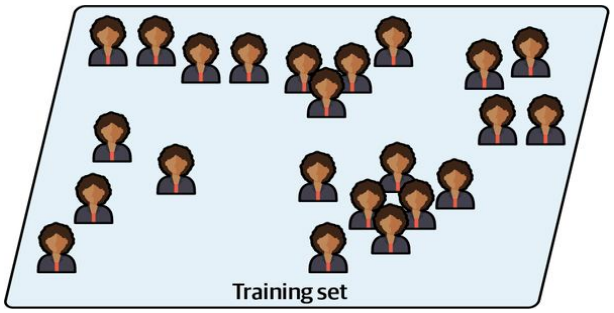


Fonte: Gerón (2020)

No caso de filtros de *spam* treinados para classificar *e-mails* como “*spam*” ou “não *spam*”. Essa abordagem é consolidada em algoritmos como *KNN*, regressão linear e logística, *SVM, Random Forest* e redes neurais. Russell e Norvig (2010) destacam que o objetivo é induzir uma função de classificação a partir de pares conhecidos. Em redes neurais, esse aprendizado ocorre pelo ajuste iterativo dos pesos com base no erro entre a previsão e o rótulo real, por meio do processo de retropropagação (Haykin, 2001).

O aprendizado não supervisionado, por sua vez, trabalha com dados não rotulados, buscando identificar estruturas ocultas ou agrupamentos de forma autônoma apresentado na Figura 6.

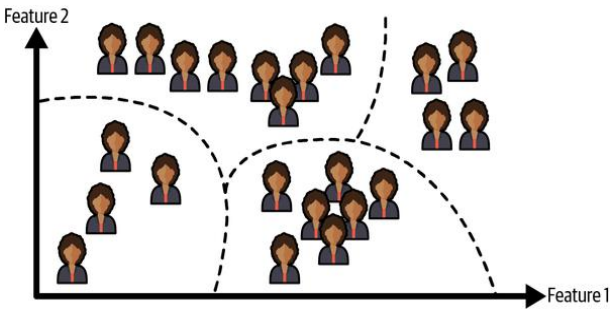
Figura 6 – Conjunto de treinamento não rotulado para aprendizagem não supervisionado.



Fonte: Gerón (2020).

Algoritmos como *Clustering, K-Means, PCA* e *t-SNE* exploram padrões sem conhecimento prévio de saídas, permitindo, por exemplo, separar indivíduos em grupos conforme suas características apresentadas na Figura 7.

Figura 7 – Agrupamento.

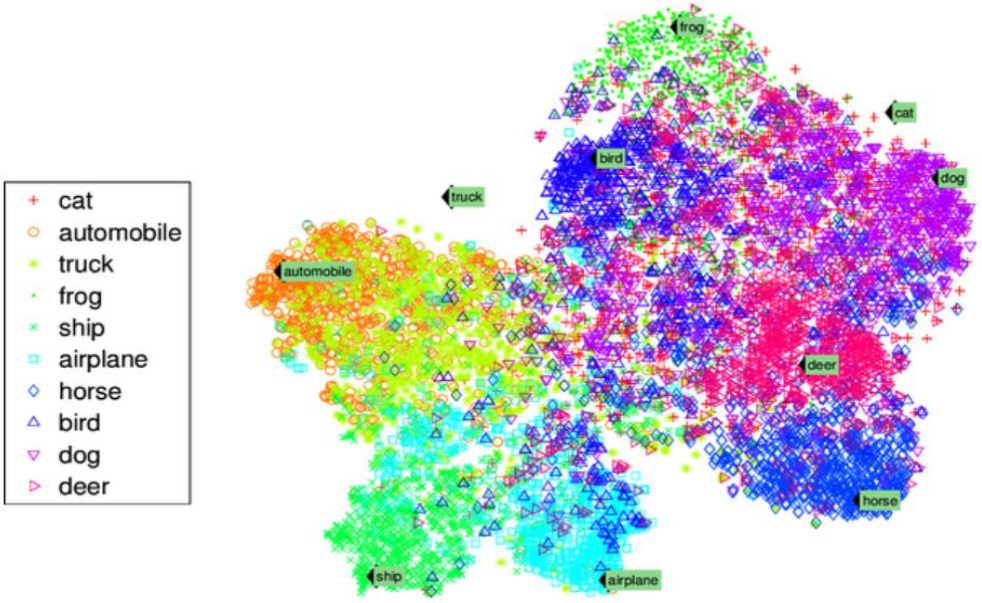


Fonte: Gerón (2020)

Os métodos de agrupamento podem ser usados para identificar grupos de dados significativos. Bruce (2019).

Visualizações como mapas de redução de dimensionalidade, também são bons exemplos de algoritmo de aprendizado não supervisionado conforme apresentado na Figura 8.

Figura 8 – Exemplo de uma visualização *t-SNE* (*(t-distributed Stochastic Neighbor Embedding)* destacando grupos semânticos



Fonte: Gerón (2020).

Esses modelos permitem identificar como os dados se organizam e revelar padrões que passariam despercebidos. Mitchell (1997) destaca sua importância em tarefas exploratórias, e Russell e Norvig (2010) reforçam sua utilidade em cenários com muitos dados e poucos rótulos. Exemplos como os mapas auto-organizáveis de Kohonen (*SOM*) (Haykin, 2001) ilustram essa capacidade. Enquanto o aprendizado supervisionado é indicado quando as respostas são conhecidas, o não supervisionado foca a descoberta de padrões e agrupamentos, sendo ambos fundamentais para aplicações de AM em PLN, visão computacional e análise preditiva.

3.2.2 TF-IDF

A representação numérica de textos é uma etapa fundamental no processamento de linguagem natural, e uma das estratégias mais utilizadas para esse fim é a abordagem *TF-IDF (Term Frequency – Inverse Document Frequen*cy). Ela combina dois componentes estatísticos complementares que buscam medir a importância de um termo dentro de um documento e dentro de um conjunto de documentos.

O primeiro componente, *Term Frequency (TF)*, expressa a frequência relativa de um termo *t* em um documento *d*. Segundo Jurafsky e Martin (2022), a frequência pode ser normalizada pela divisão entre o número de ocorrências do termo e o número total de termos no documento, conforme representado na Equação 1:

$$TF(t) = \frac{t}{N} \tag{1}$$

Onde *t* representa o número de termos e *N* a quantidade de documentos

Esse valor indica o quanto um termo é representativo dentro de um único documento.

O segundo componente, *Inverse Document Frequency* (IDF), busca reduzir a influência de termos muito frequentes no *corpus* como palavras comuns que não ajudam na diferenciação entre documentos. O IDF avalia a raridade do termo *t* na coleção de documentos e é definido pela Equação 2:

$$IDF(j) = \log\left(\frac{N}{df(j)}\right) \tag{2}$$

Onde $\log$ aplica uma escala logarítmica para suavizar valores muito altos e $df(j)$ corresponde ao número de documentos em que o termo aparece. Como destacam Manning, Raghavan e Schütze (2008), quanto mais raro for o termo, maior será seu peso no cálculo.

A partir desses dois componentes, é obtido o valor do $TF\text{-}IDF$, calculado pela Equação 3:

$$TF\text{-}IDF(j) = TF(j) \times IDF(j) \tag{3}$$

Onde $TF\text{-}IDF(j)$ representa a relevância final do termo j, $TF(j)$ quantas vezes o termo j aparece no documento e $IDF(j)$ mede o quanto o termo j é raro ou informativo no conjunto total de documentos. Palavras muito comuns têm IDF baixo, palavras raras têm IDF alto, sendo assim permite identificar quais palavras são mais relevantes para distinguir um documento dentro de um conjunto maior (Mitchell, 1997; Russell & Norvig, 2010).

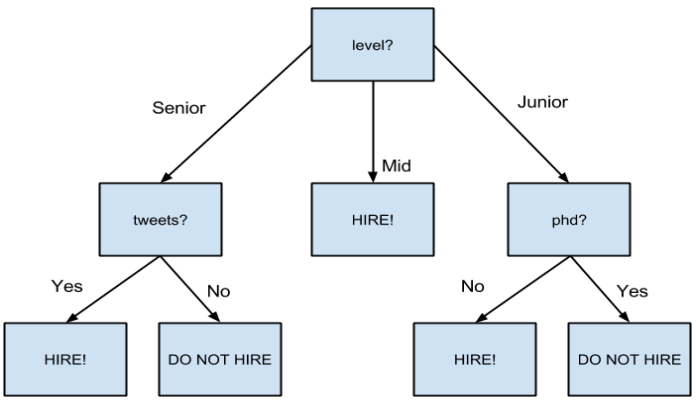
O *TF-IDF* é um indicador estatístico que mede a importância de um termo no documento e no corpus, permitindo transformar linguagem natural em vetores interpretáveis. Essa técnica é amplamente usada em classificação supervisionada, mineração de texto e análise de sentimentos (Jurafsky & Martin, 2022).

3.2.3 RANDOM FOREST

Segundo Mitchell (1997), AM ocorre quando um sistema melhora seu desempenho em uma tarefa (T) a partir de uma experiência (E), avaliado por um critério de desempenho (P). Na análise de sentimentos aplicada ao clima organizacional, a tarefa (T) consiste em classificar sentimentos em respostas abertas; a experiência (E) é o conjunto de dados rotulados manualmente, e o desempenho (P) é medido por métricas como acurácia, precisão, recall e *F1-score*.

O *Random Forest* aplica essa formulação ao gerar múltiplas árvores de decisão a partir de diferentes amostras e variáveis, aprendendo padrões linguísticos associados aos sentimentos. A combinação das previsões produz resultados mais robustos e adequados a dados de clima organizacional, que costumam ser ruidosos, subjetivos e de alta dimensionalidade. Dessa forma, o modelo permite inferir sentimentos e percepções coletivas a partir da linguagem natural. Grus (2019) apresenta o funcionamento de uma árvore de decisão, ilustrado na Figura 9, nas quais a classificação é feita por meio de perguntas binárias sobre atributos, o que favorece a transparência e a interpretação do processo decisório.

Figura 9 – Árvore de decisão para contratação



Fonte: Grus (2019).

Internamente, cada árvore do *Random Forest* é composta por nós de decisão e nós folha. Os nós de decisão representam testes binários sobre os atributos dos dados, por exemplo “a frequência da palavra ‘pressão’ é maior que 3?”. Cada divisão ocorre em um nó, separando os exemplos conforme critérios de pureza, como *gini impurity* ou entropia. Os nós folha, por sua vez, representam a decisão final da árvore, atribuindo um rótulo de classe, como positivo, neutro ou negativo.

Do ponto de vista da Inteligência Artificial, modelos baseados em árvores incluindo o *Random Forest* são classificados como IA Fraca ou estreita, pois executam apenas a tarefa específica de classificação aprendida a partir de exemplos rotulados. No contexto funcional de Russell e Norvig (2010), é considerado também de sistemas de IA Reativa, uma vez que respondem a entradas atributos textuais vetorizados, como *TF-IDF* produzindo uma saída de rótulos de sentimento sem memória, consciência ou representação interna do ambiente. Esse enquadramento reforça que, embora sejam modelos estatísticos simples, eles fazem parte das técnicas de IA aplicadas ao processamento de linguagem natural.

Em uma única árvore de decisão, há risco de *overfitting*, pois ela pode memorizar padrões específicos do conjunto de treino. O *Random Forest* reduz esse problema ao construir múltiplas árvores com diferentes subconjuntos de dados e variáveis, combinando suas previsões por votação majoritária, o que diminui a variância e aumenta a robustez do modelo (Breiman, 2001).

Proposto por Breiman (1994), o *Random Forest* é um algoritmo de *ensemble learning* que utiliza *bagging (bootstrap aggregating)*, em que cada árvore é treinada com uma amostra aleatória com reposição. Sua principal inovação é a aleatoriedade na seleção de variáveis em cada divisão, o que reduz a correlação entre as árvores e melhora a capacidade de generalização do conjunto.

A evolução do *Random Forest* se apoia em décadas de pesquisa sobre algoritmos baseados em árvores:

- 1980: surgimento dos primeiros algoritmos de árvores de decisão, como o *CART (Classification and Regression Trees),* de Breiman et al. (1984).
- 1994: Breiman propõe o *bagging*, visando reduzir variância e aumentar a estabilidade dos modelos.

- 2001: surge o *Random Forest*, combinando *bagging* e aleatoriedade na seleção de atributos, menos propenso a *overfitting*.
- Pós-2001: diversas extensões surgem, como:
    - maior aleatoriedade nos limites de divisão (Geurts et al., 2006);
    - *Quantile Regression Forests*, úteis para estimativas de intervalo (Meinshausen, 2006);
    - *Random Survival Forests*, aplicados à análise de sobrevivência (Ishwaran e Kogalur, 2007);
    - *Oblique Random Forests*, com divisões não ortogonais.

Russell e Norvig (2022) apontam que árvores de decisão são intuitivas, mas propensas ao *overfiting*. O *Random Forest* reduz essa limitação ao introduzir variabilidade entre as árvores, gerando modelos mais generalizáveis. Haykin (2009) destaca a eficácia de *ensembles* em ambientes ruidosos, como textos subjetivos, e Grus (2019) reforça a importância de modelos interpretáveis e replicáveis, características em que o *Random Forest* se destaca.

Sob a perspectiva da inteligência coletiva, Castro (2007) argumenta que sistemas como o *Random Forest* se alinham ao paradigma de cooperação entre múltiplos aprendizes, o que ressoa com o conceito de clima organizacional, construído a partir da percepção agregada de indivíduos.

Ferrari e West (2017) enfatizam que a eficácia de modelos preditivos depende tanto do algoritmo quanto da estrutura dos dados e do objetivo da previsão. Na análise de sentimentos aplicada ao clima organizacional, o *Random Forest* permite identificar padrões linguísticos, prever riscos como queda de engajamento ou aumento de turnover e apoiar decisões estratégicas de gestão de pessoas.

Na prática, a aplicação do *Random Forest* começa pela transformação dos textos em dados estruturados, utilizando técnicas como *TF-IDF*, *word embeddings* (*Word2Vec, GloVe*) ou agregações por pergunta (Jurafsky e Martin, 2022). O modelo é treinado com base rotuladas manualmente, caracterizando um cenário clássico de aprendizado supervisionado (Mitchell, 1997). A capacidade do *Random Forest* de identificar variáveis com maior poder discriminativo como termos, expressões ou padrões linguísticos fornece *insights* valiosos sobre pontos críticos (Breiman, 2001).

Mesmo em um cenário em que modelos baseados em *Transformers*, como o *BERT*, ganham destaque no PLN por capturarem contextos profundos (Tunstall, von Werra & Wolf, 2022), o *Random Forest* permanece amplamente utilizado, principalmente em organizações que priorizam interpretabilidade, rapidez e menor demanda computacional (Grus, 2019; Haykin, 2009).

3.2.4 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Processamento de linguagem natural (PLN) é o subcampo da inteligência artificial dedicado ao estudo de métodos que permitem às máquinas compreender, interpretar e manipular linguagem humana. Combina linguística computacional, estatística e aprendizado de máquina para lidar com tarefas como análise de sentimentos, classificação de texto, tradução automática, sumarização e resposta a perguntas. (Jurafsky e Martin, 2022; Géron, 2019; Grus, 2016). Para Caseli e Nunes (2024) o adjetivo “natural”, é uma referência às línguas faladas pelos humanos, se diferenciando de outras formas de linguagem, como notações matemáticas, códigos de programação, sistemas gestuais ou representações visuais. A área evoluiu ao longo das décadas, passando por diferentes paradigmas nos anos 1980 com o uso de regras linguísticas (paradigma simbólico), em 1990 com o uso de modelos estatísticos (paradigma estatístico) e atualmente a adoção de técnicas de *deep learning* (paradigma neural), marcadas por maior capacidade de generalização e eficiência. Uma linha do tempo ampliada da evolução do PLN, baseada em Jurafsky e Martin (2022) e complementada por Géron (2019), evidencia a transição entre métodos simbólicos, estatísticos e neurais:

- 1970: Era simbólica, baseada em regras linguísticas;
- 1980–1990: primeiros modelos estatísticos;
- 1990–2010: consolidação do aprendizado estatístico, supervisionado, com *TF-IDF* combinado a (*Random Forest, SVM, Naive Bayes*);
- 2013–2017: início da aplicação do *deep learning* ao PLN;
- 2017–Presente: era dos *Transformers* e dos *LLMs*, iniciada com *Attention is All You Need* (Vaswani et al., 2017);
- 2018: Lançamento do *BERT* (Devlin et al.), modelo bidirecional pré-treinado que revoluciona tarefas como *NER, QA* e *STS*;

- 2020–2023: popularização dos *LLMs (Large Language Model*s), como *GPT-3, T5* e *LLaMA*, com avanços em texto, *zero-shot learning* e raciocínio contextual.

Quadro 3 – Resumo das eras do PLN

Período	Principal Abordagem	Exemplos de Técnicas/Modelos
1970	Simbólica (regras)	Gramáticas, dicionários, heurísticas
1980 - 1990	Estatística inicial	*N-gram, HMM (Hidden Markov Models)*
1990 - 2010	Estatística supervisionada	*TF-IDF + Random Forest, SVM, Naive*
2013 - 2017	*Deep learning* com *RNNs*	*word2vec, LSTM, GRU*
2017 - Presente	*Transformers* e *LLMs*	*BERT, GPT, T5, LLaMA*

Fontes: Russell e Norvig (2010); Mitchell (1997); Haykin (1999); Jurafsky e Martin (2022).

Russell e Norvig (2010) e Mitchell (1997) discutem as abordagens simbólicas e estatísticas clássicas do processamento de linguagem natural, abordando desde regras linguísticas até modelos supervisionados tradicionais. Já Jurafsky e Martin (2022) apresentam a evolução recente do campo, marcada pela consolidação de técnicas estatísticas, vetorização e modelos baseados em *deep learning* e *Transformers*. Haykin (1999) contribui ao descrever os fundamentos das redes neurais artificiais, que servem de base conceitual para o avanço dos modelos neurais aplicados ao PLN, como *LSTM, GRU* e, posteriormente, arquiteturas inspiradas em mecanismos de atenção.

Segundo Grus (2016), o PLN envolve "*a difícil tarefa de transformar linguagem natural ambígua e rica em contexto em algo que computadores possam manipular com precisão*", o que destaca a importância de representações estatísticas e vetoriais. Dessa forma, o PLN abrange desde métodos como *TF-IDF, LDA* e classificadores supervisionados até modelos, como *BERT* e *GPT*, permitindo transformar texto livre de comentários em *insights* quantitativos úteis.

Para Sun et al. (2021), técnicas de PLN baseadas em modelos generativos e aprendizado profundo são eficazes para interpretar grandes volumes de *feedback* textual, oferecendo suporte valioso à tomada de decisão organizacional. Uma técnica simples e amplamente utilizada para visualizar padrões linguísticos é a nuvem de palavras, que representa cada termo com tamanho proporcional à sua frequência.

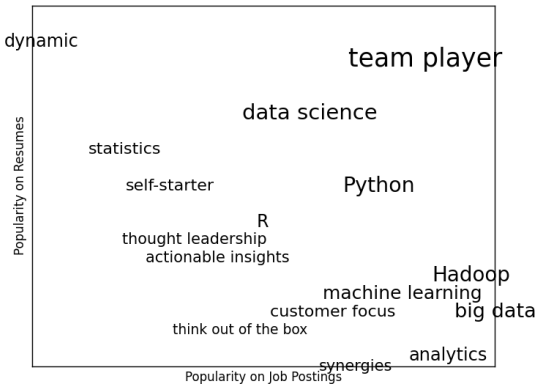
Grus (2016) exemplifica na Figura 10 o uso desse recurso ao comparar a frequência de jargões técnicos em postagens de emprego e currículos, utilizando escalas de 0 a 100 para facilitar a análise.

Figura 10 – Nuvem de jargões

Fonte: Grus (2016).

Parece interessante, mas não diz nada. Uma abordagem mais informativa seria distribuir as palavras de forma que a posição horizontal indique a popularidade nas postagens e a posição vertical indique a popularidade nos currículos, com uma visualização que transmite alguns *insights* mais claros, apresentada na Figura 11.

Figura 11 – Uma nuvem de palavras mais significativa (embora menos atraente)



Fonte: Grus (2016).

Os seres humanos captam referências sutis a emoções e sentimentos no texto, permitindo, por exemplo, compreender uma fala educada que esconde emoções negativas ou ironias. Os computadores não possuem essa capacidade de forma natural, mas contam com o PLN, um campo da ciência da computação voltado ao entendimento e à geração de linguagem entre máquina e ser humano (Massaron e Muller, 2019).

3.2.5 ANÁLISE DE SENTIMENTOS

A análise de sentimentos é uma tarefa fundamental de classificação de textos, cujo objetivo é identificar a polaridade de sentimento positivo, negativo ou neutro presente em um enunciado. Modelos baseados em *Transformers*, como o *BERT*, têm sido utilizados para essa finalidade devido à sua capacidade de capturar contexto e nuances da linguagem natural, quando aplicados com *fine-tuning* e técnicas de preparação e avaliação de dados (Tunstall, Werra e Wolf, 2022).

Contudo, a área também inclui métodos de aprendizado de máquina. Caseli e Nunes (2024) destacam que algoritmos como *SVM, Naive Bayes e Random Forest* podem ser empregados com representações de texto, como *TF-IDF*, apresentando desempenho em cenários supervisionados. Nesse contexto, a combinação *TF-IDF* com *Random Forest* constitui uma abordagem robusta e interpretável para classificação de sentimentos, pois transforma o texto em vetores de relevância estatística e utiliza o *ensemble* de árvores para identificar padrões linguísticos discriminativos (Breiman, 2001).

Além disso, técnicas de aprendizado profundo, como *LSTM*, e abordagens híbridas baseadas em *embeddings* também são utilizadas, assim como métodos que empregam léxicos de sentimentos ou redes de conceitos. A análise de sentimentos pode ser conduzida em três níveis de granularidade o documento, sentença e aspecto sendo este último particularmente desafiador para a língua portuguesa devido à escassez de recursos linguísticos especializados.

Um dos desafios desse tipo de tarefa é o desequilíbrio entre classes, comum em bases rotuladas por emoções como alegria, medo ou tristeza. Esse fenômeno pode enviesar o modelo, podendo favorecer classes mais frequentes. Para mitigar esse problema, é recomendado empregar técnicas como ajuste de pesos na função

de perda, amostragem estratificada ou aumento de dados sintéticos para classes minoritárias. No caso de *Transformers,* algumas práticas envolvem o uso do *DistilBERT* aliado à *API Trainer da Hugging Face*, monitorando indicadores como acurácia e *f1-score* ao longo do treinamento (Tunstall, Werra e Wolf, 2022).

3.3 TIPOS DE DADOS

Dados estruturados são organizados em formatos fixos, como tabelas e planilhas, o que facilita seu armazenamento e análise com ferramentas tradicionais (Kimball e Ross, 2013). Dados semiestruturados, como arquivos *JSON, XML* ou *e-mails*, não seguem uma estrutura rígida, mas contêm metadados que permitem sua interpretação (Batini, Ceri e Navathe, 1992). Já os dados não estruturados, como textos, áudios e vídeos, não possuem um formato padronizado e exigem o uso de técnicas avançadas, como IA e PLN, para uma análise eficiente (Marr, 2016; Katal et al., 2013).

