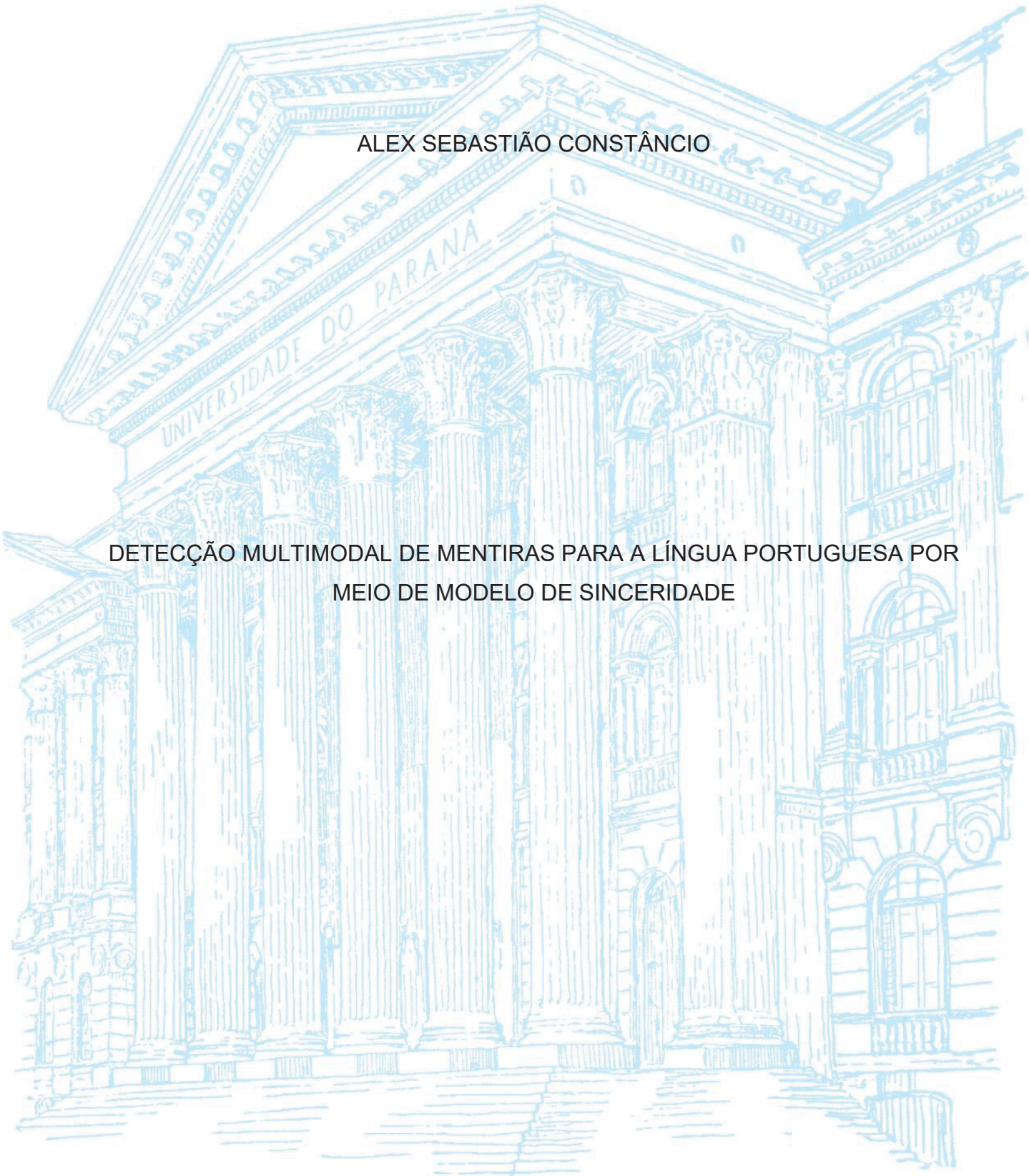
UNIVERSIDADE FEDERAL DO PARANÁ

ALEX SEBASTIÃO CONSTÂNCIO



DETECÇÃO MULTIMODAL DE MENTIRAS PARA A LÍNGUA PORTUGUESA POR MEIO DE MODELO DE SINCERIDADE

CURITIBA

2024

ALEX SEBASTIÃO CONSTÂNCIO

DETECÇÃO MULTIMODAL DE MENTIRAS PARA A LÍNGUA PORTUGUESA POR MEIO DE MODELO DE SINCERIDADE

Tese apresentada ao Programa de Pós-Graduação em Gestão da Informação, Área de Concentração de Informação, Tecnologia e Gestão do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná como requisito para a obtenção de título de doutorado.

Orientadora: Profª. Drª. Deborah Ribeiro Carvalho
Co-orientadoras: Profª. Drª. Helena de Fátima Nunes Silva e Profª. Drª. Jocelaine Martins da Silveira.

CURITIBA
2024

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

C757d Constâncio, Alex Sebastião

Detecção multimodal de mentiras para a língua portuguesa por meio de modelo de sinceridade [recurso eletrônico] / Alex Sebastião Constâncio – Curitiba: UFPR, 2024.

Tese (Doutorado) apresentada Curso de Pós-Graduação em Gestão da Informação, Setor de Sociais Aplicadas, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Gestão da Informação.

Orientadora: Profª. Drª. Deborah Ribeiro Carvalho
Co-orientadoras: Profª. Drª. Helena de Fátima Nunes Silva e Profª. Drª. Jocelaine Martins da Silveira

1. Inteligência artificial. 2. Software - Desenvolvimento. 3. Aprendizagem do computador. I. Carvalho, Deborah Ribeiro. II. Silva, Helena de Fátima Nunes. III. Silveira, Jocelaine Martins da. IV. Universidade Federal do Paraná. V. Título.

CDD 001.535

Bibliotecária: Vilma Machado CRB-9/1563

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA INFORMAÇÃO - 40001016058P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **ALEX SEBASTIÃO CONSTÂNCIO** intitulada: **Detecção multimodal de mentiras para a Língua Portuguesa por meio de Modelo de Sinceridade**, sob orientação da Profa. Dra. DEBORAH RIBEIRO CARVALHO, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua aprovado no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 06 de Dezembro de 2023.

DEBORAH RIBEIRO CARVALHO
Presidente da Banca Examinadora

RONAN ASSUMPCAO SILVA
Avaliador Interno (INSTITUTO FEDERAL DO PARANÁ)

BRUNO ANGELO STRAPASSON
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

JOCELAINE MARTINS DA SILVEIRA
Coorientador(a)

MYRIAM REGATTIERI DE BIASE DA SILVA DELGADO
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

HELENA DE FÁTIMA NUNES SILVA
Coorientador(a)

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA INFORMAÇÃO - 40001016058P1

ATA Nº412023

ATA DE SESSÃO PÚBLICA DE DEFESA DE DOUTORADO PARA A OBTENÇÃO DO GRAU DE DOUTOR EM GESTÃO DA INFORMAÇÃO

No dia seis de dezembro de dois mil e vinte e tres às 08:00 horas, na sala Salão Nobre de Ciências Contábeis, UFPR - Setor de Ciências Sociais Aplicadas, 1º Andar Jardim Botânico, foram instaladas as atividades pertinentes ao rito de defesa de tese do doutorando **ALEX SEBASTIÃO CONSTÂNCIO**, intitulada: **Detecção multimodal de mentiras para a Língua Portuguesa por meio de Modelo de Sinceridade**, sob orientação da Profa. Dra. DEBORAH RIBEIRO CARVALHO. A Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná, foi constituída pelos seguintes Membros: DEBORAH RIBEIRO CARVALHO (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ), RONAN ASSUMPCAO SILVA (INSTITUTO FEDERAL DO PARANÁ), BRUNO ANGELO STRAPASSON (UNIVERSIDADE FEDERAL DO PARANÁ), JOCELAINE MARTINS DA SILVEIRA (UNIVERSIDADE FEDERAL DO PARANÁ), MYRIAM REGATTIERI DE BIASE DA SILVA DELGADO (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ), HELENA DE FÁTIMA NUNES SILVA (UNIVERSIDADE FEDERAL DO PARANÁ). A presidência iniciou os ritos definidos pelo Colegiado do Programa e, após exarados os pareceres dos membros do comitê examinador e da respectiva contra argumentação, ocorreu a leitura do parecer final da banca examinadora, que decidiu pela Aprovado. Este resultado deverá ser homologado pelo Colegiado do programa, mediante o atendimento de todas as indicações e correções solicitadas pela banca dentro dos prazos regimentais definidos pelo programa. A outorga de título de doutor está condicionada ao atendimento de todos os requisitos e prazos determinados no regimento do Programa de Pós-Graduação. Nada mais havendo a tratar a presidência deu por encerrada a sessão, da qual eu, DEBORAH RIBEIRO CARVALHO, lavrei a presente ata, que vai assinada por mim e pelos demais membros da Comissão Examinadora.

CURITIBA, 06 de Dezembro de 2023.

DEBORAH RIBEIRO CARVALHO
Presidente da Banca Examinadora

RONAN ASSUMPCAO SILVA
Avaliador Interno (INSTITUTO FEDERAL DO PARANÁ)

BRUNO ANGELO STRAPASSON
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

JOCELAINE MARTINS DA SILVEIRA
Coorientador(a)

MYRIAM REGATTIERI DE BIASE DA SILVA DELGADO
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

HELENA DE FÁTIMA NUNES SILVA
Coorientador(a)

RESUMO

A mentira é um fenômeno psíquico-social complexo e pervasivo. É complexo porque pode ter muitas motivações e desencadear múltiplas variações de expressão no emissor. É pervasivo porque pesquisas apontam para uma incidência média de duas mentiras ao dia para cada pessoa. Em alguns casos a comunicação não sincera não oferece risco ao interlocutor ou ao coletivo, mas existem casos em que a descoberta precoce de uma mentira pode representar a diferença entre a vida e a morte. Dados empíricos demonstram que a mentira pode ser, em muitos casos, detectada a partir da observação de pistas involuntariamente expressas pelo emissor. Em resposta aos cenários mais dramáticos que envolvem a mentira, esta pesquisa apresenta um "Modelo de Sinceridade" capaz de identificar padrões de comunicação sincera para então identificar exemplares de narrativas mentirosas. O Modelo de Sinceridade é um modelo de Aprendizado de Máquina, mais especificamente de Aprendizado Profundo, chamado de Autoencoder. Por incorporar caracteres acústicos, verbais e visuais, trata-se de um modelo multimodal. Por ser um modelo treinado com narrativas expressas em português do Brasil, trata-se do primeiro estudo deste tipo dedicado à língua portuguesa. Após 3.290 experimentos com diferentes arquiteturas de Autoencoder, o modelo multimodal final empregando o mecanismo de Atenção *multi-head* atingiu a acurácia balanceada de 0,714 na detecção de mentiras em 49 narrativas de 12 sujeitos, pertencentes a um conjunto de dados denominado "*Multimodal Deception Detection Dataset for Brazilian Portuguese*" (MMDDD-PtBr), elaborado especialmente para este estudo, também o primeiro no mundo dedicado ao português. O Modelo de Sinceridade inova ao utilizar o aprendizado autossupervisionado para seu treinamento, potencialmente pavimentando o caminho para a construção de um modelo de detecção independente de dados rotulados, que ainda hoje são raros. Inova também por abordar o problema da detecção de mentiras como um problema de descoberta de anomalias, para o qual a pesquisa desenvolveu uma nova métrica para quantificação da confiança de detecção, denominada de "Escore de Sinceridade", que também oportunizou um novo modelo de fusão de modalidades. O resultado alcançado supera por 17 pontos percentuais a linha de base de acurácia de 0,540, frequentemente apontada como a probabilidade de um indivíduo não treinado detectar uma mensagem não sincera. A margem de ganho alcançada aponta para os efeitos positivos da abordagem e das técnicas e métricas empregadas.

Palavras-chave: detecção de mentiras; aprendizado autossupervisionado; aprendizado profundo; autoencoders; detecção de anomalias; mecanismo de atenção; conjunto de dados multimodal.

ABSTRACT

Deception is a complex and pervasive psychosocial phenomenon. It's complex because it may have many motivations and trigger multiple expression shifts at the emitter. It is pervasive because research shows that the average person tells two lies a day. In some cases, insincere communication poses no risk to the interlocutor or to the population, but there are cases in which the early discovery of a lie can mean the difference between life and death. Empirical data shows that lies can often be detected by observing clues involuntarily expressed by the emitter. In response to the most dramatic scenarios involving deceptions, this research presents a "Sincerity Model", capable of identifying patterns of sincere communication and then identifying examples of lying narratives. The Sincerity Model is a Machine Learning model, more specifically Deep Learning, called Autoencoder. Because it includes acoustic, verbal and visual features, it is a multimodal model. Since it is a model trained with narratives expressed in Brazilian Portuguese, it is the first study of its kind dedicated to Portuguese. After 3,290 experiments with different Autoencoder architectures, the final multimodal model that uses the multi-head Attention mechanism achieved a balanced accuracy of 0.714 in deception detection of 49 narratives from 12 subjects, belonging to a dataset called the "Multimodal Deception Detection Dataset for Brazilian Portuguese" (MMDDD-PtBr), especially developed for this study, also the first in the world dedicated to Portuguese. The Sincerity Model innovates by using self-supervised learning for its training, potentially paving the way for the construction of a detection model independent of labeled data, which are still rare nowadays. It also breaks new ground by approaching the problem of lie detection as an anomaly detection problem, for which this research has developed a new metric for quantifying the detection confidence, called the "Sincerity Score" that leveraged a novel mechanism for modality fusion. The result achieved exceeds by 17 percentage points the accuracy baseline of 0.540, often referred to as the probability of an untrained individual to detect a deceptive message. The achieved margin of gain points to the positive effects of the approach, as well as the techniques and metrics used.

Keywords: deception detection; self-supervised learning; deep learning; autoencoders; anomaly detection; attention mechanism; multimodal dataset.

DEDICATÓRIA

Dedico esta tese a meus pais, Adelino e Arlete, que colocaram a minha educação e meus valores acima de todas as suas prioridades, dando-me o que eles mesmos não tiveram. E também à minha esposa Denise, pelo estímulo e paciência constantes que soube demonstrar durante as muitas horas em que estive me dedicando a esta conquista.

