Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

MINISTÉRIO PÚBLICO DO ESTADO DA BAHIA

ESPECIALIZAÇÃO LATO SENSU EM DATA SCIENCE E ANALYTICS - Turma MPBA

Machine Learning e BERT: Inovação na Compatibilização de Cargos e Funções com os Postos de Trabalho no MPBA[1]

Fernando Antônio Alves da Cunha Junior
Central de Apoio Técnico – CEAT
fernando.cunha@mpba.mp.br

Rodrigo da Silva Nunes
Diretoria de Tecnologia da Informação – DTI
rodrigo.nunes@mpba.mp.br

Tiago Miranda de Magalhães
Coordenadoria de Gestão Estratégica – CGE
tiago.magalhaes@mpba.mp.br

Resumo

Este trabalho apresenta uma abordagem inovadora para a compatibilização de cargos e funções no Ministério Público do Estado da Bahia (MPBA) utilizando técnicas de Machine Learning, com destaque para o modelo BERT. A metodologia desenvolvida explora a aplicação de Processamento de Linguagem Natural (NLP) para automatizar a análise de compatibilidade entre atribuições legais e atividades reais dos servidores. Os resultados obtidos mostram a eficácia do BERT na classificação de textos e indicam que o uso de Inteligência Artificial pode aprimorar significativamente a gestão de recursos humanos no setor público.

Palavras-chave: Machine Learning, BERT, Gestão de Pessoas na Esfera Pública

Abstract

This paper presents an innovative approach to job-role compatibility in the Public Prosecutor's Office of the State of Bahia (MPBA) using Machine Learning techniques, with a focus on the BERT model. The methodology developed explores the application of Natural Language Processing (NLP) to automate the analysis of compatibility between legal job descriptions and the actual tasks performed by public servants. The results demonstrate the effectiveness of BERT in text classification and indicate that the use of Artificial Intelligence can significantly enhance human resource management in the public sector.

Keywords: Machine Learning, BERT Public Sector Human Resource Management

Orientador: Prof. Dr. Oberdan Pinheiro, SENAI CIMATEC
Salvador, Bahia, Brasil, 30 de agosto de 2024

[1]MPBA: Ministério Público do Estado da Bahia, com o registro de agradecimento pelo incentivo e plena cooperação, enfatizando os votos de sucesso e crescimento na área da automatização de tarefas e ampliação da aplicação da inteligência artificial responsável.

Machine Learning e BERT: Inovação na Compatibilização de Cargos e Funções com os Postos de Trabalho no MPBA

Sumário

1 Contexto

As relações de trabalho no serviço público sempre foram muito debatidas, com foco, em especial, no binômio "ATRIBUIÇÕES" legais *versus* "ATIVIDADES" a serem cumpridas.

Com a transformação nas estruturas de trabalho, impactadas pela rápida e constante evolução da tecnologia, surgiram demandas que antes não existiam, ainda mais no contexto das transformações e tecnologias disruptivas, como as Inteligências Artificias (IA).

Nesse mote, perguntando para IAs Generativas : "Qual o limite de atuação de um servidor público?", a resposta é basicamente a mesma, que as **atribuições** dos integrantes de órgãos públicos são regidas por lei, e essa delimita tudo que pode e o que está além das atribuições dos cargos ou funções, sendo esse o raciocínio correto. Dois exemplos de conversas sobre esse tema podem ser encontrados nesse link do ChatGPT ou nesse outro do Gemini.

As regulamentações gerais do servidores público federais estão na Lei nº 8.112/1990 (Estatuto do Servidor Público Federal) e na Lei nº 6.677/1997 (Estatuto dos Servidores Públicos do Estado da Bahia), para o servidores estaduais, com desdobramentos, complementos ou especificações, nos atos normativos dos órgãos públicos com a listagem das **atribuições** dos servidores de cada ocupação.

Acima de tudo, há o princípio da **Legalidade**, como ensina MEIRELLES (2003)[2] pontuando que "o administrador público está, em toda a sua atividade funcional, sujeito aos mandamentos da lei e às exigências do bem comum, e deles não se pode afastar ou desviar" ou ainda "na Administração Pública só é permitido fazer o que a lei autoriza. A lei, define até onde o administrador público poderá atuar de forma lícita, sem cometer ilegalidades, define como ele deve agir" GASPARINI (2001)[3].

Por outro lado, a práxis de trabalho muitas vezes não está contemplada nas regras vigentes de modo *ipsis litteris*, seja pela modernização das demandas ou pelo uso de novas ferramentas de trabalho. Assim, encaramos o perigo do desvio de função que pode ocorrer no bojo da dialética inicialmente citada entre as **atribuições** dos servidores *versus* as **atividades** dos **postos de trabalho**.

[1](...) tecnologias que chegam e alteram muitos dos pressupostos Tecnocientíficos a partir de suas aplicações. As Tecnologias Disruptivas, em especial, criam sérios impactos nas relações sociais, gerando incertezas em razão de seu potencial inovativo (...)[a]

[a]Tecnologia Disruptiva e Direito Disruptivo: Compreensão do Direito em um Cenário de Novas Tecnologias

IA Generativa ou Inteligência Artificial Generativa se refere ao uso de IA para criar novos conteúdos, como texto, imagens, música, áudio e vídeos. Ela usa modelos de fundação (modelos de IA grandes) capazes de realizar várias tarefas ao mesmo tempo, além de resumos, perguntas e respostas, classificações e muito mais. Além disso, com o mínimo de treinamento necessário, os modelos de fundação podem ser adaptados para casos de uso específicos com poucos dados de exemplo.[a]

[a]Google em "Casos de Uso da IA Generativa".

O desvio de função ocorre quando o servidor passa a exercer atribuições diversas daquelas que correspondem ao cargo para o qual ele foi nomeado e empossado, isto é, o exercício de atividades ou serviços estranhos à competência de um cargo caracteriza desvio de função.[a]

[a]Orientações sobre Desvio de Função de Servidor no Serviço Público - TCE/SC

Conceituado como a menor divisão de atividades comuns dentro de um setor, grupo de pessoas que trabalham com as mesmas atividades.

[2]MEIRELLES, Hely Lopes. Direito Administrativo brasileiro. 28 ed. atual. por Eurico de Andrade Azevedo, Délcio Balestero Aleixo e José Emmanuel Burle Filho. São Paulo: Malheiros, 2003. 792 p.

[3]GASPARINI, Diógenes. Direito Administrativo. 6ª Ed. São Paulo: Saraiva, 2001.

Afinal, por mais que o legislador tenha buscado contemplar de modo genérico as possibilidades de atuação, as atividades reais de trabalho são, ao mesmo tempo, fluidas e dinâmicas, como ensina DE ARAUJO *et a.*(2020) [2]:"As inovações tecnológicas que ocorreram nas últimas décadas tiveram um enorme impacto no cotidiano das pessoas(...), as interações interpessoais mudaram, e a dimensão exata dessa influência vem sendo cada vez mais estudada.", porém a estrutura de gestão pública avança de modo pouco visto pela sociedade e, como discorreu, desde o final do século passado, CASTOR (1998)

> "repetidamente, ano após ano, iniciativa após iniciativa, o conflito entre as forças conservadoras e as forças modernizantes têm trazido resultados que estão longe de dar uma resposta efetiva, adequada, ao que a sociedade almeja de suas organizações públicas. Essa dinâmica perversa e ineficaz acentuou no seio da comunidade um sentimento de desamparo e frustração em relação à máquina pública, tornando urgente a busca de soluções concretas e eficientes, sob pena de comprometer o próprio sistema democrático e a filosofia liberal de governo."

Essas **atividades** dos órgãos e das unidades de trabalho são exercidas de modo subdividido em diversos **postos de trabalho**, que, por sua vez, têm integrantes com formações e ocupações plurais.

No caminho por corrigir as distorções que foram se acumulando ao longo do tempo, em função desse dilema prático contextualizado acima, o Ministério Público do Estado da Bahia, no ano de 2020, instituiu um grupo de trabalho, que buscou analisar formas de compatibilização das carreiras, funções, cargos e gratificações existentes com as atividades reais exercidas nos postos de trabalho, doravante denominado GT Compatibilização.

Como resultado inicial, o grupo apresentou uma metodologia analítica e comparativa, agrupando cada **atribuição** normativa dos servidores públicos, em todas as variações possíveis da carreira, das funções, dos cargos e das gratificações no MPBA, em três categorias: gestão (GES), assessoramento (ASS) e operacional (OPE).

A atividade de trabalho ('trabalho real') pode ser definida, então, como um processo de regulação e gestão das variabilidades e do acaso. Compreender a atividade de trabalho é compreender os compromissos estabelecidos pelos trabalhadores para atender a exigências freqüentemente conflitivas e muitas vezes contraditórias. Esses compromissos se vinculam a dois pólos de interesses: os relativos aos próprios trabalhadores (saúde, desenvolvimento de competências, prazer) e os relativos à produção. [a]

[a] Fiocruz: Trabalho Real

Portaria 1399/2020

A PROCURADORA-GERAL DE JUSTIÇA DO ESTADO DA BAHIA, no uso de suas atribuições legais, (...) Institui o Comitê de Gestão de Pessoas do Ministério Público do Estado da Bahia

(...)

Art. 2° Fica criado o Grupo de Trabalho denominado Grupo de Trabalho (GT) de Compatibilização de Cargos e Funções com o fim específico de analisar, propor e discutir matérias inerentes à compatibilização de atribuições de cargos e funções ocupados por servidores efetivos e comissionados existentes no Ministério Público, às atividades dos postos de trabalho de cada órgão/unidade (...)

De forma análoga, as **atividades** reais dos **postos de trabalho** foram categorizadas nas mesmas três áreas. Isso permitiu que a equipe de trabalho realizasse uma análise comparativa minuciosa, estabelecendo pares (**atribuição**, **atividade**) para cada uma das categorias.

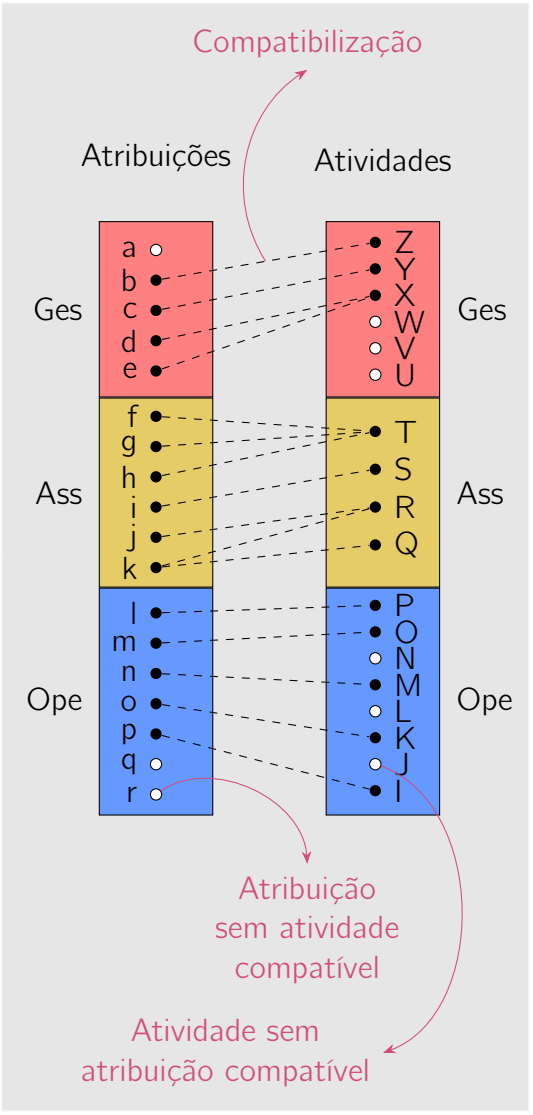
Os conjuntos "ATRIBUIÇÕES" e "ATIVIDADES" foram particionados em seis subconjuntos por categoria, como ilustrado no diagrama, ou seja, cada atribuição categorizada como de gestão foi analisada individualmente com cada atividade de gestão, e o mesmo para assessoramento e operacional.

O resultado foi a criação de uma base de conhecimento abrangente, que não só categorizou 484 **atribuições** e 585 **atividades**, mas também evidenciou a existência de relações entre elas, denominada "**compatibilização**", fornecendo um panorama claro das interações e sobreposições no contexto das **atividades** diárias dos servidores públicos. Este processo envolveu a leitura e análise de todos os pares possíveis, totalizando 7931 **compatibilizações** distintas.

Os resultados obtidos pelo GT Compatibilização fornecem um framework para a análise e gestão das **atividades** dos servidores públicos. Esta base de conhecimento pode evoluir para a identificação de áreas de redundância, lacunas nas **atribuições** e oportunidades para otimização dos processos de trabalho, culminando em possibilitar a implantação de modelos de gestão por competências.

A metodologia aplicada pode ser utilizada como modelo para outros órgãos e unidades de trabalho que enfrentam desafios similares. Além disso, a análise detalhada das relações entre **atribuições** e **atividades** pode servir como um guia para futuras reformas administrativas, ajudando a alinhar as práticas de trabalho com as demandas modernas e as inovações tecnológicas.

A adoção dessa abordagem também promove uma maior transparência e eficiência na gestão pública, alinhando-se com os princípios de legalidade e boa governança. Em última análise, a **compatibilização** efetiva entre **atribuições** e **atividades** pode melhorar significativamente a prestação de serviços públicos, respondendo de forma mais ágil e precisa às necessidades da sociedade, gerindo melhor e avançando sobre o desenvolvimento e aplicações corretas das competências dos servidores.



A gestão por competência é um modelo de acompanhamento da equipe que tem o objetivo de desenvolver ao máximo as habilidades técnicas e comportamentais dos profissionais. Isto é, compreender os pontos fortes e as oportunidades de desenvolvimento de cada colaborador. Assim, é possível traçar estratégias para melhorar o desempenho individual e, consequentemente, aumentar os resultados alcançados pela empresa.[a]

[a] ARTIGO – Gestão por competência: o que é e como implantar na empresa.

2 Problema

Os grandes problemas (ou soluções) a serem enfrentados podem ser separados nos eixos jurídico, computacional e humano *versus* complexidade *versus* tempo.

Começando pelo Direito, ALEXANDRINO (2010) [3] ressalta que "a administração pública promove o denominado **desvio de função**, vale dizer, o dirigente da unidade administrativa de lotação do servidor impõe a este o exercício de atribuições de outro cargo, diversas daquelas que correspondem ao cargo para o qual ele foi nomeado e empossado" (grifo nosso), essa situação fundante passa a ser resolvida em modo judicial, com destaque curto para duas jurisprudências consolidadas sobre a necessidade de equidade salarial e a proibição de desvios de funções na administração pública: uma do Tribunal de Justiça do Estado da Bahia (TJBA) - 2019; e outra do Supremo Tribunal Federal (STF) - 2020.