Em projetos de *people analytics*, a integração desses diferentes tipos de dados é essencial para gerar diagnósticos mais completos e profundos. Conforme apresentado no Quadro 4.

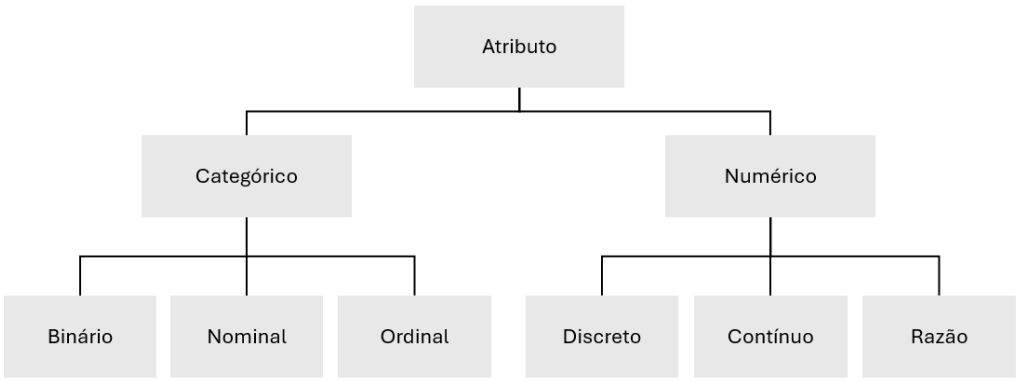
Quadro 4 – Tipos de dados com resumo comparativo

Tipo de dado	Estrutura	Exemplo	Facilidade de Análise
Estruturado	Rígida	Tabela *SQL*, planilhas	Alta
Semiestruturado	Flexível	*JSON, XML, e-mails*	Moderada
Não estruturado	Nenhuma	Texto livre, vídeo, áudio, *PDFs*	Baixa (IA/PLN)

Fonte: Elaborado pela autora (2025).

Além dos tipos de dados estruturados, semiestruturados e não estruturados, também se atribuem valores de atributos a esses dados. O valor de um atributo corresponde a uma medida associada a um determinado aspecto do dado, podendo ser numérica ou categórica, conforme ilustrado na Figura 12 (Castro e Ferrari, 2016).

Figura 12 – Tipos de atributos



Fonte: Castro e Ferrari, (2016).

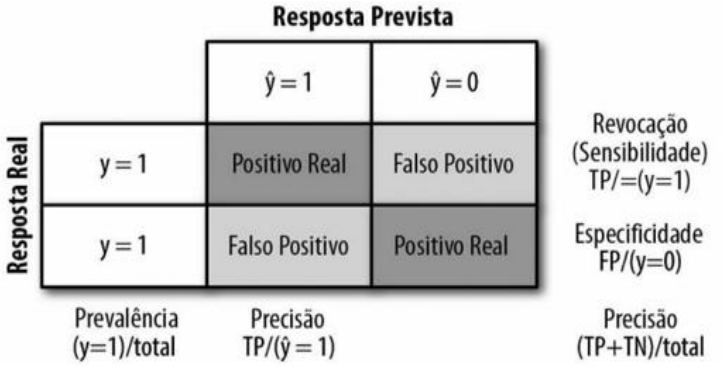
Os atributos em um conjunto de dados podem ser numéricos ou categóricos. Os atributos numéricos incluem valores discretos (inteiros) ou contínuos (reais), enquanto os categóricos assumem símbolos distintos. Entre os categóricos, há os atributos binários, que admitem apenas dois valores 0 ou 1, os nominais, com categorias sem ordem (ex.: estado civil), e os ordinais, que permitem ordenação, embora sem definir a distância entre as categorias (ex.: nível educacional). Já os atributos do tipo razão são numéricos contínuos com ponto zero natural e permitem todas as operações matemáticas usuais, como nos casos de peso, distância, velocidade e salário (Castro e Ferrari, 2016; Batini, Ceri e Navathe, 1992).

3.4 MÉTRICAS

Existem termos-chave para avaliar modelos de classificação. A precisão é o correspondente ao porcentual ou proporção de caso classificados corretamente. Já a matriz de confusão apresenta as contagens de registros segundo status de classificação previsto e real (Bruce e Bruce, 2019). Também conhecida como matriz de contingência ou matriz de erro, ela organiza nas linhas as classes originais e, nas colunas, as classes preditas (Castro e Ferrari, 2016).

Segundo Bruce e Bruce (2019), a matriz de confusão é uma tabela que apresenta o número de previsões corretas e incorretas por categoria de resposta. A Figura 13 exemplifica essa relação para uma variável binária Y e destaca as principais métricas derivadas dessa matriz.

Figura 13 – Matriz de confusão para uma resposta binária e diversas métricas



Fonte: Bruce e Bruce (2019).

Segundo Castro e Ferrari (2016), o desempenho de um classificador pode ser avaliado por meio da matriz de confusão, que relaciona as classes reais e previstas e evidencia quatro resultados: verdadeiro positivo e verdadeiro negativo, quando o modelo acerta as classificações e falso positivo e falso negativo, quando erra, correspondendo aos chamados erros tipo 1 (alarme falso) e tipo 2, respectivamente.

Existem duas taxas derivadas dessas quantidades, expressas em valores percentuais. As fórmulas correspondentes são apresentadas nas equações (4) e (5) (Castro e Ferrari, 2016).

$$TVP = \frac{VP}{VP + FN} \tag{4}$$

Sendo TVP = Taxa de verdadeiros positivos, VP = Verdadeiro Positivo, FN = Falso Negativo.

$$TFP = \frac{FP}{FP + VN} \tag{5}$$

Sendo TFP = Taxa de falsos positivos, FP = Falso positivo, VN = Verdadeiro Negativo

A TVP corresponde ao percentual de objetos positivos classificados corretamente, e a TFP corresponde ao percentual de objetos negativos classificados

como positivos. A precisão ou acurácia (ACC) é uma medida de erro total, é também conhecida como taxa global de sucesso do algoritmo, é o número de classificações corretas dividido pelo número total de classificações, apresentada na Equação (6). (Castro e Ferrari, 2016); (Bruce e Bruce 2016).

$$ACC = \frac{VP + VN}{VP + FP + VN + FN} \tag{6}$$

Sendo ACC = Acurácia, TFP = Taxa de falsos positivos, VP = Verdadeiro Positivo, VN = Verdadeiro Negativo, FP = Falso positivo, FN = Falso Negativo.

Outras duas medidas são a precisão (*precision-Pr*) apresentada na Equação (7) e a revocação (*recall-Re*) na Equação (8), também conhecida como sensibilidade, ambas associadas ao conceito de relevância, podem ser expressas também em termos VP, FP e FN, sendo que a precisão mede entre as previsões positivas, quantas estão corretas, e a exatidão do algoritmo e o *recall* mede entre as amostras realmente positivas, quantas foram corretamente identificadas. Castro e Ferrari (2016), Bruce e Bruce (2016).

$$Pr = \frac{VP}{FP + VP} \tag{7}$$

$$Re = \frac{VP}{FN + VP} \tag{8}$$

Sendo Pr = Precisão, Re = *Recall*.

Outra medida importante é a medida *f1-score* apresentada na Equação (9), amplamente utilizada na avaliação de desempenho de ferramentas de busca e modelos de classificação. Essa métrica combina precisão e o *recall*, em um único indicador, variando no intervalo [0,1], e representa a média harmônica entre essas duas medidas. O *f1-score* expressa o equilíbrio entre evitar falsos positivos (FP) e falsos negativos (FN):

$$F = \frac{2 \times Pr \times Re}{(Pr + Re)} \tag{9}$$

Para avaliar o desempenho dos modelos PLN, é necessário utilizar métricas quantitativas específicas como as citadas neste estudo. A Tabela 1 descreve as principais métricas utilizadas para avaliar modelos de PLN em tarefas de classificação e que estão incluídas neste estudo.

Tabela 1 – Métricas de avaliação de modelos PLN

Métrica	Matemática	Interpretação
Acurácia	$\mathrm{ACC} = \frac{VP + VN}{VP + FP + VN + FN}$	Proporção de previsões corretas em relação ao total
Precision	$Pr = \frac{VP}{FP + VP}$	Entre as previsões positivas, quantas estão corretas.
Recall	$Re = \frac{VP}{FN + VP}$	Entre as amostras realmente positivas, quantas foram corretamente identificadas.
f1-score	$F = \frac{2 \times Pr \times Re}{(Pr + Re)}$	Média entre precisão e *recall*. Equilíbrio entre evitar falsos positivos e falsos negativos.

Fonte: Elaborado pela autora (2025).

No contexto da análise de sentimentos para compreensão do clima organizacional, a escolha das métricas de avaliação é fundamental, sobretudo em cenários com classes desbalanceadas. Castro e Ferrari (2016) destacam que *precisão, recall, f1-score* e acurácia oferecem visões complementares do desempenho e devem ser interpretadas junto à matriz de confusão, que evidência falsos positivos e falsos negativos, erros particularmente críticos quando respostas negativas são classificadas como positivas.

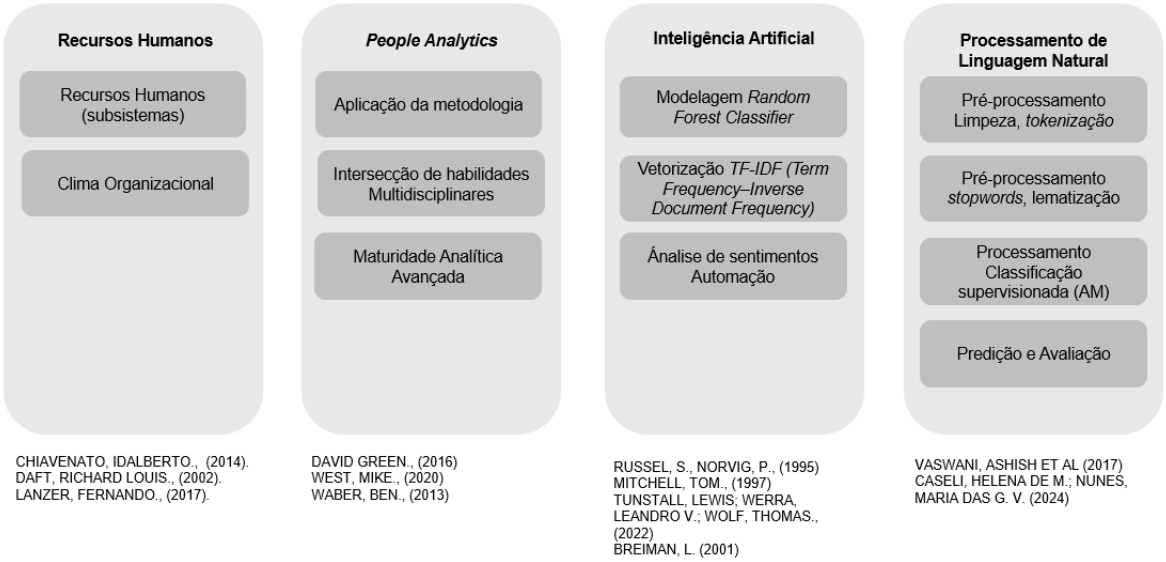
Bruce e Bruce (2016) ressaltam que a interpretação das métricas deve considerar o impacto prático dos erros, pois decisões baseadas em dados enviesados podem prejudicar nas decisões estratégicas. Provost e Fawcett (2013) reforçam que métricas isoladas, como a acurácia, podem ser insuficientes, tornando essencial uma

avaliação ampla fundamentada na matriz de confusão e nas implicações da classificação.

3.5 PRINCIPAIS AUTORES

A Figura 14 apresenta um exemplo de autores seminais para o tema de recursos humanos, *people analytics*, inteligência artificial e processamento de linguagem natural.

Figura 14 – Autores seminais para recursos humanos, *people analytics*, inteligência artificial e processamento de linguagem natural



Fonte: Elaborado pela autora (2025).

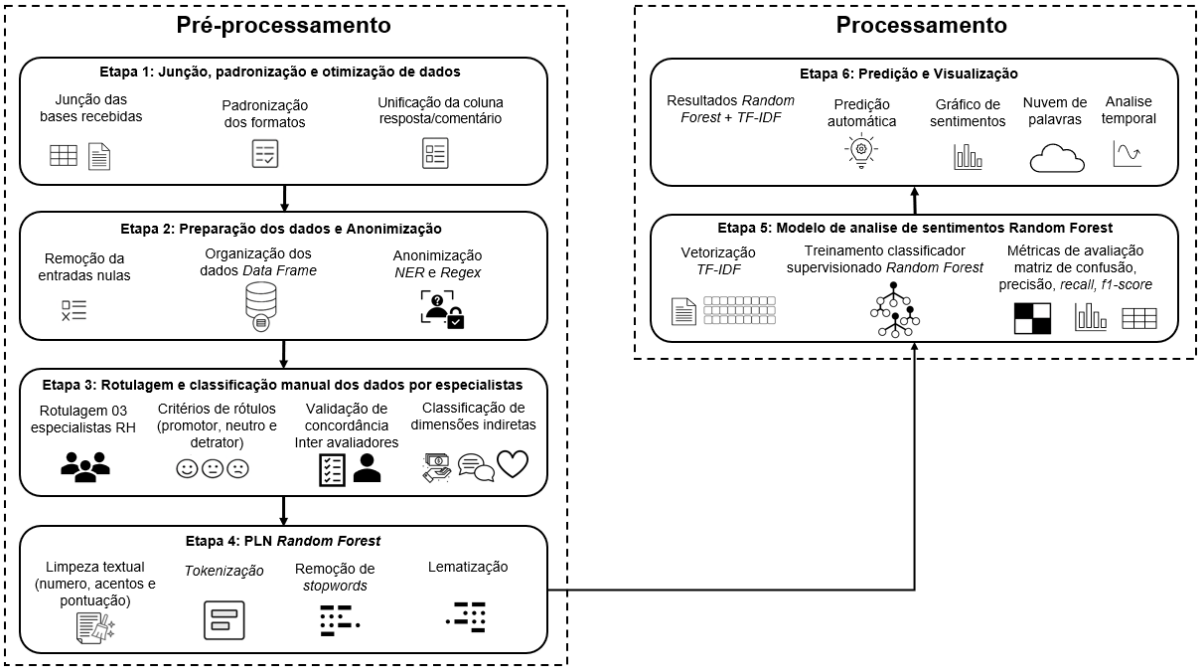
4 MATERIAIS E MÉTODOS

Esta seção descreve os materiais e métodos empregados na condução deste estudo, bem como o detalhamento dos experimentos computacionais realizados.

4.1 CARACTERIZAÇÃO DA PESQUISA

Este estudo é uma pesquisa aplicada que utiliza técnicas de PLN e IA para analisar entrevistas de desligamento e classificar sentimentos com *Random Forest*, oferecendo aplicação direta em RH, experiência do empregado e gestão de pessoas (Devlin et al., 2019). Com base no aprendizado supervisionado (Russell & Norvig, 2022), o modelo prevê categorias automaticamente e apoia decisões, contribuindo para melhorar o clima organizacional e elevar a maturidade do *people analytics* (West, 2020). A Figura 15 apresenta o fluxo metodológico adotado.

Figura 15 – Processo metodológico e experimental



Fonte: Elaborado pela autora (2025).

4.2 BASE DE DADOS

Para atender ao tema e escopo desta pesquisa, a primeira etapa foi identificar dados que tivessem estrutura e potencial para extrair características de sentimentos dos empregados, e sendo assim, os dados de resposta de pesquisas de desligamento foram mapeados com tais características, por conter dados de texto em respostas espontâneas e abertas, no idioma português do Brasil.

A empresa pesquisada pertence ao segmento financeiro, atuando como banco digital, possui 152 empregados e deste 2018. Os dados deste estudo foram disponibilizados no ambiente *Microsoft SharePoint,* no formato *.xlsx*, coletadas pela empresa por meio de formulário *Google* entre 2022 e 2023 e, posteriormente, pela plataforma de gestão de pessoas *Elofy*, no período de junho de 2023 a junho de 2025. As bases contêm variáveis de texto aberto relativas a motivo da saída, dificuldades encontradas, sugestões e pontos positivos. Do total de entrevistas, 36 (31%) foram conduzidas por um profissional de RH juntamente com o empregado.

O conjunto de dados inclui uma amostra de 117 respondentes, entre 166 empregados desligados no período, indicando que 70% dos desligados responderam a entrevista. O número de perguntas variou de 5 a 16 ao longo dos anos, sendo posteriormente padronizado para 18 perguntas, uma vez que algumas possuíam o mesmo contexto. A base de dados contém 866 registros de texto validos e 127 registros vazios, totalizando 739 registros (linhas) e 14 variáveis (colunas): com ID empregado desligado, Cargo, Time, Data desligamento, Data envio pesquisa, Status, Data resposta, Categoria desligamento (voluntário ou involuntário), Time gestor, Modo envio pesquisa, ID Pergunta, Pergunta, Resposta e Comentário. Após a anonimização, a variável “pergunta” foi convertida em *pergunta_anon*, enquanto “resposta” foi concatenada ao comentário, originando *resposta_anon*.

Posteriormente, foram criados os campos de ID da Pergunta para todas as questões, a escala de sentimento Likert (1-5), escala geral (0-10) e os sentimentos classificados em detrator, neutro e promotor. Também foram definidos critérios de desempate, consenso, nota média (sem consenso) e dimensões indiretas. As variáveis não apresentadas na amostra do Quadro 5 foram excluídas por apresentarem entre 37% e 78% de dados nulos e por não serem relevantes para os objetivos desta pesquisa.

Quadro 5 – Amostragem dos dados utilizados

ID empregado desligado	Data desligamento	ID Pergunta	pergunta_anon	resposta_anon	Comentário
275060	02/05/2025	74199	Fique à vontade para colocar aqui mais esclarecimentos sobre suas respostas acima, caso ache necessário	Complementou dizendo que a empresa nunca gostou da área de dados, se sentiu bem quando o pessoa foi o líder direto e direcionou o time.	
71739	02/05/2025	74194	Qual é o motivo de sua saída da empresa?	Proposta mais atrativa.	Como dito em minha carta de demissão, recebi uma proposta que se alinha bem com questões pessoais, além de representar certo ganho financeiro
71739	02/05/2025	40957	O quanto você recomendaria a empresa	É uma empresa com potencial de crescimento e boas pessoas para trabalhar com bons benefícios	
71739	02/05/2025	40958	Como você avalia a sua relação com seu líder nos últimos meses?	2	Faz algum tempo que não tenho líder direto, mas o pessoa que ficou nessa posição temporariamente é uma boa referência. A nota é mais sobre não ter tido líder nesse tempo do que sobre ele.
71739	02/05/2025	40959	Como você avalia a sua relação com seus colegas nos últimos meses?	5	São pessoas incríveis, com muito potencial e vontade de crescer junto com a empresa.

Considerando as características formais segundo a categorização padrão em ciência de dados, a partir da amostragem de dados, a base possui múltiplas variáveis como notas de dimensões diretas escala Likert, sentimentos, texto das respostas entre outras configurando um conjunto de dados multivariado, caracterizando um problema de classificação supervisionada. Embora a base tenha sido adaptada para a análise de consenso entre especialistas, o foco central permanece na classificação de sentimentos. Os atributos incluem predominantemente textos de perguntas e respostas, notas inteiras nas escalas 1–5 e 0–10 e a variável de sentimento categórica, combinando assim dados reais, categóricos, inteiros e textuais. A base deriva de avaliações de empregados no contexto de RH, clima organizacional e *eNPS*, sendo assim se alinha a estudos de ciências sociais aplicadas e comportamento organizacional. O formato dos dados é tabular, com estrutura matricial composta por linhas representando cada observação (resposta) e colunas com valores numéricos, textuais ou categóricos, sem presença de dados faltantes após tratamento. Resumo apresentado no Quadro (6).

Quadro 6 – Tipos de dados com resumo comparativo

Característica	**Valor**
Tipo de dado	Multivariado
Tarefa (*task*)	Classificação
Tipos de atributos	texto, inteiro, categórico
Área	ciência sociais e negócios
Formato	Tabular (*matrix*)
Valores faltantes	Não

Fonte: Elaborado pela autora (2025).

4.3 INSTRUMENTOS DE PESQUISA

Este estudo utilizou uma combinação de *softwares* e linguagens de programação para a organização, o tratamento, a anonimização, a análise e modelagem preditiva dos dados.

O *Google Collaboratory*, em *Python*, foi empregado para estruturar as respostas, realizar a engenharia de atributos e aplicar a anonimização automatizada por meio do *spaCy (pt_core_news_lg)* e de expressões regulares, removendo nomes, organizações e dados sensíveis presentes nos formulários.

Após anonimização o *Microsoft Excel (.xlsx)* foi utilizado para a inspeção visual, identificação de padrões e criação manual das dimensões, com apoio de 03 especialistas de RH. Também foi essencial na conferência das perguntas e respostas, estruturação das planilhas com escalas (1–5 e 0–10), categorização em sentimentos (detrator, neutro e promotor), e na curadoria inicial, incluindo limpeza preliminar e geração de relatórios tabulares.

O *Collaboratory,* novamente com *Python* e bibliotecas como *pandas, numpy, matplotlib* e *scikit-learn*, foi o principal ambiente de processamento e análise, foram normalizados os sentimentos (E1, E2 e E3), definidas regras de concordância para o sentimento final (E4) e conduzidas análises descritivas e visuais.

A vetorização *TF-IDF* gera representações textuais de alta dimensionalidade e elevada esparsidade, nas quais o significado surge da combinação contextual entre termos. Nesse contexto, o *Random Forest* mostra maior adequação conceitual ao capturar relações não lineares e interações complexas entre palavras, sem assumir independência entre atributos.

Embora modelos como *Naive Bayes* e Regressão Logística sejam referência em classificação textual, ambos apresentam limitações conceituais diante da natureza semântica dos dados e da esparsidade do espaço vetorial, além disso este estudo não teve como objetivo realizar uma comparação entre algoritmos, mas avaliar o impacto metodológico do uso de dados sintéticos em um mesmo modelo.

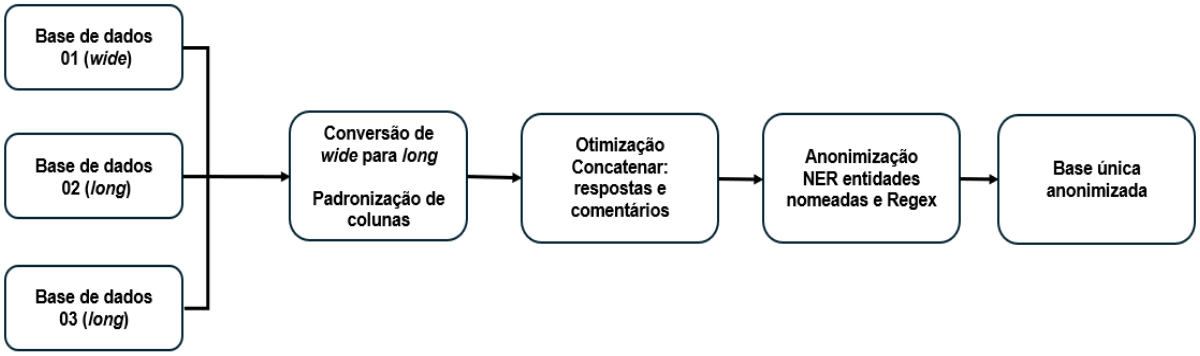
Por fim, o modelo preditivo de análise de sentimentos foi desenvolvido em *Python* com *Random Forest* e *TF-IDF*, sendo avaliado por métricas como acurácia, precisão, *recall* e *f1-score*, consolidando o rigor aplicado ao tratamento e interpretação dos dados coletados pelos instrumentos de pesquisa.