Dedico também a todos aqueles que praticam a ciência em favor do bem comum, compreendendo que, em sua essência, trata-se da busca pela verdade que explica, demonstra e, em última análise, liberta.

AGRADECIMENTOS

Primeiramente a Deus, pois tudo o que foi possível, o foi por Sua vontade e infinita misericórdia.

À minha família que por vezes recebeu menos atenção do que merecia para que eu pudesse exercer a dedicação necessária a esta realização.

À minha orientadora, Prof.ª. Drª. Deborah Ribeiro Carvalho, que acreditou no potencial da minha proposta, empreendendo seus melhores esforços e compartilhando seus melhores valores no sentido de vê-la finalmente realizada.

Às minhas co-orientadoras Prof.ª. Drª. Helena de Fátima Nunes Silva e Prof.ª. Drª. Jocelaine Martins da Silveira por seus prestimosos conselhos, sugestões e recomendações ao longo da jornada.

Ao Programa de Pós-graduação em Gestão da Informação (PPGGI), que dispendeu esforços e recursos para o apoio financeiro na matrícula dos cursos da Data Science Academy e na publicação do artigo no periódico internacional PloS ONE.

Aos demais professores do PPGGI, assim como à secretária Simone Batista que, por seu empenho diário, torna possível a continuidade do programa, como pináculo do saber.

Aos meus superiores na cadeia de comando, Prof.°. Dr. Marco Antônio Ribas Cavalieri, Valmir Antunes Pereira e Leonardo Melo, pela oportunidade do afastamento remunerado que foi fator decisivo no desenvolvimento e qualidade desta pesquisa.

A todos que de alguma forma, por atos ou omissões, palavras ou silêncio, contribuíram para que os meus esforços pudessem frutificar.

EPÍGRAFE

“Então conhecerás a verdade e a verdade vos libertará.”
Jesus Cristo (João 8:32)

“A verdade nunca se torna uma mentira. É por isso que digo que a verdade é Deus.”
Mahatma Gandhi

“Um computador mereceria ser chamado de inteligente se pudesse enganar um ser humano, fazendo-o acreditar que é humano.”
Alan Turing

“Quando uma pessoa encontra uma verdadeira resposta, as contradições desaparecem.”
Albert Einstein

“Às vezes, as pessoas não querem ouvir a verdade, porque não desejam que suas ilusões sejam destruídas.”
Friedrich Nietzsche

“A verdade é para o homem o que o antimônio é para o ourives: ele a estende sem cessar e nunca a esgota.”
Voltaire

“A arte é a mentira que nos permite conhecer a verdade.”
Pablo Picasso

“A verdade só é encontrada quando se liberta da preocupação de ser o que os outros pensam.”
Ralph Waldo Emerson

“A verdadeira felicidade está na descoberta da verdade.”
Leo Tolstoy

“A verdade é a única arma que um homem precisa para vencer o mal.”
Atticus Finch (O Sol é para Todos, Harper Lee)

“Quando você eliminou o impossível, o que resta, por mais improvável que pareça, deve ser a verdade.”
Sherlock Holmes (Arthur Conan Doyle)

“A verdade vai despir uma alma tão nuamente quanto uma espada.”
Gandalf (O Senhor dos Anéis, J.R.R. Tolkien)

“Ou tu sabes ou não sabes, ‘eu acho’ não existe!”
Adelino Juvenal Constâncio

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

LISTA DE SIGLAS

AE	Autoencoder
AAE	Attention Autoencoder
CSV	Comma-separated value
EQM	Erro quadrático médio
ES	Escore de Sinceridade
FACS	Facial Action Coding System
FAU	Facial Action Unit
GPT	Generative Pre-trained Transformer
GPU	Graphic Processing Unit
IA	Inteligência Artificial
JSON	JavaScript Object Notation
KLd	Kullback-Liebler divergence
KLn	Kullback-Liebler divergence gaussiana
LSTM	Long Short-Term Memory
MAAE	Multi-head Attention Autoencoder
MLP	Multi-layer Perceptron
MMDDD-PtBr	Multimodal Deception Detection Dataset for Brazilian Portuguese
MP4	MPEG-4 Part 14
MSE	Mean Squared Error
NAS	National Academy of Science
PCA	Principal Component Analysis
PLN	Processamento de Linguagem Natural
RNA	Rede neural artificial
RLTDDD	Real-life Trial Deception Detection Dataset
SVM	Support Vector Machines
WAV	Wave-form audio format

SUMÁRIO

1 INTRODUÇÃO

No contexto deste estudo, entende-se por “mentira” toda mensagem proferida cujo conteúdo é considerado falso pelo emissor (Zuckerman; DePaulo; Rosenthal, 1981). É um ato intencional e consciente, visando a um objetivo específico. Se o conteúdo é verdadeiro, mas reconhecido como falso pelo emissor, a mensagem continua sendo considerada como mentira.

Ao longo deste documento as palavras “mentira” e “não sinceridade” são utilizadas como sinônimos, assim como as palavras “verdade” (no sentido de ser o oposto de “mentira”) e “sinceridade” também recebem a mesma acepção.

A detecção de mentiras é um processo que avalia informações de diversas fontes (comportamentais, linguísticas, emocionais e fisiológicas, além das investigativas) para determinar se uma dada mensagem é sincera ou não. Trata-se de um julgamento inferencial guiado por evidências (Levine, 2014). No contexto deste estudo, evidências são informações suficientemente confiáveis para permitir a tomada de uma decisão segura. As evidências investigativas (aquelas provenientes de processos que validam uma informação a partir de dados externos ao emissor) não foram exploradas.

As evidências em foco nesta pesquisa foram baseadas na presença, ausência ou variação de sinais (ou pistas), como por exemplo a alteração na frequência de piscadas ou a modificação no tom da voz durante uma narrativa. Tais sinais são espontaneamente demonstrados pelo emissor da mensagem e se assume que têm poder discriminante. Uma vez percebidos, servem como critério para decidir se uma mensagem é ou não sincera.

Algumas mentiras não têm efeitos danosos, coletiva ou individualmente, quando proferidas para evitar situações embaraçosas ou desconfortáveis (atritos pessoais), objetivando a manutenção de relacionamentos (Turner; Edgley; Olmstead, 1975; Vrij, 2008) e são consideradas como lubrificantes sociais (Vrij, 2008).

Porém, há mentiras cuja detecção pode prevenir efeitos de mais elevada gravidade. A importância de se detectar as mentiras em tais circunstâncias, como em casos de natureza policial, estimulou o desenvolvimento de aparatos tecnológicos para auxiliar na atividade. O exemplo mais famoso de um dispositivo com esta finalidade é o polígrafo, que foi introduzido por John Larson no Departamento de Polícia de Berkeley, em 1921 (Ball, [s.d.]).

O polígrafo é um equipamento incapaz de detectar mentiras. O que o dispositivo faz é registrar alterações fisiológicas experimentadas por um sujeito durante um interrogatório. Dada alteração observada é considerada evidência de sinceridade, ou não, a partir do julgamento do operador (Ekman, 1992).

Portanto, não é a máquina que detecta a mentira (embora o polígrafo seja muitas vezes chamado de “detector de mentiras”). O dispositivo apenas fornece dados que são lidos pelo operador, que efetivamente os interpreta para categorizar um testemunho como sincero ou não. Como consequência, a detecção de mentiras baseada em polígrafo permanece como um processo de interpretação humana. É pertinente observar que há casos catalogados de suspeitos inocentados pela apresentação de provas, mesmo estes tendo sido considerados não sinceros pelo uso do polígrafo (Ekman, 1992).

Dadas as limitações existentes na aplicação do polígrafo como tecnologia para secundar a detecção de mentiras, outras estratégias passaram a ser experimentadas, como o Aprendizado de Máquina (*Machine Learning*). A adoção do Aprendizado de Máquina introduz uma clara diferença em relação ao polígrafo. Neste cenário, o que se procura é transferir todo o processo de detecção para a máquina.

Ao se aplicar o Aprendizado de Máquina à detecção de mentiras, modelos matemático-computacionais são construídos com base em diversas pistas de origem fisiológica e comportamental, de natureza verbal e não-verbal. Alguns desses modelos são construídos combinando pistas de diversas origens (voz, rosto e linguagem, por exemplo), consistindo no que se chama abordagem multimodal. Outros aprofundam-se em uma única categoria de sinais, constituindo a abordagem monomodal.

Abordagens multimodais tiram proveito do estado tecnológico que permite alto volume de processamento e consumo de memória. A combinação de várias modalidades está em alinhamento com relatos, em entrevistas, de indivíduos com alto grau de precisão para detectar mentiras (O’Sullivan; Ekman, 2004; Vrij, 2008).

Os modelos construídos (mono ou multimodais) operam como classificadores, recebem como entrada as pistas coletadas durante a emissão de uma mensagem, as processam e medem o grau de similaridade com os padrões de sinceridade (verdade) e não sinceridade (mentira) previamente identificados, para fornecer a probabilidade de presença de mentira. Estes modelos são grandemente dependentes de conjuntos de dados com narrativas já classificadas como sinceras e não sinceras (dados que

foram coletados e manualmente classificados por meio de esforço humano). Tais conjuntos de dados são frequentemente chamados de conjuntos de dados rotulados.

Embora os esforços estejam em crescimento, os resultados ainda podem ser considerados preliminares. Muitos experimentos estão baseados em dados coletados em laboratório, pois há escassez de conjuntos de dados rotulados.

Diversos estudos fazem uso de uma modalidade de operação de Aprendizado de Máquina chamada de “aprendizado supervisionado”, necessariamente dependente de dados rotulados que, como já colocado, são escassos para o problema em questão (Constancio et al., 2023).

A superação desta situação pode passar por abordagens outras de Aprendizado de Máquina, que sejam capazes de prescindir de dados rotulados. Dessa forma, o panorama de aplicação de Aprendizado de Máquina ao problema de detecção de mentiras mostra-se amplo e diverso, com inúmeras opções técnicas, assim como com diversas incertezas e dificuldades.

Dado o arsenal de ferramental oferecido pelo campo de Aprendizado de Máquina e as inúmeras decisões decorrentes para se iniciar uma pesquisa a respeito, uma revisão sistemática de literatura foi realizada no ano de 2022 para compreender o estado da ciência a respeito da detecção de mentiras suportada por Aprendizado de Máquina. A revisão forneceu estatísticas a respeito das técnicas adotadas ao longo de uma década (período de 2011 a 2021), assim como o grau de sucesso de diversas abordagens para a solução do problema (Constancio et al., 2023).

O estudo foi conduzido sobre 81 artigos recuperados dos portais de periódicos ACM Digital Library, IEEE *Xplore*, Scopus e Web of Science e mostrou que diversos países promovem pesquisas a respeito da aplicação de Aprendizado de Máquina ao problema de detecção de mentiras. Além do inglês, outras oito diferentes línguas foram exploradas em tais estudos, nenhum deles por pesquisadores brasileiros ou para a língua portuguesa.

Também foi realizada uma análise de desempenho por modalidades individuais e combinadas. Ficou evidenciado o aumento na precisão de detecção pela utilização de mais de um modal de pistas, o que acompanha pressupostos teóricos (O’Sullivan; Ekman, 2004; Vrij, 2008).

Outra conclusão evidente é a dependência dos experimentos em consumir e processar conjunto de dados rotulados. De forma geral, os autores relatam a

dificuldade de encontrar dados rotulados, precisando eles mesmos coletá-los, em situações várias que são elaboradas especificamente para seus estudos.

A quase totalidade destes estudos fez uso de aprendizado supervisionado. Apenas dois dos artigos recuperados buscaram evitar o problema de escassez de dados rotulados por meio de uma abordagem baseada em aprendizado não supervisionado, que independe de dados rotulados. Nenhum dos estudos fez uso do que se chama “aprendizado autossupervisionado”, outra possibilidade do Aprendizado de Máquina, que faz uso das técnicas de aprendizado supervisionado sem a necessidade de rotulação dos dados (Goodfellow; Yoshua; Courville, 2016), restando esta modalidade de aprendizado como terreno virgem para pesquisa.

Fatores situacionais e idiossincráticos podem afetar o comportamento de uma pessoa e, portanto, o que a mesma deixa transparecer ao enganar, o que torna essas pistas ainda mais específicas, com detecção mais desafiadora (Vrij, 2008). Não levar tais fatores em consideração pode reduzir a precisão da detecção.

Assim, a exploração dos aspectos idiossincráticos e situacionais pode promover maior precisão aos preditores para detectar mentiras, representando uma oportunidade de pesquisa. Tais fatores pessoais e situacionais servem como uma explicação razoável para a variedade de resultados relatados na revisão de literatura (alguns excelentes e outros apenas regulares). São fatores de confusão que operam como forte estímulo para mais pesquisas e esforços na produção de conjuntos de dados rotulados a partir de dados reais em circunstâncias mais diversificadas, para que mais pistas possam ser identificadas e relacionadas a ocasiões e situações particulares.

1.1 Problema, hipótese e questão de pesquisa

Dado o contexto apresentado, considerando especificamente o campo de Aprendizado de Máquina, é possível compreender que um problema hoje enfrentado por interessados em detectar a mentira é: **distinguir a sinceridade da não sinceridade em uma narrativa de um sujeito durante uma comunicação**.

Diante das lacunas de pesquisa reveladas pela revisão sistemática de literatura conduzida, este estudo foi efetuado com vistas à validação da seguinte hipótese de pesquisa: **o Aprendizado de Máquina autossupervisionado é capaz de viabilizar um modelo apto a distinguir uma narrativa sincera de uma não sincera proferida por um indivíduo específico.**

O problema e a hipótese de pesquisa rementem, então, à seguinte questão de pesquisa: **qual modelo de Aprendizado de Máquina autossupervisionado é capaz de utilizar pistas multimodais de um indivíduo específico para distinguir uma narrativa sincera de uma não sincera expressa em língua portuguesa?**

Em resumo, o que esta pesquisa propõe é a elaboração de um modelo de Aprendizado de Máquina alimentado com dados multimodais, para utilizá-los na distinção entre sinceridade e não sinceridade, em narrativas em vídeos expressas em língua portuguesa do Brasil.