O desafio de mitigar **desvios de função** perpassa pela pré-análise de minutas de textos normativos a serem propostos para criação de novos **postos de trabalho** de modo a comparar as **atividades** pensadas com as **atribuições** normativas disponíveis. Esse é o componente inicial do segundo problema, pois a análise plural de atos jurídicos sobre carreiras, funções, cargos e gratificações implica numa ramificação de possibilidades de difícil manipulação humana, tanto pela memória, quanto pela rede de relações complexas entre artigos, parágrafos, incisos e os dilemas práticos do avanço tecnológico e as novas demandas das emergentes.

Assim, a abordagem feita pelo GT Compatibilização, pensada por um grupo plurissetorial, que agrupou integrantes do Gabinete da Procuradoria-Geral de Justiça, Superintendência de Gestão Administrativa, Coordenadoria de Gestão Estratégica – CGE e Diretoria de Gestão de Pessoas – DGP, com ênfase para a participação de especialistas gestão de pessoas da Coordenação de Provimento e Desenvolvimento de Pessoal – CPDP/DGP e de servidores com conhecimento em mapeamento de processos de trabalho da Unidade de Gerenciamento e Suporte a Processos, buscou consolidar todos os documentos do MPBA sobre o tema.

Daí, construiu-se uma metodologia de trabalho pelo GT Compatibilização, de modo criativo para reduzir essa pluralidade de dados, agrupando-os em categorias, com desenvolvimento dado pelo *know how* dos seus integrantes e, ao longo das leituras, incorporou detalhes técnicos que dependiam do grupo da época, o que evidencia, em parte, a demonstração dos argumentos do parágrafo anterior.

EIXOS

Jurídico

- Desvio de Função;
- Elaboração normativa das **atividades** de novos postos de trabalha alinhados com as **atribuições** legais dos servidores.

H×C×T

- (H)umano: *know-how* dos integrantes do GT;
- (C)omplexidade: pluralidade de atos normativos relativos a **atribuições** e mapeamento complexo das **atividades** efetivas;
- (T)empo: a leitura classificatória e comparativa de toda base de dados consome períodos dilatados e tende a aumentar a cada novo órgão/unidade.

Computacional

- Novo olhar sobre a base de dados;
- Emulação do modelo do GT com Machine Learning

TJBA 2019 - (...) é inaceitável a administração pública fixar vencimentos diferentes para servidores que ocupam o mesmo cargo e função, uma vez que tal prática fere o princípio da isonomia. O tribunal determinou a equiparação salarial, corrigindo a ilegalidade perpetrada pela administração pública.[a]

STF 2020 - (...) a equiparação salarial é necessária quando há igualdade de cargos e funções(...) a majoração dos vencimentos deve refletir a correta aplicação da legislação vigente, garantindo tratamento isonômico entre servidores.[b]

[a] TJBA. Apelação nº 0300249-35.2014.8.05.0271.

[b] STF. Recurso Extraordinário 1.265.825.

Toda construção foi documentada de modo extenso e foi desenvolvida uma planilha Excel como ferramenta de trabalho auxiliar. Por outro lado, muitos detalhes foram incluídos durante as sessões remotas de trabalho, dependendo da memória e atenção dos servidores, evidenciando o custo, relativo ao tempo, desse trabalho.

Para enfrentar os dilemas práticos da normatividade frente à atuação nos desafios cotidianos, isto é, o que realmente é desenvolvido no trabalho do servidor, o GT Compatibilização iniciou a compilação de regramentos e planilhas sobre **atribuição** dos diversos tipos de Carreiras, Funções, Cargos e Gratificações ainda vigentes; e listagem com as **atividades** feitas *in loco* por cada posto de trabalho em todos os setores ministeriais, declaradas pelos servidores e validadas pelos chefes imediatos, gerada por um estudo institucional realizado pela Diretoria de Gestão de Pessoas.

Toda essa base de dados foi organizada em planilha Excel, sendo cada atividade e atribuição classificada como de gestão, de assessoramento ou operacional, de modo cíclico e comparativo. A forma de vinculação entre atividade e possíveis atribuições compatíveis seria feita por leitura e comparação item a item e assim o trabalho foi iniciado.

Na sequência, em cada posto de trabalho, as atividades do tipo OPE buscariam ser compatibilizadas com atribuições do tipo OPE, o mesmo para as outras duas categorias. Sendo assim, esse é o elemento chave para a sequência da metodologia e caso alguma **atribuição** tivesse sua **categoria** rotulada equivocadamente, isso impactaria com quais atividades ela poderia ser comparada, e vice versa, deixando de produzir análises e/ou alinhamentos existentes.

Como ponto adicional, o perfil do **posto de trabalho** foi observado para categorização das **atividades**, em função do tipo de atuação ali desempenhada. De modo análogo, a análise da natureza do cargo influenciou nos rótulos das **atribuições**, por exemplo, os cargos eminentemente administrativos tiveram todas as suas tarefas legais **categorizadas** como operacionais.

A complexidade do estudo se deu na construção dos baremas classificatórios de cada **categoria**. Além disso, a polissemia de palavras da língua portuguesa acarretou desafios adicionais. Como por exemplo, algumas expressões do conjunto das **atribuições** foram mais repetidas no subconjunto de assessoramento, mas no dentre as **atividades**, foram dominantes na gestão.

Fazendo agora um salto temporal, em novembro de 2022, aconteceram dois fatos concomitantes que permitiram retomar o olhar sobre esse estudo de **compatibilização**: o MPBA, em parceria com o SENAI/CIMATEC, concebeu

Todo o trabalho transcorreu no época mais rigorosa da Pandemia da COVID, com reuniões diárias de duração média de 6 horas, entre março de 2020 e janeiro de 2021. Sendo aí um duplo desafio, primeiro na adaptação do que se tornou o "novo normal" por quase 2 anos e na confecção metodológica adequada.

Criando uma biblioteca com 687 documentos ao longo do trabalho.

Coleta de dados

- Atribuições dadas na normatividade vigente;
- Atividades dos Postos de Trabalho.

Classificação dos dados

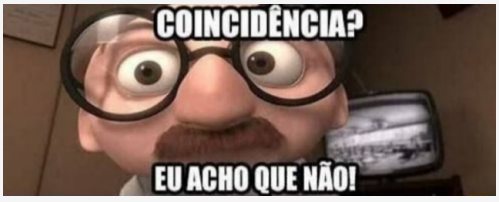
- Cada **atribuição** e **atividade** foi classificada como de Gestão, Assessoramento ou Operacional.

Análise comparativa

- Os conjuntos de mesma **categoria** foram comparados de modo a analisar a **compatibilidade**.

Consolidação

- Cada **posto de trabalho** teve suas **compatibilizações** entre atividades e atribuições resumidas em relatório próprio.



Filme "Os Incríveis", Pixar, Walt Disney Pictures,2004.

uma turma, em nível de pós-graduação, para especializar um grupo de 30 (trinta) servidores em Data Science & Analytics, um curso com ênfase em técnicas de machine learning e inteligência artificial e; a OPEN.AI, de modo disruptivo, lançou o CHAT-GPT 3, uma evolução dos chatbots, com uso de inteligência artificial, que permite conversas sobre temas genéricos com enorme taxa de acertos.

Esses dois fatores permitiram compreender que boa parte do que fora feito pelo grupo de trabalho, de modo manual, assemelha-se ao trabalho de aprendizagem de uma inteligência artificial. Sendo assim, poder-se-ia tentar replicar o que foi feito em 2020, mas com o advento de reconhecimento de padrões e técnicas de aprendizagem de máquina.

Ao longo do tempo, a metodologia que o GT Compatibilização aplicou assemelhou-se em muitos pontos com um processo rudimentar de aprendizagem de máquina. Com propostas iniciais de categorização fazendo ponderações sobre palavras, expressões e sentidos (conotativo e denotativo) para definir uma base inicial de rotulação das atribuições em cada categoria. A leitura, item a item, rotulou os textos em função de um conjunto de interpretações produzido pela expertise do grupo de trabalho.

Portanto, esse passa a ser o último problema, ou a solução, explorar modelos de reconhecimento de padrões de modo a aplicar o aprendizado supervisionado.

A base de dados categorizada será particionada em **atribuições** e **atividades**. Cada base será estudada com seus elementos intrínsecos e os resultados expostos ao final, com o aplicação do BERT, sigla em inglês cuja tradução fica Representações Codificadoras Bidirecionais de Transformadores. Basicamente, de modo simplista, é uma tecnologia de rede neural que se resume a ser uma inteligência artificial, mas com grandes ganhos, como ensina DEVLIN *et al* (2022) [4]: "Uma característica salutar do BERT é sua arquitetura unificada para diferentes tarefas. Há uma diferença mínima entre a arquitetura pré-treinada e a arquitetura downstream final.", em tradução livre.

Sem adentrar, portanto, na base de análise de todos os pares possíveis de **atribuições** e **atividades**, que ficará como objeto de futuros trabalhos.

ChatGPT é um protótipo de chatbot de IA baseado em diálogo, capaz de compreender a linguagem humana natural e gerar texto escrito semelhante ao humano, impressionantemente detalhado.
É a mais recente evolução da família GPT – ou Generative Pre-Trained Transformer – de IAs geradoras de texto.[a]

[a]Traduzido de What is AI chatbot phenomenon ChatGPT and could it replace humans?

a) (...) aprendizado de máquina, é definido pelo uso de conjuntos de dados rotulados para treinar algoritmos para classificar dados ou prever resultados com precisão.[a]
b) O machine learning (aprendizado de máquina, em português) é um termo cada vez mais em evidência. Trata-se do modo como os sistemas utilizam algoritmos e dados para simular a maneira de aprender dos seres humanos, com melhora gradual e contínua por meio da experiência.[b]

[a]IBM: O que é aprendizado de máquina (ML)?
[b]INSPER: Aprendizagem de Máquina segue etapas e tem aplicação no dia a dia

Um problema de reconhecimento de padrão consiste de uma tarefa de classificação ou categorização, onde as classes são definidas pelo projetista do sistema (classificação supervisionada) ou são "aprendidas" de acordo com a similaridade dos padrões (classificação não supervisionada).[a]

[a]BIANCHI, Marcelo Franceschi de, Extração de características de imagens de faces humanas através de wavelets, PCA e IMPCA, Dissertação de Mestrado, 2006, USP - SP

3 Justificativa

A proposta deste trabalho, que integra o uso de Machine Learning e BERT para a compatibilização de cargos e funções no Ministério Público do Estado da Bahia (MPBA), está em três pilares: princípio da impessoalidade; conveniência e oportunidade; e o avanço da Inteligência Artificial.

No setor público, o princípio da impessoalidade é essencial para garantir decisões objetivas e imparciais, baseadas em critérios técnicos, sem dependência ou vinculações de pessoas específicas ou grupos.

Conforme contextualizado, a proposta metodológica do GT Compatibilização, mesmo em fase embrionária, estava atrelada a um grupo e, dada a especificidade da tarefa, pelo conhecimento dos seus integrantes. A aplicação de técnicas de Processamento de Linguagem Natural (NLP), como o BERT, permite a sistematização dessa sapiência para conclusões mais precisas e organizadas das **atribuições** e **atividades** dos servidores, reduzindo a pessoalidade da análise. A IA assegura que grandes volumes de dados sejam processados de forma uniforme, aplicando as mesmas ponderações em todos os casos, promovendo transparência e equidade na gestão.

A emulação do método proposto com o uso de tecnologias avançadas como o BERT pode auxiliar fortemente para o cumprimento das leis e regulamentos, prevenindo desvios de função. A conformidade legal é registrada por meio de uma análise detalhada das **atividades** dos servidores, alinhando-as com suas **atribuições** legais. Essa abordagem baseada em dados reforça a impessoalidade e refina a prestação de contas, proporcionando clareza sobre os critérios e processos de decisão adotados.

O uso de machine learning e modelos como o BERT para automatizar a **compatibilização**, garante isonomia e eficiência no tratamento de dados, mitigando o risco de vieses humanos e otimizando o tempo gasto em análises manuais. Essa é uma oportunidade, que precisa se alinhar à conveniência, de aplicar tecnologias de ponta para necessidades da gestão.

O momento é propício para a implementação deste projeto, em função do contexto de transformação digital acelerada e a permanente demanda por eficiência e transparência na administração pública. No caso do MPBA, a parceria com o SENAI/CIMATEC para a capacitação de servidores em Data Science & Analytics, evidencia uma oportunidade para modernizar os processos de gestão. sendo a utilização de inteligência artificial facilitadora da adaptação às demandas e desafios emergentes, isto é, a IA é uma solução estratégica.

> São princípios básicos da Administração Pública: (...) impessoalidade, que exige que a atuação do administrador público seja voltada ao atendimento impessoal e geral, ainda que venha a interessar a pessoas determinadas, não sendo a atuação atribuída ao agente público, mas à entidade estatal a que se vincula (...)
> O princípio da impessoalidade compreende a igualdade de tratamento que a administração deve dispensar aos administrados que estejam na mesma situação jurídica. Exige, também, a necessidade de que a atuação administrativa seja impessoal e genérica, com vistas a satisfazer o interesse coletivo. Esta é a razão pela qual deve ser imputada a atuação administrativa ao órgão ou entidade estatal executora da medida, e não ao agente público, pessoa física.
> Forçoso convir que, em decorrência do princípio da impessoalidade, é vedado tratamento discriminatório aos administrados que se encontrem nas mesmas situações. [a]
>
> [a] Os Princípios mais Relevantes do Direito Administrativo

> Há conveniência sempre que o ato interessa, convêm ou satisfaz ao interesse público. Há oportunidade quando o ato é praticado no momento adequado à satisfação do interesse público. [...] A oportunidade diz respeito com o momento da prática do ato. [...] A conveniência refere-se à utilidade do ato.[a]
>
> [a] GASPARINI (2009) *apud* O Exercício da Discricionariedade Administrativa no Contexto do Estado Social e Democrático de Direito: Limites e Possibilidades a partir da Constituição Federal de 1988

A constante evolução tecnológica, impulsionada pela IA, exige adaptação e soluções inovadoras para os processos de trabalho. A aplicação de machine learning na análise de dados textuais, como será demonstrado neste trabalho, representa um passo importante nessa direção, permitindo que o MPBA se posicione na utilização de IA para aprimorar a gestão de seus recursos humanos.

A otimização de processos e a alocação eficiente de recursos humanos no MPBA, proporcionada pela metodologia proposta, têm impacto direto na qualidade dos serviços prestados à sociedade. A redução de custos e a agilidade na tomada de decisões, resultantes da automação de tarefas, liberam servidores que poderão ser melhor aproveitados em outras áreas prioritárias, como o aprimoramento da atuação do órgão na defesa dos direitos dos cidadãos.

A introdução do BERT impulsionou uma série de pesquisas e desenvolvimentos em IA e NLP, inspirando novos modelos e abordagens, como BERTimbau, que é uma versão do BERT treinada para o português brasileiro, e tem mostrado resultados superiores em NLP, sendo muito recomendado para o processamento e análise de grandes volumes de dados textuais, em suas versões Base e Large. Sua precisão em capturar nuances sutis da linguagem, garante que as análises sejam confiáveis e possam fundamentar decisões informadas.