4.4 TRATAMENTO DE DADOS

Foram consolidadas três fontes: a base de dados 01, estruturada em formato largo (*wide*), e as bases de dados 02 e 03, organizadas em formato longo (*long*). A etapa de integração envolveu a transformação da base 01 de *wide* para *long*, de modo a alinhar seu padrão de estrutura com as demais, a criação de colunas ausentes, como ID da pergunta, comentário e fonte, para garantir consistência, a padronização dos nomes das variáveis, unificando campos como data de desligamento, pergunta e

resposta, e a concatenação vertical dos registros das três bases, com a adição de uma coluna de origem para rastreabilidade. Além disso, o campo de comentário foi incorporado às respostas para preservar a granularidade da análise qualitativa. Essa preparação permitiu a criação de uma base única (*df_unificado*) contendo 866 registros, sendo 127 vazios totalizando 739 registros válidos, adequada para análises quantitativas, mineração de texto, análises de sentimento, e construção de visualizações analíticas.

Figura 16 – Diagrama de unificação das bases



Fonte: Elaborado pela autora (2025).

Como o processo reúne informações sensíveis dos instrumentos de pesquisa, o estudo segue a Lei Geral de Proteção de Dados (LGPD), que define direitos dos titulares e obrigações para quem realiza o tratamento de dados pessoais (BRASIL, 2018).

Assim os dados textuais foram submetidos a um processo automatizado de anonimização. Para isso, foi utilizado o modelo pré-treinado *spaCy (pt_core_news_lg)* para reconhecimento de entidades nomeadas *(Named Entity Recognition – NER)*, com foco nas categorias *PER* (Pessoa) e *ORG* (Organização) com substituição por termos genéricos (“pessoa” e “organização”).

Adicionalmente, foi empregado expressões regulares (*regex*) para identificar e remover padrões sensíveis presentes nos instrumentos de pesquisa, incluindo endereços, *e-mail*, CPF e números de telefone, preservando a confidencialidade dos respondentes e da empresa envolvida, sem comprometer o sentido semântico dos textos. Essa etapa além de garantir a conformidade com os princípios da LGPD (Lei

nº 13.709/2018) e com boas práticas de governança de dados, também atende às disposições do *NDA (Non-Disclosure Agreement)*, firmado entre a organização participante e a pesquisadora, que autoriza o uso dos dados para fins acadêmicos desde que totalmente anonimizados, conforme previsto na cláusula 3.2 do documento.

Para preservar a origem das informações das variáveis resposta e comentário, assegurar maior granularidade na análise textual e evitar a mistura de contextos distintos, o estudo unificou respostas e comentários em um único campo. Nos casos de respostas numéricas por escala, essas notas foram registradas em novas colunas para avaliação dos especialistas, enquanto os comentários foram tratados como respostas textuais, conforme ilustrado nos exemplos (Quadro 7, Quadro 8, Quadro 9, Quadro 10 e Quadro11).

Quadro 7 – *Data frame* resposta (escala) e comentário

ID empregado desligado	**Pergunta**	**Resposta (escala)**	**Comentário**
275060	De uma escala de 0 - 10 qual nota você dá para a empresa?	“6”	“Não gostei da cultura da empresa e nem da gestão”

Quadro 8 – *Data frame* concatenado a resposta com comentário e escala em novas colunas

ID empregado desligado	**Pergunta**	**Resposta**	**Comentário**	**E1 (0-10)**	**E2 (0-10)**	**E3 (0-10)**
275060	De uma escala de 0 - 10 qual nota você dá para a empresa?	“Não gostei da cultura da empresa e nem da gestão”	...	6	6	6

Quadro 9 – *Data frame* resposta (alternativas) e comentário

ID empregado desligado	Pergunta	Resposta (alternativas)	Comentário
71739	Voltaria a trabalhar conosco?	(X) Sim ( ) Não	“Com certeza voltaria, ótima empresa!”

Quadro 10 – *Data frame* concatenado a resposta (alternativas) e comentário

ID empregado desligado	Pergunta	Resposta (alternativas)	Comentário
71739	Voltaria a trabalhar conosco?	“Sim. Com certeza voltaria, ótima empresa!”	...

Quadro 11 – *Data frame* final concatenado com resposta alternativas, comentários e novas colunas de escala (sem a coluna vazia “comentários)

ID empregado desligado	Pergunta	Respostas (escala e alternativas)	E1 (0-10)	E2 (0-10)	E3 (0-10)
275060	De uma escala de 0 - 10 qual nota você dá para a empresa?	“Não gostei da cultura da empresa e nem da gestão”	6	6	6
71739	Voltaria a trabalhar conosco?	“Sim. Com certeza voltaria, ótima empresa!”			

A unificação das bases na etapa de pré-processamento permitiu formar um conjunto de dados mais robusto, coerente e adequado às análises, centralizando informações, ampliando a amostra e possibilitando uma visão longitudinal para comparar padrões ao longo do tempo. Além disso, evitou a duplicidade de esforços

ao concentrar em um único *pipeline* as etapas de limpeza, modelagem e visualização, ao mesmo tempo em que facilitou a padronização de variáveis, perguntas, formatos de data e tipos de resposta fatores essenciais para análises de PNL e construção de indicadores.

Entre 2022 e 2025, foram identificadas 28 perguntas no total. Destas, cinco eram exclusivamente de escala (0–10) sem campo de comentário, e uma continha apenas nomes como resposta, motivo pelo qual foram excluídas da análise. Outro aspecto identificado foi que quatro perguntas apresentavam similaridade, o que levou à sua padronização para evitar redundâncias processo detalhado no Quadro 12. Após esse tratamento, houve uma redução de dez itens, resultando em 18 perguntas válidas e otimizadas. Também foram consideradas quatro perguntas de escala com campo de comentário duas no formato 0–10 e duas no formato 1–5.

Quadro 12 – Otimização de perguntas

Perguntas	**Otimização das Perguntas**
Nos conte o motivo pelo qual você está deixando a empresa	Qual é o motivo de sua saída da empresa?
Motivo da sua saída	
Motivo do desligamento	
Você voltaria a trabalhar conosco futuramente?	Você voltaria a trabalhar conosco?

As características das respostas com e sem comentários foram determinantes para manter as perguntas ou excluir da análise, alguns exemplos de perguntas ilustram os diferentes critérios. Perguntas apenas de escala sem comentário, não foram consideradas devido as características metodológicas de dados de textos necessárias para esta pesquisa. Já perguntas de escalas acompanhadas de comentários, foram incluídas na base de dados.

Para concluir o Quadro 13 apresenta as 18 perguntas finais validas e otimizadas com ID final da pergunta.

Quadro 13 – Versão final das 18 perguntas validas e otimizadas

ID Pergunta	Pergunta_anon
1A	Com uma palavra, consegue nos dizer qual sentimento em trabalhar aqui.
2B	Como você avalia a sua relação com seu líder nos últimos meses?
3C	Como você avalia a sua relação com seus colegas nos últimos meses?
4D	Conte-nos três coisas boas que você encontrou aqui na empresa
5E	Espaço aberto para sugerir, trazer feedbacks ou comentar algo que não tenha sido perguntado!
6F	Existem coisas que você gostaria de saber antes de ter começado seu trabalho aqui na empresa?
7G	Fique à vontade para colocar aqui mais esclarecimentos sobre suas respostas, caso ache necessário
8H	Há mais alguma opinião que você gostaria de compartilhar antes de encerrarmos a nossa conversa?
9I	Houve algum empecilho que possa ter impedido seu crescimento na empresa?
10J	Qual é o motivo de sua saída da empresa?
11K	O quanto você recomendaria a empresa
12L	Você voltaria a trabalhar conosco?
13M	O que você faria para melhorar a situação que o(a) está levando a sair?
14N	Quais os pontos positivos que você encontrou em trabalhar na empresa?
15O	Quais os pontos que você pensa que poderíamos melhorar?
16P	Você encontrou dificuldades nas ferramentas para a realização do seu trabalho, como softwares ineficientes ou processos mal elaborados?
17Q	Você acha que a comunicação interna era eficiente?
18R	Você acha que poderíamos melhorar a estrutura remota da empresa?

O tratamento de dados teve como objetivo transformar respostas textuais brutas de entrevistas de desligamentos em uma base estruturada, segura e utilizável por modelos de IA e PLN, o processo ocorreu em três etapas principais:

Pré-processamento e Limpeza: preparação dos dados e a rotulagem manual realizada por três especialistas de RH, que classificou as respostas nas escalas 1–5 e 0–10, atribuindo os sentimentos promotor, neutro, detrator. A concordância entre avaliadores garantiu a qualidade dos rótulos, e o consenso final foi usado como

verdade de campo (*ground truth)* no treinamento do modelo, adicionalmente foi realizado a padronização textual (letras minúsculas, remoção de acentuação e caracteres especiais), *tokenização*, *lematização* com uso de bibliotecas como *NLTK*. Remoção de *stopwords* palavras irrelevantes como “sim”, “para”, etc., filtro de palavras com menos de quatro caracteres, de forma reduzir ruído e concentrar o vocabulário em palavras mais informativas para a classificação.

Representação vetorial com *TF-IDF*: transformação dos textos pré-processados em vetores numéricos por meio da técnica *TF-IDF (Term Frequency–Inverse Document Frequency)*, utilizando ferramentas como *TfidfVectorizer* do *scikit-learn*. Essa etapa gera uma matriz esparsa em que cada linha representa um documento (resposta textual) e cada coluna representa um termo do vocabulário, ponderado por sua importância relativa no *corpus*, servindo como entrada para o algoritmo *Random Forest*.

Codificação e preparação para modelos supervisionados: conversão dos rótulos textuais de sentimento em variáveis numéricas por meio de ferramentas como *LabelEncoder*. Cálculo de pesos de classe para mitigar o desbalanceamento entre categorias detrator, neutro, promotor durante o treinamento do modelo *Random Forest*. Essa etapa inclui ainda a divisão dos dados em conjuntos de treino, validação e teste, garantindo que o desempenho do modelo seja avaliado de forma robusta.

4.5 LIMITAÇÕES DA PESQUISA

Apesar dos avanços proporcionados por esta pesquisa, algumas limitações devem ser reconhecidas. Primeiramente, o estudo foi conduzido com base em dados provenientes de uma única organização do setor financeiro, o que pode restringir a generalização dos resultados para outros setores ou culturas organizacionais. Além disso outra limitação diz respeito à natureza dos dados analisados, restritos ao conteúdo textual de entrevistas abertas, sem a inclusão de outros elementos de comunicação digital frequentemente presentes em ambientes corporativos, como *emojis, GIFs*, imagens ou áudios, os quais podem carregar significados importantes no contexto de sentimentos, emocional e intencional das mensagens. Essas restrições metodológicas apontam para a necessidade de ampliar o escopo da coleta e aprofundar questões regulatórias e multimodais em pesquisas futuras.

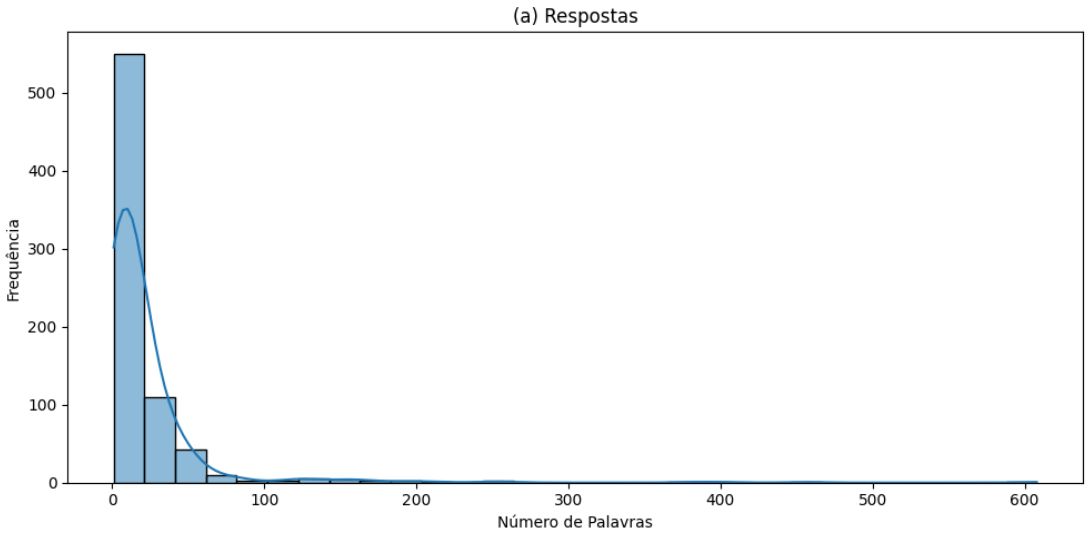
5 RESULTADOS E DISCUSÕES

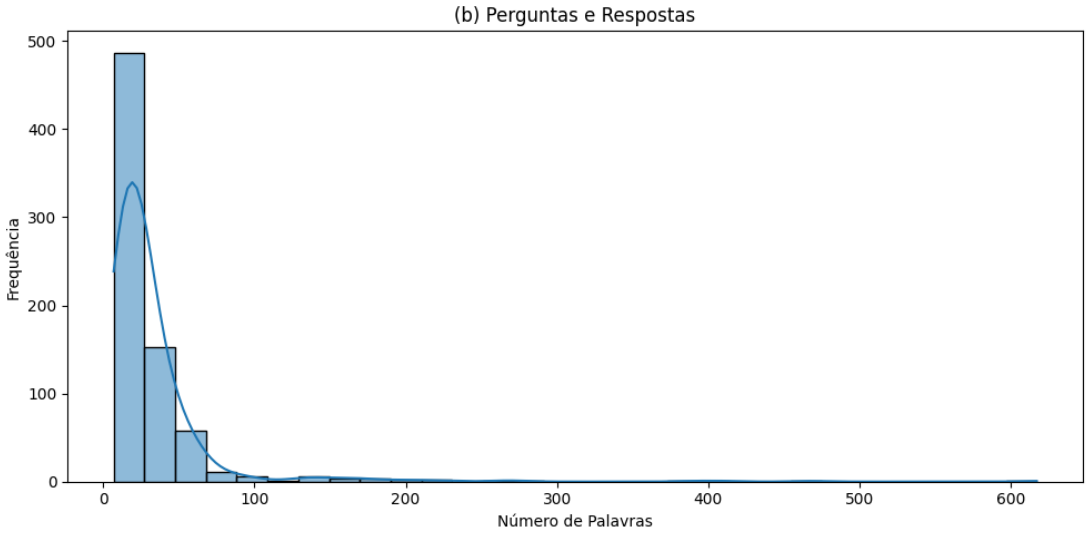
De acordo com o objetivo desta pesquisa, os resultados apresentados a seguir é sobre as análises exploratória dos dados, do processo de concordância entre as classificações realizadas por especialistas humanos e o modelo preditivo *Random Forest*, permitindo uma apresentação de resultados baseadas em processamento de linguagem natural.

Para garantindo legibilidade a pessoas com deficiências na percepção de cores e redução de vieses, foi adotada a paleta cromática nos gráficos está baseada nos princípios de acessibilidade visual da *Color Universal Design (CUD).* (OKABE; ITO, 2008). A visualização de dados cumpre seu papel analítico quando acompanhada de explicação clara e contextualizada, alinhada aos princípios de *storytelling* com dados de Knaflic (2015), que compreende a visualização como um meio de explicação estruturada.

A Figura 17 apresenta a distribuição do tamanho dos textos no conjunto de dados, comparando somente as (a) Respostas e (b) Perguntas e Respostas. Os histogramas mostram o número de palavras por registro e a frequência com que cada faixa de tamanho ocorre. Para destacar a tendência geral e suavizar a assimetria típica desse tipo de dado, foi aplicada uma curva de densidade *KDE (Kernel Density Estimation)*, permitindo visualizar melhor o comportamento da distribuição.

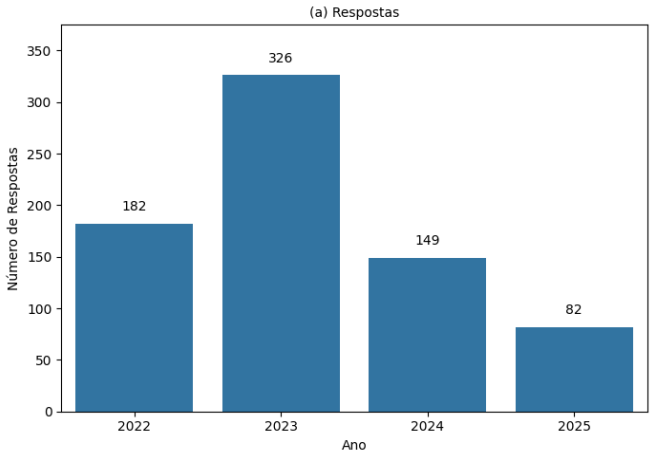
Figura 17 – Distribuição do tamanho: (a) Respostas (b) Perguntas e Respostas

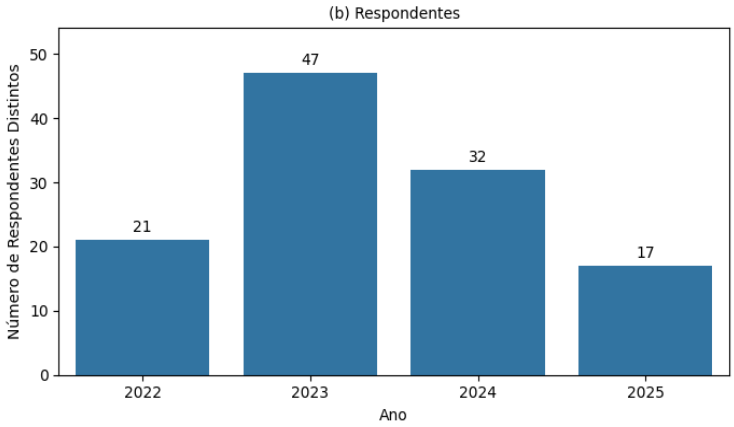




As respostas, isoladas ou combinadas com as perguntas, predominam abaixo de 50 palavras, com poucas ocorrências extensas, o que favorece a vetorização via TF-IDF ao reduzir ruído e dispersão lexical. Além disso, 139 respostas (19%) correspondem a perguntas alternativas com textos muito curtos como “sim” e “não”, “outro”, nesse cenário, a concatenação da pergunta amplia o contexto semântico sem comprometer a concisão e fortalece o sinal informativo para o modelo. Dessa forma, os gráficos indicam um cenário positivo para a análise de sentimentos com *Random Forest*, que tende a performar melhor quando alimentado por textos curtos, objetivos e contextualizados. Na Figura 18 os gráficos mostram que as variações entre anos refletem o número distinto de perguntas disponíveis e o fato de que nenhuma é obrigatória.

Figura 18 – Número por ano: (a) Respostas (b) Respondentes





Em 2022, com 9 perguntas, cada respondente respondeu quase todas as perguntas, resultando na maior média proporcional. Em 2023, apesar de haver 15 perguntas, a média caiu para cerca de sete respostas por pessoa. Em 2024, com 11 perguntas, observa-se o menor valor (4,6), indicando respostas incompletas. Já em 2025, mesmo com apenas 5 perguntas, a média de (4,8) mostra alta completude. Assim, as variações parecem refletir o número de perguntas por ano, e não mudanças no interesse dos participantes. Esse comportamento é complementado pela Tabela 2, que mostra o percentual de adesão às entrevistas em cada ano, contextualizando os volumes analisados.

Tabela 2 – Cenário anual de saídas e entrevistas (respondentes)

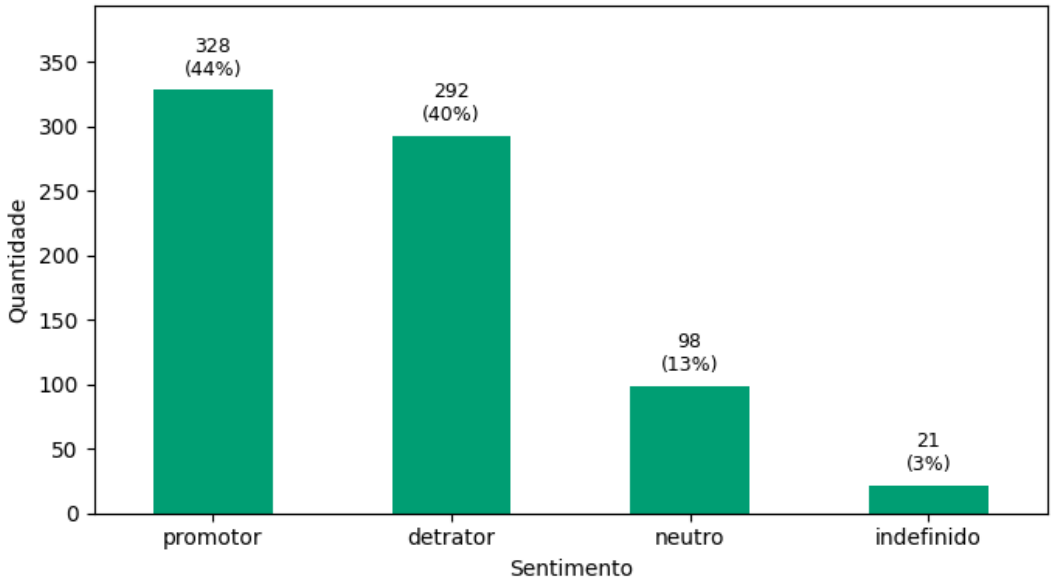
Ano	Empregados	Desligamentos	Entrevistas	Adesão
2022	168	21	21	100%
2023	158	83	47	57%
2024	144	42	32	76%
2025	152	20	17	85%

Em 2022, houve 100% de adesão às entrevistas, com as 21 saídas. Em 2023, embora tenha ocorrido o maior número de desligamentos (83), apenas 57% participaram da entrevista, representando a menor adesão do período. Em 2024, aumenta para 76% (32 entrevistas entre 42 saídas), enquanto em 2025, mesmo com 20 desligamentos, tem 85% de participação um dos índices mais elevados. Esses resultados evidenciam variações anuais na participação dos desligados, com

diferenças na adesão e na execução do processo de entrevistas de saída. O pico de desligamentos em 2023 está alinhado ao cenário no mercado brasileiro, onde a taxa de mudança de emprego aumentou de 26% em 2020 para 40% em 2025, segundo a *PMENEWS* (2025), atingindo recordes históricos conforme dados do Cadastro Geral de Empregados e Desempregados (CAGED). Além disso, pesquisas da *EQUIDAM* E *NORTHERN* (2024) mostram que *startups* podem crescer 268% nos primeiros anos e entre 144% no segundo ano e 71% no terceiro ano um ritmo que intensifica reestruturações e impacta diretamente as equipes. Assim, o volume elevado de saídas em 2023 reflete um contexto de maior mobilidade do segmento e dinamismo no mercado de trabalho.