A abordagem multimodal foi escolhida com a pretensão de explorar um maior número de pistas de não sinceridade, tentativamente procurando elevar a precisão do modelo proposto. A escolha por aprendizado autossupervisionado se deu em resposta à raridade de conjuntos de dados rotulados voltados à detecção de mentiras. A ideia de construir modelos individuais, dedicados a sujeitos específicos veio em resposta ao efeito situacional e idiossincrático que pode influenciar a interpretação das pistas durante uma comunicação não sincera. A exploração da língua portuguesa do Brasil em particular se deveu à noção de que língua e cultura podem apresentar padrões de sinceridade e não sinceridade próprios.

O problema, a hipótese e a pergunta de pesquisa delinearam toda a consecução da pesquisa, descrita nas seções e nos capítulos a seguir.

1.2 Objetivos

Visando contribuir para a solução do problema de pesquisa, validar a hipótese de pesquisa e finalmente responder à questão de pesquisa, o seguinte objetivo geral foi estabelecido: **elaborar modelos de Aprendizado de Máquina autossupervisionados para a detecção individual e multimodal de mentiras para a língua portuguesa**.

Como meio para o atingimento do objetivo geral foram definidos os seguintes objetivos específicos:

A. coletar narrativas e organizar um conjunto de dados rotulado voltado à detecção de mentiras para a língua portuguesa do Brasil;
B. elaborar, avaliar e identificar os modelos autossupervisionados de melhor desempenho para cada modalidade;
C. elaborar um modelo de fusão das modalidades individuais em um modelo integrado multimodal.

Os objetivos específicos representaram marcos dentro da pesquisa e contribuíram para o atingimento do objetivo geral em um processo progressivo.

1.3 Justificativas

Esta pesquisa apresenta justificativas para a ciência, para o Ministério Público e órgãos de segurança, para a sociedade, para o Programa de Pós-graduação em Gestão da Informação e para o autor.

1.3.1 Para a Ciência

A revisão sistemática de literatura que foi realizada com o propósito de identificar o estado da ciência no período de 2011 a 2021 permitiu compreender e correlacionar abordagens e resultados até então atingidos dentro do tema.

Dentre inúmeras descobertas, a revisão transpareceu a maciça preferência pela modelagem do problema de detecção de mentiras em uma forma que requer dados rotulados (na terminologia do Aprendizado de Máquina, tarefa de classificação). Dos 81 estudos, apenas dois exploraram o tema com o aprendizado não supervisionado.

Esta pesquisa, ao propor a abordagem pela detecção de anomalias, explorou o problema por outro ângulo, até então não experimentado, fornecendo uma nova oportunidade de entendimento e contribuição.

1.3.2 Para o Ministério Público e órgãos de segurança

A mentira é um fenômeno pervasivo, havendo estudos que sugerem que cada pessoa mente em média duas vezes ao dia. Embora diversas destas mentiras não representem risco social, existem ocasiões em que a mensagem não sincera pode ter implicações críticas para o interesse coletivo ou mesmo de indivíduos sob algum tipo de vulnerabilidade.

Um exemplo de aplicação é durante interrogatórios policiais. Embora os resultados de uma ferramenta de Detecção de Mentiras não possam servir como evidência jurídica, poderão servir como critério para a determinação de linhas investigativas para a coleta de evidências, elevando a eficiência policial.

Entrevistas para fornecimento de vistos de permanência e imigração também podem se valer deste tipo de tecnologia, o que pode ter impacto na manutenção da segurança nacional.

Na segurança aeroportuária, em pontos de controle de segurança para identificar possíveis passageiros com más intenções ou informações falsas, além de procedimentos de imigração, para verificar a autenticidade das informações fornecidas por viajantes sobre seu propósito de visita e histórico criminal.

1.3.3 Para a sociedade

A mentira não é objeto de atenção unicamente no contexto policial ou investigativo. Circunstâncias outras podem demandar a detecção de não sinceridade, tais como:

a) **terapias de saúde**: identificar a falta de sinceridade em entrevistas, que podem servir para estabelecer desde linhas de atuação até a identificação de riscos de morte, como tentativas de suicídio, depressão (Ekman, 1992) e sujeição à violência;
b) **entrevistas de emprego**: identificar a não sinceridade em concursos ou processos seletivos de natureza profissional, de interesse privado ou público (Ding et al., 2019);
c) **debates e comunicações políticas**: identificar a falta de sinceridade por parte de candidatos a cargos eletivos ou agentes públicos (Kopev et al., 2019), proporcionando assim a elevação dos valores democráticos e protegendo os interesses coletivos;
d) **empresas de seguros**: verificar a veracidade das alegações feitas por segurados em casos de sinistros;
e) **condução de pesquisas e estudos**: validar as respostas dos participantes em pesquisas, garantindo a qualidade dos dados coletados;
f) **negociações comerciais**: garantir acordos justos e honestos;
g) **casos de fraude**: identificar possíveis mentiras em solicitações de empréstimos, transações suspeitas etc;
h) **serviços de atendimento ao cliente**: avaliar a autenticidade das reclamações dos clientes e resolver disputas de forma justa;
i) **monitoramento de mídias sociais**: identificar informações falsas ou contas fraudulentas, além de tentativas de aliciamento e abuso de crianças e adolescentes.

As justificativas apresentadas representam alguns cenários de uso em que o objeto de estudo, ou suas derivações, poderão ser empregadas no intuito de proteger a sociedade em geral ou indivíduos em fragilidade dos efeitos nocivos da mentira.

1.3.4 Para o Programa de Pós-graduação em Gestão da Informação

As pistas de expressão que foram utilizadas como entrada para o treinamento do algoritmo de Aprendizado de Máquina são os dados que demonstram processos emocionais e psíquicos de um sujeito sob as condições em estudo.

Durante uma sessão de entrevista, as pistas oferecidas pelo sujeito podem ser interpretadas como informações reconhecidas e processadas por um Detector de Mentiras humano, que forma o seu julgamento a respeito da sinceridade das mensagens analisadas. Esta mesma interpretação é válida para o caso de um Detector de Mentiras baseado em Aprendizado de Máquina.

Trata-se, portanto, de um processo de aquisição de informação psíquico-emocional, seu processamento e a produção de uma nova informação, que é a probabilidade da presença de não sinceridade em mensagens.

Nesse sentido, a pesquisa representa uma cooperação entre tecnologia da informação, ciência da computação e psicologia, visto que opera interdisciplinarmente nestes domínios com vistas a construir um conjunto de dados e um método de processamento e operação que contribuísse para o campo da Detecção de Mentiras por Aprendizado de Máquina.

A revisão de literatura conduzida para a construção dos fundamentos teóricos deste estudo evidenciou que a absoluta maioria dos estudos vêm se dedicando à língua inglesa. Nesse sentido, até onde se pôde levantar, foi este o primeiro estudo mundial a explorar o tema de Detecção Automatizada de Mentiras para a língua portuguesa.

1.3.5 Para o autor

A tecnologia de Aprendizado de Máquina é de interesse pessoal do pesquisador, que observa sua aplicação cada vez mais frequente em diversos campos do conhecimento.

A aplicação específica sobre o tema de Detecção de Mentiras exerceu atração pelos seus valores potenciais para benefício coletivo, o que é um outro fator de atração do tema pelo pesquisador.

Finalmente, por tratar-se de pesquisa ainda não aplicada à língua portuguesa, o estudo foi entendido pelo pesquisador como oportunidade valiosa para contribuir simultaneamente com a ciência e com a sociedade.

1.4 Originalidade e não trivialidade da pesquisa

A revisão sistemática de literatura que foi realizada em 2022, no período compreendido entre 2011 e 2021, revelou que foram publicados diversos artigos que relataram avanços e contribuições ao tema. Os critérios de inclusão amplos de tal revisão permitiram identificar contribuições, estratégias e lacunas. As contribuições trazidas por esta pesquisa compreendem:

a) a elaboração e disponibilização do primeiro conjunto de dados rotulados do mundo, específico para detecção de mentiras para a língua portuguesa, que poderá ser utilizado em pesquisas posteriores;
b) a construção de um modelo de Aprendizado de Máquina que aprende os caracteres idiossincráticos e contextuais de uma narrativa para prevenir o efeito de fatores de confusão individuais, abordagem ainda não relatada na literatura;
c) a inclusão (como pista para detecção de narrativa não sincera) das oscilações linguístico-emocionais (pistas verbais) específicas para a língua portuguesa, ainda não relatada na literatura;
d) a modelagem do problema de detecção de mentiras como um problema de detecção de anomalias, estratégia ainda não experimentada na literatura científica.

No que tange à última contribuição, foi incluída no bojo desta pesquisa a proposta do que se chamou de **Modelo de Sinceridade** e de uma métrica específica para sua avaliação, o **Escore de Sinceridade**, ambos figurando como inovação trazida pelos estudos aqui encaminhados.

A detecção de anomalias procura desenvolver métodos para identificar eventos que, de alguma forma, discrepam daquilo que é considerado como o comportamento normal (Goodfellow; Yoshua; Courville, 2016). Por exemplo, a detecção de tentativas de invasão em redes e sistemas computacionais online (ataques cibernéticos) é um exemplo de detecção de anomalias (Chen et al., 2018). Outro exemplo é a detecção de operações fraudulentas em transações por cartão de

crédito (Wang; Bah; Hammad, 2019). Ainda outro é a detecção de patologias ou doenças raras (Lu; Xu, 2018; Pratella et al., 2021).

Em todos os casos, algum processo é utilizado para identificar o comportamento normal (e esperado) e métricas são empregadas para perceber quando eventos parecem divergir significativamente daquele comportamento.

Neste estudo, a abordagem de deteção foi considerar as expressões acústicas, verbais e visuais sinceras como sendo a “expressão normal” do indivíduo e a mentira como uma anomalia. As variações percebidas constituem as pistas para detecção e, quando divergiram suficientemente da expressão normal, foram caracterizadas como mentira.

Em matéria de Aprendizado de Máquina, a detecção de anomalias frequentemente faz uso de modelos de Aprendizado autossupervisionados (Goodfellow; Yoshua; Courville, 2016). Assim, o “Modelo de Sinceridade” é um modelo de Aprendizado de Máquina autossupervisionado especializado na detecção de anomalias, que objetivou capturar as expressões multimodais próprias da sinceridade (expressões de normalidade). Quando uma narrativa não sincera foi submetida ao modelo, a expectativa era de que esta não fosse compatível com a sinceridade capturada anteriormente e pudesse ser caracterizada como uma anomalia, portanto, como mentira.

Entendeu-se que as variações de expressão apresentadas por um indivíduo não necessariamente tinham sua gênese na mentira. Assim, uma anomalia de expressão (por exemplo, a presença de uma expressão facial incomum) poderia existir ainda que a mensagem fosse verdadeira, pois outros processos (internos ou externos) poderiam estar operando. No entanto, a proposta do Modelo de Sinceridade figura como uma abordagem tentativa para explorar aspectos peculiares do indivíduo e de sua situação, diferindo do grande volume de outras abordagens que procura padrões coletivos para a expressão não sincera. A expectativa foi de que a abordagem pelo viés individual pudesse revelar novos padrões e, assim, contribuir para o conhecimento da área.

A sinceridade foi escolhida como caso normal porque entendeu-se que eventos verdadeiros são mais facilmente confirmáveis por aquele que deseja coletá-los, tornando mais fácil de construir o modelo de referência (sinceridade como referência). Por exemplo, um interrogador pode solicitar ao interrogado para descrever uma situação que já saiba de antemão que é verdadeira e assim coletar expressões

de sinceridade (como ocorre com o polígrafo). Adicionalmente, não é esperado que o interrogado sofra variações de expressão para relatar verdades confortáveis, outro aspecto de normalidade.

Finalmente, embora haja pesquisas relatando que uma pessoa em média mente duas vezes ao dia, entendeu-se que na maior parte do tempo o indivíduo atua sinceramente, o que reforçou a noção de sinceridade como atitude normal e mentira como anomalia.

1.5 Delimitação da pesquisa

A revisão de literatura realizada permitiu perceber que um variado conjunto de ferramentas tem sido utilizado para apoiar os pesquisadores em seus experimentos. Em particular, ficou evidente uma reincidência de três ferramentas em específico.

O OpenSMILE é a preferência quase absoluta quando se trata de extração de características acústicas, por este motivo foi adotado também nesta pesquisa. Outro fator de adoção é que o OpenSMILE é um produto de acesso gratuito para uso acadêmico.

O LIWC (*Linguistic Inquiri and Word Count*) foi frequentemente adotado em estudos de natureza verbal. Embora tenha uma versão para o português, não é um produto gratuito, exigindo a compra de uma licença de uso por tempo determinado (assinatura) e por tal, não foi adotado.

Finalmente, os estudos de natureza visual mais recentes tenderam a adotar maciçamente o OpenFace para extrair características visuais diversas a partir de imagens e vídeos digitais. Por suas capacidades e por seu acesso gratuito, foi adotado nesta pesquisa. No entanto, o OpenFace apresenta também limitações. As característica faciais extraídas limitam-se a 17 ações faciais das 46 existentes no FACS, o *Facial Action Coding System* (Ekman; Friesen; Hager, 2002).

O OpenFace tem o objetivo de oferecer a extração de características em tempo real. Para tanto, requer vídeos com uma qualidade mínima (para evitar a necessidade de pré-processamento) e eventualmente pode reduzir a precisão em favor do tempo de resposta.

Experimentos preliminares também mostraram que o OpenFace não está preparado para lidar com a oclusão, que é quando parte do rosto do sujeito está encoberto, por exemplo, por uma das mãos.

Como característica adicional, o OpenFace não oferece dados em nível de músculo (por exemplo, o músculo zigomático maior), mas da composição desses para a formação das ações faciais, o que significa dizer que a análise de músculos específicos não foi possível.

Os vídeos utilizados para a construção do conjunto de dados utilizado nos experimentos realizados apresentam pessoas mentindo e sendo sinceras em um programa de TV. Naquele contexto, os participantes foram orientados a falar a verdade ou mentira com o objetivo de convencer observadores a respeito de suas narrativas.

Por ser um programa de TV, as consequências da eventual descoberta da mentira não tinham caráter punitivo, o que significa que as emoções envolvidas são diferentes de, por exemplo, um interrogatório policial onde existe o risco de consequências legais. Entendeu-se que este fato influenciaria a presença e a intensidade das pistas de não sinceridade, mas a expectativa foi de que ainda assim as expressões de cada caso seriam suficientemente discrepantes para serem percebidas pelo Modelo de Sinceridade.