A análise precisa das **atribuições** e **atividades** pode levar a uma melhor gestão de talentos e à utilização mais eficiente das habilidades dos servidores.

O boom da IA, no qual o BERT está incluso, trouxe uma nova era de análises inteligentes. Essas tecnologias avançadas melhoram a eficiência e a precisão das operações, proporcionando uma base sólida para decisões baseadas em dados. Assim, este trabalho não só está alinhado com os princípios fundamentais da administração pública, mas também aproveita uma oportunidade estratégica para inovar e melhorar os processos de gestão no MPBA, utilizando as mais recentes tecnologias em IA.

A aplicação de IA no MPBA também prepara a instituição para futuras reformas administrativas, que podem exigir uma reestruturação das funções e atividades dos servidores. A metodologia baseada em IA facilita a adaptação a novos cenários legais e administrativos, permitindo uma resposta rápida e eficaz às mudanças na legislação ou nas políticas públicas.

A evolução da inteligência artificial (IA) ao longo das décadas tem sido marcada por avanços significativos que têm moldado a maneira como vivemos, trabalhamos e interagimos com a tecnologia. Desde suas primeiras concepções e os marcos históricos como a Conferência de Dartmouth em 1956, a IA passou de uma ideia teórica a uma parte integral de nosso cotidiano. Inicialmente focada em tarefas específicas, como jogar xadrez ou resolver teoremas matemáticos, a IA expandiu suas capacidades para incluir aprendizado de máquina, processamento de linguagem natural, visão computacional e robótica avançada.[a]

[a] A Evolução da Inteligência Artificial e seus Impactos ao Longo dos Anos

O BERTimbau melhora o estado da arte nessas tarefas em relação ao BERT multilíngue e às abordagens monolíngues anteriores, confirmando a eficácia de grandes modelos de linguagem pré-treinados para o português.[a]

[a] Tradução de Bert: Pretraining of deep bidirectional transformers for language understanding

A administração pública direta e indireta de qualquer dos Poderes da União, dos Estados, do Distrito Federal e dos Municípios obedecerá aos princípios de legalidade, impessoalidade, moralidade, publicidade e eficiência (...) [a]

[a] Constituição Federal de 1988, Artigo 37 *caput*.

4 Objetivos

4.1 Objetivo Geral

As bases de dados "Atribuições" e "Atividades" são coleções de textos classificados, o que pode ser trabalhado no campo do Processamento de Linguagem Natural (NLP). Com o avanço da tecnologia, especialmente na área de aprendizado profundo, modelos como o BERT (Bidirectional Encoder Representations from Transformers) têm revolucionado a maneira como lidamos com textos, proporcionando resultados significativamente melhores em várias tarefas de NLP. Este trabalho explora a aplicação do modelo BERT para a tarefa de classificação de texto, detalhando todas as etapas envolvidas no processo, desde a preparação dos dados até a avaliação do modelo. O objetivo central é emular o modelo de **compatibilização** desenvolvido pelo Grupo de Trabalho.

O modelo BERT, desenvolvido pelo Google, representa um avanço significativo no campo do NLP. Com a arquitetura baseada em Transformers que, como escreve HAN *et al.* (2021) [4] "é um tipo de arquitetura neural que codifica os dados de entrada como recursos poderosos por meio do mecanismo de **atenção**", por sua vez, "(...) evita a recorrência e, em vez disso, confia, inteiramente em um mecanismo de atenção para estabelecer **dependências** globais entre entrada e saída", como ensina VASWANI *et al.*(2017)[5], ambos com tradução e grifos nossos.

A rede neural que aprende, cria memória e, assim, o significado, aplicando o monitoramento de relações em dados sequenciais, como as palavras desta frase, permitindo a captação do contexto bidirecional em uma sentença, algo que, de maneira eficaz, não conseguiam fazer, modelos anteriores (escrevendo com o estilo de um personagem famoso de filmes, mas garantindo a compreensão).

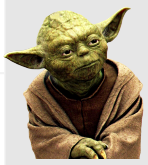
A bidirecionalidade, especialmente para o português, garante a captura de nuances e relações semânticas complexas, melhorando a precisão e a relevância das previsões. Afinal, nossa língua materna tem construções lógicas que podem ter sentidos quebrados, por exemplo, com hipérbatos, como no Hino Nacional "Ouviram do Ipiranga as margens plácidas / De um povo heroico o brado retumbante" é uma escrita rebuscada para "As margens plácidas do Ipiranga ouviram o brado retumbante do povo heroico". Na leitura bidirecional, construções assim são captadas e guardadas em memória, com possibilidade de conexão com outras partes do texto para classificação ou outra forma de trabalho.

[4] HAN et al., Transformer in Transformer, 2021

[5] VASWANI *et al.*, Attention Is All You Need, 2017

> À análise de texto e linguagem por meio computacional é dado o nome de Processamento de Linguagem Natural (Natural Language Processing – NLP). (...)
> Os recentes avanços em NLP por meio de redes neurais têm proporcionado várias pesquisas com tarefas próprias de NLP, como classificação de texto, análise semântica, extração de informação e (...), pode-se citar a organização e recuperação da informação.[a]
>
> [a] Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI

> Argumentamos que as técnicas atuais restringem o poder das representações pré-treinadas, especialmente para as abordagens de ajuste fino.
> A principal limitação é que os modelos de linguagem padrão são unidirecionais, e isso reduz a escolha de arquiteturas que podem ser usadas durante o pré-treinamento(...)
> Nossa maior contribuição é generalizar ainda mais essas descobertas para arquiteturas bidirecionais profundas, permitindo que o mesmo modelo pré-treinado resolva com sucesso um amplo conjunto de tarefas de PNL.[a]
>
> [a] Tradução de BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding



"Difícil de ver. Sempre em movimento está o futuro."
Yoda, Star Wars

4.2 Objetivos Específicos

Os objetivos específicos foram:

(a) analisar a base de dados com o intuito de gerar conhecimento aprofundado sobre a complexidade do problema enfrentado na Compatibilização de Cargos e Funções com os Postos de Trabalho no MPBA, explorando nuances e padrões ocultos na conexão dos dados textuais.;

(b) desenvolver e aplicar modelos de Machine Learning para prever a **compatibilidade** entre cargos e funções com base nas **atividades** reais, começando com modelos mais simples, porém instáveis, para o tipo de problema, como árvore de decisão e florestas aleatórias, formando um quadro comparativo para avaliar o desempenho de técnicas mais avançadas, permitindo a compreensão inicial dos padrões presentes nos dados para, na sequência, seguir para um modelo preditivo mais robusto;

(c) implementar e refinar um modelo BERT, explorando suas principais versões, **BERT Base** e **BERT Large**, enfocando na otimização dos hiperparâmetros, para aprimorar a capacidade de previsão e classificação de textos e análise semântica, comparando os resultados;

(d) utilizar o modelo BERT para emular o processo de **compatibilização** do Grupo de Trabalho, buscando automatizar e aprimorar a análise de compatibilidade entre as **atribuições** legais e as **atividades** reais dos servidores, registrando as métricas gerais (precisão, eficácia, f1 score e recall);

(e) criar tabelas e gráficos para apresentar os resultados, comparando o desempenho dos diferentes modelos e técnicas utilizados, e entender a qual a melhor métrica de avaliação para o processo.

Em suma, o fluxo do trabalho permeou a revisitação do modelo do GT Compatibilização e potencializando-o por meio da aplicação do BERT.

A análise dos dados entregou argumentos para o uso da IA mitigar os problemas listados nas sessões anteriores, como o trabalho humano e manual, de modo a alcançar um novo patamar analítico, buscando contribuir estrategicamente para o avanço da gestão de pessoas no MPBA.

Por fim, a visualização clara dos resultados, com comparação evolutiva dos modelos entregou uma avaliação técnico com foco no modelo BERT debatendo os impactos dos hiperparâmetros no redução de *overfitting* e na melhoria das *métricas* de avaliação.

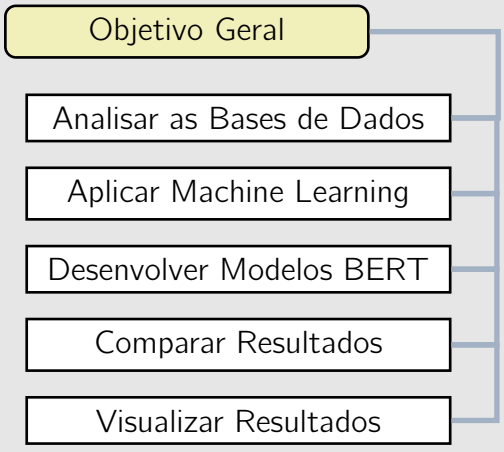
[5] A análise dos dados é um processo complexo que envolve retrocessos entre dados pouco concretos e conceitos abstratos, entre raciocínio indutivo e dedutivo, entre descrição e interpretação.[a]

[a] A Análise de Dados na pesquisa Científica: importância e desafios em estudos organizacionais

[6] Árvores de decisão são extremamente sensíveis a pequenos desvios nos dados: alterações mínimas podem resultar em uma árvore significativamente diferente. Elas podem facilmente ficar em *overfitting* e, mesmo que haja métodos de validação e poda, essa é uma área cinzenta.[a]

[a] Traduzido de "Main Steps for Doing Data Mining Project Using Weka"

Os principais hiperparâmetros são: Batch Size, que é o "número de exemplos de treinamento usados em uma iteração."[7], " hiperparâmetro chave no treinamento de modelos de aprendizado profundo"[8]; e Learning Rate que "controla o quanto você está ajustando os pesos da rede em relação ao gradiente de perda"[9]. Traduzido das referências destacadas



5 Método Proposto

De início, os dados foram recebidos três em arquivos Excel, um para cada órgão ministerial destacado para o piloto, separados em diversas abas, cada uma contendo uma diversidade de informações. Sendo assim, foi necessário estudar a estrutura das planilhas para identificar onde encontrar as **atribuições** normativas organizadas, as **atividades** em cada posto de trabalho e a **compatibilização** proposta.

Com essa visão geral, foram produzidos dois arquivos, um para atribuições e outro para as atividades, contendo apenas as dimensões dos respectivos textos e categorias. Um olhar objetivo identificou que a terceira parte, especificamente sobre a compatibilização, necessitaria de mais tempo e, por isso, não foi abordada no estudo.

O primeiro passo foi o pré-processamento dos dados recebidos para observar o tipo de (des)balanceamento dos dados em cada dataset.

[10](...) balanceamento de dados refere-se ao processo de equiparar, para um determinado conjunto de dados, a quantidade de amostras de cada classe disponível para análise, tornando-as igualmente representadas diante do processo de aprendizagem.(...)
Entretanto, ao se trabalhar com conjuntos de dados e aplicações reais, o mais provável é que os dados disponíveis para análise estejam desbalanceados.[a]

[a] Balanceamento de dados com base em oversampling em dados transformados.

Depois, com o intuito de criar um parâmetro comparativo, os ajustes dados seguiram o roteiro para o processamento em modelos de árvore. Na abordagem, foram retiradas, as stopwords e outras que aparecem em múltiplas classificações.

Foram aplicados dois modelos: árvore de decisão e floresta randômica.

[11] Stopwords são palavras irrelevantes e insignificantes que aparecem em uma linguagem para ajudar a construir sentenças, mas que não representam nenhum conteúdo nos documentos.
[12] Normalmente, as palavras que aparecem em muitos documentos não são indexadas pois sua utilização compromete a precisão e a eficiência de um sistema de busca. O prejuízo semântico dessa estratégia é perder a busca exata por compostos como "máquina de lavar", onde a preposição "de" não pode ser buscada. [a]

[a] Análise de Dados de Saúde: Mineração de Texto com a Utilização do Orange Canvas para Exploração da Informação

Seguimos com a preparação dos dados pela codificação da variável alvo CATEGORIA, que originalmente estava em formato categórico e precisava ser convertido de dados textuais para valores numéricos, facilitando o uso em algoritmos de aprendizado de máquina.

Para lidar com as diferentes naturezas dos dados (categóricos e textuais), foi usado um ColumnTransformer. A principal transformação focou-se na coluna ATRIBUIÇÕES processadas, onde aplicamos várias técnicas de mineração de texto, como TfidfVectorizer, CountVectorizer e HashingVectorizer, com as principais características de cada "Vectorizer", listadas a seguir, sem aprofundamento, pois não é o foco do presente trabalho:

Característica	Tfidf	Count	Hashing
Representação dos termos	TF-IDF	Contagem de frequência	Hashing
Dimensionalidade	Variável	Variável	Fixa
Vocabulário	Criado automaticamente	Criado automaticamente	Não armazenado

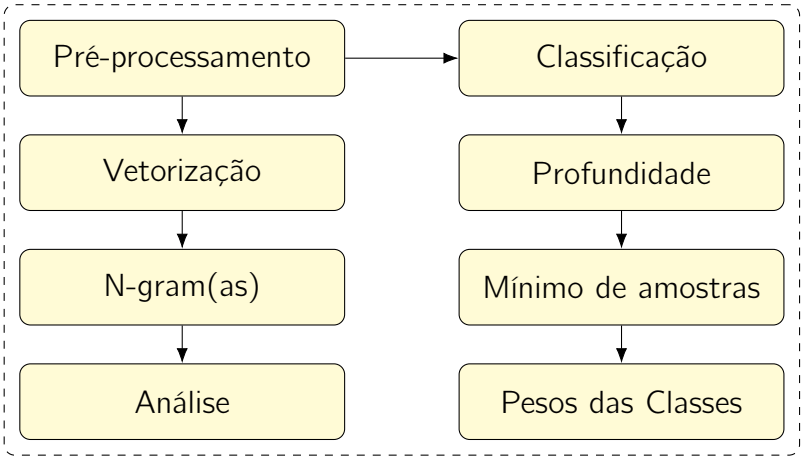
Tabela 1: Comparação entre os principais vectorizers.