Na etapa da Figura 19 de análise de concordância entre especialistas, foi utilizada a base unificada contendo os dados de texto e as avaliações dos três especialistas de RH (E1, E2, E3), registradas tanto na escala 0–10 quanto na escala Likert de 1–5. A partir dessas avaliações foram geradas as notas e sentimentos detrator, neutro e promotor que compõem o rótulo consensual E4. Embora a escala 0–10 tenha sido usada para cálculos internos, a pesquisa adota a escala Likert como referência principal, por sua maior aderência ao perfil do público de RH e aos objetivos do estudo.

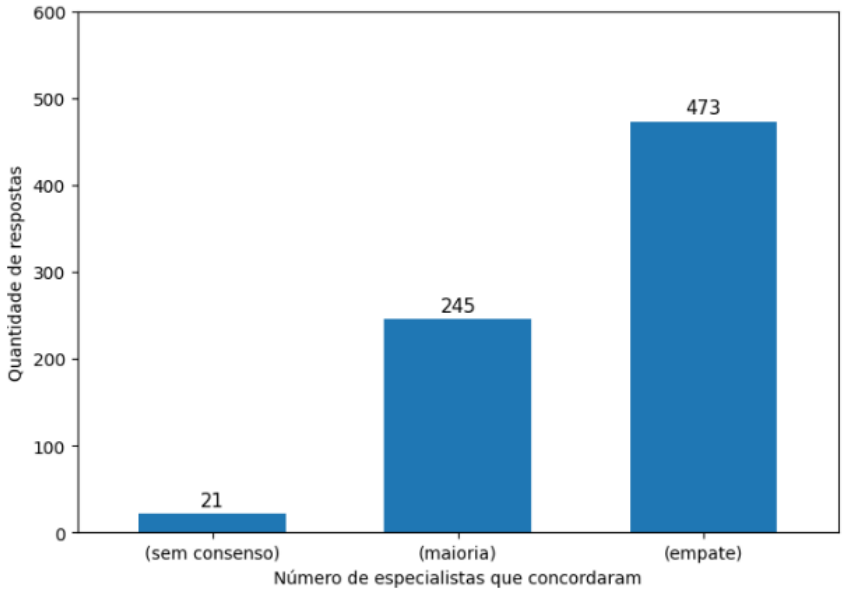
Figura 19 – Distribuição dos sentimentos (E4)



Dentre 739 respostas válidas, 21 dessas, ou seja, 3% não representavam uma concordância entre os 03 especialistas. A maioria das respostas foi classificada como promotora, seguida por detratora, o que indica uma predominância de sentimentos polarizados. O grupo neutro representa uma proporção menor, percepções intermediárias são menos frequentes nas entrevistas. Já o grupo indefinido é composto por respostas cuja interpretação variou entre os especialistas, demonstrando uma possível subjetividade presente em alguns relatos.

A Figura 20 representa a distribuição da concordância de sentimento entre especialistas na classificação de entrevistas de desligamento. Foram consideradas três situações, quando os três especialistas atribuíram o mesmo sentimento (concordância total 3/3), quando dois especialistas concordaram (maioria 2/3), e quando não houve consenso mínimo (sem consenso).

Figura 20 – Distribuição de concordância de sentimentos entre especialistas



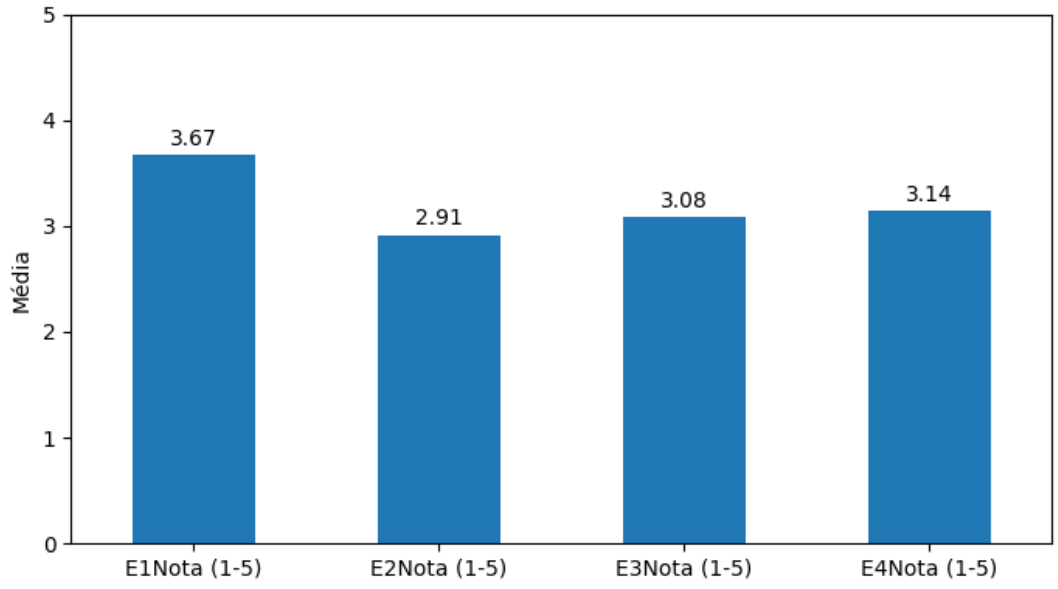
Dos 739 casos analisados, 473 respostas (64%) apresentaram concordância total entre os três especialistas, indicando consistência na avaliação dos sentimentos. Em 245 dos casos (33%), houve concordância entre dois avaliadores, sendo aplicado o critério da maioria para definir o sentimento final. Já 21 respostas (3%) registraram divergência completa entre os especialistas, sem consenso. Nesses casos, foi

utilizada a regra de cálculo médio entre escalas para estabelecer a classificação final, resultando na categoria indefinido.

O gráfico evidencia confiabilidade do processo de anotação manual, uma vez que a maior parte das classificações é sustentada por acordos entre os avaliadores. A adoção de regras para resolver divergências permitiu consolidar um conjunto de dados consistente e adequado para as etapas posteriores de análise de sentimentos.

A Figura 21 apresenta a média das notas atribuídas pelos especialistas E1, E2 e E3, bem como a nota média agregada representada por E4. Cada especialista avaliou as respostas de desligamento em uma escala de 1 a 5, onde valores mais altos indicam percepções mais favoráveis em relação ao clima organizacional. O objetivo do gráfico é demonstrar a consistência entre os avaliadores e a construção de um indicador de consenso (E4).

Figura 21 – Média de notas (1-5) por especialistas



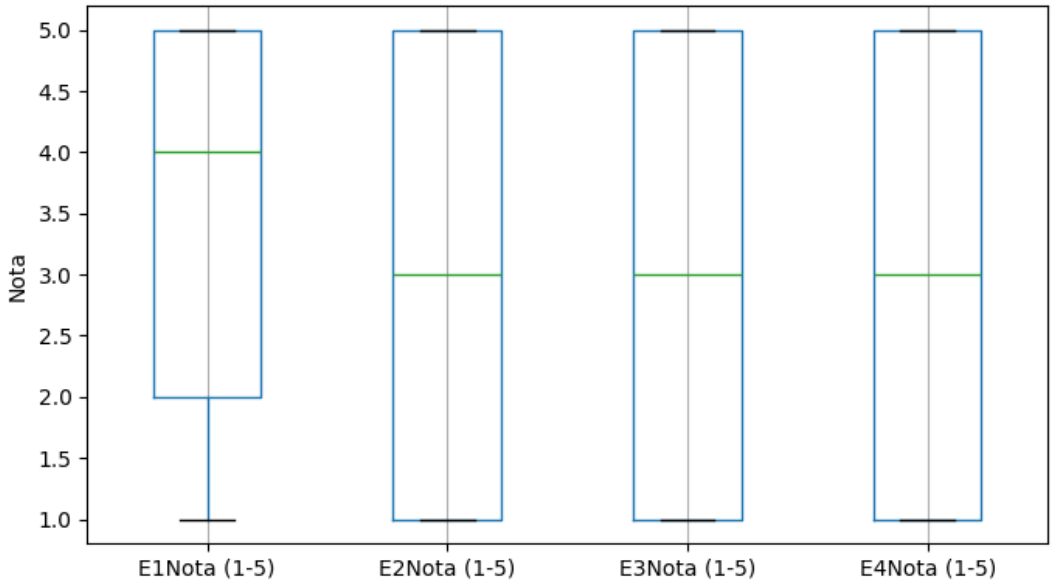
As médias das avaliações mostram diferenças no padrão de avaliação entre os especialistas. O E1 apresenta a média mais alta (3,67), indicando uma postura mais otimista na atribuição das notas. Em contraste, o E2 registra a menor média (2,91), com um perfil mais crítico nas avaliações. O E3 apresenta uma média intermediária (3,08), entre os extremos representados por E1 e E2. A nota E4, que representa o modelo consensual entre os especialistas, tem a média 3,14, entre os

valores de E1 e E2 e próxima da média de E3. Esse comportamento demonstra que a E4 atua como ponto de equilíbrio, sintetizando as avaliações individuais sem favorecer de forma desproporcional nenhum avaliador.

De forma geral, a E4 mostra uma representação adequada, reduzindo vieses individuais, uma medida estável e imparcial dos especialistas uma característica que corrobora com o gráfico 20 e reforça a importância do rótulo consensual E4 para estabilizar a base antes da modelagem.

A Figura 22 apresenta a distribuição das notas de 1 a 5 atribuídas por quatro especialistas (E1 a E4) para a dimensão E4, utilizando *boxplots*.

Figura 22 – Distribuição de notas (1-5) por especialistas



A Figura 22 mostra que todos os especialistas utilizaram toda a faixa da escala (1 a 5), com medianas entre 3 e 4, indicando variação ampla e ausência de concentração em um único nível de nota. O E1 apresenta a mediana mais alta (4), confirmando a postura mais positiva nas avaliações. Já E2 e E3 exibem medianas em 3, refletindo perfis mais equilibrados ou críticos. A nota E4, derivada das avaliações dos três especialistas, também apresenta mediana 3 e amplitude total, demonstrando que sintetiza a diversidade de percepções individuais. Assim, a distribuição confirma que o E4 atua como uma medida representativa e consistente, adequada para apoiar decisões baseadas em consenso ou em média ponderada.

A Tabela 3 apresenta a distribuição do número de palavras nas respostas classificadas por sentimento.

Tabela 3 – Estatística de palavras por sentimento (E4)

Sentimento	Quantidade Respostas	Mínimo Palavras	Média Palavras	Máximo Palavras
detrator	291	1	32.18	608
neutro	126	1	16.89	146
promotor	322	1	13.03	214

A análise do comprimento das respostas mostra diferenças entre os sentimentos. O grupo detrator apresenta as respostas mais extensas, com média de 32.18 palavras e máximo de 608 palavras, indicando que colaboradores insatisfeitos tendem a elaborar mais suas justificativas. As respostas neutras possuem comprimento intermediário, com média de 16.89 palavras e máximo de 146, sugerindo um estilo mais contido ou objetivo. Já o grupo promotor reúne o maior número de respostas 322, porém com textos mais curtos, média de 13.03 palavras e máximo de 214, revelando uma tendência de comportamento a manifestações positivas mais diretas e sucintas. Essas diferenças no tamanho textual por sentimento são relevantes para a análise de sentimentos e para a modelagem com técnicas de PLN, pois indicam padrões distintos de detalhamento conforme a polaridade de sentimento na resposta.

A Tabela 4 apresenta estatísticas descritivas sobre a quantidade de caracteres nas respostas, segmentadas pelo sentimento atribuído na avaliação E4 (detrator, neutro e promotor).

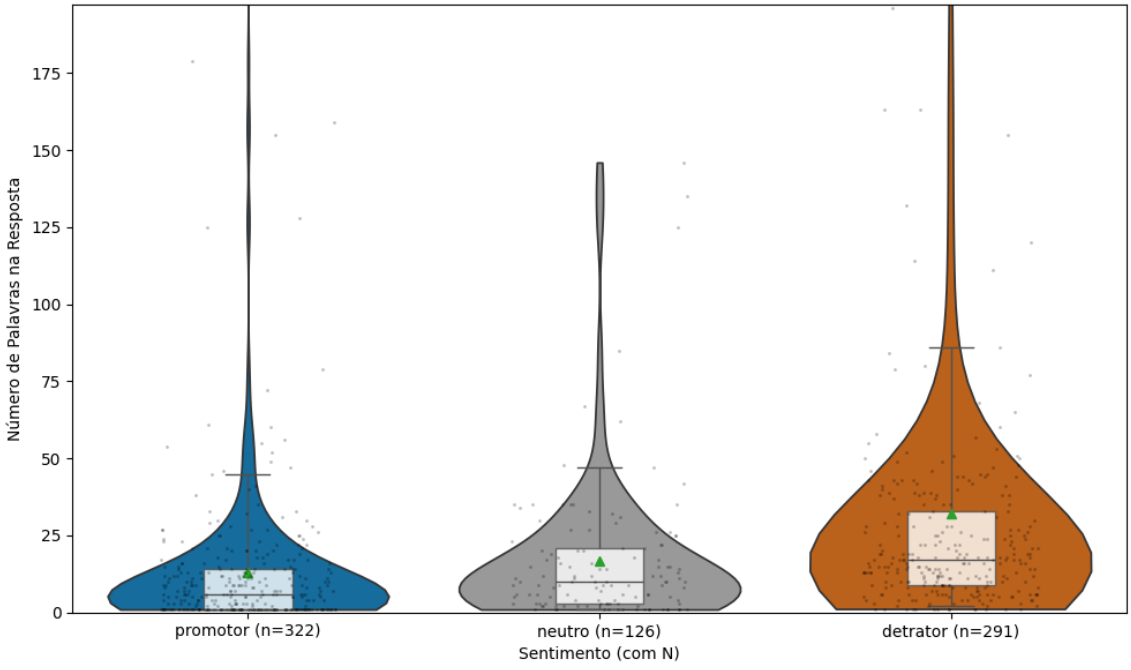
Tabela 4 – Estatística de caracteres por sentimento (E4)

Sentimento	Quantidade Respostas	Mínimo Caracteres	Média Caracteres	Máximo Caracteres
detrator	291	3	195.35	3666
neutro	126	3	99.54	808
promotor	322	3	79.37	1189

A análise do tamanho das respostas em caracteres reforça o padrão observado anteriormente na contagem de palavras. O grupo detrator apresenta os textos mais extensos, com média de 195.35 caracteres e máximo de 3666, indicando maior detalhamento nas manifestações negativas. As respostas neutras mostram comprimento intermediário média de 99.54 caracteres e as promotoras são mais curtas, com média de 79.37 caracteres. Essa diferença de densidade textual entre os sentimentos é importante para orientar ajustes de pré-processamento, como limites de truncamento, normalização do tamanho das entradas e estratégias de vetorização mais adequadas ao volume real de informação presente em cada classe.

A Figura 23 apresenta a distribuição do tamanho das respostas em número de palavras para cada sentimento (E4). Cada violino ilustra a densidade e a variação de comprimento dos textos dentro de cada grupo, combinando *violin plot, boxplot,* pontos individuais. O “n” representa o tamanho da amostra por sentimento.

Figura 23 – Tamanho de respostas por sentimento (E4)

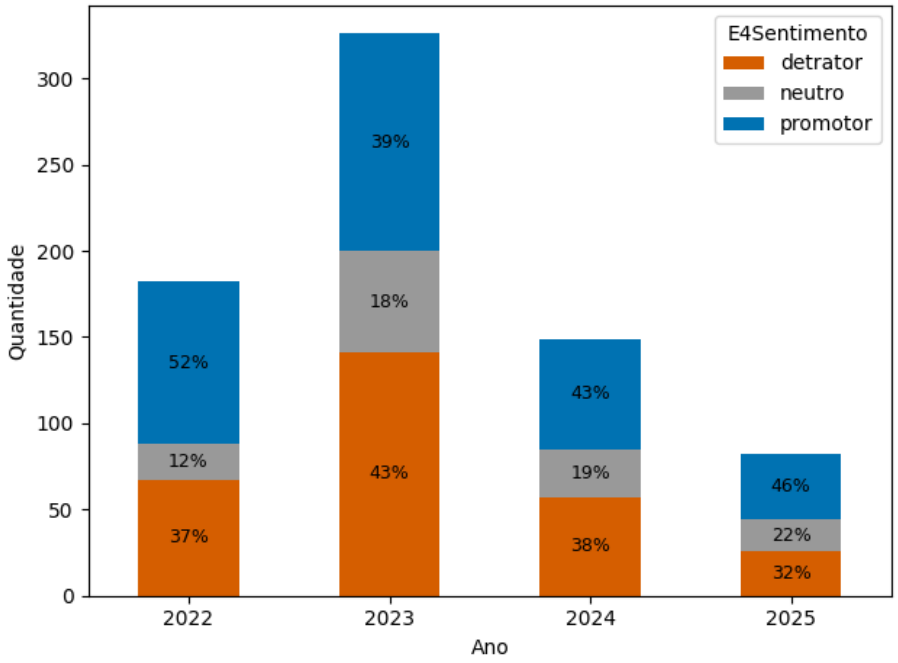


O gráfico da Figura 23 mostra que o comprimento das respostas varia conforme o sentimento detrator escrevem textos mais longos e detalhados, neutros tendem a respostas curtas e objetivas, e promotores permanecem próximos aos

neutros na mediana, mas com maior dispersão. Isso indica diferentes níveis de envolvimento emocional entre as classes.

A Figura 24 complementa essa análise com a distribuição anual dos sentimentos (E4), permitindo comparar o volume de respostas em número absoluto e a proporção percentual de cada classe ao longo do tempo.

Figura 24 – Distribuição de sentimento (E4) por ano



Em 2023 ocorre o maior volume de respostas, com predominância de sentimentos detratores, seguidos por promotores e, em menor proporção, neutros. Em 2025, o volume reduzido reflete dados parciais. Nos demais anos, as proporções entre as três categorias variam, indicando oscilações nas percepções dos colaboradores ao longo do tempo.

Essa análise temporal ajuda a identificar tendências de clima entre os desligados e apoia o RH na avaliação de impactos e mudanças internas.

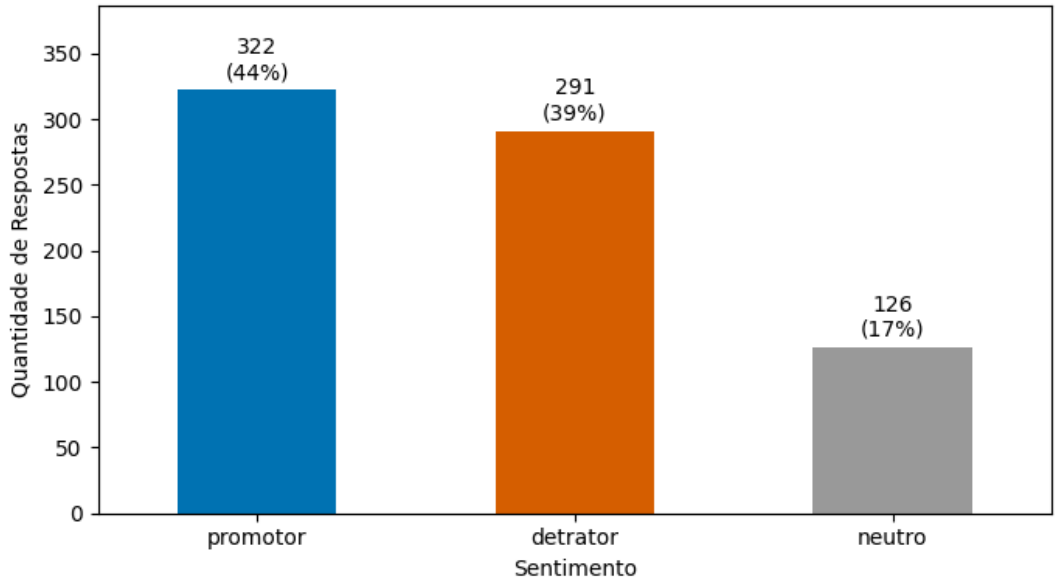
A Tabela 5 apresenta os mesmos dados do gráfico anterior, porém com valores absolutos e percentuais, permitindo uma comparação mais precisa da distribuição dos sentimentos entre os anos.

Tabela 5 – Quantidade e percentual de sentimento (E4) por ano

Ano	Detrator	Neutro	Promotor
2022	67 (37%)	21 (12%)	94 (52%)
2023	141 (43%)	59 (18%)	126 (39%)
2024	57 (38%)	28 (19%)	64 (43%)
2025	26 (32%)	18 (22%)	38 (46%)

A Figura 25 apresenta a distribuição final dos sentimentos atribuídos às respostas da pergunta do E4.

Figura 25 – Distribuição final de sentimento (E4)



A análise mostra predominância de respostas promotoras, seguidas por detratoras, enquanto as neutras representam a menor parte. Esse padrão indica uma distribuição polarizada, com pouca neutralidade e maior tendência a percepções positivas ou negativas. Essa configuração favorece modelos supervisionados de análise de sentimentos, pois há boa representatividade nas classes promotor e detrator, o que contribui para maior estabilidade e capacidade de distinção entre sentimentos opostos.

A Tabela 6 apresenta a distribuição dos sentimentos por dimensão indireta, permitindo visualizar quais temas concentram percepções mais positivas, neutras ou negativas.