1.6 Estrutura deste documento

O capítulo 2 apresenta o referencial teórico que oferece os conceitos fundamentais para a compreensão do contexto, dos experimentos e das propostas desta pesquisa. Inclui-se aqui Aprendizado de Máquina, Aprendizado Profundo e Detecção de Mentiras.

O capítulo 3 apresenta os encaminhamentos metodológicos postos em atividade para a realização da pesquisa com vistas a confirmar a hipótese de pesquisa e atender aos objetivos, geral e específicos, da pesquisa.

O capítulo 4 apresenta resultados atingidos, oriundos dos encaminhamentos realizados.

O capítulo 5 apresenta a considerações finais relativas aos resultados atingidos, limitações e propostas para futuras pesquisas.

2 REFERENCIAL TEÓRICO

Esta pesquisa discorre a respeito de estratégias para a **Detecção Automática de Mentiras** em narrativas expressadas em língua portuguesa pela extração de pistas de várias modalidades (acústicas, verbais e visuais) e seu processamento por modelos de Aprendizado de Máquina.

O tema de Detecção de Mentiras suportada por Aprendizado de Máquina vem recebendo atenção de estudiosos desde o início dos anos 2000. As estratégias vêm sendo variadas, assim como os resultados atingidos. Neste estudo, procurou-se trazer uma nova abordagem ao panorama do tema. Trata-se da modelagem do problema de detecção de mentiras como um problema de detecção de anomalias, atacado especificamente por Aprendizado Profundo, na forma de um modelo de Aprendizado de Máquina específico conhecido como Autoencoder.

Para melhor compreensão do contexto da pesquisa e da abordagem experimentada, as seções seguintes apresentam alguns conceitos fundamentais. A última seção apresenta um quadro contendo um resumo para evidenciar quais conceitos teóricos fundamentaram o percurso metodológico trilhado.

2.1 A respeito da mentira

Sob o aspecto cognitivo-emocional, a mentira é um fenômeno complexo, que pode se manifestar em múltiplas formas, como falsificação de informações, evasão de assunto, ambiguidade narrativa, omissão de detalhes, utilização de exageros ou eufemismos, por exemplo (Walczyk et al., 2014).

As motivações para tais atitudes são muitas. Algumas mentiras são proferidas com o intuito de evitar situações desagradáveis, sem que impliquem em qualquer efeito especialmente prejudicial aos envolvidos (Vrij, 2008). Outras mentiras, no entanto, podem ter sua gênese em intenções com repercussões mais sérias para o emissor. Estas são chamadas de mentiras de alto risco (do inglês, *high-stakes deception*) ou mentiras sérias (do inglês, *serious lies*). As motivações para mentiras sérias estão sujeitas a classificações diversas, como por exemplo (Walczyk et al., 2014):

1. **instrumental**: a mentira é usada para obter vantagem, poder, status social, podendo propiciar o prazer de exercer controle sobre outros ou sobre uma situação;

2. **evitar punição**: a mentira é usada para evitar uma consequência tida como negativa pelo emissor, buscada para atingir uma sensação de alívio ou a redução da ansiedade;
3. **autopreservação**: a mentira opera como um protetor psicológico ao evitar confrontos e embaraço, em resposta a inseguranças, medo e como meio de reduzir a ansiedade;
4. **autopromoção**: a mentira é utilizada como ferramenta para a projeção de uma imagem falsa do emissor, em resposta ao desejo por respeito, por orgulho ou como meio de reduzir a insegurança;
5. **proteção de terceiros**: a mentira altruística visa proteger outros de eventos indesejados, originada por sentimentos de compaixão e empatia;
6. **direito adquirido**: a mentira é utilizada para esconder uma verdade tida como injustamente proibitiva pelos receptores, originada em um senso de ressentimento ou indignação;
7. **danos a terceiros**: a mentira é empregada como meio de provocar sofrimento a terceiros, em resposta ao ódio e ao desejo de vingança;
8. **relacionamento interpessoal**: a mentira é empregada como ferramenta para manter, melhorar, reduzir ou dissolver uma interação ou para controlar o grau de intimidade, originada no ódio ou insegurança.

Sob um aspecto humano-social, a mentira é um fenômeno altamente recorrente e pervasivo, demonstrado por uma pesquisa (Turner; Edgley; Olmstead, 1975) que analisou 870 declarações em registros escritos de conversação de 130 voluntários. A conclusão foi que os participantes mentiram ao menos duas vezes ao dia. De um quarto a um terço das conversações incluíam algum tipo de mentira.

Apesar da convivência frequente com a não sinceridade, uma meta-análise (Aamondt; Custer, 2006) conduzida sobre 206 estudos (artigos científicos, teses e dissertações do período de 1970 a 2004) de 108 fontes relevantes de publicação (segundos os autores), compreendendo um total de 16.537 sujeitos concluiu que uma pessoa sem treinamento tem a probabilidade de 54% de identificar uma mensagem não sincera. Os autores relatam um intervalo de confiança de 95% em seu estudo.

Ao proferir uma mentira séria, o emissor tende a experimentar variações fisiológicas (aumento da pressão arterial e ritmo respiratório, ressecamento da garganta, aumento da sudorese, variação na atividade eletrodérmica e até mesmo dilatação da pupila) (Vrij, 2008) e de expressão (alteração na frequência e intensidade

de movimentos, pausas longas ao responder, momentos de hesitação, tom da voz, uso de gestos, expressões faciais involuntárias, supressão de alguns tipos de palavras em favor de outras) (Ekman, 1992; Vrij, 2008). Tais variações podem ser percebidas por um observador preparado, que então as utiliza como pistas de não sinceridade (Ekman, 1992; Zuckerman; DePaulo; Rosenthal, 1981). O grau de seriedade da mentira modula a intensidade das pistas produzidas (Walczyk et al., 2014).

Durante a emissão de mentiras sérias, os diversos sinais involuntariamente exibidos podem ser identificados por pessoas com capacitação específica (detectores de mentira humanos, também chamados de *lie catchers*). Assim, acredita-se que durante uma mentira de alto risco (mentira séria) as mudanças fisiológicas e de expressão experimentadas pelo emissor tendem a produzir indicadores mais evidentes de não sinceridade (DePaulo et al., 2003; Porter; Brinke, 2010a; Zuckerman; DePaulo; Rosenthal, 1981).

Há uma variedade de pistas já, de alguma forma, exploradas na detecção de mentiras. Uma meta-análise conduzida em um corpus de 116 artigos resultou em 158 pistas (DePaulo et al., 2003). Nenhuma delas constitui o que pode ser chamado de “dispositivo de Pinóquio”, ou seja, um sinal indiscutível de mentira (Vrij, 2008).

Os sinais de mentira podem ser verbais (efeitos nas construções linguísticas de no vocabulário utilizados pelo emissor) e não-verbais (efeitos na maneira como o emissor sonoriza sua fala e/ou se expressa corporal e facialmente). As diferentes fontes de sinais são normalmente chamadas de modalidades ou canais. Por exemplo, os sinais identificados na voz de um emissor são oriundos do canal vocal ou pertencentes à modalidade acústica.

Exemplos de sinais não-verbais incluem gestos como autoadaptadores (tocar o próprio corpo, rosto ou cabelo) (Vrij, 2008), manipuladores (beliscar, escolher, arranhar) (Ekman, 1992), emblemas (gestos que substituem palavras) (Ekman, 1992), ou ilustradores (gestos que acompanham a fala) (Ekman, 1992; Vrij, 2008). Existem também as expressões e microexpressões faciais (Ekman, 1992), o aumento no tom da voz e a dilatação da pupila (DePaulo et al., 2003). A alteração na frequência de piscadas ou a presença de pausas durante a fala são outros exemplos, assim como a intensificação de pausas ou a presença de hesitações (Vrij, 2008).

Pistas verbais incluem a frequência de uso de pronomes pessoais (menos autorreferências para distanciamento do incidente), o volume de palavras negativas (refletindo o estado emocional) e a presença de palavras que expressam incerteza

(como “talvez” ou “quem sabe”) (Ten Brinke; Porter, 2012; Vrij, 2008). Tais alterações tendem a ocorrer sem a percepção do interrogado, tornando-as menos passíveis de monitoração e manipulação.

As variações na expressão verbal são postuladas pela hipótese proposta pelo psicólogo alemão Udo Undeutsch da Universidade de Colônia, que iniciou seus estudos em 1957 com a publicação de um conjunto de critérios para avaliação de credibilidade de testemunhos conhecido como Análise da Realidade de Declarações (do inglês, *Statement Reality Analysis*), com publicações adicionais em 1967, 1982, 1983 e 1984 (Steller, 1989). A Hipótese de Undeutsch, como passou a ser conhecida, estabelece que as memórias de eventos experienciados (a verdade) propiciam descrições mais ricas em detalhes e com estrutura linguística mais coerente, quando comparadas com uma fabricação (a mentira) (Amado; Arce; Fariña, 2015).

Um estudo demonstrou que existem diferenças na forma de expressar verbalmente um discurso real em comparação com um fictício (Taylor et al., 2015). Neste estudo, um total de 60 indivíduos (N=60) oriundos de quatro grupos culturais diferentes (falantes de quatro línguas diferentes) foram solicitados a escrever frases sinceras e não sinceras. O objetivo do estudo foi confirmar a hipótese de que existe mudança perceptível na expressão verbal de cada circunstância (sinceridade e não sinceridade). Adicionalmente, os autores também procuraram avaliar se as mudanças são as mesmas ou variam de acordo com o grupo cultural/língua.

Os resultados do estudo apontam para a confirmação da hipótese de variação no estilo de expressão verbal nos casos de sinceridade e não sinceridade. Também sugerem que estas variações não são as mesmas em cada grupo cultural/língua, evidenciando que as pistas válidas para uma língua não necessariamente serão aplicáveis para outra.

Outro estudo (Papantoniou et al., 2021) procurou explorar os efeitos das diferentes pistas verbais em diferentes línguas e culturas (Inglês nos Estados Unidos, Inglês na Índia, Bélgica, Rússia, México e Romênia) para identificar semelhanças e diferenças no potencial preditivo de diferentes grupos de características verbais. A conclusão foi de que algumas características com evidente potencial preditivo em uma língua/cultura não apresentam o mesmo grau de importância em outra.

Os autores apontam que seus resultados são uma outra forma de demonstrar as mesmas conclusões colocadas em estudo anterior (Taylor et al., 2015).

Adicionalmente, os autores relataram o interesse em incluir no seu estudo outras línguas, como português, alemão, italiano e árabe.

2.2 Detecção de mentiras

Técnicas para detectar mentiras vêm sendo desenvolvidas há séculos. Até mesmo Charles Darwin tem escritos datados de 1.872 a respeito de pistas faciais (Denault et al., 2022). Uma revisão bibliométrica publicada em 2022, baseada em achados na *Web of Science*[1] apresenta um crescimento constante de publicações neste assunto. O estudo mostrou que houve a publicação de 3.245 artigos na década de 2010-2019, contra apenas três na década de 1900-1909 (Denault et al., 2022; Ten Brinke; Porter, 2012).

Alguns indivíduos mostram um nível de precisão destacado na detecção de mentiras, comparados ao cidadão médio. Uma meta-análise (Aamondt; Custer, 2006) concluiu que o indivíduo comum tem a **probabilidade de 54%** de detectar que uma mensagem não é sincera e surpreende ao demonstrar que agentes policiais formam, junto com estudantes de psicologia, os grupos específicos de pior desempenho (Aamondt; Custer, 2006; Salles, 2020), visto que detectar mentiras em interrogatórios parece ser uma habilidade esperada na atividade policial. Os grupos com maiores níveis de precisão na detecção de mentiras são professores, agentes sociais, criminosos e agentes do serviço secreto.

Aspectos demográficos como sexo e idade parecem não interferir na capacidade de detectar mentiras (Aamondt; Custer, 2006; O'Sullivan; Ekman, 2004). Tampouco o nível cultural ou educacional.

Os indivíduos com a capacidade distinta de identificar a mentira baseiam seu julgamento na observação de pistas manifestadas involuntariamente pelo interlocutor (O'Sullivan; Ekman, 2004). Ao utilizar um testemunho sincero como linha de base, os detectores de mentira humanos (*lie-catchers*) exploram a capacidade de perceber alterações de expressão que podem ser utilizadas como pistas para não sinceridade. O desenvolvimento da habilidade de observar pistas e detectar mentiras parece ser motivado por seus próprios interesses e por um esforço consciente de aprimoramento (O'Sullivan; Ekman, 2004).

O "Projeto *Wizard*" (O'Sullivan; Ekman, 2004) aplicou um processo de medição da capacidade de detectar mentiras em um variado conjunto de voluntários.

[1] https://www.webofknowledge.com

Aqueles que atingiram a marca de 90% de acerto (ser capaz de detectar 9 entre 10 casos de narrativas não sinceras) foram convidados para participar da continuidade da pesquisa, pois foram considerados indivíduos habilidosos na atividade, passando a receber o título de "*Wizard of deception detection*" ("Mago da detecção de mentiras", em tradução livre).

Alguns dos "Magos" identificados demonstraram alto grau de sucesso para detecção dentro de certos contextos, mas apresentaram perda de desempenho em outros. No entanto, alguns "Magos" foram submetidos a circunstâncias e temas diversificados e conseguiram manter seu nível de desempenho. Estes especialistas foram chamados de "*Ultimate Wizards*" ou "Magos Supremos", em uma tradução livre.

Uma das conclusões do "Projeto Wizard" foi que os "magos" relataram, em entrevista, avaliar pistas verbais e não-verbais combinadas. Aqueles que apresentaram desempenho especialmente alto demonstraram alta sensibilidade à nuances da expressão verbal. Os entrevistados também relataram perceber incongruências entre as duas formas de expressão e a utilização dessa percepção como uma pista de não sinceridade.

Vários dos magos revelaram que tiveram "experiências infantis incomuns", como ser filhos de alcoólatras, ser filhos de mães que trabalhavam fora ou não falar o inglês até a idade escolar. Por força das circunstâncias, os magos sofreram da necessidade de identificar flutuações emocionais em outrem por meio de pistas não-verbais.