[13] (...) permite que um transformador (...) seja aplicado apenas nas colunas definidas.[a]

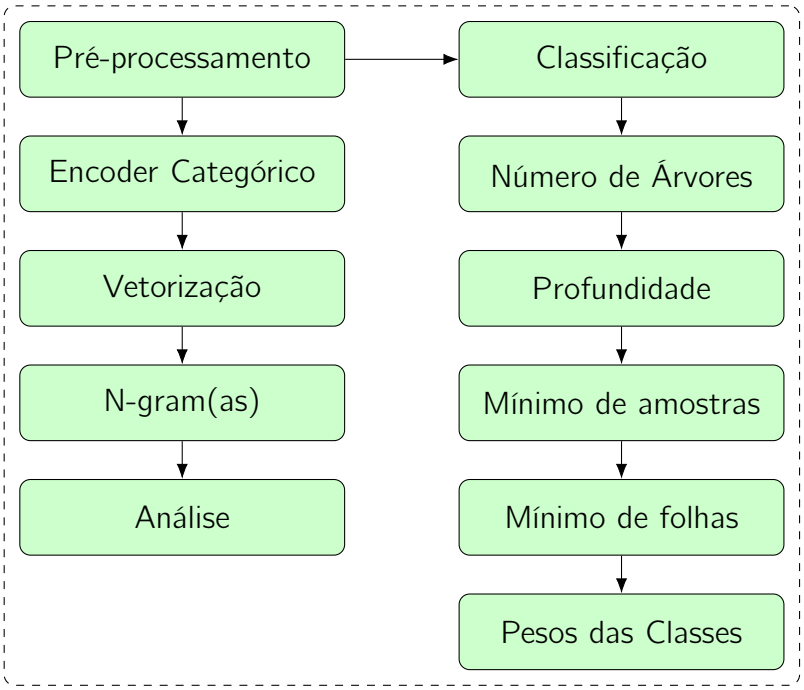
[a] Machine Learning aplicado à Saúde

Com a definição do pré-processamento, um pipeline foi construído, tanto para o pré quanto para o modelo de classificação. E, assim, a **árvore de decisão** foi o primeiro classificador base, seguindo para a **floresta randômica**, com otimização de hiperparâmetros com um gridsearch, incluindo as variações de vetorizações já citadas, alcance de n-gramas, análise de texto (palavras ou caracteres), profundidade máxima da árvore, e outras características importantes do classificador, como exposto nos fluxos a seguir:

Árvore de Decisão



Floresta Randômica



Os resultados serão expostos na próxima sessão, enfocando os parâmetros mais promissores para cada modelo, acrescentado que, no intuito de perceber o comportamento mais geral para classificação, foi incluída uma validação cruzada com cinco divisões (cv=5) para avaliar cada combinação de parâmetros, calculando as métricas e observando a estabilidade e os resultados das configurações do pipeline.

[14] (...) dividir uma tarefa completa de machine learning em um fluxo de trabalho de várias etapas. Cada etapa é um componente gerenciável que pode ser desenvolvido, otimizado, configurado e automatizado individualmente. [a]

[a] Microsoft Azure

[15] A utilização do Grid Search para otimizar modelos de machine learning pode contribuir significativamente para a redução do overfitting. Ao encontrar combinações mais adequadas de hiperparâmetros, é possível diminuir a disparidade entre a acurácia de treino e a acurácia de teste, mitigando assim o overfitting.[a]

[a] Como otimizar modelos de machine learning com grid search?

[16] (...) consiste na divisão controlada ou não controlada da amostra de dados em duas subamostras, a escolha de um preditor estatístico, incluindo qualquer estimativa necessária, em uma subamostra e em seguida, a avaliação de seu desempenho medindo suas previsões em relação às outras subamostra. Um exemplo de divisão controlada é fornecido pelo cauteloso estatístico que separa uma parte selecionada aleatoriamente de sua amostra sem olhar e então joga sem inibição com o que resta, confiante no conhecimento de que os dados relativos à retirada de terras fornecerão um julgamento imparcial sobre a eficácia da sua análise.[a]

[a] Traduzido de Cross-Validatory Choice and Assessment of Statistical Predictions

O $k = 5$ foi estabelecido em função do poderio computacional disponível.

Após a limpeza (apenas para os modelos de árvores, não para o BERT), os dados foram divididos em conjuntos de treino e validação, como pode ser vertido de GROSSE [18] "precisamos de treinamento separado e conjuntos de testes: se treinarmos nos dados de teste, não temos ideia se a rede está generalizando corretamente, ou se está simplesmente memorizando o os exemplos do treinamento", ou seja se está ocorrendo overfitting. Seguindo para YING (2019) [19], em tradução livre:

> Quando da existência de overfitting, o modelo funciona perfeitamente no conjunto de treinamento, mas se ajusta mal ao conjunto de testes. Isto é devido ao superajustado gerar dificuldade em lidar com informações do conjunto de teste, que podem ser diferentes daqueles do conjunto de treinamento. Por outro lado, modelos sobreajustados tendem a memorizar todos os dados, incluindo ruído inevitável no conjunto de treinamento, em vez de aprender a disciplina escondida atrás os dados.

Com atenção para que as bases desbalanceadas tenham a distribuição adequada das categorias existentes, pois, pode ser entendido de PROVOST [20] " que ao estudar problemas com dados desbalanceados, aplicando os classificadores padronizados por algoritmos de aprendizado de máquina, sem ajustar o limite de saída, há risco de ocorrer um erro crítico"

Com os parâmetros comparativos definidos, era hora da entrada o estado da arte, o BERT, em sua versão brasileira, o BERTimbau, com as variações Base e Large, enfocando na otimização dos hiperparâmetros, para aprimorar a capacidade de previsão e classificação de textos e análise semântica, comparando os resultados.

Agora, não mais estaríamos num preprocessamento com corte de texto e expressões, pois, por tudo já escrito, o modelo de análise bidirecional precisa de dados para construir os vínculos e captar as nuances existentes. Sendo assim, os dados dos arquivos das **atribuições** legais e das **atividades** reais contendo textos e suas respectivas **categorias** (mapeadas para um número) foram utilizados diretamente no modelo BERT, apenas com uma etapa de limpeza básica para remoção de números, caracteres especiais e espaços em branco extras, concluindo essa fase com a padronização dos textos para letras minúsculas.

[17] Os métodos de divisão de dados anteriores podem ser implementados assim que especificarmos uma divisão razão. Uma proporção comumente usada é 80:20, o que significa que 80% dos dados são para treinamento e 20% para testes. Outras proporções como 70:30, 60:40 e até 50:50 também são utilizadas na prática. Não parece haver uma orientação clara sobre qual proporção é melhor ou ideal para um determinado conjunto de dados.[a]

[a] Traduzido de Optimal Ratio for Data Splitting

Foster Provost é coautor do famoso livro Data Science Para Negócios O que você Precisa Saber Sobre Mineração de Dados e Pensamento Analítico de Dados.

[4] **BERT Base**: Possui 12 camadas de codificadores, 12 cabeças de atenção, tamanho oculto (profundidade) de 768 e 110 milhões de parâmetros. É mais rápido de treinar e usa menos recursos computacionais. **BERT Large**: Possui 24 camadas de codificadores, 16 cabeças de atenção, tamanho oculto de 1024 e 340 milhões de parâmetros. Geralmente, tem desempenho superior ao BERT Base, mas requer mais treinamento e recursos. [a]

[a] Traduzido de BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Ajustando o Batch Size, que é o número de amostras de dados processadas em cada iteração durante o treinamento; e o Learning Rate, que, de modo bastante simplista, controla o ritmo de aprendizagem do modelo.

Com os dados preparados, foi realizada a tokenização, convertendo os textos em uma sequência de tokens numéricos, compreensíveis pelo modelo BERT. Esse processo foi feito utilizando o BertTokenizer, que gerou os IDs dos tokens e as máscaras de atenção. A estruturação dos dados de entrada foi organizada com os textos e seus respectivos rótulos para o treinamento.

O modelo BERT foi configurado para a tarefa de classificação de sequências, com ajustes específicos nos hiperparâmetros, como o tamanho do batch e a taxa de aprendizado. O treinamento usou a classe Trainer da biblioteca transformers, que facilitou o gerenciamento do processo, incluindo a implementação de callbacks, isto é, métricas de retorno calculado que poderiam incluir restrições e freios, como Early Stopping, lição ensinada por DODGE *et al* (2020) [23] mostrando "que um melhor desempenho pode ser alcançado com os mesmos recursos computacionais usando Early Stopping que interrompem os testes menos promissores no início do treinamento".

Durante o treinamento, foram calculadas a acurácia, o F1 score, a precisão e o recall, para cada época, dados usados para a construção de gráficos que ilustraram a evolução das perdas de treinamento e validação, expondo as métricas de desempenho, ao longo das épocas. Isso permitiu uma análise visual do progresso do modelo e facilitou a identificação de possíveis ajustes necessários.

Os resultados obtidos mostraram que o modelo BERT teve majoritariamente uma crescente em todas as métricas guardadas, indicando sua capacidade de generalização e eficácia na tarefa proposta. A utilização de técnicas de regularização e do early stopping mostrou-se essencial para evitar o overfitting e garantir a robustez do modelo em dados de validação.

Conclui-se que a aplicação do modelo BERT para classificação de textos é altamente eficaz, especialmente combinada com uma preparação cuidadosa dos dados e uma avaliação detalhada das métricas de desempenho. Este estudo oferece uma base para futuras pesquisas e aplicações em PLN, destacando a importância do uso de ferramentas de monitoramento do treinamento e ajuste fino dos hiperparâmetros. As técnicas e métodos discutidos podem ser adaptados para outros contextos, ampliando o alcance e a aplicabilidade das soluções desenvolvidas.

[21] (...) tokenizar um texto é dividi-lo em palavras ou subpalavras, que são então convertidas em IDs por meio de uma tabela de consulta.[a] Uma explicação bastante didática pode ser assistida no vídeo [22] em inglês Word-based tokenizers.

[a] Traduzido de **Summary of the tokenizers**

A "Trainer Class" gerencia o treinamento e o ciclo de validação, com os modelos especificados, assim como os argumentos de treinamento, conjuntos de validação, dentre outras configurações.

Recurso responsável por interromper o treinamento quando o modelo não apresentava melhorias significativas após um número definido de épocas, que será exposto em gráficos na próxima sessão, mostrando que, após um certo tempo o modelo de adequada demais aos dados de treino, sem conseguir generalizar no conjunto de testes, ou seja, gerando overfitting.

[24] Regularização é qualquer técnica suplementar que visa melhorar a generalização do modelo, ou seja, produzir melhores resultados no conjunto de testes.[a]

[a] Traduzido de Regularization for Deep Learning: A Taxonomy

[4] (...) o modelo BERT pré-treinado pode ser ajustado com apenas uma nova camada de saída para criar modelos de última geração para uma ampla uma série de tarefas.[a]

[a] Voltando para BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

6 Resultados

Ao aplicarmos o processamento dos dados das **atribuições** normativas e das **atividades** reais dos servidores do MPBA, prontamente ficou evidente a complexidade do problema para os integrantes que desempenharam a missão inicial e os algoritmos triviais de machine learning.

Com o conhecimento adquirido durante o curso Data Science & Analytics, do SENAI/CIMATEC, um novo olhar foi lançado sobre os dados. Destarte, de modo didático, lançamos os conjuntos rotulados das atribuições em nuvens de palavras de modo a observar as palavras mais presentes em cada subconjunto.

De imediato, a palavra "gestão" está destacada nas atribuições de assessoramento, quando o senso comum poderia já associá-la ao conjunto homônimo, assim como "acompanhar" está no operacional, quando poder-se-ia ser um termo mais próximo da gestão ou do assessoramento, incluindo aqui a polissemia das palavras do nosso idioma.

De modo análogo, para os conjuntos categorizados das **atividades**, têm-se nuvens com certa diferença as quais geram quase um cumulonimbus.

Vencida a exposição da complexidade da missão, podemos ainda exemplificar que, no universo das **atividades**, "acompanhar" mudou de grupo e foi para gestão, demonstrando os argumentos anteriores. Com a contagem em cada subconjunto, temos a distribuição do top 15 de palavras nas **atribuições**:

Palavra	Frequência
acompanhar	52
processos	51
projetos	48
atividades	45
prestar	44
documentos	42
informações	41
elaborar	40
ações	39
gestão	38
executar	37
procedimentos	37
coordenar	36
segurança	35
assuntos	34

Tabela 2: 15 Palavras mais comuns nas **atribuições**

Nuvens das Palavras Relativas às **Atribuições**

Figura 1: Assessoramento

Figura 2: Gestão

Figura 3: Operacional

Nuvens das Palavras Relativas às **Atividades**

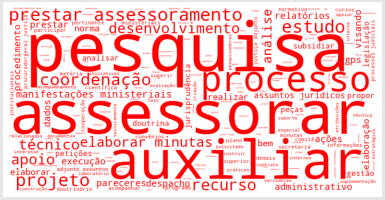


Figura 4: Assessoramento

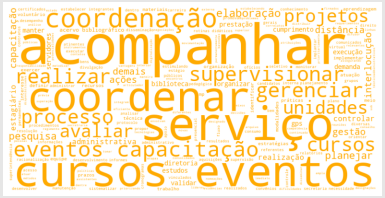


Figura 5: Gestão

Figura 6: Operacional

Para retomar o ponto geral, como já escrito anteriormente, o estudo tabulou 478 **atribuições** normativas, 103 de gestão, 200 de assessoramento e 175 operacionais. O que identifica um desbalanceamento nos dados. Ampliando esse escopo inicial, cada frase definindo uma **atribuição** foi separada em palavras e tabuladas de modo único.

O que restou ainda mais interessante, pois no universo das **atribuições**, o conjunto da gestão há 305 palavras distintas, no de assessoramento tem 462 palavras e o 559 no operacional. No entanto, há interseções entre os três conjuntos e também dois a dois, que podem ser representadas de modo didático por diagramas de Venn-Euller.

Agora, construindo um cruzamento entre a tabela 2 com os conjuntos da figura 8, a distribuição das 10 primeiras palavras nos conjuntos gestão, assessoramento e operacional fica como no gráfico a seguir:

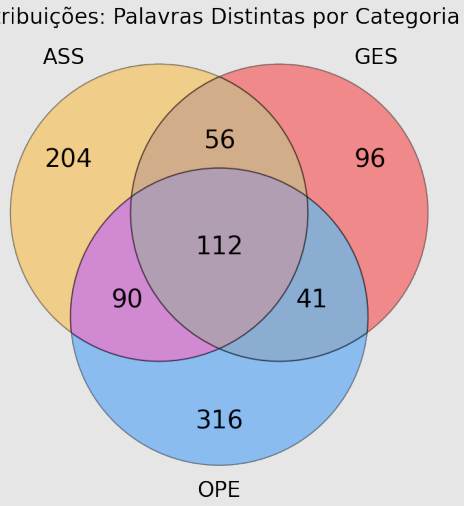


Figura 8: Representação dos conjuntos do universo das **atribuições**

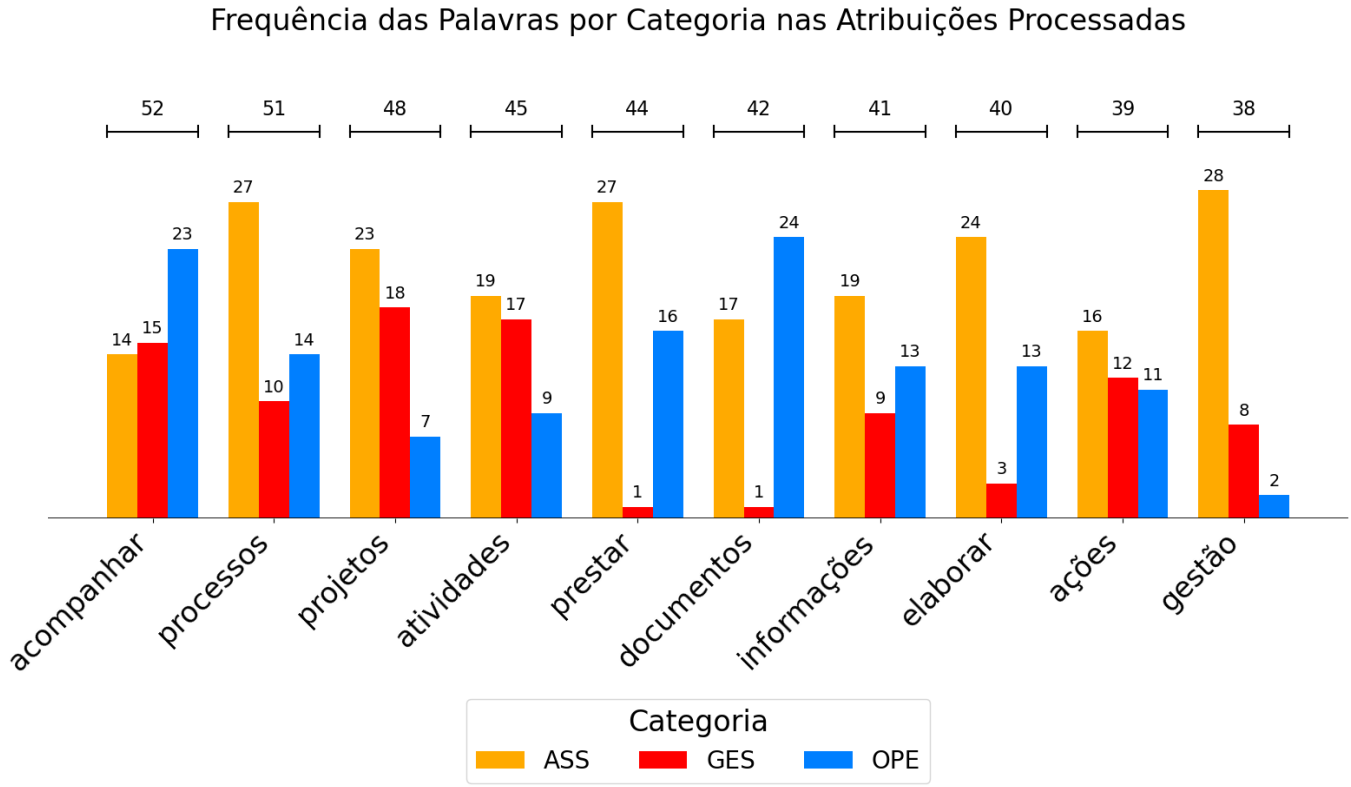


Figura 9: Top 10 das palavras dos conjuntos das **atribuições**

O universo das **atividades** reais tem cardinalidade 583, sendo 88 de gestão, 117 de assessoramento e 378 operacionais. Também com desbalanceamento nos dados, como rememorado na comparação.

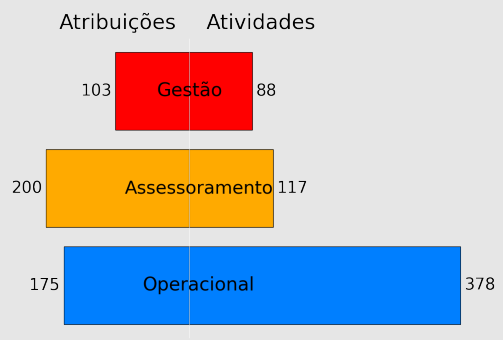


Figura 11: Comparação numérica entre **Atribuições** e **Atividades**

Seguindo o *modus operandi*, cada frase definindo uma **atividade** foi separada em palavras e tabulando-as de modo único, com 419 palavras distintas na gestão, 446 no assessoramento e 941 no operacional. Com insistente observação das interseções dois a dois e entre os três conjuntos, representadas no diagrama.

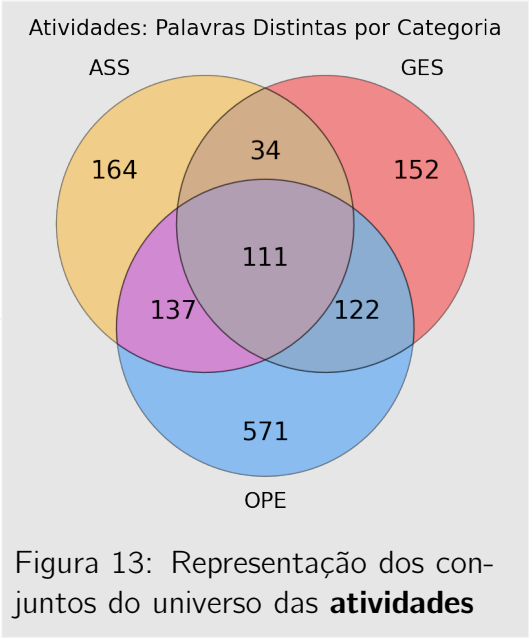


Figura 13: Representação dos conjuntos do universo das **atividades**

Palavra	Frequência
atividades	61
elaborar	51
acompanhar	50
capacitação	47
unidade	45
documentos	44
prestar	41
procedimentos	40
cursoseventos	38
processos	37
coordenação	35
relatórios	34
apoio	32
informações	32
realizar	31

Tabela 3: Palavras mais comuns nas atribuições processadas

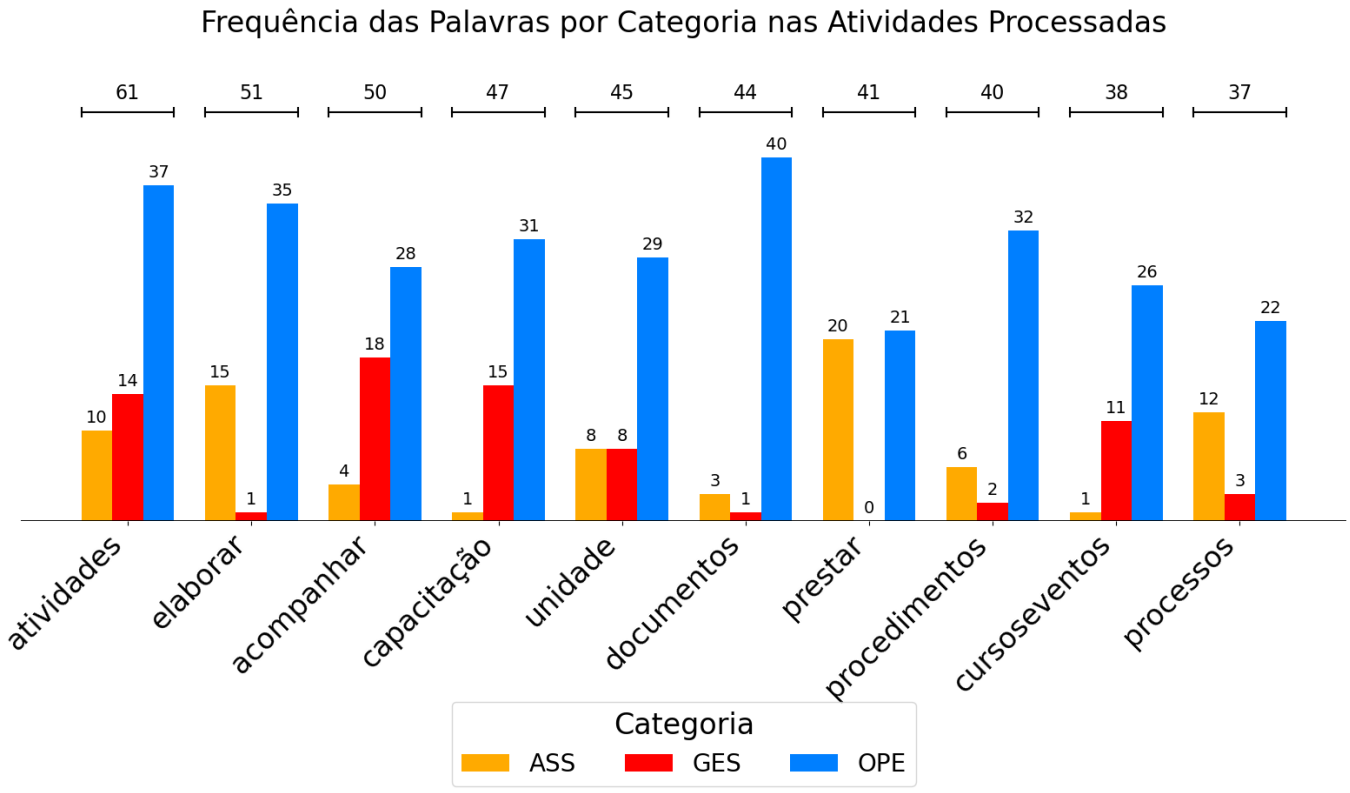


Figura 14: Top 10 das palavras dos conjuntos das **atividades**

Com a complexidade dos conjuntos exposta nas páginas anteriores, fica claro que abordagens triviais não iriam ajudar a máquina a entender o problema.

Mesmo assim, por teimosia, seguimos para entender como isso impactaria modelos de árvore de decisão e floresta randômica, de modo a criar parâmetros comparativos em um dos modelos mais gerais do Machine Learning.

No caso das **atribuições**, foram feitos cenários que incluíram (ou não, como diria Caetano) a dimensão das carreiras, funções, cargos e gratificações (CFCG); outros que retiravam as stopwords; e alguns que excluíam as palavras que figuravam em mais de uma das categorias (gestão, assessoramento e operacional).

Para as **atividades**, foram testados quadros com e sem a dimensão da unidade original dos dados; dos postos de trabalho; e os mesmo modos de exclusão (ou não) das stopwords e das palavras pertencentes a mais de uma categoria.

O pré-processamento codificou a variável alvo com o LabelEncoder, que transforma a variável categórica CATEGORIA em uma forma numérica, passo essencial para o tratamento de muitos algoritmos de aprendizado de máquina.

Em seguida, definiu-se o ColumnTransformer para processar tanto a característica categórica CFCG quanto o texto contido nas ATRIBUIÇÕES.

Utilizou-se OneHotEncoder para a característica categórica, convertendo-a em uma representação binária apropriada para algoritmos de machine learning. Para o texto, usou-se o passthrough, uma estratégia que permite a substituição posterior por diferentes técnicas de vetorização, como TF-IDF, Count Vectorizer e Hashing Vectorizer, já explicadas na sessão anterior.

Como métrica de avaliação, em função do desbalanceamento dos dados, foi usado o "F1 Score", pois, como ensina SMOLIC (2024) [26]:

> Conjuntos de dados desbalanceados são comuns no aprendizado de máquina, onde uma classe de observações supera significativamente a outra. Nesses casos, a precisão por si só pode ser enganosa. O "F1 Score" fornece uma métrica de avaliação mais robusta, especialmente quando a classe minoritária é de particular interesse. Ao considerar a precisão e o recall, o "F1 Score" representa o desempenho do modelo, apesar do desequilíbrio de classes.

[25] Codifica rótulos de destino com valores entre 0 e $n_{classes} - 1$.
Este transformador deve ser usado para codificar valores alvo, ou seja, y, e não a entrada X.[a]

[a] Tradução da documentação do LabelEncoder no Scikit-Learn

[25] Este estimador permite que diferentes colunas ou subconjuntos de colunas da entrada sejam transformados separadamente e os recursos gerados por cada transformador sejam concatenados para formar um único transformador.[a]

[a] Tradução da documentação do Column-Transformer no Scikit-Learn

[25] Isso cria uma coluna binária para cada categoria e retorna uma matriz esparsa ou matriz densa (...). Por padrão, o codificador deriva as categorias com base nos valores exclusivos de cada recurso. Alternativamente, você também pode especificar as categorias manualmente.[a]

[a] Tradução da documentação do OneHot-Enconder no Scikit-Learn

[25] Permite que os dados passem pelo transformador sem alterações. Utilizado quando se deseja incluir colunas que não precisam de pré-processamento pipeline, isto é, colunas que já estão no formato necessário para o modelo de machine learning. [a]

[a] Tradução da documentação do Pipeline no Scikit-Learn

Para fechar o giro da Árvore de Decisão, foi feito um GridSearch com uma gama variada de opções e com o Cross-Validation, seguindo o que, com tradução livre, motivou o seu uso por GANESAN (2024) [27]:

> Para otimizar o desempenho do modelo é importante ajustar os hiperparâmetros. Existem três métodos mais amplamente utilizados disponíveis, como GridSearch (...). Essas pesquisas exploram as diferentes combinações de valores de hiperparâmetros que ajudam a encontrar a configuração mais eficaz e ajustar o modelo de árvore de decisão. (...)
> GridSearch é um método fundamental para ajuste de hiperparâmetros que pesquisa exaustivamente os hiperparâmetros predefinidos. Ele avalia todas as combinações possíveis e o torna uma abordagem confiável para encontrar o hiperparâmetro ideal. O GridSearch geralmente leva a previsões altamente precisas quando recursos computacionais adequados estão disponíveis. (...)
> (...) No entanto, a desvantagem do GridSearch é o seu custo computacional, especialmente quando se trata de espaços de parâmetros de alta dimensionalidade.

param grid = { 'preprocessor text':
- TfidfVectorizer(),
- CountVectorizer(),
- HashingVectorizer (n features=2**4),
- HashingVectorizer (n features=2**8),
- HashingVectorizer (n features=2**16),

'preprocessor text ngram range':
- [(1, 1), (1, 2), (1, 3),
- (2, 2),(2, 3)],

'preprocessor text analyzer':
- ['word', 'char'],

'classifier max depth':
- [10, 20, None],

'classifier min samples split':
- [2, 3, 4, 5, 6, 7, 8, 9, 10],

'classifier class weight':
- [None, 'balanced'] }

No caso da Árvore de Decisão para as **atribuições**, o modelo resultante foi o que não utilizou a dimensão do cargos e funções e descartou as palavras que figuravam em mais de uma categoria:

[Árvore de Decisão - ATRIBUIÇÕES] Melhor configuração de parâmetros:
'classifier class weight': 'balanced', 'classifier max depth': None, 'classifier min samples split': 6,
'preprocessor text': TfidfVectorizer(analyzer='char', ngram range=(2, 3)),
'preprocessor text analyzer': 'word', 'preprocessor text ngram range': (1, 2)
Melhor score de validação cruzada: $0.7888060542994756 \pm 0.03652526362165277$

A Árvore de Decisão para as **atividades** com melhor performance não considerou os postos de trabalho e incluiu as as palavras que figuravam em mais de uma categoria:

[Árvore de Decisão - ATIVIDADES] Melhor configuração de parâmetros:
'classifier class weight': None, 'classifier max depth': None, 'classifier min samples split': 3,
'preprocessor text': HashingVectorizer(analyzer='char', n features=65536, ngram range=(2, 3)),
'preprocessor text analyzer': 'word', 'preprocessor text ngram range': (1, 3)
Melhor score de validação cruzada: $0.7352958438378824 \pm 0.04748278854064607$

O pré-processamento da Floresta Randômica foi o mesmo narrado na páginas anteriores. Por outro lado, o *RandomForestClassifier* trouxe parâmetros adicionais específicos, como o número de estimadores, profundidade máxima, número mínimo de amostras para dividir um nó, número mínimo de amostras por folha e pesos de classe.