Tabela 6 – Distribuição: sentimentos de respostas (E4) por dimensão indireta

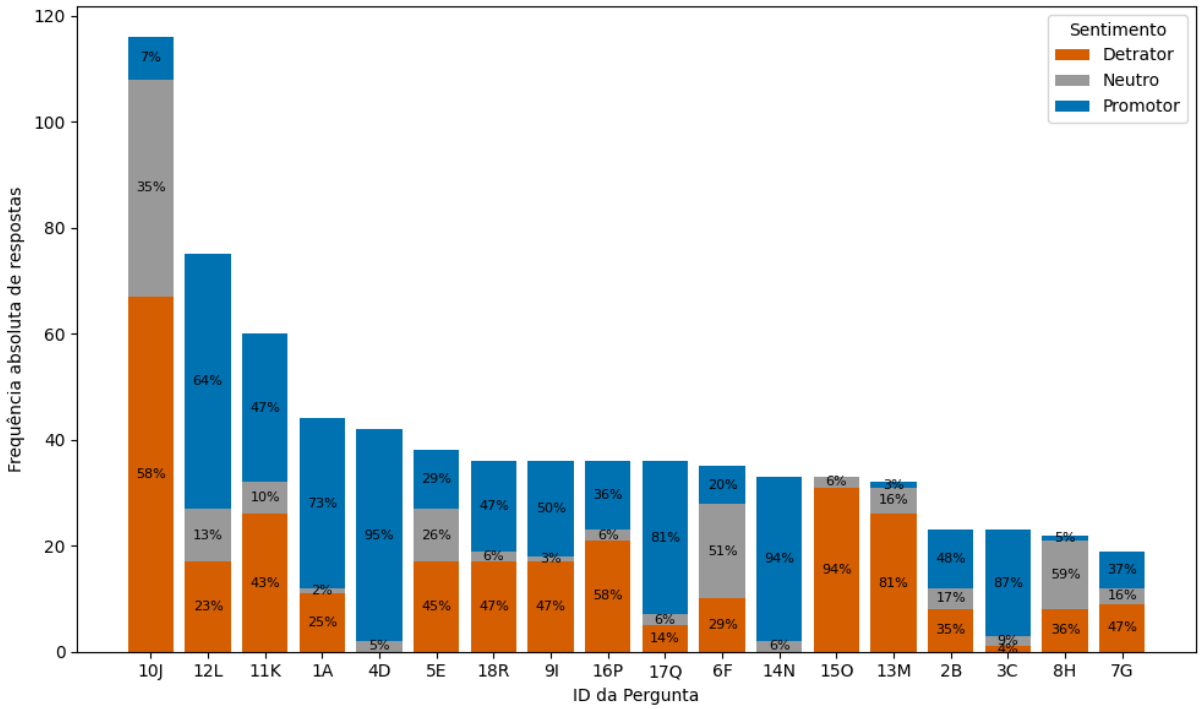
Dimensão indireta	Detrator	Neutro	Promotor
Carreira	**40**	34	35
Comunicação	19	12	**34**
Diversidade	**20**	2	1
Engajamento	40	33	**115**
Inovação	**55**	14	38
Liderança	**64**	16	25
Retenção	3	3	**24**
Saúde e bem-estar	**12**	2	6
Time	24	5	**43**
Treinamento	**14**	5	1

A análise por dimensão indireta mostra padrões distintos de sentimento. Carreira, diversidade, inovação, liderança, saúde e bem-estar e treinamento destacadas na tabela apresentam predominância de respostas detratoras, indicando maior concentração de percepções negativas. Em contraste, comunicação, engajamento, retenção e time têm maior número de respostas promotoras, refletindo avaliações mais positivas nessas dimensões, e nenhuma predominância neutra.

Cada dimensão apresenta variação interna entre detrator, neutro e promotor, o que reforça a importância de considerar a dimensão associada à resposta ou ao par pergunta mais resposta especialmente quando o texto é curto e não apenas a pergunta. Essa contextualização auxilia na interpretação mais precisa do sentimento e na identificação de nuances relevantes para o entendimento do clima organizacional.

A Figura 26 apresenta a frequência absoluta de respostas por pergunta, com indicação percentual da proporção de cada categoria de sentimento.

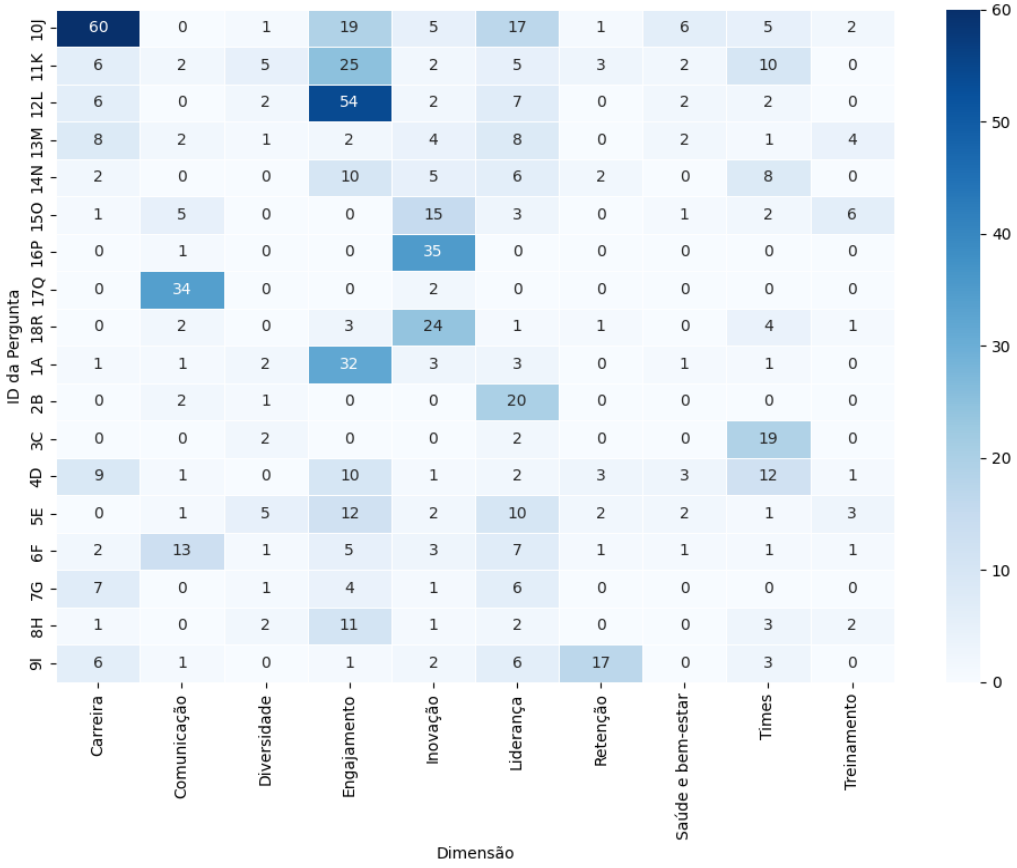
Figura 26 – Distribuição de sentimento por resposta de pergunta (E4)



As perguntas também possuem variação de sentimentos conforme as respostas. Os resultados indicam predominância de respostas detratoras nas perguntas 10J (67; 58%), 5E (17; 45%), 16P (21; 58%), 18R (17; 47%), 13M (26; 81%), 15O (31; 94%) e 7G (9; 47%). Isso significa que sete perguntas (39% do total) concentram respostas com sentimento negativo. As perguntas 6F (18; 51%) e 8H (13; 59%) apresentam predominância de sentimento neutro, enquanto nas nove perguntas restantes prevalece o sentimento promotor. Essa variação é importante para a análise de sentimentos porque indica a distribuição real das classes no conjunto de dados, orientando o tratamento de possíveis desbalanceamentos. Além disso, revela que cada grupo de perguntas gera vocabulários distintos negativos, neutros ou positivos o que enriquece as representações *TF-IDF* e melhora a capacidade do modelo de distinguir os sentimentos. Esses padrões também contribuem para validar se o modelo aprende de forma coerente com o contexto das respostas, fortalecendo a interpretação e a robustez dos resultados.

O *heatmap* apresenta a distribuição da quantidade das respostas de cada pergunta entre as diferentes dimensões avaliadas na Figura 27.

Figura 27 – Quantidade de respostas por dimensão



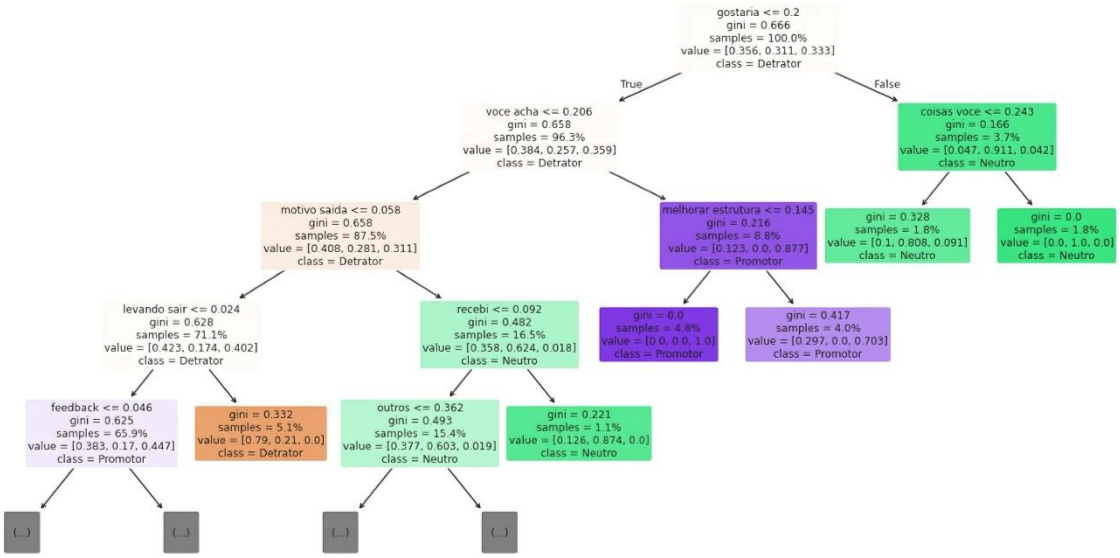
A visualização evidencia como cada pergunta se relaciona com as dimensões temáticas. Algumas perguntas apresentam forte concentração de respostas em dimensões específicas, como carreira, diversidade, engajamento e inovação, enquanto outras possuem distribuição mais transversal em áreas como comunicação, liderança, retenção e saúde e bem-estar. A maioria das perguntas, porém, mostra associação clara a uma única dimensão predominante, com poucos registros nas demais. Esse padrão indica que o questionário está estruturado e que os respondentes identificam cada pergunta com seu tema central.

5.1 AVALIAÇÃO DO MODELO

A Figura 28 apresenta a visualização da árvore de decisão do modelo *Random Forest* utilizado na classificação de sentimentos (E4), com as classes detrator, promotor e neutro. Cada nó representa uma regra de decisão baseada em valores dos vetores *TF-IDF*, que refletem a relevância ponderada de termos presentes nas

respostas. Os rótulos exibidos nos nós como “gostaria”, “recebi”, “coisas você”, “melhorar estrutura” indicam os termos que mais contribuíram para separar as classes nesse subconjunto da floresta.

Figura 28 – Árvore *Random Forest* com 3 níveis de profundidade

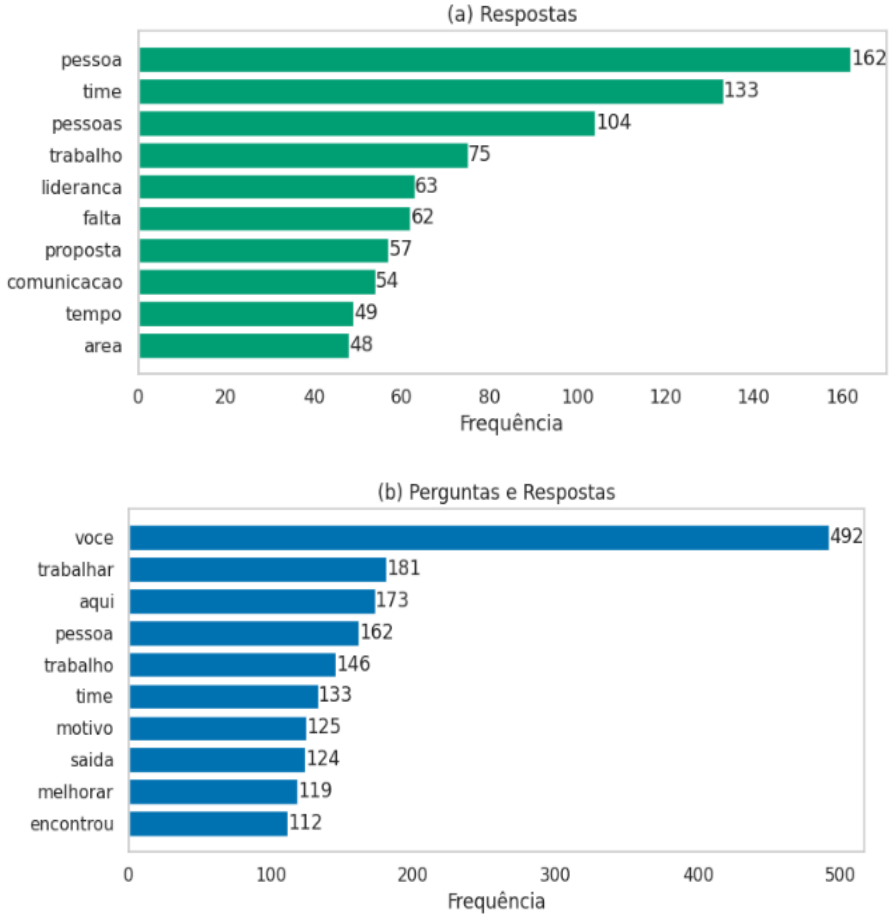


A árvore completa possui profundidade 47 e 129 nós, sua versão limitada a três níveis de profundidade facilita a interpretação. Na Figura 28, a divisão inicial (gostaria ≤ 0.206) direciona principalmente para a classe detrator, enquanto subdivisões posteriores distinguem os demais sentimentos conforme os pesos *TF-IDF*. Valores menores de *Gini* indicam nós mais puros e decisões mais consistentes.

Para o RH, essa estrutura torna visível como certas expressões influenciam a classificação, oferecendo transparência e facilitando a interpretação dos padrões linguísticos. Além de automatizar a análise de grandes volumes de texto, o *Random Forest* gera *insights* claros sobre clima, engajamento e experiência do colaborador, ampliando a capacidade analítica e apoiando decisões estratégicas.

A Figura 29 apresenta dois gráficos de barras das entrevistas de desligamento a primeira baseada apenas nas respostas e a segunda combinando perguntas e respostas. Em ambas, foram incluídas somente palavras com mais de três letras, garantindo maior relevância na identificação dos termos mais frequentes.

Figura 29 – Top 10 palavras: (a) Respostas (b) Perguntas e Respostas



Os dois gráficos revelam que os empregados utilizam um vocabulário rico e tematicamente consistente ao relatar suas percepções, especialmente sobre liderança, trabalho, área e time, além de termos emocionalmente marcados como *falta* e *melhorar*. Os dados revelam também que palavras como *trabalho* (substantivo) e *trabalhar* (verbo) aparecem separadamente mesmo após a lematização, pois seguem regras morfológicas distintas, e preserva nuances semânticas relevantes para a modelagem. Esse padrão confirma a adequação do uso de *TF-IDF* para destacar palavras discriminativas e do *Random Forest* para explorar essas diferenças na classificação de sentimentos. Assim, a análise visual reforça a coerência interna do *corpus* e sustenta a eficácia estatística do modelo adotado.

Dando continuidade a esse processo, o modelo foi treinado utilizando uma divisão de 60% dos dados para treino, 20% para teste e 20% para validação, mantendo os mesmos hiperparâmetros em todas as execuções. Para investigar o

impacto do balanceamento das classes, foi aplicado o *SMOTE* apenas no conjunto de treino, técnica que gera exemplos sintéticos no espaço vetorial *TF-IDF*. O conjunto original de 443 instâncias, 175 detratores, 76 neutros e 192 promotores foi reamostrado para 576 instâncias equilibradas, com 192 exemplos por classe, resultando em 911 características após a vetorização *TF-IDF*. Em seguida, o *pipeline* foi retreinado e avaliado sobre os conjuntos originais de teste e validação.

As métricas da Tabela 7 são resultados da avaliação de desempenho do modelo de classificação *Random Forest* sem *SMOTE* e a Tabela 8 com balanceamento *SMOTE* na fase de treinamento, considerando as três classes de sentimento.

Tabela 7 – Tabela de métricas de treinamento *Random Forest*

	Precision	***Recall***	***f1-score***	***Support***
Detrator	90%	77%	83%	175
Neutro	74%	84%	79%	76
Promotor	87%	93%	90%	192
Accuracy			86%	443
Macro avg	84%	85%	84%	443
Weighted avg	86%	85%	85%	443

Tabela 8 – Tabela de métricas de treinamento *Random Forest (SMOTE)*

	Precision	***Recall***	***f1-score***	***Support***
Detrator	89%	83%	86%	175
Neutro	84%	83%	84%	76
Promotor	87%	93%	90%	192
Accuracy			87%	443
Macro avg	87%	86%	87%	443
Weighted avg	87%	87%	87%	443

As métricas de treinamento mostram que o *Random Forest* apresenta desempenho consistente em ambos os cenários.

No modelo sem *SMOTE*, a acurácia de (86%) e os *f1-scores* para promotor (90%) e detrator (83%) indicam boa distinção entre respostas positivas e negativas, enquanto a classe neutra apresenta desempenho menor (79%) devido à sua ambiguidade e menor frequência.

Com *SMOTE*, teve um ganho na classe neutra com *f1-score* (84%) e melhora na macro e *weighted averages* (87%), mas sem mudanças relevantes nas classes promotor e detrator, que mantêm desempenho.

De modo geral, ambos os modelos capturam os padrões linguísticos, mas as melhorias trazidas pelo *SMOTE* são discretas e não se refletem de forma consistente nas etapas de validação e teste reforçando que o modelo sem *SMOTE* permanece mais adequado para representar o comportamento real dos dados.

As métricas da Tabela 9 são resultados da avaliação de desempenho do modelo de classificação *Random Forest* sem *SMOTE* e a Tabela 10 com balanceamento *SMOTE* na fase de teste, considerando as três classes de sentimento.

Tabela 9 – Tabela de métricas de teste *Random Forest*

	Precision	***Recall***	***f1-score***	***Support***
Detrator	75%	74%	75%	58
Neutro	57%	48%	52%	25
Promotor	80%	86%	83%	65
Accuracy			75%	148
Macro avg	71%	69%	70%	148
Weighted avg	74%	75%	75%	148

Tabela 10 – Tabela de métricas de teste *Random Forest (SMOTE)*

	Precision	***Recall***	***f1-score***	***Support***
Detrator	76%	74%	75%	58
Neutro	58%	44%	50%	25
Promotor	78%	86%	82%	65
Accuracy			75%	148
Macro avg	70%	68%	69%	148
Weighted avg	74%	75%	74%	148

Os resultados de teste mostram que o *Random Forest* mantém desempenho, com acurácia de (75%) em ambos os cenários. O modelo identifica as classes promotor com *f1-score* (83%) sem *SMOTE*, (82%) com *SMOTE* e detrator com *f1-score* (75%) em ambas as versões, evidenciando capacidade de distinguir sentimentos positivos e negativos. A classe neutra segue com o menor desempenho com *f1-score* (52%) sem *SMOTE*, (50%) com *SMOTE*, reflexo da frequência dessa categoria no conjunto de teste e de sua maior ambiguidade semântica. As métricas médias reforçam esse padrão o macro *f1-score* cai com *SMOTE* de (70% para 69%), enquanto o *weighted f1-score* permanece praticamente igual de (75% para 74%).

Assim, o balanceamento não produz ganhos relevantes e mantém o desempenho praticamente idêntico ao modelo original. No conjunto, ambas as versões generalizam de forma semelhante, mas o modelo sem *SMOTE* preserva melhor o comportamento observado no treinamento e se mantém mais adequado ao representar as nuances reais dos dados.

As métricas da Tabela 11 são resultados da avaliação de desempenho do modelo de classificação *Random Forest* sem *SMOTE* e a Tabela 12 com balanceamento *SMOTE* na fase de validação, considerando as três classes de sentimento.

Tabela 11 – Tabela de métricas de validação *Random Forest*

	Precision	***Recall***	***f1-score***	***Support***
Detrator	75%	60%	67%	58
Neutro	41%	56%	48%	25
Promotor	73%	76%	74%	65
Accuracy			66%	148
Macro avg	63%	64%	63%	148
Weighted avg	68%	66%	67%	148

Tabela 12 – Tabela de métricas de validação *Random Forest (SMOTE)*

	Precision	***Recall***	***f1-score***	***Support***
Detrator	70%	69%	70%	58
Neutro	52%	52%	52%	25
Promotor	73%	74%	74%	65
Accuracy			69%	148
Macro avg	65%	65%	65%	148
Weighted avg	68%	68%	68%	148

Na validação, o modelo demonstra desempenho, com acurácia de (66%) sem *SMOTE* e (69%) com *SMOTE*. Em ambas as versões, as classes promotor e detrator mantêm *f1-scores* entre (67% e 74%), indicando que o modelo consegue identificar sentimentos positivos e negativos. A classe neutra permanece com *f1-score* (48%) no modelo original e melhora para (52%) após *SMOTE*, refletindo sua baixa representatividade e maior ambiguidade semântica.

As métricas macros e *weighted* mostram que o desempenho geral cresce com *SMOTE* macro *f1-score* de (63% para 65%), *weighted f1-score* de (67% para 68%), mas sem mudanças substanciais no padrão das classes. De modo geral, os resultados indicam que o modelo generaliza adequadamente nessa etapa, mas que

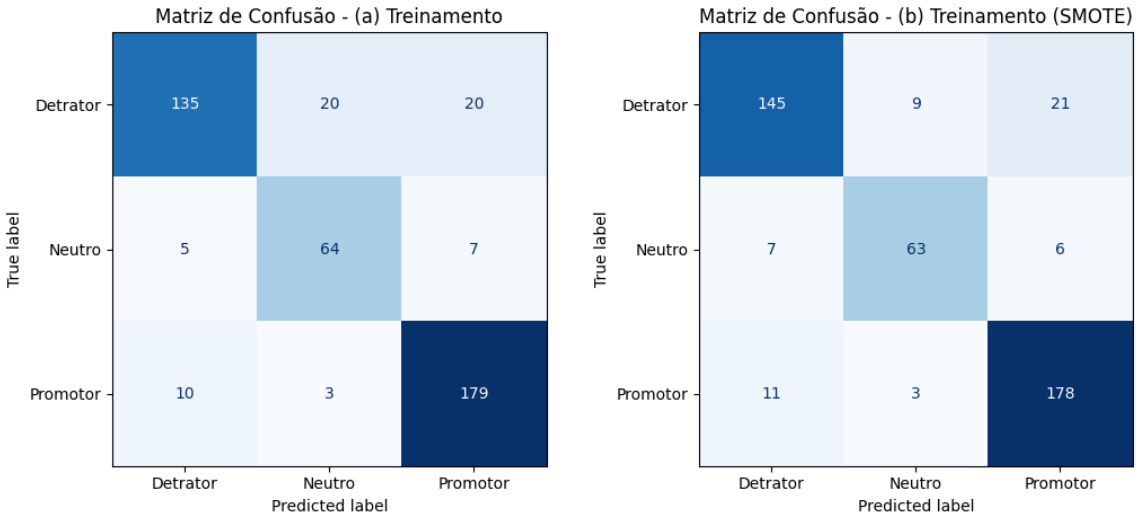
os ganhos trazidos pelo *SMOTE* são modestos e não alteram de forma significativa o comportamento observado.

As métricas mostram um modelo com bom desempenho no treinamento, uma queda na validação e uma recuperação no teste, o que indica capacidade de generalização. Para obter esses resultados, foram aplicados ajustes de profundidade da árvore, de pesos com *class_weight="balanced"*, testes de diferentes números de estimadores *n_estimators*, variações de *max_features* e diferentes proporções de divisão entre treino, validação e teste. Além disso, técnicas de balanceamento, ampliação do *n-gram range* e lematização contribuíram para a otimização do modelo.

No geral, o comportamento do modelo indica ausência de *overfitting* com impacto crítico, pois o desempenho no teste confirma que o modelo permanece robusto e adequado para aplicação prática em análises de clima organizacional via PLN.

A Figura 30 apresenta uma comparação entre a matriz de confusão de treinamento sem *SMOTE* e com *SMOTE*.