Os autores relataram que perceberam nos magos um senso de entusiasmo e consciência a respeito do experimento, reputando-os como altamente motivados e interessados no tema. Em entrevistas, os magos relataram dedicação específica ao aumento do desempenho na detecção de mentiras e por longo tempo. Em consequência, outra conclusão do "Projeto *Wizard*" foi de que, assim como um atleta olímpico, um *lie-catcher* habilidoso, um "*Wizard*", pode apresentar um talento inato, que precisa ser desenvolvido por meio de experiência e pela própria vontade do praticante.

Os "Magos Supremos da Detecção de Mentiras" parecem adquirir suas capacidades motivados por um desejo pessoal de desempenhar melhor suas funções (O'Sullivan; Ekman, 2004). É semelhante a qualquer outra habilidade ou talento profissional, melhorado pelo esforço, dedicação, interesse pessoal, conhecimento técnico e treinamento. Estes indivíduos altamente qualificados na atividade de

detectar mentiras são o resultado de seus próprios esforços intensos e conscientes. Os registros de suas capacidades representam evidências materiais de que altas taxas de precisão na detecção podem ser atingidas.

A diversidade de circunstâncias que os detectores de mentiras enfrentam melhora e generaliza suas habilidades, o que mostra a importância de ter dados da vida real rotulados, coletados de várias fontes, incluindo crianças e pessoas sob tratamento médico e psicológico ou interrogatório policial, ou em julgamento.

Um elemento complicador no processo de detectar mentiras são os fatores individuais e contextuais. Mesmo em uma mentira séria, os padrões de manifestação de pistas podem variar e até mesmo se inverter de um indivíduo para outro (Ekman, 1992; Porter; Brinke, 2010a; Vrij, 2008).

Alguns sinais podem apresentar recorrência entre diferentes indivíduos. No entanto, assumir que uma pista sempre poderá ser interpretada como mentira constitui um erro de julgamento denominado de “Perigo de Brokaw” (Ekman, 1992).

Outra situação de agravamento para a detecção é considerar que certas emoções necessariamente estarão presentes em situação de não sinceridade. Estas mesmas emoções (raiva, medo e deleite) ocorrem (Ekman, 1992), combinadas ou não, com frequência acima do acidental, especialmente em um contexto de interrogatório. Ainda assim, a presença das emoções pode ter sua origem em outros processos psicológicos do interrogado e assumir que são preditores indiscutíveis de mentira constitui um erro de julgamento descrito como “Erro de Otelo” (Ekman, 1992).

2.3 Tecnologias para detectar mentiras

A existência de mentiras em situações de segurança, com consequências críticas, naturalmente estimulou estudiosos a desenvolver formas de detecção. Uma revisão bibliométrica publicada em 2022, baseada em achados na *Web of Science*[2] demonstra o crescimento constante de publicações de estudos científicos a respeito de detecção de mentiras. O volume de **3.245 artigos** foi registrado na década de 2010-2019, contra apenas **três** na década de 1900-1909 (Denault et al., 2022; Ten Brinke; Porter, 2012).

A baixa precisão na detecção de mentiras por um indivíduo mediano (cerca de 54%) e a reconhecida relevância em detectar mentiras em circunstâncias de alto risco estimularam a criação de tecnologias auxiliares. O exemplo mais famoso é o

[2] https://www.webofknowledge.com

polígrafo, introduzido no Departamento de Polícia de Berkeley por John Larson (Ball, [s.d.]), em 1921. Os modelos atuais do polígrafo podem monitorar várias respostas fisiológicas de um sujeito e requerem uma calibração preliminar para estabelecer uma linha de base para os sensores.

É essencial afirmar que o polígrafo não detecta mentiras, mas alterações fisiológicas relacionadas às emoções (Ekman, 1992). Uma narrativa é considerada verdade ou não a depender inteiramente da interpretação do operador a respeito das saídas do dispositivo. Há relatos de falsos positivos que foram inocentados de acusações depois que investigações posteriores provaram que o teste detectou erroneamente uma mentira (Ekman, 1992).

Assim, rigorosamente, não é a máquina que detecta a mentira (embora o polígrafo seja muitas vezes chamado de “detector de mentiras”). O dispositivo apenas fornece dados que são lidos pelo operador, que os interpreta para categorizar um testemunho como sincero ou não. O processo de detectar mentiras é, nesse contexto, ainda uma **inferência humana**.

A *American Polygraph Association*[3] alega que estudos realizados pela *National Academy of Sciences*[4] (NAS), nos Estados Unidos, conduziu uma meta-análise que evidenciou a acurácia do polígrafo em 0,89 (89% de acerto). No entanto, um olhar mais atento ao estudo que a NAS realizou mostra que a mesma registrou, não a acurácia, mas sim a área sob a curva ROC (do inglês, *Receiver Operating Characteristic*) em 0,89, métrica que captura apenas a relação entre os casos positivos e não pode ser interpretada como taxa de acerto. Em outras palavras, dizer que o desempenho do polígrafo atinge 0,89 na área da curva ROC não significa dizer que o polígrafo tem 89% de percentual de acerto. Um estudo conduzido para averiguar o grau de confiabilidade do polígrafo concluiu que sua acurácia é ainda desconhecida e que os estudos que atestam sua eficácia podem ter sua validade desafiada (Iacono; Ben-Shakhar, 2019).

Estratégias de detecção fisiológica como o polígrafo ou a análise de estresse vocal sofrem de grande imprecisão porque a condição emocional do indivíduo pode de fato estar alterada sem que essa alteração seja motivada por não sinceridade (Ekman, 1992; Porter; Brinke, 2010a), em uma manifestação do “Erro de Otelo”.

---

[3] https://www.polygraph.org
[4] http://www.nasonline.org

2.4 Aprendizado de máquina

A Inteligência Artificial (IA) é o ramo da computação que se dedica a estudar e desenvolver processos que habilitem a máquina a executar tarefas de forma a imitar o “comportamento inteligente” (Jakhar; Kaur, 2020). Tais “agentes inteligentes” são dispositivos capazes de perceber o ambiente e aumentar as suas chances de sucesso em alguma tarefa graças a maior capacidade de adaptação às circunstâncias (Bini, 2018).

O Aprendizado de Máquina é um ramo da Inteligência Artificial que procura dar aos computadores a capacidade de aprender com os dados e estar apto a atuar em uma situação sem ter sido especificamente programado para ela, definição atribuída a Arthur Samuel em 1959 (Bell, 2015). O objetivo é dar às máquinas a capacidade de tomar melhores decisões ao extrair padrões intrínsecos a conjuntos de dados (Jakhar; Kaur, 2020).

O Aprendizado de Máquina oferece métodos que habilitam a solução de problemas cuja abordagem tradicional (por meio de programação específica) é muito difícil de ser realizada (Goodfellow; Yoshua; Courville, 2016).

Um exemplo simples é o reconhecimento de caracteres. Esta atividade trivial para o ser humano se mostra muitíssimo desafiadora para ser realizada por um algoritmo tradicional. Existem diferentes fontes e tamanhos de letras, a presença de ruído (como manchas), diferentes ângulos de escrita, o próprio tamanho variável das letras em uma mesma palavra. Todos estes fatores são tranquilamente processáveis pela mente humana, mas quando um programador tenta escrever um algoritmo que lide com todas estas situações concomitantemente, percebe o quão complexo é o desafio.

A abordagem por Aprendizado de Máquina se baseia na apresentação de diversas amostras de caracteres (dados de entrada) a um algoritmo que irá identificar marcadores e padrões diferenciais de cada letra de forma não estrita (padrões generalizáveis), ou seja, sendo capaz de lidar com algum grau de variância e ainda realizar a correta correspondência entre uma imagem de entrada e uma letra.

Os diversos métodos existentes constroem uma representação compacta de padrões de interesse existente nos dados, chamado de modelo.

Em termos genéricos, o processo de Aprendizado de Máquina consiste em:

1. **adquirir conjuntos de dados**, pois são a matéria-prima dos modelos;

2. **pré-processar os dados** para retirar ruídos eventuais ou alterar o formato ao necessário para o modelo;
3. **selecionar o modelo** apropriado ao problema, pois os modelos não são universalmente aplicáveis a todas as situações;
4. **treinar o modelo**, que é o processo de submeter os dados para que o modelo possa identificar e armazenar os padrões de interesse existentes;
5. **validar o modelo**, que consiste em avaliar se seus diversos parâmetros de treinamento (hiperparâmetros) estão ajustados para os dados e o problema em vista;
6. **testar o modelo**, que consiste em submeter o modelo a dados ainda não utilizados para avaliar sua capacidade de generalização;
7. **aplicar o modelo**, que é uso do modelo na situação real prevista.

Quando um modelo de Aprendizado de Máquina passa a memorizar os padrões ao invés de aprendê-los, é dito que está sofrendo de *overfitting* (Goodfellow; Yoshua; Courville, 2016). Nesta situação, o modelo tem um alto desempenho quando submetido aos dados de treinamento, mas demonstra desempenho significativamente inferior quando alimentado com dados ainda não conhecidos. Técnicas foram desenvolvidas para detectar e superar a incidência de *overfitting*, sendo recursos frequentemente utilizados pelos praticantes e estudiosos da área.

O Aprendizado de Máquina tem sido aplicado com sucesso em grande número de campos e problemas, tais como classificação de documentos, visão computacional, processamento de linguagem natural, previsão de estrutura de proteínas, detecção de fraudes, diagnóstico médico, sistemas de recomendação, entre outros (Mohri; Rostamizadeh; Talwalkar, 2012).

2.4.1 Exemplo de aplicação de Aprendizado de Máquina

O Aprendizado de Máquina oferece um vasto conjunto de técnicas, criando várias oportunidades para abordar os problemas. A matéria-prima do Aprendizado de Máquina são os dados, que podem ser imaginados como uma planilha onde cada linha corresponde a uma observação de interesse e cada coluna a um atributo descritivo desta observação. Em termos mais técnicos, cada linha corresponde ao registro de um fenômeno e cada coluna a uma das variáveis que o influenciam.

Tecnicamente, a planilha de dados é frequentemente chamada de **conjunto de dados**, ou ***dataset***. As linhas são chamadas de **exemplos** (*examples*) ou

indivíduos (*individuals*) e as colunas de **características** (*features*). Um indivíduo é, portanto, um conjunto de características (Goodfellow; Yoshua; Courville, 2016). No Quadro 1 apresenta-se um exemplo de *dataset*.

QUADRO 1 - EXEMPLO DE CONJUNTO DE DADOS NÃO-ROTULADOS

Estação do ano	Fase da lua	Direção do vento	Velocidade do vento	Chuva na véspera	Período do dia
Verão	Minguante	N-NE	40,50	Sim	Manhã
Verão	Nova	N-S	0,90	Sim	Manhã
Outono	Cheia	-	0,00	Não	Tarde
Inverno	Nova	L-NO	17,36	Não	Manhã
Primavera	Crescente	S-SO	31,00	Sim	Manhã

FONTE: O AUTOR (2023)

Neste cenário, o *dataset* descreve um conjunto de fatores que influenciam a qualidade de uma experiência de surfe em uma dada praia. Cada exemplo (linha) diz respeito a uma configuração de características (colunas) observada em um certo dia. Os dados podem ser categóricos (todas as colunas, exceto “Velocidade do vento”) ou contínuos (coluna “Velocidade do vento”).

O *dataset* poderia ser longo o suficiente para descrever todos os dias de vários anos. As características listadas foram as consideradas relevantes para descrever o indivíduo de forma útil, em resposta ao problema que se deseja resolver. Para outros problemas poderia haver outras características.

Imaginando um sistema de recomendação de surfe, o Aprendizado de Máquina seria alimentado com o *dataset* do Quadro 1 para aprender uma representação processável por máquina. Esta representação é chamada de **modelo**.

Para o cenário em foco, seria necessária a rotulação dos exemplos, relacionando dados e conceito de experiência de surfe, como mostrado no Quadro 2.

QUADRO 2 - EXEMPLO DE CONJUNTO DE DADOS ROTULADOS

Estação do ano	Fase da lua	Direção do vento	Velocidade do vento	Chuva na véspera	Período do dia	Nível de surfe
Verão	Minguante	N-NE	40,50	Sim	Manhã	Bom
Verão	Nova	N-S	0,90	Sim	Manhã	Moderado
Outono	Cheia	-	0,00	Não	Tarde	Péssimo
Inverno	Nova	L-NO	17,36	Não	Manhã	Moderado
Primavera	Crescente	S-SO	31,00	Sim	Manhã	Excelente

FONTE: O AUTOR (2023)

A coluna mais à direita do Quadro 2 ("Nível de surfe”) foi adicionada para registrar rótulos que classificam cada uma das configurações de características dentro de um nível de satisfação de surfe. A rotulação foi feita por um ou mais agentes humanos, sob critérios aplicáveis ao problema em questão (recomendação de surfe).

O algoritmo de Aprendizado de Máquina, ao receber o conjunto de dados rotulados, é capaz de gerar um modelo que o descreva dentro de algum nível de acurácia. Quando as características de um novo dia são fornecidas a este modelo, o mesmo produz um dos rótulos que conheceu anteriormente, realizando uma **predição** da experiência de surfe para aquele dia. Por isso, o modelo gerado para este cenário é frequentemente chamado de **modelo preditivo**.

É importante compreender que para esta mesma tarefa existem diferentes algoritmos que produzem diferentes modelos, sendo alguns deles mais precisos que os outros. A indicação de uso de um modelo é determinada pelo tipo e volume dos dados envolvidos, além de características intrínsecas do problema.

Retornando à definição de Aprendizado de Máquina, se percebe que o sistema de recomendação de surfe foi concebido sem que um programador tenha que escrever as regras para categorizar um conjunto de fatores, como seria em uma abordagem tradicional de programação. Ao invés disso, o sistema recebeu dados, aprendeu os padrões existentes neles e relacionou estes padrões a rótulos. A recomendação se faz pela similaridade de um novo dia (um novo indivíduo) com os padrões aprendidos.

2.4.2 Aprendizado supervisionado

O processo de aprendizado descrito no caso do sistema de recomendação de surfe é conhecido como **aprendizado supervisionado**. Esta modalidade de aprendizado requer que os dados brutos das observações sejam enriquecidos com rótulos (Mahesh, 2018), como foi feito no exemplo do sistema de recomendação de surfe. É como uma criança, que ao realizar inferências procura confirmar suas conclusões com um professor (ou supervisor).