A gama mais ampla de parâmetros e opções proporcionou uma exploração amplificada do espaço de hiperparâmetros com o uso do GridSearch. Por fim, a validação cruzada também foi realizada, utilizando a métrica f1_weighted para avaliar a performance de cada combinação de parâmetros.

Para a Floresta Randômica das **atribuições**, o modelo resultante também foi o que não utilizou a dimensão do cargos e funções, mas não descartou as palavras que figuravam em mais de uma categoria.

[28] A Floresta Randômica ou Aleatória, Random Forest, consiste em "n", número de árvores de decisão que são treinadas usando o subespaço do treinamento dos dados e o resultado de todas as árvores de decisão são usados coletivamente para a previsão final do rótulo "y".
Devido ao grande número de árvores utilizadas para a previsão de "y", o erro de uma única árvore é superado ou compensado por outra árvore de decisão, disponível na floresta.[a]

[a] Traduzido livremente de **A Detailed Review on Decision Tree and Random Forest**

[Floresta Randômica - ATRIBUIÇÕES]
Melhor configuração de parâmetros:
'classifier class weight': None, 'classifier max depth': None, 'classifier min samples leaf': 1,
'classifier min samples split': 5, 'classifier n estimators': 300, 'preprocessor text':
CountVectorizer(analyzer='char', ngram range=(2, 3)),
'preprocessor text analyzer': 'word',
'preprocessor text ngram range': (1, 1)
Melhor score de validação cruzada:
$0.7847559473143848 \pm 0.05356827616232403$

Nesse momento, observando o ponto de parâmetro tão próximo das **atribuições** na Decision Tree e na Random Forest, ficou decidido já seguir para o BERT, sem fazer o mesmo estudo para as **atividades**, e usar logo o estado da arte.

O BERT tem uma coleção de variações, inspiradas no modelo, mas com ajustes personalizados que pode buscar redução do tamanho do modelo ou refazer o treinamento com sequências mais longas, algumas com nomes interessantes em referências aos substantivos próprios como o ALBERT ou o ROBERTA, DistilBERT, entre outros. Para o presente estudo, foi usado o BERTimbau, em suas versões Base e Large.

Alguns modelos BERT foram analisados no interessante trabalho de ROSA JUNIOR (2021) [30] que cita " o RoBERTa (...) obteve resultados superiores aumentando o número de parâmetros e da base de dados." e segue dizendo que "o DistilBERT (...) para gerar vários benefícios como a redução do tempo de treinamento."

param_grid = {
[repetidos do Decision Tree]
'preprocessor text':
'preprocessor text ngram range':
'preprocessor text analyzer':
'classifier class weight':
'classifier min samples split':
• [2, 5, 10], [redução na variedade]
[adicionados no Random Forest]
'classifier n estimators':
• [100, 200, 300],
'classifier max depth':
• [10, 20, None],
'classifier min samples leaf':
• [1, 2, 4], }

[29] A Lite BERT (ALBERT) usa estratégias de otimização para reduzir o tamanho do BERT.
Elimina *embeddings* do one-hot na cada inicial, estruturando as palavras para uma baixa dimensionalidade.[a]

[a] Traduzido de "The Evolution of Embeddings"

Robustly Optimized BERT Pretraining Approach (RoBERTa)

O treinamento foi feito com um tipo de GridSearch parrudo para o modelo BERT, variando o batch size, a learning rate e os modelos. Para a avaliação, foram gravadas as métricas: acurácia; **F1 Score**; Precisão; e Recall, além das perdas de teste e treino, construindo assim uma grande base que registrou o comportamento do treinamento na progressão das épocas.

Cumpre ressaltar que tamanhas variações de processamento não foram alcançadas por um processador i7 de 7ª geração com 16gb de RAM e 256 gb de HD SSD. O batch size de 128 não foi atingido por ele. Para seguir, na verdade, refazendo todo o estudo, foi aplicada uma máquina mais moderna, um o i7 de 11 geração com 64 gb de RAM e 4 TB de disco M.2 NVMe.

Por fim, todos os resultados foram colocados em gráficos de barras e de retas de modo a tornar a análise visual, com especial atenção para as perdas, em função da percepção do ajuste dos dados ao treinamento, sem generalização efetiva no conjunto de teste, isto é, o *overfitting*.

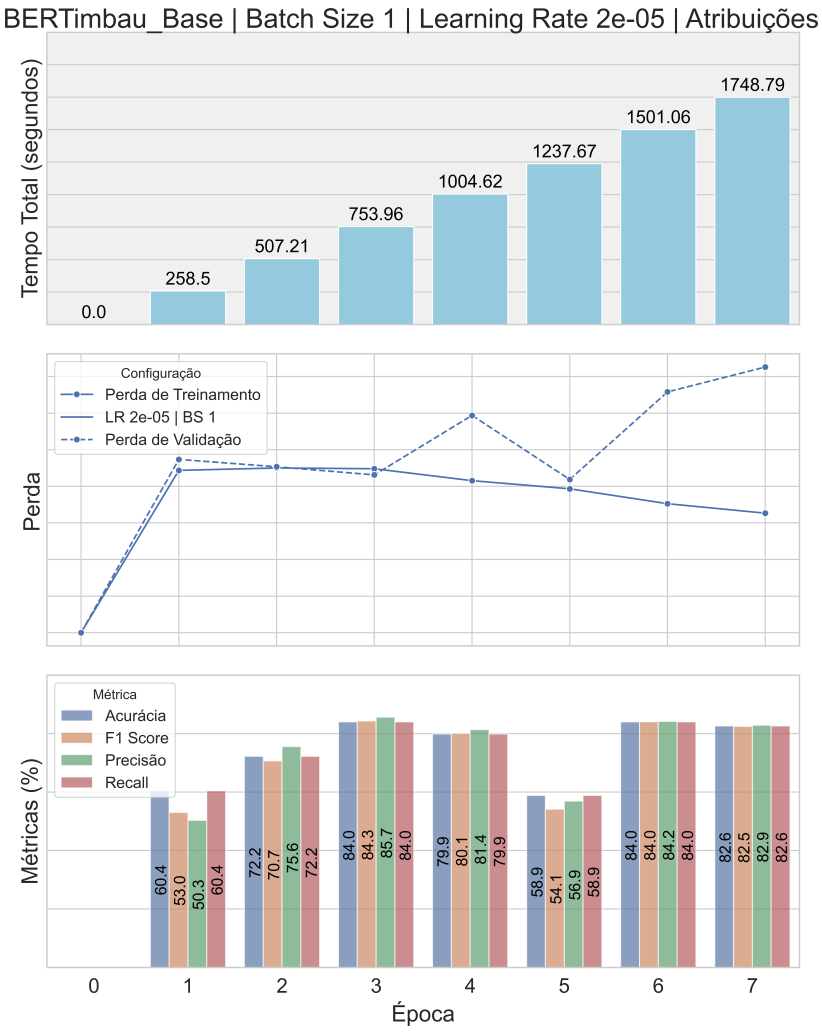


Figura 15: Gráfico das **Atribuições** no modelo BERT Base, batch size 1, learning rate 2e-5

Os parâmetros registrados para teste foram:
batch_sizes = [1, 2, 4, 8, 16, 32, 64, 128]
learning_rates = [2e-5, 3e-5, 4e-5]
modelos = ['BERTimbau_Base', 'BERTimbau_Large']

[31] Quando a perda de treinamento decresce com as épocas, enquanto a perda de validação crescem há forte indicativo de overfitting.ª

ª Traduzido livremente de Machine Learning Students Overfit to Overfitting

No gráfico ao lado, que combina três visualizações dos dados, há a informação que todo o processo levou, aproximadamente, 30 minutos. A partir da 3ª época, a perda de validação começou a oscilar, não mais voltando à sua menor marca. Os dados seguiram se ajustando na validação, com sua curva monótona decrescente a partir do ponto citado, ou seja, exemplo concreto de quando o modelo parou de generalizar.
Em relação às métricas, o comportamento das quatro (acurácia, precisão, F1 Score, Recall) foram muito semelhantes na progressão e, em função do desbalanceamento dos dados e da melhor comunicação visual, seguiremos abordando apenas o "F1 Score".
Para o exemplo específico no universo das **atribuições**, BERT Base, Batch Size 1 e Learning Rate 2e-5, no binômio perdas e métricas, a melhor época foi a 3ª. Além disso, o treinamento parou em função do "Early Stop", que é uma regularização para evitar *overfitting* em métodos iterativos, a finalização se deu após quatro épocas sem melhoria da perda de validação.

Para a análise visual de cada giro do gridsearch, foi desenvolvida uma ferramenta de observação dos resultados, cumprindo um monitoramento rigoroso para garantir argumentos para cumprir os objetivos geral e específicos, em busca de um modelo performático na tarefa do processamento da linguagem natural para a classificação das **atribuições** e das **atividades**, nas categorias gestão, assessoramento e operacional.

Com a ferramenta, foi possível observar o resultados como o da combinação de parâmetros a seguir:

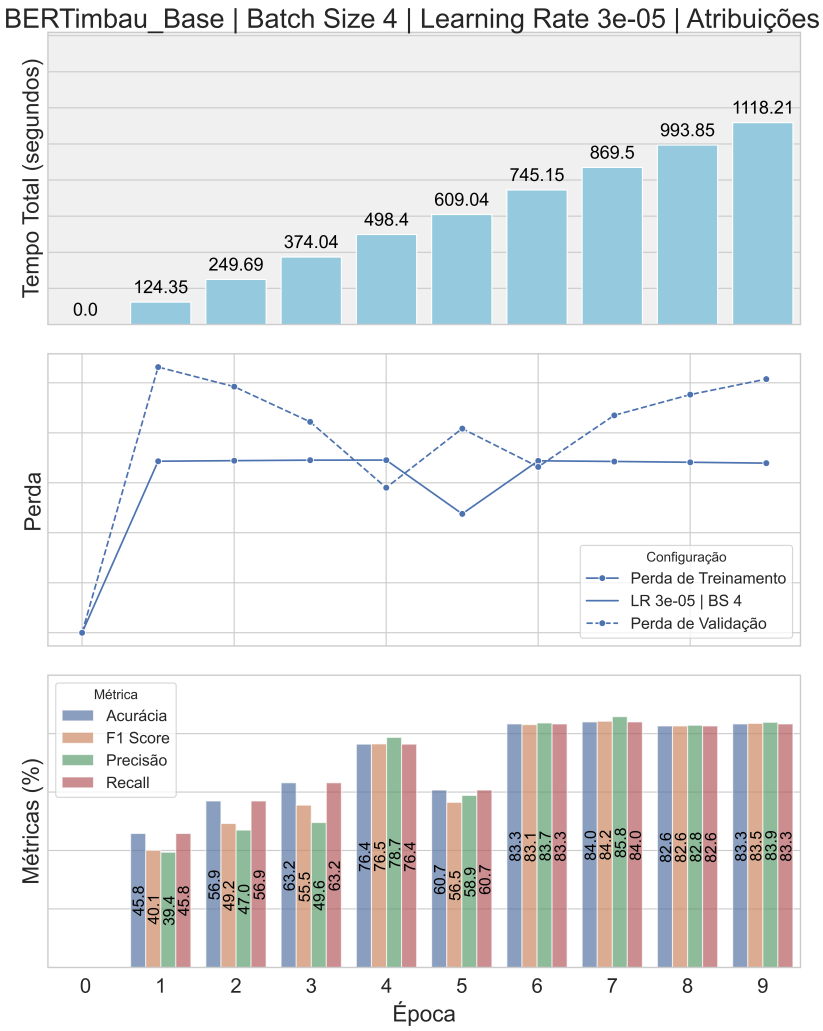


Figura 18: Gráfico das **Atribuições** no modelo BERT Base, batch size 4, learning rate 3e-5

Uma lupa nos dados finos da 4ª até a 6ª época, revela que houve um queda drástica no rendimento do modelo na época 5. As configurações do modelo podem ter colaborado para essa queda, mas, outros atributos inseridos, e o estado da arte do BERT, o trouxeram de volta a bom termo na 6ª época.

O gráfico da figura 15 foi construído por essa ferramenta, que é um dos módulos do gerador de gráficos customizáveis com definição: do modelo do BERT (Base ou Large); dos conjunto batch size e learning rate; do tipo do universo dos dados (**atribuições** ou **atividades**); das épocas mínima e máxima; e das métricas a serem exibidas, tudo isso escolhidas através de widgets do Python, com o controle inicial exposto abaixo:

Figura 17: Seletor do gerador de gráficos

O comportamento das métricas vinha numa crescente até piorar na 5ª época, e depois se recuperar na época 6.

Nas configurações, há o parâmetro "warmup_steps=500", que pode causar flutuações iniciais, quando o modelo ainda está ajustando seu aprendizado.

Outro atributo, a "max_grad_norm" pode ter limitado o ajuste fino, resultando em uma perda de validação maior e medições de referência menores.

Por outro lado, o "load_best_model_at_end=True" permitiu recarregar os melhores parâmetros, recuperando o modelo logo na 6ª época.

Com todos os dados catalogados, era hora de comparar os resultados, num gráfico de calor (heatmap) confrontando Batch size com Learning Rate, nos BERTs Base e Large, aquecendo o chart em função do "F1 Score".

Contudo, a topologia do gráfico dependia da localização da melhor configuração em cada conjunto de parâmetros, ou seja, qual a **menor perda de validação** com o maior "F1 Score".

Mas os dados continham oscilações, para cima e para baixo. Como visto nos dois gráficos anteriores. O problema residia nessa segunda hora, quando a oscilação era decrescente, mas sem trazer um bom momento de perda de validação, isto é, uma flutuação incerta do rendimento do modelo, destoando da tendência de comportamento e da métrica do "F1 Score", em regra, reduzida em relação aos resultados interessantes antes obtidos, em resumo, um tipo de *outlier*.

Nesse caso, não seriam valores que no conjunto geral dos dados estariam fora de contexto, mas sim numa sequência de passos (variável, deslizante, móvel), estariam fora de um intervalo aceitável (e personalizável) de predição, frente à regressão linear escolhida.

Isso foi visto em um dos mapas de calor produzidos, em miniatura na figura 22. Pois o "F1 Score" trazido para o BERT Base, Batch Size 128, Learning Rate 2e-05, foi de 8,08%, o que chamou atenção e pediu uma outra abordagem, pois, se isso estava ali para um valor, poderia estar para muitos outros.

A abordagem utilizada neste trabalho para a desconsideração automática de *outliers* para localizar os menores valores na perda de validação se baseia em conceitos da análise de séries temporais e da detecção de anomalias, buscando, como ensinaram HUBER & RONCHIETTI (2009) [33], estimadores robustos "projetados para serem resistentes a outliers, ou seja, eles não devem ser influenciados desproporcionalmente por alguns poucos pontos extremos nos dados", de modo que pontos extremos distorçam a escolha da melhor configuração do modelo.

Para analisar para configuração e estimar tendências, foi aplicada uma janela deslizante e regressão linear, juntamente com a definição percentual de limites de variação, de modo a identificar pontos que se desviam significativamente da tendência geral, similarmente a técnicas de detecção de anomalias baseadas em distância. Essa abordagem se assemelha ao uso de estimadores robustos em estatística, que descartam os valores mais extremos para evitar que eles influenciem desproporcionalmente o resultado. Ressaltando que o objetivo é identificar a menor perda com maior "F1 Score", não refazer a estrutura de aprendizagem do BERT.