Figura 30 – Matriz de confusão *Random Forest*: (a) Treinamento (b) Treinamento (*SMOTE*)



No treinamento sem *SMOTE*, o modelo apresenta bom desempenho para promotor com (179 acertos) e detrator (135 acertos), enquanto a classe neutra

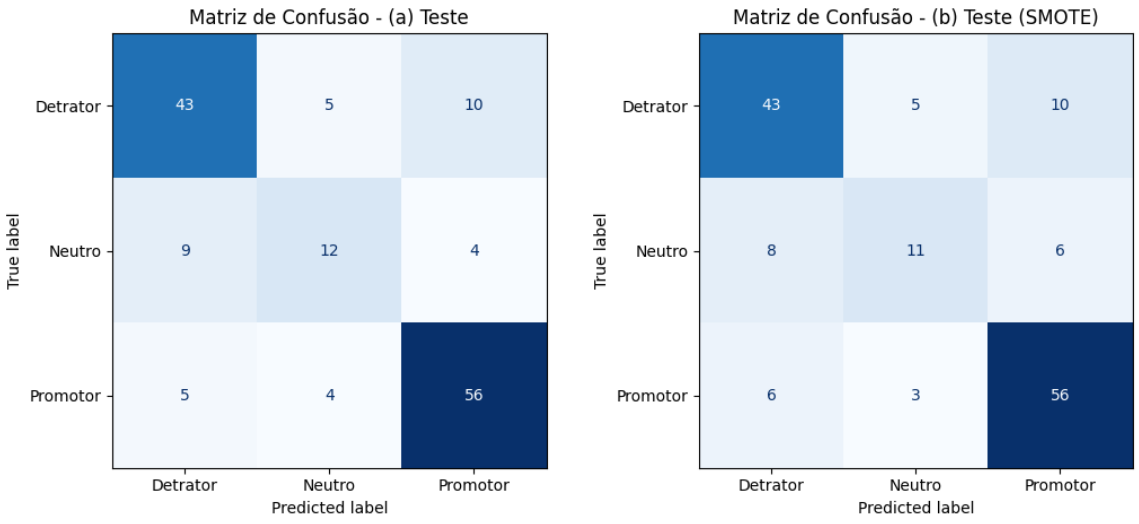
apresenta (64 acertos), há alguma confusão entre as classes detrator e promotor por exemplo, 20 detratores foram classificados como neutros e 20 como promotores sugerindo uma possível sobreposição semântica em certas respostas textuais.

Com *SMOTE*, a classe neutra praticamente não melhora com (63 acertos), e a redistribuição dos erros não altera a qualidade geral do aprendizado. Embora haja leve redução da confusão entre detrator e neutro, surge aumento em detrator e promotor. Já promotor mantém desempenho estável, mostrando que os dados sintéticos não alteram substancialmente seu reconhecimento.

O *SMOTE* não agrega ganho real ao modelo, para a classe neutra. Considerando que os dados originais refletem linguagem real de pessoas e apresentam melhor fidelidade semântica, o modelo sem *SMOTE* se mantém como a escolha mais adequada para representar os padrões textuais do conjunto de entrevistas.

A Figura 31 apresenta uma comparação entre a matriz de confusão de teste sem *SMOTE* e com *SMOTE*.

Figura 31 – Matriz de confusão *Random Forest*: (a) Teste (b) Teste (*SMOTE*)



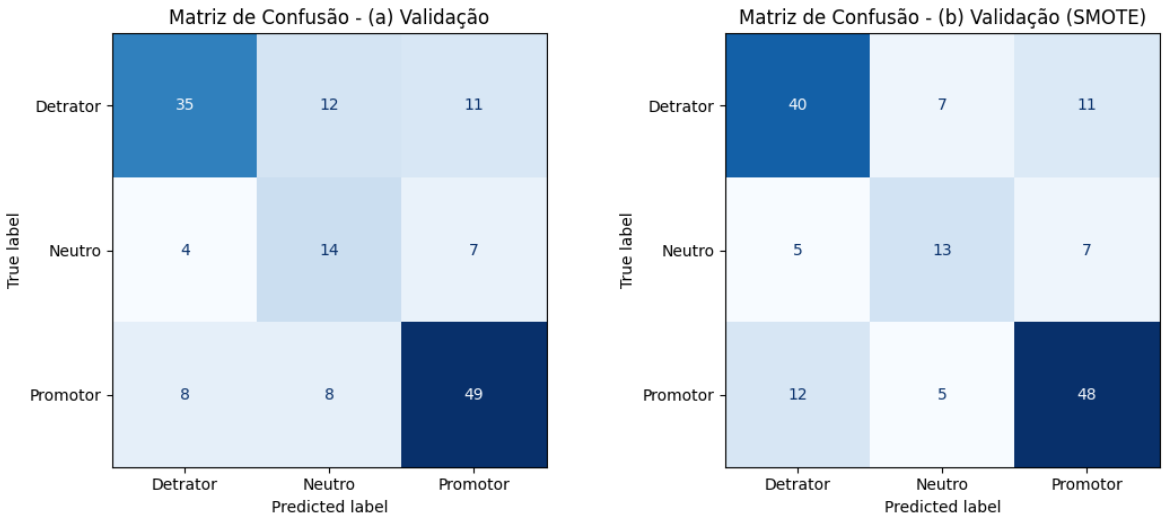
No conjunto de teste sem *SMOTE*, o modelo mantém identificação da classe promotor, com (56 acertos), e desempenho para detrator (43 acertos), embora apresente (15 erros) distribuídos entre neutro (5) e promotor (10), indicando perda

moderada de *recall*. A classe neutra apresenta (12 acertos), sendo confundida com detrator (9) e, em menor proporção, com promotor (4), reforçando sua maior ambiguidade linguística.

Na versão com *SMOTE*, o desempenho se mantém praticamente inalterado promotor e detrator preservam os mesmos acertos, evidenciando que o balanceamento não afeta essas classes no teste. A classe neutra varia para (11 acertos), com redistribuição dos erros entre detrator e promotor, sem melhoria na separação dessa categoria. No teste, o *SMOTE* não produz ganhos de desempenho o padrão de acertos e erros permanece praticamente o mesmo nas três classes. Isso confirma que, em dados reais, o modelo generaliza melhor quando treinado com as amostras originais, preservando as nuances textuais e evitando ruído semântico introduzido pelos exemplos sintéticos.

A Figura 32 apresenta uma comparação entre a matriz de confusão de validação sem *SMOTE* e com *SMOTE*.

Figura 32 – Matriz de confusão *Random Forest*: (a) Validação (b) Validação (*SMOTE*)



Na validação sem *SMOTE*, o modelo mantém desempenho na classe promotor com (49 acertos), e desempenho para detrator com (35 acertos). A classe neutra segue com (14 acertos), sendo confundida com detrator (4) e promotor (7).

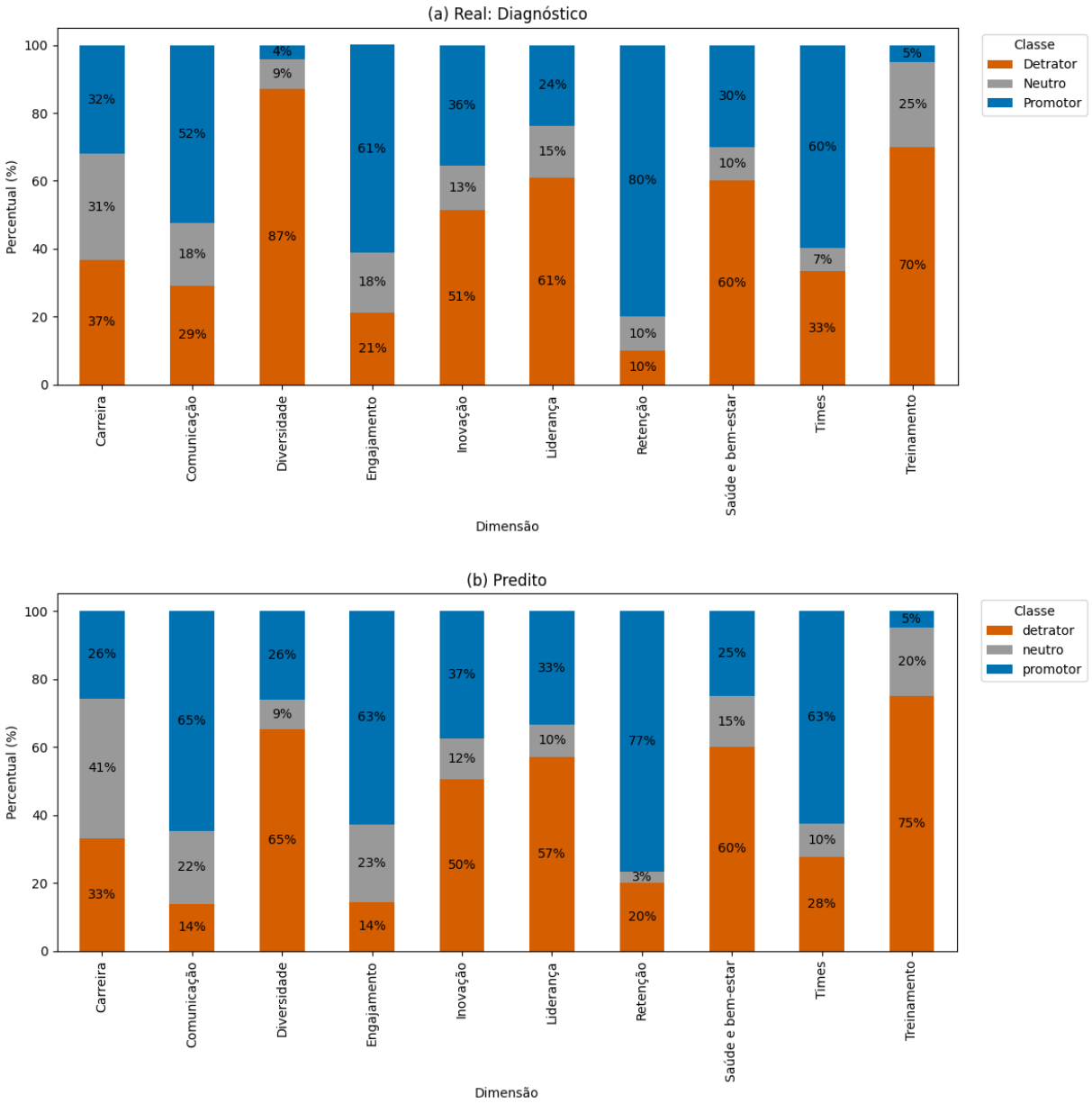
Esses padrões de erro, semelhantes aos observados no teste, mostram consistência no comportamento do modelo. Com *SMOTE*, há um aumento na identificação da classe detrator para (40 acertos), enquanto promotor diminui para (48 acertos) e neutro diminui para (13 acertos) mantendo o desempenho praticamente igual. A classe neutra permanece a mais suscetível a confusão, especialmente com detrator (5) e promotor (7), indicando que o balanceamento com *SMOTE* não resolve suas ambiguidades linguísticas. Embora o *SMOTE* apresente uma melhora na classe detrator, ele não gera avanços relevantes para neutro e promotor, mantendo o desempenho geral praticamente inalterado. A consistência entre treino, validação e teste é maior no modelo sem *SMOTE*, que preserva melhor as nuances linguísticas presentes nos textos reais. Ainda, por serem dados sensíveis de pessoas, utilizar apenas exemplos reais preserva a autenticidade dos relatos e evita distorções semânticas. Assim, o *Random Forest* treinado sem a aplicação de *SMOTE* foi selecionado como modelo final, por apresentar melhor desempenho e maior adequação ao contexto da análise de sentimentos em dados textuais humanos.

Essa decisão também se fundamenta em aspectos conceituais e metodológicos relacionados à natureza dos dados e às técnicas empregadas. O método *SMOTE*, ao gerar amostras sintéticas por interpolação, isto é, pela combinação de valores intermediários entre vetores existentes no espaço de características, demonstra ser conceitualmente incompatível com representações textuais baseadas em *TF-IDF*, caracterizadas por alta dimensionalidade e elevada esparsidade. Nesses cenários, a interpolação entre vetores tende a produzir instâncias artificiais que não correspondem a combinações semanticamente válidas de termos, comprometendo a coerência linguística dos dados gerados. Além disso, a interpolação artificial de textos pode introduzir distorções semânticas, ao combinar pesos de termos que, no discurso humano real, não coexistiriam de forma natural. Esse fenômeno pode afetar negativamente a representação do significado e do sentimento expressos nas respostas, reduzindo a fidelidade do conjunto de dados sintéticos em relação ao corpus original. Do ponto de vista do algoritmo de aprendizagem, o uso de dados sintéticos em espaços *TF-IDF* pode também impactar a construção das árvores do *Random Forest,* uma vez que divisões (*splits*) passam a ser realizadas com base em padrões artificiais e pouco interpretáveis. Isso tende a aumentar o risco de *overfitting* a estruturas não semânticas, prejudicando a

capacidade de generalização do modelo. Dessa forma, a opção pelo *Random Forest* sem *SMOTE* mostra mais coerência tanto sob a perspectiva empírica quanto conceitual, preservando a integridade semântica dos dados e assegurando maior robustez analítica ao modelo de análise de sentimentos.

A Figura 33 apresenta uma análise de distribuição de sentimentos por dimensão indireta para dados diagnósticos e predito.

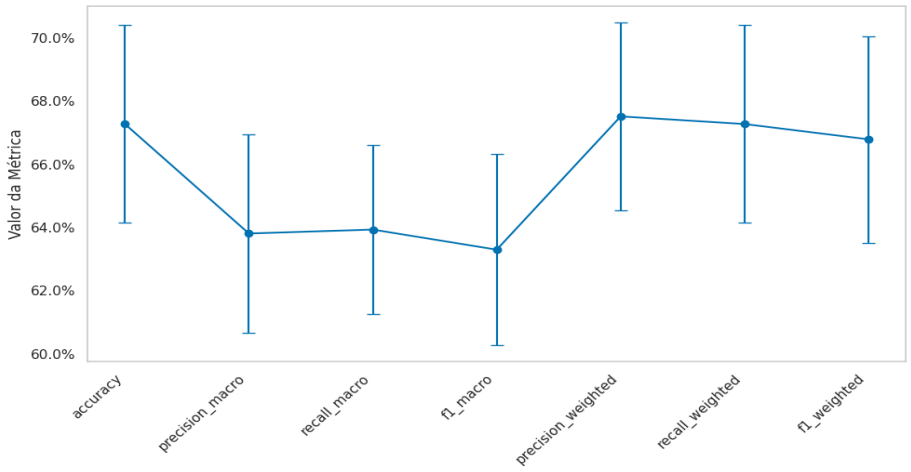
Figura 33 – Distribuição de sentimentos por dimensão indireta: (a) Real: Diagnóstico e (b) Predito



A comparação entre os cenários real diagnóstico e predito revela dois comportamentos consistentes do modelo: (1) respostas neutras tendem a ser reclassificadas como detratoras, sobretudo em dimensões mais sensíveis como retenção e treinamento, indicando que o modelo interpreta nuances críticas ou ambíguas como sinal de insatisfação, e (2) quando há redução de detratores, essa migração ocorre prioritariamente para a classe neutra, e não diretamente para promotor, como observado em carreira e saúde e bem-estar. Esses deslocamentos revelam o comportamento da classificação automática frente às nuances textuais de cada área, evidenciando onde o modelo tende a intensificar avaliações negativas retenção e treinamento ou suavizá-las em carreira. As diferenças entre real diagnóstico e predito mostram que algumas respostas neutras tendem a ser classificados como detratores em dimensões sensíveis, enquanto reduções de detratores podem migrar primeiro para neutro. Os resultados foram organizados do mais amplo ao mais específico. Primeiro, é apresentado a validação cruzada (5 *folds*), que oferece uma visão geral da estabilidade do modelo. Em seguida, são exibidos os intervalos de confiança gerados por *bootstrapping*, comparando 100 e 1000 reamostragens para evidenciar o ganho de precisão. Por fim, as distribuições acumuladas das métricas, permitindo visualizar a variabilidade real do desempenho. Essa sequência favorece uma interpretação progressiva, coerente e estatisticamente fundamentada.

A Figura 34 apresenta como o modelo se comporta em múltiplas partições dos dados, evidenciando estabilidade e capacidade de generalização.

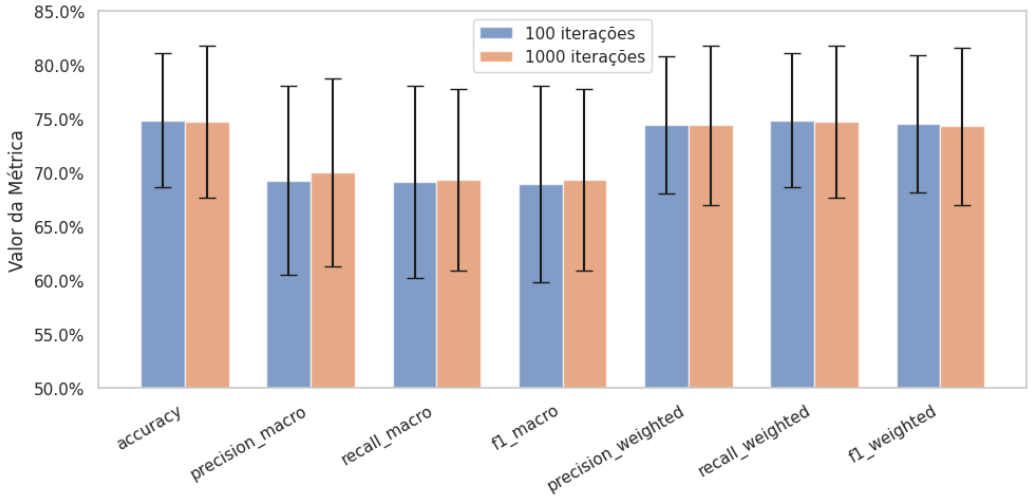
Figura 34 – Validação cruzada (5 *folds*): Panorama geral de estabilidade



O gráfico mostra que o modelo apresenta desempenho ao longo dos cinco *folds*, com acurácia média de (75%). As métricas macros *precision, recall* e *f1-score*, mais sensíveis ao desbalanceamento das classes, estão próximas de (70%), refletindo o comportamento da classe neutra. Já as métricas *weighted*, que ponderam o desempenho pelo tamanho real das classes, alcançam valores mais altos (74%–75%) e menor variação, indicando maior estabilidade e capacidade de representar o comportamento global das respostas. No conjunto, os resultados confirmam que o *Random Forest* final mantém desempenho, variabilidade controlada e coerência com os padrões observados nas etapas de treino, validação e teste, reforçando a escolha do modelo como solução final.

A Figura 35 compara os intervalos de confiança obtidos com 100 e 1000 reamostragens no conjunto de teste. O gráfico evidencia que, apesar de pequenas diferenças iniciais, as estimativas convergem e se estabilizam com 1000 iterações, demonstrando robustez dos resultados e coerência com o comportamento observado na validação cruzada.

Figura 35 – Comparação dos intervalos de confiança (100 *vs* 1000 *Bootstrap*)

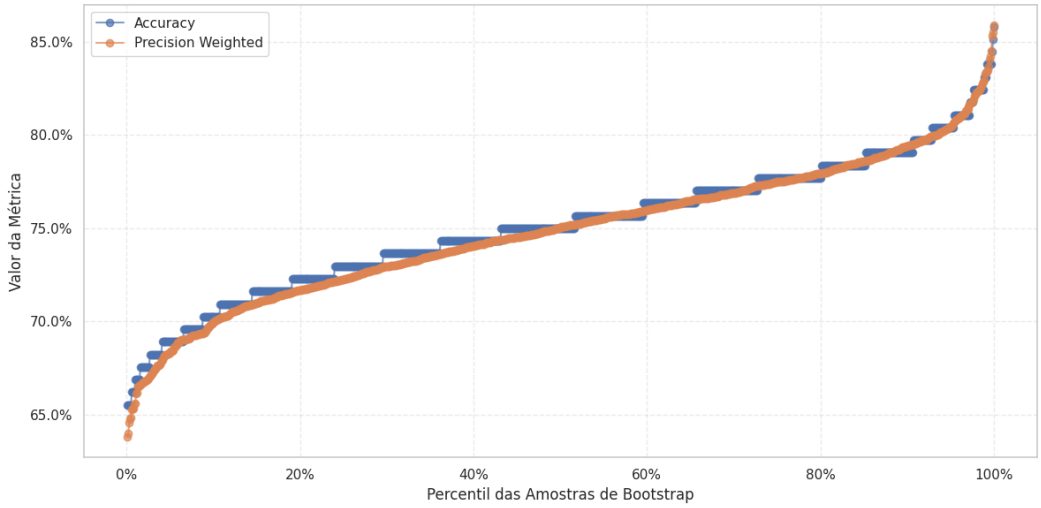


O gráfico mostra que 100 iterações já produzem estimativas estáveis, pois as médias das métricas praticamente não se alteram quando aumentadas para 1000 reamostragens. A diferença está nos intervalos com 1000 iterações, os IC95% tornam-se mais estreitos e consistentes, reduzindo a variabilidade e aumentando a precisão estatística. Assim, aumentar o número de *bootstraps* não melhora o desempenho do

modelo, mas aprimora a confiança nas estimativas ao diminuir o ruído da amostragem. Após demonstrar que 1000 iterações produzem estimativas mais estáveis e com menor variabilidade, a próxima análise aprofunda essa evidência ao observar a distribuição cumulativa das métricas obtidas via *bootstrap*. Essa visualização permite verificar não apenas a média e o IC95%, mas o comportamento completo da distribuição, revelando como a acurácia e a *precision weighted* se concentram em torno de valores estáveis, com baixa dispersão.

Diante disso, a Figura 36 reforça a robustez estatística do modelo no conjunto de teste, consolidando a escolha do *bootstrap* como método adequado para estimar a incerteza das métricas.