Os rótulos não fazem parte das observações, mas de uma pré-classificação que visa guiar o processo de aprendizado. Por este motivo, a atividade solucionada por um modelo preditivo é clamada de **classificação**. Se os rótulos não forem valores discretos, mas sim contínuos (por exemplo, a probabilidade de uma boa experiência de surfe), então a atividade recebe o nome de **regressão**.

Alguns algoritmos famosos para atividades de classificação/regressão incluem (Mahesh, 2018):

a) **Árvore de Decisão** (*Decision Tree*), que produz como modelo final uma árvore de decisão interpretável ao ser-humano;

b) **Naïve-Bayes**, que entrega uma probabilidade (calculada a partir do Teorema de Bayes) de que um exemplo pertença a uma dada classe;
c) ***Support Vector Machines*** (SVM), que calcula e utiliza um hiperplano ótimo para dividir o espaço em dois e então classificar os indivíduos;
d) **Rede Neural Artificial** (RNA), modelo inspirado na estrutura de neurônios biológicos, organizado em camadas, que é capaz de encontrar uma função não-linear para mapear um conjunto de valores de entrada em outro conjunto de saída, não necessariamente de mesma cardinalidade.

A lista acima não é exaustiva, existindo diversas alternativas de modelos para classificação/regressão.

2.4.3 Aprendizado não supervisionado

Outra modalidade é o **aprendizado não supervisionado**, que não requer dados rotulados. Neste caso, o algoritmo é imediatamente alimentado pelo *dataset* mostrado no Quadro 1, que trabalha para organizar os indivíduos em grupos (*clusters*) segundo algum critério de similaridade (ou distância). Por este motivo, esta atividade é frequentemente chamada de **agrupamento** (*clustering*).

Este tipo de aprendizado é importante porque não exige o esforço humano para a rotulação dos conjuntos de dados (que pode ser um processo muito dispendioso para conjuntos de dados enormes). *Datasets* não rotulados são, portanto, mais abundantes (Wani et al., 2019).

Alguns algoritmos famosos que fazem uso de aprendizado não-supervisionado incluem (Mahesh, 2018):

a) **K-means**, que produz diversas subdivisões do espaço de indivíduos a partir de ponto de referência chamados de centroides, gerando grupos de indivíduos similares;
b) **PCA (*Principal Component Analysis*)**, que é utilizado para a redução de dimensionalidade na análise de dados multivariados, pela captura de vetores ortogonais que apresentam a maior variância;
c) ***Clustering* hierárquico**, que divide o espaço em grupos e subgrupos, gerando uma árvore de agrupamentos.

A lista acima também não é exaustiva, existindo diversas alternativas de modelos para agrupamento e outras atividades não supervisionadas.

2.4.4 Aprendizado autossupervisionado

Neste modelo de aprendizado também não há necessidade de rotular os dados, motivo que faz com que muitas vezes seja classificado como não supervisionado. No entanto, os algoritmos de treinamento são **os mesmos utilizados no aprendizado supervisionado**, motivo pelo qual alguns autores denominam esta modalidade de aprendizado de autossupervisionado (Goodfellow; Yoshua; Courville, 2016).

Um exemplo são as **Redes Neurais Artificiais**, cuja proposta original foi baseada em aprendizado supervisionado, mas que hoje apresentam variantes específicas que não requerem dados rotulados.

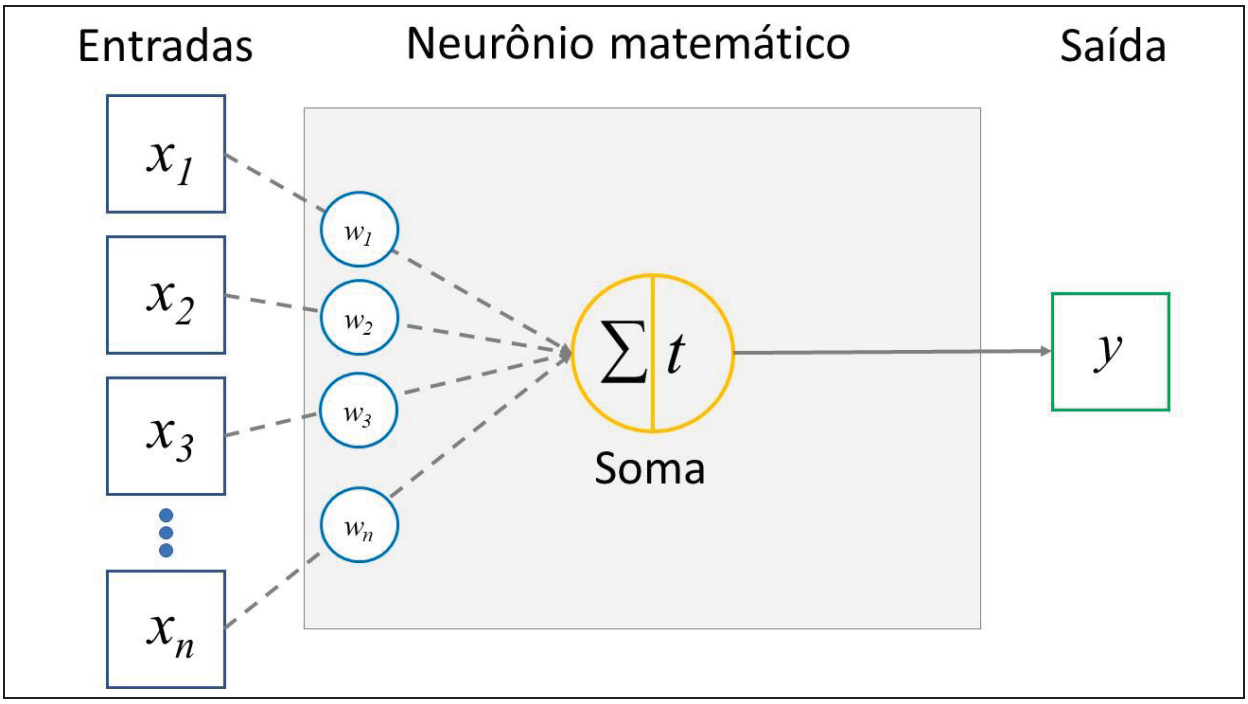
2.4.5 Redes neurais artificiais

As redes neurais artificiais, primeiramente propostas em um artigo seminal (McCulloch; Pitts, 1943), modelaram matematicamente, pela primeira vez, o comportamento de um neurônio biológico. Os autores (McCulloch era neurocientista e Pitts físico) informaram que o modelo proposto não expunha todas as características operacionais de um neurônio biológico, capturando apenas alguns de seus diversos aspectos. O sistema em questão era um modelo de cálculo onde operações matemáticas e lógicas podiam ser resolvidas universalmente por neurônios artificiais corretamente configurados.

Neste modelo, o neurônio é uma unidade processadora que recebe diversos sinais de entrada e produz um sinal de saída. Os sinais de entrada são recebidos pelos dendritos e ponderadamente somados segundo um conjunto de fatores de intensificação ou atenuação dos sinais de entrada, chamados de pesos sinápticos. Quando sinais de entrada são fornecidos ao neurônio e este os processa para produzir sua saída, diz-se que ocorreu uma **inferência**.

O sinal de saída não é emitido imediatamente. Apenas quando o somatório ponderado das entradas (chamado de potencial de ativação) atinge ou supera um limiar de ativação é que o sinal de saída é emitido. Diz-se, neste caso, que o neurônio está **ativado**. Na Figura 1 está colocado um modelo abstrato do neurônio matemático.

FIGURA 1 - MODELO DE UM NEURÔNIO MATEMÁTICO



FONTE: O AUTOR (2023)

O modelo na Figura 1 representa o neurônio matemático (unidade processadora) que recebe um conjunto de "*n*" entradas (x_1, x_2, x_3, ... x_n) para produzir uma saída (y). O neurônio matemático é composto pelos componentes w_1, w_2, ... w_n, $\sum$ e t. Os componentes $\sum$ e t combinados são algumas vezes chamados de "Soma".

A expressão matemática do neurônio artificial pode ser vista na equação (1).

$$y = \begin{cases} 1 \; se \; X.W^T \geq t \\ 0 \; se \; X.W^T < t \end{cases} \quad (1)$$

Com a equação sendo formada pelos seguintes componentes:

a) y (sinal de saída do neurônio) pode assumir os valores 1 (ativado) ou 0 (não ativado);

b) X é um vetor de sinais de entrada ($X = \{x_1, x_2, x_3 \ldots x_n\}$); por exemplo, se o neurônio recebe três sinais, $X = \{x_1, x_2, x_3\}$;

c) W^T é um vetor de pesos sinápticos que modula a influência de cada sinal de entrada no neurônio; cada sinal tem um peso, por isso a cardinalidade de W é a mesma que de X;

d) t é o limiar de ativação, que determina quando o neurônio irá ou não emitir sua saída.

O artigo seminal apresentou as bases do modelo matemático e demonstrou como as unidades neuronais podem ser utilizadas para diversos fins quando **pesos sinápticos corretos** são atribuídos. No entanto, os autores não apresentaram qualquer processo específico para o estabelecimento destes pesos.

2.4.6 Aprendizado de redes neurais

A primeira proposta de um algoritmo de aprendizado automatizado dos pesos sinápticos foi desenvolvida pelo psicólogo americano Frank Rosenblatt (Rosenblatt, 1958). Este algoritmo foi aplicado a um modelo nervoso artificial batizado de **Perceptron**, vindo a demonstrar como um neurônio poderia aprender autonomamente a partir de um conjunto de dados.

O algoritmo de aprendizado aplicado ao Perceptron é um processo iterativo baseado em sucessivas evoluções de uma solução inicial (não-ótima) candidata, entendendo-se aqui que tal solução corresponde a um conjunto de pesos sinápticos e um limiar de ativação. Um conjunto de dados rotulados é necessário para o treinamento, ou seja, o processo de melhorar as configurações do neurônio até que o mesmo passe a produzir as respostas desejadas (até que o neurônio aprenda as relações matemáticas existentes nos dados de entrada). Por este motivo, o conjunto de dados utilizado é frequentemente chamado de conjunto de treinamento. Os dados são submetidos ao neurônio e os rótulos são utilizados para validar a resposta que este produz. Trata-se, portanto, de aprendizado supervisionado.

O neurônio é inicializado com pesos aleatórios, sem qualquer critério específico, assim como o limiar de ativação. Por questões matemáticas, apenas evita-se o zero. Uma inferência é realizada para cada item no conjunto de treino e a quantidade de classificações corretas é contada. Esta contagem constitui o erro da rede, que é multiplicado por um fator arbitrariado “α”, chamado de taxa de aprendizado. Este produto é o fator de correção, que é somado aos pesos e subtraído do limiar de ativação. Após a correção, uma nova inferência é feita (ou seja, o neurônio executa nova classificação). O processo é repetido até que os erros cheguem a zero ou um limite de repetições seja atingido. Essas repetições são chamadas de **épocas** (do inglês, “*epochs*”).

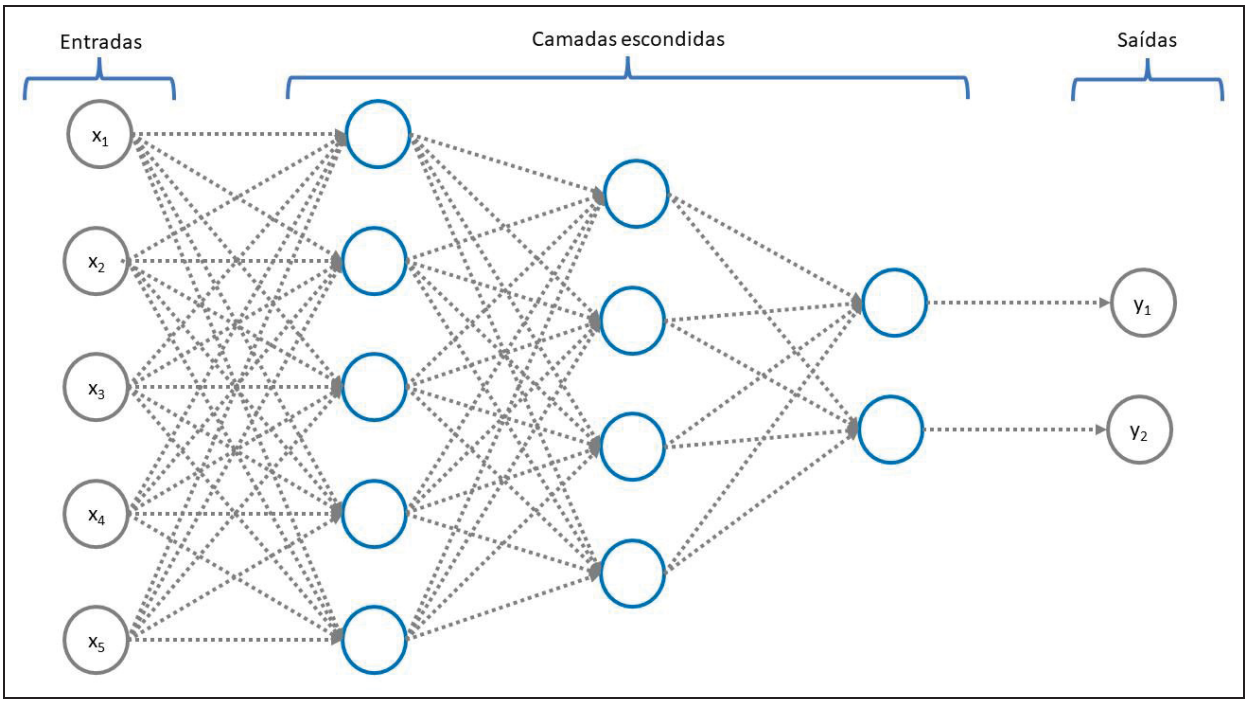
Estudos posteriores mostraram que o Perceptron apresenta limitações severas para problemas considerados “interessantes” e soluções alternativas foram propostas, modificando o treinamento de uma rede para um processo de otimização

matemática que frequentemente utiliza o método conhecido como “gradiente descendente”.