Em AGGARWAL (2017) [32], os *Outliers* são definidos como "objetos que são significativamente diferentes do resto dos dados"
E ele segue: "A detecção de outliers é um passo importante em muitos aplicativos de mineração de dados, pois outliers podem representar ruído, erros ou eventos interessantes que merecem investigação adicional."

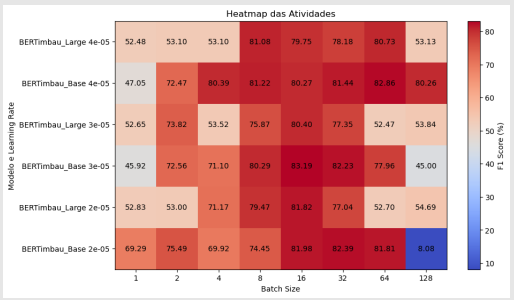


Figura 20: Heatmap das **Atividades**

No contexto do ajuste de hiperparâmetros, a presença de anomalias pode levar a escolhas subótimas da melhor configuração do modelo, especialmente em cenários de dados desbalanceados.
A desconsideração dessas predições ditas aqui como *outliers* , busca controlar seu impacto na análise, garantindo que a escolha da melhor configuração do modelo não seja distorcida por pontos extremos.

CAO *et al* (2024) [34] relata que a "estratégia de aprendizagem implícita pode falhar em identificar com precisão relações anormais através do contexto global. Em contraste, a estratégia explícita pode delinear melhor as relações e produzir resultados de detecção mais precisos", em tradução livre, isto é, implantar um limite direto, pode auxiliar na separação das anomalias e resultar numa localização mais eficiente dos pontos ótimos.

Com esse ponto crítico, foi necessária uma abordagem criativa, para evitar a leitura humana de todos os 96 modelos gerados. Para tal, foi desenvolvido um método de análise das sequências de perdas de validação de modo a estruturar múltiplos parâmetros preditivos para destacar pontos que não estavam gerando conhecimento ao modelo, e por isso, não poderiam ser considerados ótimos, mesmo que estivessem com a menor das perdas de validação.

Os valores de tolerância para ampliação ou redução das perdas a cada época buscaram perceber automaticamente movimentos bruscos nos gráficos, em especial os decrescentes, para aferir corretamente a época de melhor performance. No caso do gráfico da figura 21, os pontos bruscos só foram crescentes, mesmo destacados, não ajudaram a avaliar a melhor época, a 13ª, com um bom "F1 Score" de 83, 19%, para $P1 = P2 = P3 = 15\%$.

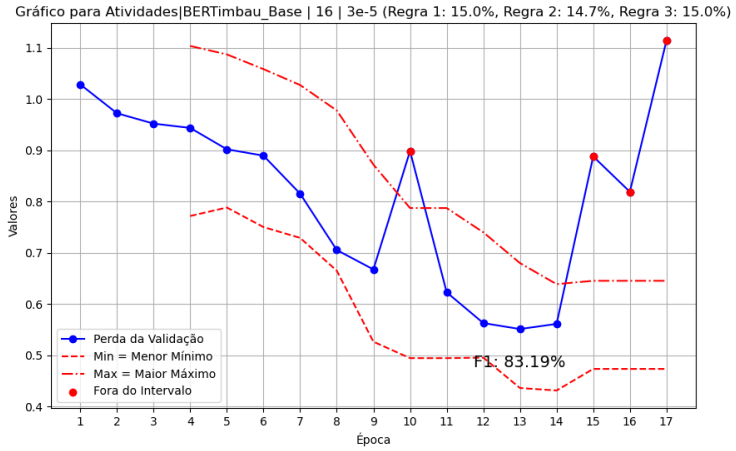


Figura 21: Gráfico do modelo **Atividades**|Base|16|3e-5

Segue outro exemplo de modelo que expôs bem o intuito do localizar a melhor época com menor perda de validação.

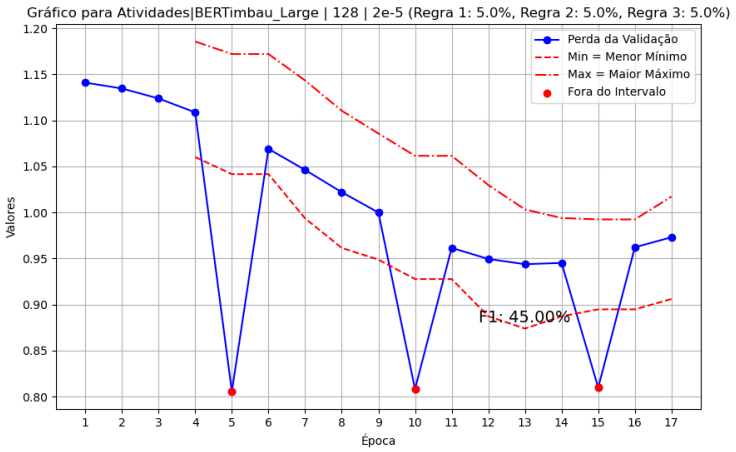


Figura 22: Gráfico do modelo **Atividades**|Large|128|2e-5

Os valores das perdas de validação em cada época foram validados com quatro parâmetros P comparativos considerados:
$P0$**:** se as perdas de validação das duas épocas subsequentes foram menores;;
$P1$**:** $\pm x$ % em relação ao valor da época anterior;
$P2$**:** $\pm y$ % em relação à média dos valores das duas épocas anteriores (com valores aceitos);
$P3$**:** $\pm z$ % do valor predito em relação a uma regressão linear de n épocas anteriores escolhidas por intervalos de tamanho variável, móvel, fluído, cambiante, a partir de uma quantidade fixa inicial de épocas (chamadas de canônicas).
O $P0$ é objetivo, binário, ou o ponto cumpre ou não, já para os demais valores abaixo do mín$\{P1, P2, P3\}$ e acima do máx$\{P1, P2, P3\}$ existiu a marcação em vermelho no gráfico e não foram considerados na busca pelas melhores configurações em cada modelo.

Atividades | BERTimbau Large | Batch Size 128 | Learing Rate 2e-5, porém esse não teve um bom "F1 Score" ficando com 45, 00%. No entanto, fica claro como os thresholds (limiares) definidos excluíram os pontos de grande oscilação. A porcentagem escolhida buscava uma métrica balanceada que não explodisse para a entrada de falsos positivos e nem reprimisse demais, excluindo os falsos negativos.

Por outro lado, deve-se analisar o *trade off* entre a redução da Perda de Validação e o aumento do "F1 Score", com isso foram construídos dois mapas de calor (heatmaps), um para as **atribuições** e outro para as **atividades**, de modo a buscar o melhor modelo, com base no "F1", para cada cenário.

Os *heatmaps* fizeram o cruzamento entre os batch sizes e o modelo BERT (Base e Large) acoplado com o learning rate, e foi aquecido do "F1 Score".

De modo geral, o resultado foi amplamente satisfatório, com melhores números para as **atribuições**. O que pode ser explicado pela construção mais genérica e pasteurizada dos textos. Findando em um top aproximado de "F1 Score" de 87,5%.

SINGH (2018) [35] ensina: "(...) precisamos encontrar o equilíbrio certo/bom sem *overfitting* e *underfitting* dos dados. Esse *trade-off* em complexidade é a razão pela qual existe uma compensação entre viés e variância."[a]

[a] Traduzido de Understanding the Bias-Variance Tradeoff

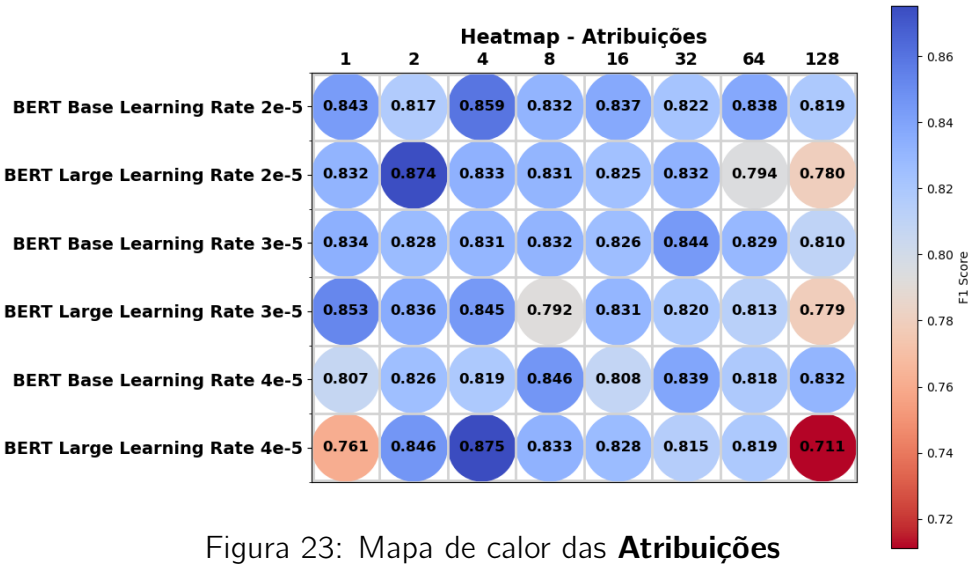


Figura 23: Mapa de calor das **Atribuições**

No caso das **atividades** a amplitude dos resultados foi maior, também pela não uniformidade de trabalho dos setores piloto usados pelo GT Compatibilização. Nesse caso, o topo foi de de "F1 Score" de 86,4%.

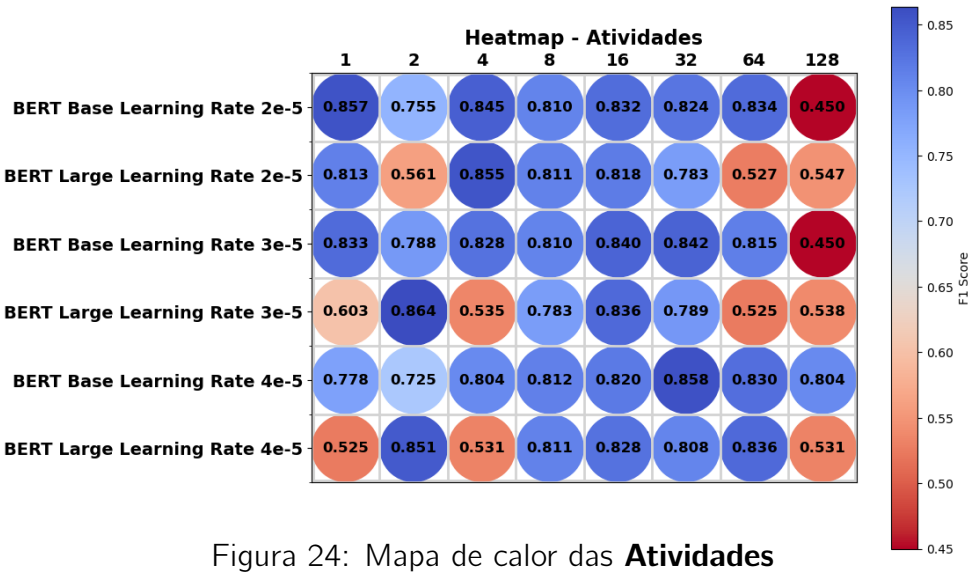


Figura 24: Mapa de calor das **Atividades**

7 Conclusão

A aplicação do modelo BERT para classificação passou por muitas etapas que garantiram novas descobertas frente aos dados existentes para o trabalho do GT Compatibilização. Desnudou, de modo matemático, a complexidade do trabalho para o trato unicamente humano. Garantiu uniformidade com o uso do algoritmo. Produziu diversos tipos de gráficos com o resultado das observações. Por fim, implementou um método inovador, mesmo que ainda empírico, para busca pela otimização do modelo computacional, expondo, minimamente, que ainda há muito o que explorar no campo do *Machine Learning*.

Os dados desbalanceados não foram majorados com *Data Augmentation* ou Aumento de Dados ou Enriquecimento dos Dados que, segundo FENG *et al.* (2021) [36] "refere-se a estratégias para aumentar a diversidade de exemplos de treinamento sem coletar explicitamente novos dados". Sendo que, os novos textos criados, complementa ANDRADE (2022) [37], "por serem sintéticos, têm seus rótulos conhecidos, e podem ser utilizados para aumentar a quantidade de dados rotulados para treino", nem outras técnicas, pois, o intuito estava em simular ao máximo os dados reais. Foram buscadas bibliotecas para tal tarefa, com funcionamento local, pela sensibilidade dos dados, mas optou-se por tentar perceber o poder do BERT com o enredamento dos dados originais, já exposto em várias momentos desse texto, e o BERT mostrou o porquê de ser tratado como o estado da arte.

O treinamento teve múltiplas abordagens, com resultados satisfatórios, contudo, dado tempo necessário para cada giro do BERT nos hiperparâmetros escolhidos e as múltiplas épocas aplicadas[6], optou-se por não usar o Cross Validation, ficando para abordagens futuras.

O salvamento de Checkpoints do modelo durante o treinamento evitou a perda do progresso nos casos de interrupções. Sendo permitido, como poucas linhas de código, a retomada do treinamento do ponto de pausa, garantindo que as melhores versões fossem sejam preservadas.

A análise das melhores configurações deu-se pelo *Trade Off* entre a suavização Perda de Validação e a majoração do "F1 Score", desafiando a criação do método de busca automatizado com controle deslizante (similar a média móvel) para garantir que os parâmetros acompanhavam o comportamento mais recente do modelo, o que, humildemente, foi a cereja do bolo do trabalho.

Poder-se-ia seguir para a parte três da missão do GT Compatibilização, e desenvolver o estudo das 484**atribuições** versus 585 **atividades**. O que não é muito complexo, pois, derradeiramente, seria uma análise de do produto cartesiano de cada **atividade** em relação a cada **atribuição**, resultando, em $484 \times 585 = 283.140$ pares de relação de dados de entrada, com validação binária de compatibilidade. Mas a ideia foi fechar o elemento chave para o método, a classificação correta das **categorias**, de modo a permitir o cruzamento correto entre as classes homônimas em cada universo.

Para esse elemento cartesiano da missão, sugere-se a ampliação, com o uso do Retrieval Augmented Generation (RAG), uma ferramenta de guarda e vinculação de dados destacados treinados, como escreve LIN (2023) [38], "através de técnicas de busca e recuperação de informações que tem como objetivo identificar algumas similaridades presentes no texto que possam conter informações expressivas", podendo ser acoplada a modelos abertos (por exemplo, Llama e versões antigas e disponíveis do ChatGPT), para criar e gerar as respostas.