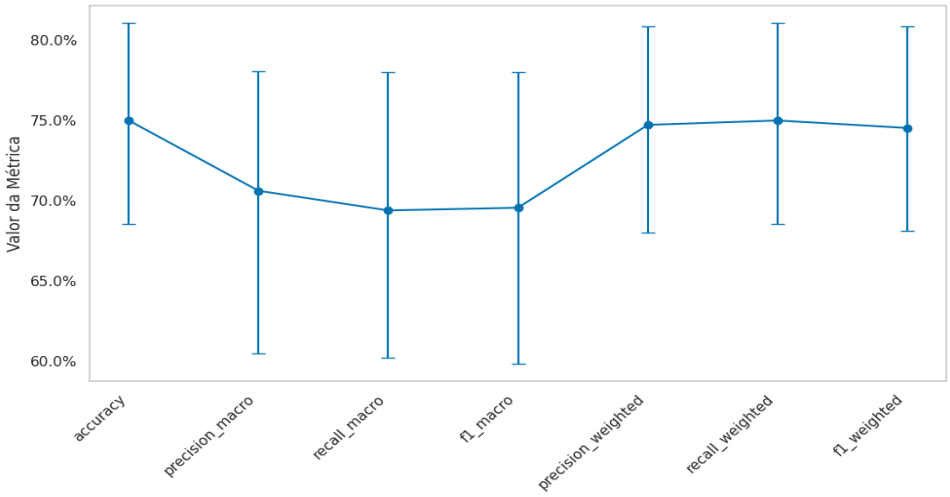
Figura 36 – Distribuição cumulativa das métricas (*accuracy* e *precision* - 1000 *Bootstrap*)



Em modelos de análise de sentimentos, o desbalanceamento entre classes pode levar a acurácia a superestimar o desempenho, favorecendo as classes majoritárias. A métrica *precision weighted* corrige esse viés ao ponderar cada classe pelo seu tamanho, fornecendo uma estimativa mais representativa da qualidade das previsões. A Figura 37 reforça essa interpretação ao mostrar a distribuição acumulada das métricas com 1000 reamostragens de *bootstrap*. As curvas de acurácia e *precision weighted* permanecem próximas e concentradas entre 70% e 78%, o que indica consistência do modelo mesmo sob diferentes subconjuntos do teste. A suavidade das curvas e a ausência de flutuações evidenciam baixa instabilidade,

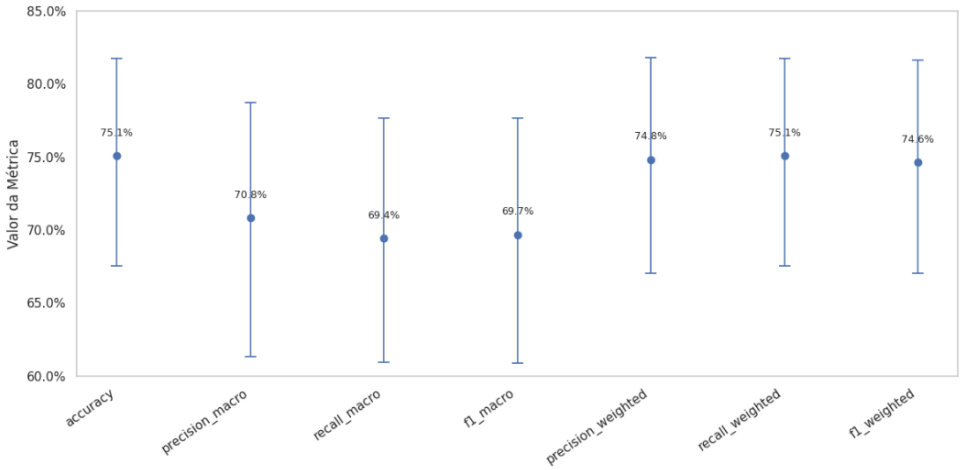
confirmando os achados anteriores o modelo mantém desempenho previsível e alinhado aos intervalos de confiança e à validação cruzada, demonstrando robustez estatística no *Random Forest* final. A Figura 37 destaca a amplitude dos intervalos de confiança (IC95%), evidenciando a variação das métricas no conjunto de teste. A linha conecta as médias apenas para facilitar a leitura, enquanto as barras de erro representam a incerteza associada a cada métrica.

Figura 37 – Intervalo de Confiança (95%) das métricas no conjunto teste



A Figura 38 apresenta os valores médios das métricas acompanhados de seus respectivos IC95%, enfatizando o ponto central estimado para cada indicador.

Figura 38 – Médias com IC95% no conjunto de teste



Ambos os gráficos revelam o mesmo padrão geral a acurácia e as métricas ponderadas *weighted* permanecem próximas de 75%, com intervalos moderados e sem sinais de instabilidade. As métricas macros, mais sensíveis ao desbalanceamento pela influência da classe neutra apresentam valores médios inferiores entre 69% e 71% e amplitudes maiores, comportamento já observado na validação cruzada. De forma conjunta, os dois gráficos reforçam que o modelo apresenta desempenho estável, previsível e coerente com os resultados obtidos no treino e validação, demonstrando capacidade de generalização em um cenário multiclasses baseado em dados textuais.

No geral, o *Random Forest* final apresenta desempenho estável e coerente os resultados de validação cruzada, *bootstrapping* e IC95% convergem para métricas em torno de 75%, com variação controlada. Isso confirma a robustez estatística e a capacidade de generalização do modelo.

6 CONCLUSÃO

O objetivo desta pesquisa é identificar sentimentos em entrevistas de desligamento por meio de técnicas de PLN, transformando narrativas espontâneas em dados úteis para compreender o clima organizacional. O estudo incluiu validação da base, rotulação manual por três especialistas de RH e verificação de concordância entre avaliadores. Após o pré-processamento textual, as análises exploratórias e temporais foram realizadas para contextualizar os padrões linguísticos. Em seguida, foi treinado o modelo supervisionado *Random Forest*, comparando versões com e sem *SMOTE* para definir a abordagem mais adequada à classificação de sentimentos.

A metodologia está estruturada em duas fases o pré-processamento e processamento para garantir a qualidade, integridade e validade analítica dos dados textuais. Cada etapa contribuiu para transformar relatos subjetivos em evidências quantitativas ao preparar, anonimizar e rotular os dados segundo critérios consistentes, possibilitando que o *pipeline* de PLN normalize o texto e maximize a informação dos termos. A vetorização *TF-IDF* atua como fundamento estatístico ao ponderar frequência local e global das palavras, aumentando a separação entre classes, enquanto o *Random Forest* ofereceu robustez computacional ao reduzir variância e capturar padrões linguísticos não lineares, preservando a interpretabilidade por meio de métricas e importâncias de atributos. Assim, o conjunto metodológico foi essencial para garantir um processo rigoroso, eficiente e replicável, capaz de gerar predições confiáveis de sentimentos e apoiar decisões estratégicas em *people analytics*. Embora o modelo treinado com *SMOTE* tenha apresentado acurácia de 75% e métricas *weighted* próximas com *precision* 74%, *recall* 75%, *f1-score* 74%, e macro de 69% a 70%, o modelo desenvolvido com dados originais demostrou desempenho superior e maior fidelidade às nuances linguísticas, sendo a opção mais adequada para análises de clima organizacional via PLN. Na prática, modelos de análise de sentimentos em contexto organizacional costumam variar entre 65% e 80% de acurácia. (Ribeiro et al., 2016; Sun et al., 2021). Estar em 75% indica que o modelo está dentro de um patamar competitivo e confiável para uso corporativo.

O modelo *Random Forest* final, combinado ao *TF-IDF*, apresentou desempenho consistente em um cenário multiclasses, com acurácia de 75%, ou seja, o modelo apresenta um desempenho sólido já que acerta 3 em cada 4 interpretações

de sentimentos, as métricas *weighted* equivalentes (precision, *recall* e *f1-score* de 75%), evidencia boa capacidade de generalização. As métricas macros, mais sensíveis à classe neutra, permanecem entre 69% e 71%, refletindo maior desafio na discriminação dessa categoria, em alinhamento com a matriz de confusão. Na validação cruzada *5-fold*, a acurácia média foi de 75%, com baixa variação entre os *folds* enquanto as métricas macros permaneceram próximas de 70% e as *weighted* entre 74% e 75%, indicando estabilidade. Os intervalos de confiança de 95% obtidos via *bootstrapping* (1000 reamostragens) mostraram acurácia média de 75%, variando entre 68% e 82%, e as métricas *weighted* na mesma faixa, evidenciando consistência. A distribuição cumulativa revelou curvas de acurácia e *precision weighted* paralelas e sem oscilações abruptas, confirmando estabilidade estatística. Em conjunto, esses resultados demonstram variabilidade controlada e robustez do modelo *Random Forest*, justificando a escolha como solução final.

Assim, esta pesquisa não apenas válida a eficácia técnica do modelo, como também contribui para elevar e incentivar a maturidade analítica nas áreas de RH no contexto de *people analytics*, fortalecendo práticas orientadas por dados e ampliando o entendimento sobre a experiência e o clima organizacional a partir de textos em linguagem natural. A análise de sentimentos poderá beneficiar um ambiente organizacional mais produtivo e humanizado com ações mais direcionadas pelas dimensões e sentimentos, com impacto direto para o negócio, sendo um modelo que pode ser generalizado para diversas áreas além do RH de maneira geral e de diversos contextos.

A pesquisa apresenta limitações, uma vez que foi realizada em apenas uma empresa de um único segmento, com dados restritos ao português e compostos por texto, sem considerar elementos de comunicação como *emojis, GIFs*, imagens ou áudios. Futuras pesquisas podem ampliar a coleta para outras organizações, segmentos e formatos de expressão, uma vez que técnicas de PLN tendem a alcançar melhor desempenho com grandes volumes e diversidade de dados. Com essa ampliação, torna-se viável empregar modelos mais avançados, incluindo arquiteturas de IA generativa como *BERT, GPT* e *T5,* capazes de capturar nuances semânticas mais complexas.

7 REFERÊNCIAS

AIRH – ACADEMY TO INNOVATE HR. **Guia de Métricas de Pessoas**. Holanda, 2025. Disponível em: https://www.aihr.com/blog/people-analytics/. Acesso em: 08 dez. 2025.

ÁLVAREZ-GUTIÉRREZ, Nadya; HERNÁNDEZ-MOGOLLÓN, Ricardo; CAMPOS-MARTÍNEZ, José Alberto. ***Human resources analytics: a systematic review from a sustainable management approach.*** *Sustainability, Basel*, v. 14, n. 15, 2022.

BAR-GIL, Boaz; EYAL, Nir; KATZ, Guy. ***AI for the people: embedding AI ethics in HR and people analytics projects****. AI and Ethics,* Israel*,* 2024.

BATINI, Carlo; CERI, Stefano; NAVATHE, Shamkant B. ***Conceptual Database Design: An Entity-relationship Approach****.* Redwood City: Benjamin/Cummings, 1992.

BENIN, K. R. A. **Processamento de Linguagem Natural e a Ciência da Informação: inter-relações e contribuições**. Dissertação (Mestrado em Ciência da Informação), Universidade Estadual de Londrina, Londrina, 2019.

BERENGUER, Fernando Sanchis. ***Análisis de Sentimiento y relaciones en la Comunicación Interna para la Mejora del Clima Laboral.*** 2024. Tese (Doutorado em Administração de Empresas), Universidad Pontificia Comillas, ICADE *Business School*, Madrid, 2024.

BISHOP, Christopher M. ***Pattern Recognition and Machine Learning****.* New York: Springer, 2006.

BLOOMAERT, **Jan.** ***Discourse: A Critical Introduction****.* Cambridge: Cambridge University Press, 2005.

BOEN, Vitor Oliveira. ***People Analytics:*** **aprendizado de máquina na gestão estratégica de pessoas, aplicando modelo preditivo de *turnover.*** *2022.* Dissertação (Mestrado em Ciências), Universidade de São Paulo (USP), São Paulo, São Paulo, 2022.

BOSCH. Bosch Tech Compass 2024. Stuttgart: Robert Bosch GmbH, 2024. Disponível em: https://assets.bosch.com/media/en/global/stories/technology_report_tech_compass_2024/bosch-tech-compass-2024.pdf. Acesso em: 08 dez. 2025.

BOSTROM, Nick. ***Superintelligence: Paths, Dangers, Strategies****.* Oxford: Oxford University Press, 2014.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Dispõe sobre a proteção de dados pessoais. *Diário Oficial da União*: seção 1, Brasília, DF, 15 ago. 2018.

BRASIL. **Ministério do Trabalho e Emprego. CAGED – Estatísticas do Cadastro Geral de Empregados e Desempregados.** Brasília, DF, 2025. Disponível em: https://www.gov.br/trabalho-e-emprego. Acesso em: 8 dez. 2025.

BREIMAN, Leo; FRIEDMAN, Jerome; OLSHEN, Richard A.; STONE, Charles J. ***Classification and Regression Trees***. Belmont: CRC Press, 1984.

BREIMAN, Leo. ***Bagging predictors. Machine Learning***, Dordrecht, v. 24, n. 2, p. 123-140, 1996.

BREIMAN, Leo. ***Random forests. Machine Learning***, Dordrecht, v. 45, n. 1, p. 5-32, 2001.

CASTRO, Leandro Nunes de. **Inteligência computacional: conceitos, técnicas e aplicações**. Rio de Janeiro: LTC, 2007.

CASTRO, Leandro Nunes de; FERRARI, Daniel Guimarães. **Mineração de dados: conceitos, técnicas, algoritmos, orientações e aplicações**. São Paulo: Saraiva, 2013.

CHIAVENATO, Idalberto. **Comportamento Organizacional**. Rio de Janeiro: Elsevier, 2005.

CHIAVENATO, Idalberto. **Gestão de Pessoas: o novo papel dos recursos humanos nas organizações**. 4. ed. Rio de Janeiro: Elsevier, 2014.

CLARK, Timothy R. ***The 4 stages of psychological safety: defining the path to inclusion and innovation*.** Oakland: Berrett-Koehler Publishers, 2023.

COULTHARD, Malcolm; SINCLAIR, John McHardy. ***Towards an Analysis of Discourse*.** Oxford: Oxford University Press, 1992.

DAFT, Richard L. **Organizações: teoria e projetos**. São Paulo: Cengage Learning, 3ª ed. 2013.

DAWES, John. ***Do data characteristics change according to the number of scale points used?*** *International Journal of Market Research*, v. 50, n. 1, p. 61–77, 2008.

DELOITTE. ***Global Human Capital Trends***. [S.l.]: Deloitte, 2024. Disponível em: https://www2.deloitte.com/. Acesso em: 08 dez. 2025.

DEVLIN, Jacob. et al. ***BERT: Pre-training of deep bidirectional transformers for language understanding.*** *arXiv preprint*, 2018. Disponível em: https://arxiv.org/abs/1810.04805. Acesso em: 8 dez. 2025.

DINA, Nasa Zata; JUNIARTA, Nyoman. ***Aspect-based Sentiment Analysis of Employee's Review Experience****. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi),* v. 4, n.1, p. 1-8, 2020.

EQUIDAM. ***Startup Growth Statistics Report 2024*.** Valência, 2024. Disponível em: https://www.equidam.com. Acesso em: 01 dez. 2025.

EDMONDSON, Amy C. **A organização sem medo: criando segurança psicológica no local de trabalho para aprendizagem, inovação e crescimento.** São Paulo: Autêntica Business, 2020.

EXPLOSION AI. *spaCy: Industrial-Strength Natural Language Processing in Python*. [S.l.], 2017. Disponível em: https://spacy.io. Acesso em: 16 out. 2025.

FERRARI, Daniel Guimarães; WEST, Mike. **Estatística para ciência de dados: modelos lineares e preditivos.** São Paulo: [s.n.], 2017.

FINSTAD, Knut. ***Response Interpolation and Scale Sensitivity: Evidence Against 5-Point Scales.*** *Journal of Usability Studies,* v. 5, n. 3, p.104-110, 2010.

FITZ-ENZ, Jac. ***The New HR Analytics: Predicting the Economic Value of Your Company's Human Capital Investments*.** New York: AMACOM, 2010.

FRONZETTI COLLADON, Andrea; SCETTRI, Giacomo. ***Look inside: predicting stock prices by analysing an enterprise intranet social network and using word co-occurrence networks.*** *International Journal of Entrepreneurship and Small Business*, v. 36, n. 4, p. 378–391, 2019. DOI: 10.1504/IJESB.2019.098986.

GALLUP. ***State of the Global Workplace* 2019**. Washington, DC: Gallup Inc., 2019. Disponível em: *https://www.gallup.com/.* Acesso em: 08 dez. 2025.

GARTNER. ***Analytics Ascendancy Model.*** *Stamford: Gartner,* 2017. Disponível em: *https://www.gartner.com/.* Acesso em: 08 dez. 2025.

*GARTNER. **Data & Analytics Maturity Model.** Stamford: Gartner, 2024. Disponível em: https://www.gartner.com/en/data-analytics/research/data-analytics-maturity-score.* Acesso em: 08 dez. 2025.

GAYE, Babacar; ZHANG, Dezheng; WULAMU, Aziguli***. Sentiment classification for employees' reviews using regression vector stochastic gradient descent classifier (RV-SGDC).*** *PeerJ Computer Science*, v. 7, e712, 2021. DOI: 10.7717/peerj-cs.712.

GREEN, David. ***What is People Analytics?*** Insight222, 2023. Disponível em: https://www.myhrfuture.com/blog/2023/5/22/what-is-people-analytics. Acesso em: 16 jun. 2025.

GREEN, David. ***The Basic Principles of People Analytics: Learn how to use HR data to drive better outcomes for your business and employees.*** *Analytics in HR B.V.*, 2016.

GÉRON, Aurélien. ***Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*.** 2. ed. Sebastopol: O'Reilly Media, 2019.

GEURTS, Pierre; ERNST, Damien; WEHENKEL, Louis. **Extremely randomized trees. Machine Learning**, v. 63, n. 1, p. 3–42, 2006.

GLOBAL WORKPLACE ANALYTICS. ***Global Workforce Trends Report****. [S.l.]:Global Workplace Analytics*, 2024. Disponível em: https://globalworkplaceanalytics.com/. Acesso em: 08 dez. 2025.

GREAT PLACE TO WORK. ***Measuring Employee Net Promoter Score****.* [S.l.]:GPTW, 2023. Disponível em: https://www.greatplacetowork.com/resources/blog/measuring-employee-net-promoter-score. Acesso em: 08 dez. 2025.

GRUS, Joel. ***Data Science*** **do Zero**. 2. ed. Sebastopol: O'Reilly, 2019.

GUENOLE, Nigel; FERRAR, Jonathan; FEINZIG, Sheri. ***The Power of People: Learn How Successful Organizations Use Workforce Analytics to Improve Business Performance*****.** London: Pearson Education, 2017.

HARVARD BUSINESS REVIEW. ***Artificial Intelligence and the Future of Work.*** *Boston: Harvard Business Review*, 2024. Disponível em: https://hbr.org/. Acesso em: 08 dez. 2025.

HAYKIN, Simon. ***Neural Networks and Learning Machines*****.** 3. ed. Upper Saddle River: Pearson, 2009.

HEARNE, Joshua. ***Strategic Integration of HR Data Analytics in Human Resource Offices*****.** 2023. 92 f. Trabalho de Conclusão de Curso (Bacharelado), *University of Maryland University College*, Maryland, 2023.

HENRIQUES, Laura de Oliveira Pedroso. ***When AI meets HR:*** **Um estudo sobre a Intenção de Uso da IA pelos Profissionais de RH.** Dissertação (Mestrado em Gestão de Recursos Humanos), Instituto Universitário de Ciências Psicológicas, Sociais e da Vida (ISPA), Lisboa, 2024.

ISHIKAWA, Edison; JOAQUIM, Carlos Eduardo de Lima; BARBOSA, Carlos Henrique Medeiros. **Análise de sentimentos da população brasileira durante a pandemia de COVID-19 como ferramenta de exploração da expressão psicossocial no espaço cibernético**. Revista Ibérica de Sistemas e Tecnologias de Informação, v. 41, p. 1–15, 2021.

ISHWARAN, Hemant; KOGALUR, Udaya B. ***Random survival forests for R. R News***, v. 7, n. 2, p. 25–31, 2007.

IRIGARAY, Hélio Arthur Reis; STOCKER, Fabricio. ***Leadership with data: improving people management through People Analytics****.* Livro eletrônico, 2023.

INSIGHT222. ***People Analytics Trends 2024****. London: Insight222*, 2024. Disponível em: https://publications.insight222.com/peopleanalyticstrends2024. Acesso em: 08 dez. 2025.

ISSON, Jean Paul; HARRIOTT, Jesse. **People analytics in the era of big data**. New Jersey: Wiley, 2016.

JIM, Cheng et al. ***Recent advancements and challenges of NLP-based sentiment analysis: a state-of-the-art review***. *Applied Sciences*, v. 14, n. 2, 2024.

JURAFSKY, Daniel; MARTIN, James H. ***Speech and language processing***. 3. ed. Versão preliminar. [S.l.]: [s.n.], 2022. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 8 dez. 2025.

JURAFSKY, Daniel; MARTIN, James H. ***Speech and language processing***. 3. ed. [S.l.]: [s.n.], 2022.

KATAL, Ashish; WAZID, Mohammed; GOUDAR, R. H. **Big data: issues, challenges, tools and good practices**. In: *IEEE International Conference on Emerging Trends and Applications in Computer Science*. Índia, 2013.

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit: the definitive guide to dimensional modeling**. Hoboken: Wiley, 2013.

KNAFLIC, Cole Nussbaumer. **Storytelling com dados: um guia sobre visualização de dados para profissionais de negócios**. Rio de Janeiro: Alta Books, 2018.

LANZER, Fábio. ***Take Off:* estratégias para cultura organizacional**. São Paulo: Elsevier, 2017.

LABOV, William. **Sociolinguistic patterns**. Philadelphia: University of Pennsylvania Press, 1972.

LIKERT, Rensis. ***A Technique for the Measurement of Attitudes*.** Tese (Doutorado) Columbia University, 1932. Publicado em: *Archives of Psychology*, v. 140, p. 1–55, 1932.

LINKEDIN. ***Global Talent Trends Report.*** 2024. Disponível em: *https://www.linkedin.com/.* Acesso em: 08 dez. 2025.

LITWIN, G. H.; STRINGER, R. ***A. Motivation and Organizational Climate***. *Boston: Harvard University Press*, 1968.

LOPES, A. C. ***Otimização de aplicações de Processamento de Linguagem Natural para Análise de Sentimentos*.** 2024. Dissertação (Mestrado Profissional em Ciência da Computação), Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São Paulo, 2024

LOURES, S. A.; RODRIGUES JÚNIOR, M.; KALILI, R. M. **Análise de sentimentos: como a inteligência artificial pode auxiliar no processamento de grandes volumes de dados?** *Revista Contemporânea*, v. 4, Rio de Janeiro: Universidade Federal do Estado do Rio de Janeiro (UNIRIO), 2024.

MATOS, F. *et al.* **Análise de sentimentos em conversas de suporte técnico via WhatsApp e Spark**. *Accounting Information Systems and Information Technology Business Enterprise*, v. 6, Maringá: Centro Universitário de Maringá (UniCesumar), 2023.

MARR, B. ***Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results***. Hoboken: Wiley, 2016.

MARR, B. ***Data-driven HR: How to use analytics and metrics to drive performance.*** 1. ed. Kogan Page, 2018.

MARTIN, S.; DAVILA-GONZALEZ, S. **Human digital twin in Industry 5.0: a holistic approach to worker safety and well-being through advanced AI and emotional analytics**. *Sensors*, v. 24, 2024.

MAYFIELD, B. J. ***Análisis emocional con integración de IA en proyectos de cambio organizacional en empresas familiares.*** Madri, 2024. Tese (Doutorado em Psicologia Organizacional), Universidade Autónoma de Madrid, Madri, 2024.

MCKINSEY & COMPANY. ***The Future of Work after COVID-19.*** *McKinsey Global Institute,* 2021. Disponível em: *https://www.mckinsey.com/.* Acesso em: 08 dez. 2025.

MEINSHAUSEN, Nicolai. **Quantile regression forests**. *Journal of Machine Learning Research*, v. 7, p. 983–999, 2006.

MINARELLI, A.; OLIVEIRA, R. ***People Analytics*** **na Prática**. São Paulo: DVS Editora, 2021.

MITCHELL, T. M. ***Machine Learning***. New York: McGraw-Hill, 1997.

MOREIRA, César Antônio Ciuffo. ***People Analytics Um estudo de caso de ciência de dados na Gerência de Recursos Humanos da Apex-Brasil***. Dissertação (Mestrado Profissional em Computação Aplicada), Universidade de Brasília, Brasília, 2025.