Ainda outros avanços mostraram que os neurônios artificiais podem ser combinados em redes cuja quantidade é determinada pela complexidade do mapeamento que se deseja fazer entre entradas e saídas. A organização dos neurônios em camadas e suas quantidades determina o que se chama de a arquitetura da rede.

O treinamento de uma rede de múltiplas camadas manteve-se como um desafio até a publicação de um artigo revolucionário que propôs o modelo **Multi-layer Perceptron** (Rumelhart; Hinton; Williams, 1985). Este modelo tornou possível a construção de redes neurais com diversas camadas, abrindo o campo para a aplicação de problemas mais complexos e de maior interesse. Na Figura 2 está colocado um exemplo do modelo Multi-layer Perceptron.

FIGURA 2 - MODELO MULTI-LAYER PERCEPTRON



FONTE: O AUTOR (2023)

Este modelo organiza a rede neural em uma camada de neurônios de entrada (que equivalem à quantidade de características presentes no conjunto de dados, no caso do exemplo com cinco características), uma camada de neurônios de saída (com a quantidade de neurônios igual às saídas esperadas da rede, no caso do exemplo de dois valores) e um conjunto de camadas escondidas (no caso do exemplo, três

camadas com cinco, quatro e dois neurônios, respectivamente), que separam as duas camadas anteriores.

Os neurônios das camadas escondidas constituem aqueles que serão treinados, ou seja, aqueles cujos pesos sinápticos serão ajustados para que o erro da rede seja minimizado.

As linhas tracejadas entre as camadas de neurônios representam a distribuição dos valores produzidos por neurônios de uma camada que atuam como entradas dos neurônios da camada seguinte. O processo é similar ao do Perceptron, valores de um conjunto de dados são fornecidos e as saídas computadas pelo modelo são subtraídas dos valores esperados, produzindo o erro da rede. O erro é multiplicado pelo “α” (taxa de aprendizado) e utilizado como fator de correção dos pesos sinápticos das várias camadas, quando então uma nova época é experimentada para avaliar se a rede já atingiu o grau de aprendizado desejado ou não. A diferença está em que a correção dos pesos ocorre da última para a primeira camada escondida em um processo chamado de “retropropagação do erro” (do inglês, “*error back-propagation*”).

Modelos com múltiplas camadas são capazes de aprender relações complexas e não-lineares entre grande número de variáveis, condição que caracteriza problemas de maior interesse e que por vezes não podem ser solucionados por outras técnicas de Aprendizado de Máquina.

2.4.7 Aprendizado profundo

O Aprendizado profundo (*Deep Learning*) é uma modalidade de Aprendizado de Máquina baseado na representação do conhecimento na forma de muitos níveis de abstração, como o que ocorre no cérebro de entes biológicos. Neste modelo, as representações com mais alto nível de abstração são construídas sobre as representações com menor nível de abstração (Wani et al., 2019).

Esta hierarquia de conceitos dá ao computador a capacidade de aprender uma variedade de relacionamentos complexos a partir de outros mais simples. Por exemplo, padrões simples poderiam ser as palavras, que podem ser combinadas em frases, que podem ser combinadas em parágrafos, em uma sequência de padrões cada vez mais complexos e abstratos. Tal hierarquia, quando expressa na forma de um diagrama, tem a forma de uma árvore de muitos níveis, ou seja, uma árvore

profunda. Esta é a razão pela qual tal abordagem é chamada de Aprendizado Profundo (Goodfellow; Yoshua; Courville, 2016).

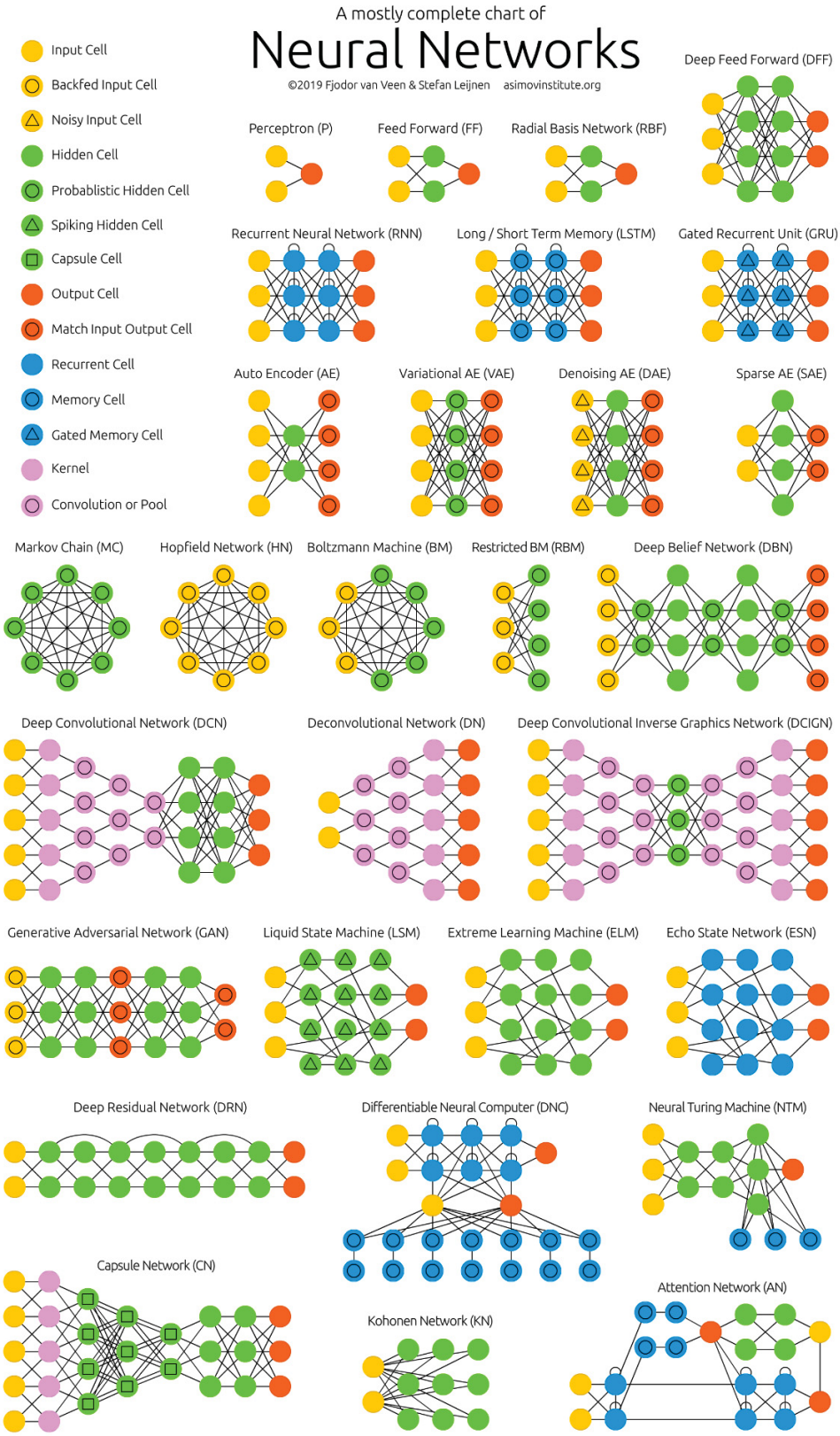
Partindo de uma arquitetura *Multi-layer Perceptron* (MLP), inúmeros modelos diferentes foram propostos para superar dificuldades particulares de problemas específicos, formando uma miríade de diferentes alternativas de arquiteturas. Na Figura 3 pode ser vista a imagem do *Neural Network Zoo*.

No *Neural Network Zoo*, disponível na página do The Asimov Institute[5], estão colocadas as arquiteturas mais populares de redes neurais. Aponta-se que não se trata de um diagrama absolutamente completo, visto que novas arquiteturas são propostas frequentemente. O que este mapa deixa claro é que existem diferentes tipos (arquiteturas) de redes neurais, alguns voltados a problemas bem específicos, outros de uso mais geral. Escolher a arquitetura mais apropriada para um determinado problema está entre as primeiras decisões tomadas em um projeto de Aprendizado de Máquina.

Como parte do Aprendizado de Máquina, o Aprendizado Profundo (*Deep Learning*) tem sido utilizado com sucesso em diversas áreas, como veículos autoguiados, visão computacional, reconhecimento de fala e diagnósticos médicos, dentre alguns exemplos (Mohri; Rostamizadeh; Talwalkar, 2012).

[5] https://www.asimovinstitute.org/neural-network-zoo/

FIGURA 3 - MAPA DE REDES NEURAIS COM UMA VARIEDADE DE ARQUITETURAS



FONTE: THE ASIMOV INSTITUTE (2019)

Uma característica diferencial do Aprendizado Profundo frente a outros processos de Aprendizado de Máquina é a seleção automática de características

(Alom et al., 2018). Em um cenário sem Aprendizado Profundo as características que serão processadas para a extração dos padrões são escolhidas a priori por diversos critérios combinados (análise exploratória dos dados, conhecimento do problema, objetivos do modelo, qualidade e volume dos dados, dentre outros) e então submetidas aos algoritmos. A presença de características inapropriadas pode operar como ruído nos dados e comprometer o desempenho do modelo. Com o Aprendizado Profundo, por outro lado, as características que realmente importam são selecionadas de forma automática e organizadas hierarquicamente.

Campos onde o Aprendizado Profundo proporcionou resultados diferenciados incluem classificação automática de imagens e reconhecimento de linguagem natural, falada ou escrita, identificação de emoções faciais, identificação de pose, escrita de poemas, pintura de quadros e muitas outras. Todo este resultado, no entanto, é trazido com o custo de muita capacidade computacional e um enorme volume de dados (Alom et al., 2018; Wani et al., 2019).

Em função destes variados avanços, a hipótese de que o Aprendizado de Máquina (profundo ou não) possa ser aplicado para detectar mentiras já vem sendo testada há mais de 20 anos.

A premissa que fundamenta a aplicação de Aprendizado de Máquina/ Profundo ao problema de detectar mentiras é a de que as diferenças fisiológicas e de expressão demonstradas entre narrativas sinceras e não sinceras podem ser descritas por dados. Espera-se que tais dados, quando processados por métodos de Aprendizado de Máquina/Profundo, produzam modelos de padrões com capacidade discriminante, podendo assim diferenciar um caso do outro.

2.4.8 Generalização

O desejado é que o modelo treinado apresente desempenhos similares tanto na etapa de treinamento, quanto na etapa de operação, situação na qual se diz que a rede generaliza bem. Porém, como as redes neurais profundas apresentam muitas camadas e muitos neurônios, existe o risco de sofrerem de *overfitting* (memorização dos padrões exatos ao invés de padrões genéricos).

O processo de utilizar estratégias diversas para evitar o *overfitting* nas redes neurais é chamado de **regularização**. Uma técnica que apresenta bons resultados é a inclusão de um fator de ***dropout*** (Srivastava et al., 2014). O *dropout* é uma taxa

aplicada às camadas do modelo para evitar que alguns neurônios recebam a correção em seus pesos sinápticos em dada época.

Por exemplo, se a taxa de *dropout* especificada para uma camada é de 0,2, então 20% dos neurônios daquela camada não participam do treinamento em dada época. Os neurônios são escolhidos aleatoriamente, portanto nas épocas seguintes existe a tendência de não repetição do mesmo grupo de neurônios excluídos. Experimentos mostram que o *dropout* pode reduzir a acurácia da rede durante o treinamento, mas aumenta durante a operação com dados novos, portanto, elevando a generalização do modelo.

2.4.9 Hiperparâmetros

Os diversos algoritmos de Aprendizado de Máquina recebem, além dos dados, parâmetros que controlam sua execução, mas que não são dados de entrada, pois não são utilizados para a descoberta dos padrões de aprendizado (Goodfellow; Yoshua; Courville, 2016).

Tais configurações são chamadas de hiperparâmetros e cada modelo de Aprendizado de Máquina pode apresentar um conjunto próprio deles. Por exemplo, o algoritmo K-Means, para agrupamento, recebe como hiperparâmetro a quantidade de grupos que se deseja alcançar.

Para o caso de redes neurais artificiais, os seguintes hiperparâmetros são aplicáveis:

1. **taxa de aprendizado** é um fator que dá a escala com que o processo de correções dos erros sinápticos é aplicado; números muitos grandes podem fazer com que a rede nunca encontre uma boa solução e número muito pequenos podem fazer com que o aprendizado evolua muito lentamente ou mesmo pare de evoluir (Patterson; Gibson, 2017);
2. **quantidade de épocas** determina a quantidade de iterações de treinamento que serão realizadas para melhorar o aprendizado da rede; números muito pequenos não permitem que o aprendizado atinja maturidade para identificar os padrões mais importantes e número muito grandes podem levar ao problema de *overfitting*;
3. **quantidade de neurônios** estabelece a quantidade de unidades processadoras, frequentemente arranjadas em camada, que aprenderão partículas dos padrões existentes nos dados;

4. **quantidade de camadas** determina a quantidade dos conjuntos de neurônios que recebem entradas e produzem saídas; a quantidade de camadas tem relação com a complexidade dos padrões que se deseja aprender a partir dos dados;
5. **tamanho do lote (*minibatch*)** controla a quantidade de indivíduos do conjunto de entrada que são processados antes de uma atualização de pesos; números menores implicam em atualizações mais frequentes o que requer mais processamento e aumenta o tempo de treinamento, mas pode oferecer maior precisão e números maiores implicam em tempos menores de processamento, mas podem interferir negativamente na precisão do modelo (Goodfellow; Yoshua; Courville, 2016);
6. **função de perda, custo ou erro** opera para medir quão longe dos valores ideais está a saída da rede neural e é utilizada para modular os fatores de correção de pesos durante o processo de treinamento (Goodfellow; Yoshua; Courville, 2016);
7. **função de ativação** introduz a não-linearidade que permite aos modelos modernos o aprendizado de relações não-lineares que frequentemente caracterizam os problemas de maior interesse; as funções de ativação também habilitam a modelagem do treinamento como um processo de otimização e atuam para superar problemas matemáticos que tendem a surgir em redes com muitas camadas, como o problema da explosão ou supressão dos gradientes (*exploding and vanishing gradiente problem*) (Goodfellow; Yoshua; Courville, 2016);
8. **algoritmo de aprendizado** estabelece o processo pelo qual os pesos são ajustados ao longo do processo de treinamento, a partir do erro calculado pela função de perda e taxa de aprendizado.