A literatura recomenda ainda o uso de otimizadores e, no caso desse trabalho, foi o AdamW, fundamental para ajustar os pesos do modelo de maneira eficiente que funciona para, aprendendo de VIEIRA (2022) [39], "realizar ajustes em alguns parâmetros durante o treinamento para melhorar o desempenho do modelo". Além disso, a utilização de um scheduler de taxa de aprendizado, lição aprendida de TSANGOURI (2024) [40] "usado para ajustar o *learning rate* em intervalos regulares", em tradução livre, ou seja, REFINAR dinamicamente a taxa de aprendizado ao longo das épocas, melhorando a convergência do modelo.

[6] O modelo foi configurado para ir até 500 épocas, mas com um Early Stop, um freio de controle, caso não obtivesse melhora após 3 ou 4 épocas (ambos testados), no intuito de evitar o sobreajuste dos dados, o famoso *Overfitting*.

Outras técnicas implantadas, como o uso de *callbacks*, *dropout* e a *normalização do gradiente*, além do registro permanente das melhores épocas, foram essenciais para que o modelo fosse bem testado e performasse bem e com estabilidade no treinamento.

Pela ciência de que: a precisão avalia a proporção de verdadeiros positivos em relação ao total de positivos previstos, enquanto o recall mede a proporção de verdadeiros positivos em relação ao total de positivos reais. Em um cenário de desbalanceamento, o "F1 Score", que **combina as duas métricas** previamente citadas, forneceu a visão equilibrada dentre as métricas clássicas.

As bases de comparações com os modelos iniciais de árvore foram vencidas rapidamente, mesmo sem tunagem de parâmetros, pois, para a tarefa específica, o BERT, de fato, é o mais adequado. Destacando, mais uma vez, a sua capacidade de entender o contexto bidirecionalmente, ponto forte em comparação com modelos mais simples, além do registro de nuances e polissemias típicas do nosso idioma materno.

Em relação aos modelos BERTimbau Base e Large, talvez pelo volume pequeno de dados, não houve diferença significativa, sendo a escolha de um dos dois mais em função do poderio computacional. Com a majoração da base de **atividades**, com a entrada de cada novo setor no estudo e, visto que a base de **atribuições** só cresce quando da criação ou alteração de cargos e funções, os parâmetros e as camadas a mais do BERT Large possam fazer diferença. Ou seja, a escolha do modelo depende do balanço entre desempenho e recursos computacionais disponíveis.

Por outro lado, é ressalta-se que os dados foram usados sem ajustes e, se eles vieram com viés, não proposital, resultaram em previsões injustas ou discriminatórias ou desproporcionais. Além disso, é fundamental garantir a alta qualidade das informações, pois dados ruidosos ou mal anotados podem afetar negativamente o desempenho do modelo. Por isso, avaliar, mitigar vieses e garantir a qualidade dos dados de treinamento faz-se necessário para afiançar a equidade do modelo. Sem perder de vista a ética necessária para usá-los e aplicá-los.

Nesse cenário, surge a demanda pela regulação e controle das formas de uso. A inteligência artificial responsável é um dos caminhos possíveis, pois foi "projetada para ajudar a reconhecer, preparar e mitigar potenciais efeitos nocivas da IA, traduzido de [41]. Reforçando que garantir que a IA seja desenvolvida e utilizada de forma ética requer um esforço de diversos setores da sociedade.

$$\text{IA Responsável} = \text{Práticas Éticas} + \text{Políticas Responsáveis}$$

A construção de uma IA Responsável implementa princípios éticos, citados ao longo desse texto em outras palavras, com os argumentos inspirados nos "Princípios para inteligência artificial da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) [7] "[42]

a) A inteligência artificial deve beneficiar as pessoas e o planeta, impulsionando o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar;

b) Os sistemas de inteligência artificial devem ser projetados de maneira a respeitar o estado de direito, os direitos humanos, os valores democráticos e a diversidade, e devem incluir salvaguardas apropriadas – por exemplo, possibilitando a intervenção humana sempre que necessário – para garantir uma sociedade justa e leal;

c) Deve haver transparência e divulgação responsável em torno dos sistemas de inteligência artificial para garantir que as pessoas entendam quando estão envolvidas com eles e possam desafiar os resultados;

d) Os sistemas de inteligência artificial devem funcionar de maneira robusta, segura e protegida durante toda a vida útil, e os riscos potenciais devem ser continuamente avaliados e gerenciados;

e) As organizações e indivíduos que desenvolvem, implantam ou operam sistemas de inteligência artificial devem ser responsabilizados pelo seu bom funcionamento, de acordo com os princípios acima.

[7] Traduzido da Recommendation of the Council on Artificial Intelligence.

Para fechar esse item de extrapolação, não se deve perder de vista que a IA pode cometer erros, equívocos ou alucinações. Sendo assim, faz-se necessário formar/instruir os usuários de modo a atentarem-se a essas possibilidades e terem ferramentes para identificá-las, contorna-las, complementa-las e corrigi-las, quando necessário.

Para a proteção contra erros e a "estupidez artificial", devemos reconhecer as pontos nos quais a Inteligência Artificial precisa avançar, e onde requer a supervisão humana adequada e ética. Isso envolve a implementação de medidas de segurança, verificação e validação para identificar e mitigar potenciais falhas ou comportamentos indesejados dos sistemas de IA. Assim como, educação e conscientização sobre os limites da IA de modo a ajudar usuários a entenderem suas capacidades e restrições, reduzindo assim o potencial de erros, consequências negativas e usos irresponsáveis.

$$\begin{aligned} &\text{Proteção contra Erros} = \\ &\quad \text{Supervisão Humana} \times \text{Limites da IA} = \\ &\text{Medidas de Segurança} + \text{Educação Consciente} \end{aligned}$$

Voltando ao cerne do trabalho, prestando contas dos objetivos planejados, cumpridos pela emulação do modelo com o BERT, a análise das bases de dados, aplicação de outros modelos de Machine Learning, comparação e visualização dos resultados.

Ao fim, ao cabo, reforçasse a importância da continuidade dos estudos na área do Processamento da Linguagem Natural NLP e a necessidade de explorar novas técnicas e abordagens para enfrentar desafios emergentes com as tecnologias disruptivas, mesmo para os desafios cotidianos simples. Cada vez mais se fala em um momento único da história da humanidade, com algo que pode ser usado para acelerar a compressão de muitas áreas.

Referências

[1] Paulo Junior Trindade dos Santos, Cristhian Magnus de Marcus, and Gabriela Samrsla Möller. Tecnologia disruptiva e direito disruptivo: Compreensão do direito em um cenário de novas tecnologias, 2019. Acesso em: 26 jul. 2024.

[2] Rodrigo Vieira de Araujo, Jurandir Zullo Jr, and Carlos Vinícius Alves Torres. Data governance: a critical analysis of big data exploitation in the public sector. *Revista de Administração Pública*, 54(4):775–791, julho/agosto 2020.

[3] Marcelo Alexandrino and Vicente Paulo. *Direito Administrativo Descomplicado*. Método, São Paulo, 18ª ed. edition, 2010. Apud Tribunal de Contas de Santa Catarina, Orientações sobre Desvio de Função de Servidor no Serviço Público.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[5] Enise Barth Teixeira. A análise de dados na pesquisa científica: importância e desafios em estudos organizacionais. *Desenvolvimento em Questão*, 1(2):177–201, outubro 2011.

[6] Dalia Jasim. Main steps for doing data mining project using weka. fevereiro 2016.

[7] Data Science Academy. O efeito do batch size no treinamento de redes neurais artificiais, 2024. Acesso em: 29 jul. 2024.

[8] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016.

[9] Sayak Paul. Introduction to cyclical learning rates. `https://www.datacamp.com/tutorial/cyclical-learning-neural-nets`, outubro 2018. Acesso em: 29 jul. 2024.

[10] Camila Maione. Balanceamento de dados com base em oversampling em dados transformados. Dissertação de mestrado, Universidade Federal de Goiás, Instituto de Informática, Goiânia, 2020. 135 f.: il.

[11] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, Chicago, second edition, 2007.

[12] Fernanda Fernandes Matos, Renato Rocha Souza, and Zilma Silveira Nogueira Reis. Análise de dados de saúde: Mineração de texto com a utilização do orange canvas para exploração da informação. In *XX Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB 2019*, Florianópolis, SC, outubro 2019.

[13] André Filipe de Moraes Batista and Alexandre Dias Porto Chiavegatto Filho. Machine learning aplicado à saúde. In *Livro de Minicursos do XIX Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2019)*, chapter 1. Sociedade Brasileira de Computação, Brasil, junho 2019. Categorias: Computação Aplicada à Saúde.

[14] Microsoft. Conceitos de pipelines de machine learning no azure machine learning. Acessado em: 30 de julho de 2024.

[15] Escola DNC. Como otimizar modelos de machine learning com grid search?, janeiro 2024. Acessado em: 30 de julho de 2024.

[16] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[17] V. Roshan Joseph. Optimal ratio for data splitting. *arXiv*, 2202(03326v1), February 2022. H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

[18] Roger Grosse. Lecture 9: Generalization. https://www.cs.toronto.edu/~lczhang/321/notes/notes09.pdf. Accessed: 2024-08-02.

[19] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022, 2019.

[20] Foster Provost. Machine learning from imbalanced data sets 101: Extended abstract. https://pages.stern.nyu.edu/~fprovost/Papers/skew.PDF. New York University, fprovost@stern.nyu.edu.

[21] Hugging Face. Tokenizer summary. https://huggingface.co/docs/transformers/tokenizer_summary. Accessed: 2024-08-02.

[22] Hugging Face. Hugging face transformers tokenizer summary. https://www.youtube.com/watch?v=nhJxYji1aho, 2021. Accessed: 2024-08-02.

[23] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv*, 2002(06305v1), fevereiro 2020.

[24] Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for deep learning: A taxonomy. *arXiv*, 1710.10686, 2017. Subimissão: 20/10/2017.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *scikit-learn: Machine Learning in Python*. scikit-learn, 2011. Acesso em: 2024-08-05.

[26] Hrvoje Smolic. Understanding the importance of f1 score in machine learning, 2024. Acesso em: 2024-08-05.

[27] Tejashree Ganesan. How to tune a decision tree in hyperparameter tuning, 2024. Acesso em: 2024-08-05.

[28] Bhushan Talekar and Sachin Agrawal. A detailed review on decision tree and random forest. *Biosc.Biotech.Res.Comm.*, 13(14):245–248, 2020. Special Issue.

[29] Avi Chawla. The evolution of embeddings: Looking back to the pre-transformer times, 2024. Acesso em: 2024-08-05.

[30] Marcos Yuri Rosa Junior. Comparando bert, roberta e distilbert para análise de sentimentos em texto, dezembro 2021. Data de aprovação: 10/dezembro/2021.

[31] Matias Valdenegro-Toro and Matthia Sabatelli. Machine learning students overfit to overfitting. In *Proceedings of the 2nd Teaching in Machine Learning Workshop*, Groningen, The Netherlands, 2022. PMLR. arXiv:2209.03032.

[32] Charu C. Aggarwal. *Outlier Analysis*. Springer, Cham, Switzerland, 2nd ed. edition, 2017.

[33] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, Hoboken, NJ, USA, 2nd ed. edition, 2009.

[34] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv*, 2401.16402, janeiro 2024. Submissão: 20/01/2024.

[35] Seema Singh. Understanding the bias-variance tradeoff. *Towards Data Science*, Maio 2018.

[36] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp, 2021.

[37] Ivan Caramello de Andrade. Geração de sentenças para classificação semissupervisionada de textos. Monografia de especialização, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil, 2022. Orientador: Prof. Dr. Ricardo Marcacini.

[38] Felipe Jun Ting Lin. A utilização de modelos de llm para geração de bulas automatizadas, 2023. Orientador: Prof. Dr. Luciano de Andrade Barbosa.

[39] Juliano Correia Passos Vieira. Análise comparativa de modelos de redes neurais para análise de sentimentos em português coloquial, 2022.

[40] Christina Tsangouri. Multibert: Enhanced multi-task fine-tuning on minbert. Stanford CS224N Default Project, Department of Computer Science, Stanford University, 2024.

[41] Comptroller and Auditor General of India. Compendium on responsible artificial intelligence. `https://sites.tcu.gov.br/SAI20/documents/India/SAI20-Summit/Compendium-on-Responsible-AI.pdf`, 2023.

[42] OECD. Recommendation of the council on artificial intelligence. OECD Legal Instruments, OECD/LEGAL/0449, 2024. © OECD 2024.

CENTRO UNIVERSITÁRIO SENAI CIMATEC

CURSO DE ESPECIALIZAÇÃO LATO SENSU EM DATA SCIENCE E ANALYTICS

ATA DE APRESENTAÇÃO DE PROJETO FINAL DE CURSO

Ata de apresentação do Projeto Final de Curso **"MACHINE LEARNING E BERT: INOVAÇÃO NA COMPATIBILIZAÇÃO DE CARGOS E FUNÇÕES COM OS POSTOS DE TRABALHO NO MPBA"**, submetido pelo aluno **Tiago Miranda de Magalhães,** como parte dos requisitos para obtenção do Certificado de Especialista em *Data Science e Analytics* pelo Centro Universitário SENAI CIMATEC, às 19:30h do dia 22 de agosto de 2024. Reuniu-se no CIMATEC, a Banca Examinadora designada pela Coordenação de curso, constituída por Prof. Dr. Oberdan Rocha Pinheiro e Msc. Adroaldo Santos Soares.

A coordenadora do curso Patricia Freitas Tourinho deu início aos trabalhos e a exposição foi realizada pelo estudante dentro do prazo de tempo estabelecido. Ao final da apresentação a banca reuniu-se atribuindo a seguinte nota: 10,0 **(dez).**

A banca de avaliadores decidiu pela:

(x) Aprovação do trabalho
Caberá ao aluno apresentar em no máximo em 30 (trinta) dias a contar da data de assinatura desta Ata, uma cópia do trabalho em PDF com restrição de edição. A Ata de Apresentação do Projeto Final de Curso deve ser digitalizada e inserida na terceira página do PFC.

( ) Reprovação do trabalho
O aluno terá que se matricular novamente no TCC – Trabalho de Conclusão de Curso e ser submetido a uma banca avaliadora no semestre seguinte**.**

As ações consequentes ao status de Aprovação deverão obedecer ao prazo proposto acima sob pena do parecer final ser modificado para o status de Reprovado automaticamente e sem possibilidade de recurso.

Para constar, lavrou-se a presente ata que vai assinada por todos os membros da Banca. Por estarem cientes de suas obrigações estão de acordo com os termos desse documento:

Salvador, 22 de agosto de 2024

Assinado eletronicamente por:
Patricia Freitas Tourinho
CPF: ***.733.265-**
Data: 28/08/2024 16:21:57 -03:00

Profª. Esp. Patricia Freitas Tourinho
Coordenadora do Curso

Assinado eletronicamente por:
Oberdan Rocha Pinheiro
CPF: ***.073.705-**
Data: 28/08/2024 08:22:17 -03:00

Prof. Dr. Oberdan Rocha Pinheiro
Professor Orientador

Assinado eletronicamente por:
Adroaldo Santos Soares
CPF: ***.731.145-**
Data: 28/08/2024 08:58:00 -03:00

Prof. Msc. Adroaldo Santos Soares
Professor convidado