NORTHERN. ***Startup Growth Benchmarks 2024: Global Scaling and Expansion Patterns.*** 2024. Disponível em: https://blog.northern.com.br/taxa-de-crescimento-startup/. Acesso em: 08 dez. 2025

OKABE, Masataka; ITO, Kei. **Color universal design (CUD): how to make figures and presentations that are friendly to colorblind people**. Tokyo: J*FLY, 2008. Disponível em: https://jfly.uni-koeln.de/color/. Acesso em: 8 jan. 2026.

PEDROSA, I. **Talent and the essential role of people analytics**. *Revista de Gestão e Pessoas*, v. 10, 2022.

PENNEBAKER, J. W.; MEHL, M. R.; NIEMEIER, R. A. ***Psychological Aspects of Natural Language Use: Our Words, Our Selves***. *Annual Review of Psychology,* v. 54, Artigo de periódico científico internacional, EUA, 2003.

PEETERS, M. C. W.; VOORDE, K. V. de. ***The Psychology of Organizational Change***. London: Routledge, 2020.

POLZER, T. **The rise of people analytics and the future of organizational research**. *Organizational Research Methods*, v. 25, 2022.

PRESTES, L. M. **Mineração de textos aplicada à *People Analytics*: classificação e análise de sentimento em comentários na plataforma Glassdoor**. 2024. Monografia (Graduação em Data Science e Big Data), Universidade Federal do Paraná, Curitiba, 2024.

PRESTON, C. C.; COLMAN, A. M. **Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences**. *Acta Psychologica*, v. 104, p. 1–15, 2000.

PROVOST, F.; FAWCETT, T. ***Data Science* para negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados**. Sebastopol: O'Reilly Media, 2013.

PWC. ***Global Workforce Hopes and Fears Survey* 2024**. *PwC Global,* 2024. Disponível em: *https://www.pwc.com/.* Acesso em: 08 dez. 2025.

REICHHELD, Frederick F. **The one number you need to grow**. *Harvard Business Review*, v. 81, n. 12, p. 46–54, 2003.

REICHHELD, F.; MARKEY, R. ***The Ultimate Question 2.0*.** *Harvard Business Review Press*, 2011.

REIS, J. C.; LOPES, F. M. **Análise de conteúdo textual com PLN: limites e possibilidades**. *Revista Brasileira de Métodos de Pesquisa*, [S.l.], 2021.

RIBEIRO, F. N. *et al.* **SentiBench: a benchmark comparison of state-of-the-practice sentiment analysis methods**. *EPJ Data Science*, v. 5, 2016.

RIGAMONTI, E.; CHIARELLO, F.; GASTALDI, L. **HR analytics state of the art and future directions: a scoping review based on natural language processing (NLP)**. *Information*, v. 14, 2023.

ROBBINS, S. P.; COULTER, M. **Administração**. 10. ed. São Paulo: Pearson Prentice Hall, 2011.

RUSSELL, S.; NORVIG, P. ***Artificial Intelligence: a modern approach***. 4. ed. Boston: Pearson, 2022.

SAJID, M. *et al.* **Using text mining and crowdsourcing platforms to build employer brand in the US banking sector**. *Global Business and Organizational Excellence*, v. 41, 2022.

SAMUEL, Arthur L. **Some studies in machine learning using the game of checkers**. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, 1959.

SANTOS, M. N. F. B. **Clima Organizacional**. 1. ed. São Paulo: Saint Paul Editora, 2021.

SATMETRIX. ***Employee Net Promoter Score: A Practical Guide***. 2020. Disponível em: *https://www.satmetrix.com/wp-content/uploads/2020/07/Employee-NPS-Practical-Guide.pdf*. Acesso em: 08 dez. 2025.

SAXENA, Surbhi; DEOGAONKAR, Anant; PAIS, Rupesh; PAIS, Reshma. **Workplace productivity through employee sentiment analysis using machine learning**. *International Journal of Professional Business Review*, v. 8, 2023.

SOUZA, Aline da Silva; FERRÃO, Carla; SOUZA, Cristiane de Oliveira; ALVES, José Erlan Dias. **Inteligência artificial aplicada à gestão de pessoas: projeto chatbot**. São Paulo: Fundação Dom Cabdal, 2022. Estudo técnico institucional.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. In: *BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS)*, 9., 2020, Rio Grande. *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Proceedings, Part I*. Cham: Springer International Publishing, 2020. (*Lecture Notes in Computer Science*, v. 12319).

SILVA, A. C.; OLIVEIRA, L. A. **Aplicação do *eNPS* em diferentes escalas de resposta**. Anais do Congresso Brasileiro de Psicologia Organizacional, Brasil, 2021.

SOUZA, Fernando Lanzer Pereira de. **Clima e cultura organizacional: entender, manter e mudar**. 1. ed. Amstelveen: LCO Partners BV; CreateSpace Independent Publishing Platform, 2017

LEIDNER, Jochen L.; STEVENSON, Mark. **Challenges and opportunities of NLP for HR applications**. *arXiv*, 2024. Disponível em: https://arxiv.org/abs/2405.07766. Acesso em: 8 dez. 2025.

SUN, T.; GUO, F.; GALLAGHER, C. M. *et al.* **Smarter people analytics with organizational text data: demonstrations using classic and advanced NLP models**. *Human Resource Management Journal*, 2021. DOI: 10.1111/1748-8583.12426.

SUNDMARK, Lyndon. ***Doing HR Analytics – A Practitioner's Handbook with R Examples*.** Analytics in HR, 2018.

TUNSTALL, L.; VON WERRA, L.; WOLF, T. ***Natural Language Processing with Transformers: Building Language Applications with Hugging Face***. Sebastopol: O'Reilly, 2022.

VASWANI, A. *et al.* **Attention is all you need**. In: *Advances in Neural Information Processing Systems*, v. 30, p. 5998–6008, 2017.

ZANGARO, M. B.; SZLECHTER, D. F. **Big data y people analytics: intimidad y emociones en la gestión de los recursos humanos**. *Revista Latinoamericana de Estudios sobre Cuerpos, Emociones y Sociedad*, v. 12, 2020.

WABER, Ben. ***People analytics: how social sensing technology will transform business and what it tells us about the future of work***. Upper Saddle River: FT Press, 2013.

WEST, M. ***People Analytics: How Social Sensing Technology Will Transform Business and What It Tells Us about the Future of Work***. FT Press, 2020.

WEST, Mike. ***People Analytics para Leigos***. Rio de Janeiro: Alta Books, 2020

APÊNDICE I – Declaração sobre o uso de Inteligência Artificial

Ferramentas de Inteligência Artificial foram utilizadas como apoio à revisão textual, organização da escrita e aprimoramento da clareza linguística. O uso dessas ferramentas não interferiu na formulação dos objetivos, na definição do método de pesquisa, na coleta e análise dos dados, na interpretação dos resultados ou nas conclusões do estudo, sendo tais etapas integralmente de responsabilidade da autora.

APÊNDICE II – Termo de consentimento e confidencialidade (NDA)

O documento firmado entre a empresa participante e a pesquisadora estabelece disposições de consentimento e confidencialidade. Na cláusula 3.2, a empresa declara ciência de que a pesquisadora poderá realizar um trabalho de pesquisa acadêmica, podendo ter acesso e utilizar informações necessárias para tal finalidade. Tais informações poderão ser usadas e divulgadas exclusivamente de forma anonimizada, garantindo a proteção das pessoas naturais e jurídicas envolvidas.

As páginas do contrato, devidamente anonimizadas, estão disponíveis neste Apêndice I. Para assegurar a anonimização, a pesquisadora suprimiu os dados pessoais sensíveis e as informações institucionais antes da disponibilização do documento. O procedimento utilizou o *software PDF24 Toolbox*[1]*,* executado em ambiente local, que converte cada página em imagem, aplica tarjas diretamente sobre o conteúdo visual e salva o arquivo em formato *PDF*, impedindo a recuperação das informações originais, inclusive por reconhecimento óptico de caracteres, assegurando a confidencialidade dos dados.

[1] Disponível em: https://tools.pdf24.org. Acesso em: 08/01/2026

[redacted]

TERMO DE CONFIDENCIALIDADE

Pelo presente instrumento particular e na melhor forma de direito,

I. [redacted], sociedade limitada sediada na Avenida [redacted], na cidade de São Paulo/SP, devidamente inscrita no CNPJ sob o nº [redacted], neste ato representada na forma de seus atos constitutivos, doravante denominada simplesmente [redacted], e,

II. **SABRINA DA SILVA VASCONCELLOS**, pessoa física, brasileira, solteira, data de nascimento 12/04/1979, Pesquisadora Especialista em People Analytics, residente e domiciliada na Rua [redacted] São Paulo/SP, CEP [redacted] inscrita no CPF sob o nº [redacted] doravante denominada simplesmente "Pesquisadora".

RESOLVEM, as partes ("Partes" ou individual e indistintamente, "Parte"), celebrar o presente Termo de Confidencialidade ("Termo de Confidencialidade") como condição para a transmissão de informações pelas envolvidas, que se obrigam a manter o mais absoluto sigilo com relação às informações por elas trocadas.

1. **Objeto**. O presente Termo de Confidencialidade objetiva a proteção às Informações Confidenciais (conforme definido abaixo) relacionadas a possível transação comercial entre as Partes, a serem formalmente disponibilizadas pela [redacted] à "Pesquisadora" e vice-versa, em razão de pesquisa acadêmica ("Projeto").

2. **Informações Confidenciais**. Todas as informações relacionadas a este Termo de Confidencialidade ou adquiridas em seu curso, reveladas pela [redacted] à "Pesquisadora" ou vice-versa, serão consideradas informações confidenciais ("Informações Confidenciais"), conforme definidas abaixo, devendo ser protegidas por ambas as Partes, abrangendo, inclusive, as respectivas empresas controladoras, controladas, coligadas e/ou sob controle comum ("Afiliadas") que eventualmente tenham acesso às informações confidenciais.

2.1. As Partes se comprometem a manter as Informações Confidenciais em sigilo, utilizando os mesmos mecanismos de cuidado e discrição para evitar a divulgação, publicação ou disseminação dessas informações a qualquer terceiro estranho à relação mantida entre as Partes.

2.2. Informação Confidencial no âmbito deste Termo de Confidencialidade significa, sem se limitar, toda e qualquer informação, patenteada ou não, de natureza técnica,

[redacted]

operacional, financeira, comercial, jurídica, *know-how*, invenções, processos, fórmulas e designs, patenteáveis ou não, planos de negócios (*business plans*), métodos de contabilidade, técnicas e experiências acumuladas, documentos, contratos, papéis, estudos, pareceres e pesquisas, transmitidas entre as Partes:

(i) por qualquer meio físico (e.g., documentos impressos, manuscritos, fac-símile, mensagens eletrônicas via *e-mail*, fotografias, etc.);
(ii) por qualquer forma registrada em mídia eletrônica, tal como CD's, DVD's, *pen drive* ou qualquer outro meio magnético;
(iii) oralmente; ou
(iv) aquelas cujo conteúdo da informação torna óbvia a natureza confidencial.

2.3. As Partes reconhecem que são confidenciais, e de propriedade da Parte que as disponibilizou, todas as informações adquiridas sobre o Projeto.

2.4. As obrigações contidas no presente Termo de Confidencialidade não se aplicarão as Informações Confidenciais divulgadas pelas Partes nas seguintes hipóteses: (i) informações que encontram-se disponíveis ao público em geral ou tornaram-se, após a sua divulgação, parte do domínio público por meio de publicação ou por outro meio qualquer, sem ter havido culpa por parte de alguma das Partes; (ii) informações que já eram do conhecimento da Parte, antes de sua divulgação; (iii) informações que foram, após sua divulgação, adquiridas de boa-fé, sem infringir o presente Termo de Confidencialidade ou qualquer outro instrumento de confidencialidade com terceiros; e/ou (iv) informações que já não são mais tratadas como confidenciais pelas Partes.

3. **Utilização Informações Confidenciais**. As Partes obrigam-se, direta ou indiretamente, a manter sigilo, bem como a limitar a utilização das informações disponibilizadas para elaboração de acordos ou contratos comerciais, as quais serão consideradas confidenciais consoantes à definição de Informações Confidenciais constante do presente Termo de Confidencialidade, não utilizando tais Informações Confidenciais em proveito próprio ou alheio, principalmente quanto ao que se refere às condições comerciais negociadas entre as partes.

3.1. As Informações Confidenciais não poderão ser utilizadas para práticas anticoncorrenciais, desleais ou de competição desleal, permanecendo a obrigação válida mesmo após o término do prazo deste Termo de Confidencialidade.

3.2. A [redacted] tem ciência de que a Pesquisadora poderá realizar um trabalho de pesquisa acadêmica, poderá ter acesso e utilizar as Informações Confidenciais para a

[redacted]

sua realização. As Informações Confidenciais poderão ser utilizadas para a realização dessa pesquisa e divulgadas, desde que preservado o anonimato das pessoas naturais e jurídicas envolvidas na coleta das informações.

4. **Vigência**. O presente termo torna-se válido a partir da data de sua efetiva assinatura por ambas as Partes e perdurará por 5 (cinco) anos após o encerramento da relação comercial havida entre as Partes, independentemente da forma na qual a relação se encerrou.

5. **Infração**. A inobservância de quaisquer das disposições de confidencialidade estabelecidas no presente Termo de Confidencialidade sujeitará a Parte infratora, bem como o agente causador ou facilitador, por ação ou omissão de quaisquer daqueles relacionados no referido termo, a reparação de todas as perdas e danos materiais sofridos pela outra Parte, inclusive as de ordem moral bem como as de responsabilidades civil e criminal respectivas, as quais serão apuradas em regular processo judicial ou administrativo.

6. **Uso da Marca**. Nenhuma das Partes divulgará, utilizará ou permitirá o uso de nomes, logotipos, marcas comerciais ou outros dados de identificação da outra Parte ou, de qualquer outra forma, discutirá ou fará referências à outra Parte em questão, em nenhum aviso, website, material promocional/de marketing, em qualquer comunicado à imprensa ou outro anúncio/propaganda, sem o consentimento prévio e por escrito da outra Parte.

7. **Não Vínculo Trabalhista**. O presente termo não estabelece entre as Partes nenhuma forma de sociedade, vínculo trabalhista ou previdenciário, relação de emprego ou responsabilidade solidária ou conjunta, inclusive devendo as Partes manter indene a outra Parte caso esta seja inserida no polo passivo de ações trabalhistas, inclusive melhores esforços para que a Parte seja excluída do polo passivo.

8. **LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS.** As Partes, suas Afiliadas, seus conselheiros, sócios, diretores, prepostos, funcionários, representados ou terceiros contratados, em comunhão de esforços, se comprometerão a observar a Lei 13.709 de 2018 ("Lei Geral de Proteção de Dados Pessoais - LGPD")

8.1. As Partes, na qualidade de agentes de tratamento, adotarão todas as medidas necessárias para que as operações realizadas durante a vigência deste Termo de Confidencialidade respeitem as diretrizes estipuladas pela LGPD, bem como os seus seguintes princípios: da finalidade; adequação; necessidade; livre acesso; qualidade dos dados; transparência; segurança; prevenção; responsabilização; e, prestação de contas.

[redacted]

9. **Independência**. Se alguma disposição deste Termo de Confidencialidade for considerada inválida em virtude de qualquer lei aplicável ou de decisão judicial, tal invalidade não afetará qualquer outra disposição deste instrumento, à qual se possa dar eficácia independentemente da disposição invalidada. Qualquer disposição do presente Termo de Confidencialidade que seja inválida, ilegal ou inexequível, na medida permitida por lei, deverá ser substituída por disposição válida, legal ou exequível, capaz de produzir efeitos que sejam tão similares quanto possível aos efeitos da disposição substituída.

10. **Irrevogabilidade e Alterações**. Este instrumento é celebrado entre as Partes em caráter irrevogável e irretratável, obrigando-se as Partes e seus sucessores a qualquer título, sendo que qualquer alteração do presente instrumento somente produzirá efeitos jurídicos se efetuada por escrito e assinada por ambas as Partes.

11. **Assinatura Eletrônica**. As Partes desde já acordam que o presente instrumento será assinado eletronicamente, nos termos do artigo 10º, § 2º, da Medida Provisória 2.200-2 de 24 de agosto de 2001, conforme alterada. As Partes declaram e reconhecem a validade, para todos os fins, da assinatura eletrônica deste instrumento, de tal forma, que uma vez assinado eletronicamente, o presente instrumento produzirá todos os seus efeitos de direito.

12. **Interpretação e Foro**. O presente Termo de Confidencialidade será regido e interpretado de acordo com as leis do Brasil, e todas as Partes elegem o foro da Comarca de São Paulo, Estado de São Paulo, para dirimir quaisquer controvérsias ou dúvidas oriundas deste termo, com renúncia a qualquer outro, por mais privilegiado que seja.

E, por estarem assim justas e acordadas, as Partes assinam o presente Termo de Confidencialidade na presença de 02 (duas) testemunhas abaixo assinadas.

São Paulo/SP, 16 de julho de 2024

[redacted] **SABRINA DA SILVA VASCONCELLOS**

Testemunhas:

Nome:
CPF:

Nome:
CPF:

Clicksign

Datas e horários em GMT -03:00 Brasilia
Log gerado em 13 de agosto de 2024. Versão v1.41.0.

Termo de Confidencialidade (NDA) - [redacted] x Sabrina Vasconcelos (v2 Jurídico[redacted]).pdf

Documento número #eb4a4844-4c23-4961-a24c-2433df48af92
Hash do documento original (SHA256): d30b4f8f48e276c6e1d2a164057eff3a40e5191c36d05cc675efc767664e92d0
Hash do PAdES (SHA256): 1f5effc1172596cf866640c3873e4e80d84355d81bb05ea0ba741ebfeb23384c

Assinaturas

1 assinatura digital e 4 assinaturas eletrônicas

[redacted]
CPF: [redacted]
Assinou como testemunha em 16 jul 2024 às 18:36:17

[redacted]
CPF: [redacted]
Assinou como testemunha em 20 jul 2024 às 12:10:59

[redacted]
CPF: [redacted]
Assinou como representante legal em 26 jul 2024 às 12:11:07
Emitido por AC SERASA RFB v5- com Certificado Digital ICP-Brasil válido até 01 jan 2026

[redacted]
CPF: [redacted]
Assinou como representante legal em 07 ago 2024 às 11:53:27

SABRINA DA SILVA VASCONCELLOS
CPF: [redacted]
Assinou como parte em 13 ago 2024 às 18:15:46

Log

16 jul 2024, 18:34:41	Operador com emai[redacted]na Conta 5ebfae7f-c860-4e50-b4ab-c9826f3476d6 criou este documento número eb4a4844-4c23-4961-a24c-2433df48af92. Data limite para assinatura do documento: 15 de agosto de 2024 (18:34). Finalização automática após a última assinatura: habilitada. Idioma: Português brasileiro.

Clicksign

Datas e horários em GMT -03:00 Brasilia
Log gerado em 13 de agosto de 2024. Versão v1.41.0.

16 jul 2024, 18:35:51 Operador com email [redacted] na Conta 5ebfae7f-c860-4e50-b4ab-c9826f3476d6 alterou o processo de assinatura. Data limite para assinatura do documento: 25 de setembro de 2024 (15:35).

16 jul 2024, 18:35:51 Operador com email [redacted] na Conta 5ebfae7f-c860-4e50-b4ab-c9826f3476d6 adicionou à Lista de Assinatura: [redacted] para assinar como representante legal, via E-mail, com os pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo [redacted].

16 jul 2024, 18:35:51 Operador com email [redacted] na Conta 5ebfae7f-c860-4e50-b4ab-c9826f3476d6 adicionou à Lista de Assinatura: [redacted] para assinar como representante legal, via E-mail, com os pontos de autenticação: Certificado Digital; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo [redacted] e CPF [redacted].

16 jul 2024, 18:35:51 Operador com email [redacted] na Conta 5ebfae7f-c860-4e50-b4ab-c9826f3476d6 adicionou à Lista de Assinatura: [redacted] para assinar como testemunha, via E-mail, com os pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo [redacted] e CPF [redacted].

16 jul 2024, 18:35:51 Operador com email [redacted] na Conta 5ebfae7f-c860-4e50-b4ab-c9826f3476d6 adicionou à Lista de Assinatura: contato@universosava.com.br para assinar como parte, via E-mail, com os pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo SABRINA DA SILVA VASCONCELLOS e CPF [redacted].

16 jul 2024, 18:35:51 Operador com email [redacted] na Conta 5ebfae7f-c860-4e50-b4ab-c9826f3476d6 adicionou à Lista de Assinatura: [redacted] para assinar como testemunha, via E-mail, com os pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo [redacted] CPF [redacted].

16 jul 2024, 18:36:17 [redacted] assinou como testemunha. Pontos de autenticação: Token via E-mail [redacted] CPF informado: [redacted] IP: [redacted] Componente de assinatura versão 1.918.0 disponibilizado em https://app.clicksign.com.

20 jul 2024, 12:10:59 [redacted] assinou como testemunha. Pontos de autenticação: Token via E-mail [redacted] CPF informado: [redacted] IP: [redacted] Localização compartilhada pelo dispositivo eletrônico [redacted] URL para abrir a localização no mapa: https://app.clicksign.com/location. Componente de assinatura versão 1.922.0 disponibilizado em https://app.clicksign.com.

26 jul 2024, 12:11:07 [redacted] assinou como representante legal. Pontos de autenticação: certificado digital, tipo A3 e-cpf. CPF informado: [redacted] IP: [redacted] Localização compartilhada pelo dispositivo eletrônico: [redacted] URL para abrir a localização no mapa: https://app.clicksign.com/location. Componente de assinatura versão 1.930.0 disponibilizado em https://app.clicksign.com.

07 ago 2024, 11:53:27 [redacted] assinou como representante legal. Pontos de autenticação: Token via E-mail [redacted] CPF informado: [redacted] IP: [redacted] Localização compartilhada pelo dispositivo eletrônico: latitude [redacted] longitude [redacted] URL para abrir a localização no mapa: https://app.clicksign.com/location. Componente de assinatura versão 1.943.0 disponibilizado em https://app.clicksign.com.

Clicksign

Datas e horários em GMT -03:00 Brasilia
Log gerado em 13 de agosto de 2024. Versão v1.41.0.

13 ago 2024, 18:15:46	SABRINA DA SILVA VASCONCELLOS assinou como parte. Pontos de autenticação: Token via E-mail contato@universosava.com.br. CPF informado: [redacted] IP: [redacted] Componente de assinatura versão 1.949.0 disponibilizado em https://app.clicksign.com.
13 ago 2024, 18:15:47	Processo de assinatura finalizado automaticamente. Motivo: finalização automática após a última assinatura habilitada. Processo de assinatura concluído para o documento número eb4a4844-4c23-4961-a24c-2433df48af92.

ICP Brasil

Documento assinado com validade jurídica.
Para conferir a validade, acesse https://www.clicksign.com/validador e utilize a senha gerada pelos signatários ou envie este arquivo em PDF.
As assinaturas digitais e eletrônicas têm validade jurídica prevista na Medida Provisória nº. 2200-2 / 2001

Este Log é exclusivo e deve ser considerado parte do documento nº eb4a4844-4c23-4961-a24c-2433df48af92, com os efeitos prescritos nos Termos de Uso da Clicksign, disponível em www.clicksign.com.