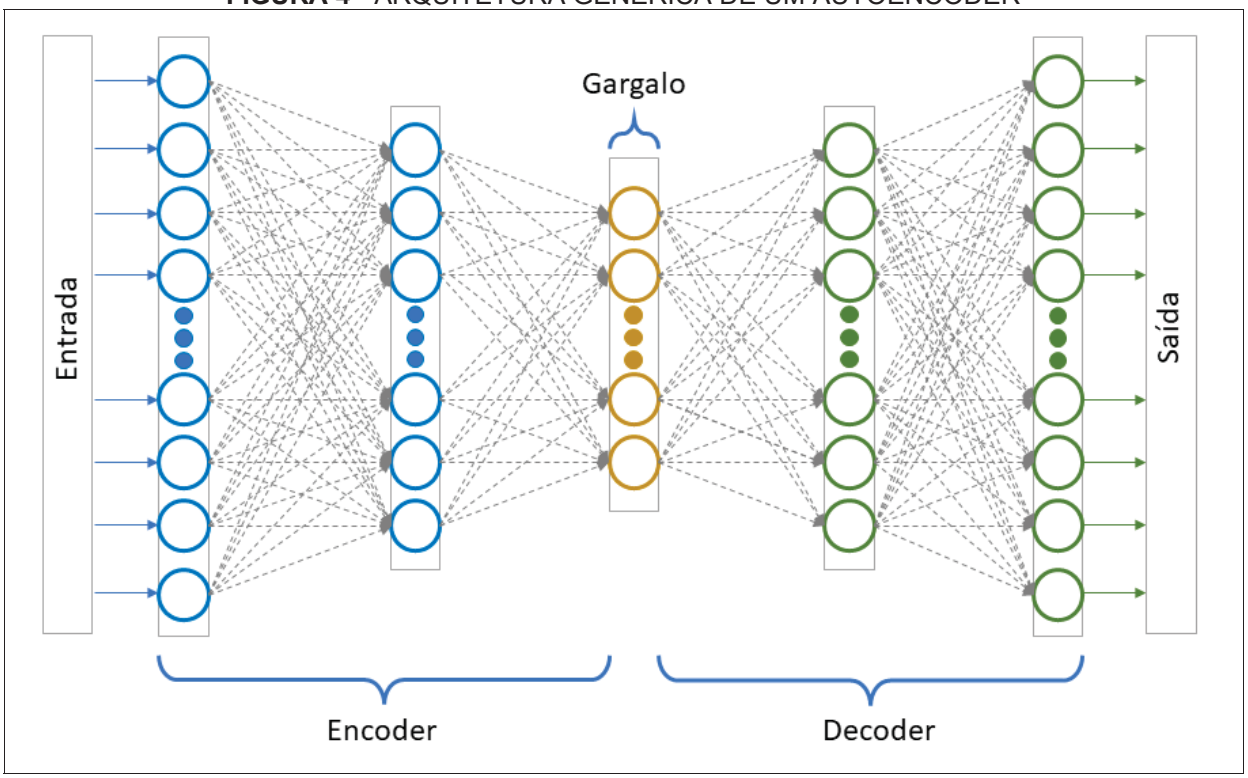
A escolha dos hiperparâmetros passa por um processo chamado ajuste de parâmetros (do inglês, "*parameter tunning*"). É uma etapa de exploração do efeito dos hiperparâmetros ao problema e implica em uma etapa de experimentações para a descoberta do conjunto de configurações mais adequado para o caso em estudo.

2.4.10 Autoencoders

Autoencoder (AE) é um tipo específico de rede neural artificial que faz uso do **aprendizado autossupervisionado**. A arquitetura deste modelo foi primeiramente

proposta na tese de doutorado de LeCun (Zhai et al., 2019) e, embora se trate de um único modelo, está virtualmente dividida em duas partes bem distintas, uma chamada de *encoder* (ou codificador) e outra chamada de *decoder* (ou decodificador). Na Figura 4 está apresentada uma representação gráfica de um Autoencoder.

FIGURA 4 - ARQUITETURA GENÉRICA DE UM AUTOENCODER



FONTE: O AUTOR (2023)

A arquitetura genérica de um Autoencoder contém um conjunto de camadas para codificar a entrada em uma camada específica chamada de “gargalo”, “vetor comprimido” ou “representação no espaço latente”. O vetor resultante da codificação ("gargalo”) será decodificado pelas camadas de decodificação, em uma distribuição simétrica ao codificador.

Eventualmente, o vetor central (gargalo) pode ter dimensionalidade maior do que a entrada, caso conhecido como *Overcomplete Autoencoder* (Goodfellow; Yoshua; Courville, 2016). Nesta configuração, a rede apresenta uma quantidade tal de neurônios que pode levar ao fenômeno da memorização (*overfitting*), fazendo com que haja uma tendência em favor dos Autoencoders com gargalos (*Undercomplete Autoencoders*).

Um Autoencoder é treinado para reconstruir (ou copiar) seus sinais de entrada em sua saída (Goodfellow; Yoshua; Courville, 2016), por este motivo a rotulação dos dados é desnecessária. Os próprios sinais de entrada são utilizados como valores

esperados de saída. Nesse sentido, o que um Autoencoder faz durante seu treinamento é registrar as relações entre as características dos dados de entrada apenas para reconstruí-los.

Durante o treinamento, os sinais de entrada são representados em uma camada escondida que os codifica, tipicamente com dimensionalidade mais baixa. Esta camada escondida consiste na saída do encoder e operará como entrada para o decoder, que então produzirá uma aproximação do que foram os sinais de entrada originais do modelo.

2.4.11 Detecção de anomalias com Autoencoders

Uma anomalia é descrita como uma variação suficientemente grande em uma amostra de dado para torná-la não pertencente à distribuição dos dados entendidos como normais. A anomalia (ou *outlier*), portanto, apresentará um conjunto de características diferente do esperado para os indivíduos tidos como normais (Wang; Bah; Hammad, 2019). Exemplos de áreas que fazem uso da detecção de anomalias são detecção de fraudes, monitoração automática de processos, diagnóstico médico, sistemas de vigilância, dentre outras (Chalapathy; Chawla, 2019).

No contexto da detecção de anomalias, os Autoencoders figuram como um modelo de bons resultados, estando no centro dos modelos autossupervisionados para este fim (Chalapathy; Chawla, 2019). Estes modelos são essencialmente treinados com dados que representam a normalidade, de forma a aprender as relações entre as variáveis submetidas que são próprias deste grupo de dados. Ao final do treinamento, tipicamente os Autoencoders são capazes de produzir uma **aproximação** da entrada. A diferença entre os dados fornecidos como entrada e aqueles produzidos como saída do Autoencoder é chamada de **erro de reconstrução**.

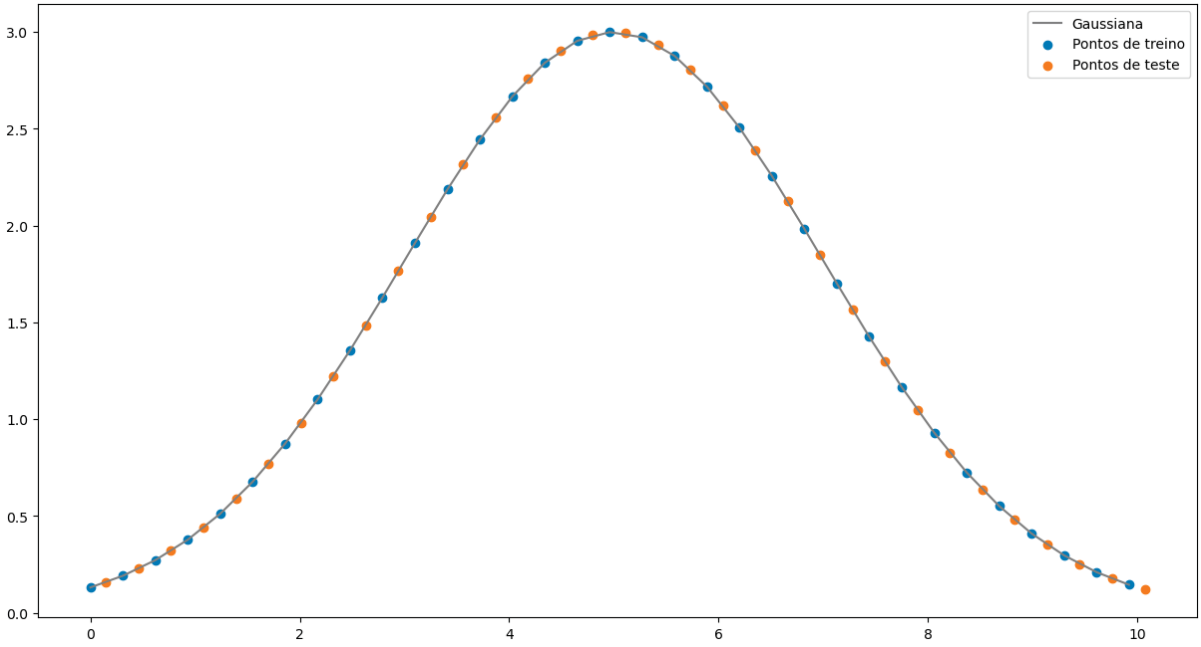
Neste campo, é frequente o uso dos *Undercomplete Autoencoders*, visto que se deseja evitar o problema de *overfitting*. No entanto, foi demonstrado que a adição de alguma técnica de regularização tem efeitos benéficos para o treinamento de *Overcomplete Autoencoders* (Yong; Brintrup, 2022), dando a esta arquitetura o potencial de aprender nuances mais complexas nos dados, levando a bons resultados diante do problema de detecção de anomalias.

O uso dos Autoencoders para detectar anomalias se baseia no estudo do erro de reconstrução. Quando “muito grande” (superação de um limiar de decisão), o erro

de reconstrução opera como indicador de anomalia (Bank; Koenigstein; Giryes, 2021; Borghesi et al., 2022; Gong et al., 2019), pois após treinado, o Autoencoder procurará distorcer os dados de entrada para conformá-los às relações aprendidas anteriormente (dados de normalidade)

Um exemplo pode tornar o princípio de operação do Autoencoder mais fácil de compreender. Na Figura 5 é possível observar um conjunto de 66 pontos pertencentes a uma curva gaussiana, que neste caso constitui a “normalidade”, ou seja, dados que se comportam de forma similar à esta curva são entendidos como casos “normais” de um dado evento.

FIGURA 5 - CONJUNTO DE 66 PONTOS QUE REPRESENTAM NORMALIDADE



FONTE: O AUTOR (2023)

Na intenção de utilizar um Autoencoder para detectar eventuais anomalias, um conjunto de 33 pontos é extraído para treinamento (observar legenda). Os demais pontos pertencem à mesma curva (também são dados normais), mas não são usados para treinamento e sim para testes, com o objetivo de aferir a qualidade do modelo treinado. Por isso, os dois conjuntos não apresentam intersecção.

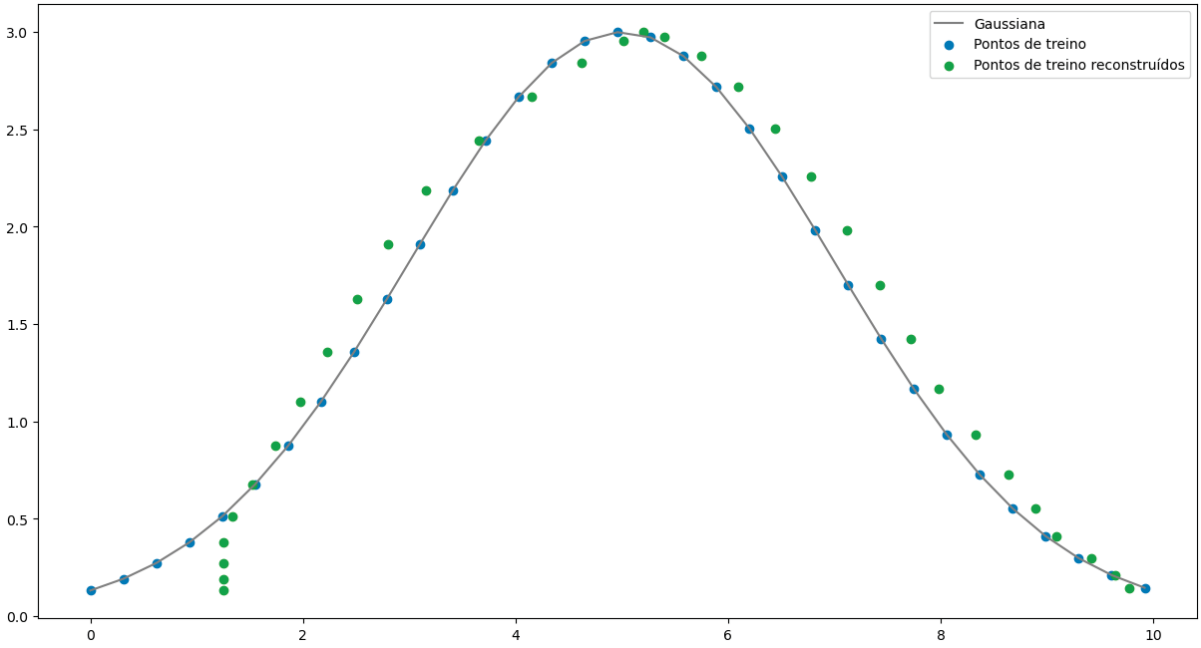
Um Autoencoder de cinco camadas pode ser treinado com os 33 pontos do conjunto de treinamento. A arquitetura deste modelo é:

- duas camadas para *encoder*, com quatro e três neurônios, respectivamente;
- uma camada para gargalo com dois neurônios;

- duas camadas para *decoder*, com três e quatro neurônios;
- camada de entrada e camada de saída com dois neurônios, as coordenadas X e Y dos pontos do conjunto.

Após o treinamento, o Autoencoder atingiu um nível de acurácia de 87,8% na reconstrução. Na Figura 6 podem ser vistos tanto o conjunto de 33 pontos de treinamento quanto o conjunto de 33 pontos reconstruídos.

FIGURA 6 - CONJUNTO DE TREINAMENTO E SUA RECONSTRUÇÃO APÓS O TREINAMENTO DO AUTOENCODER

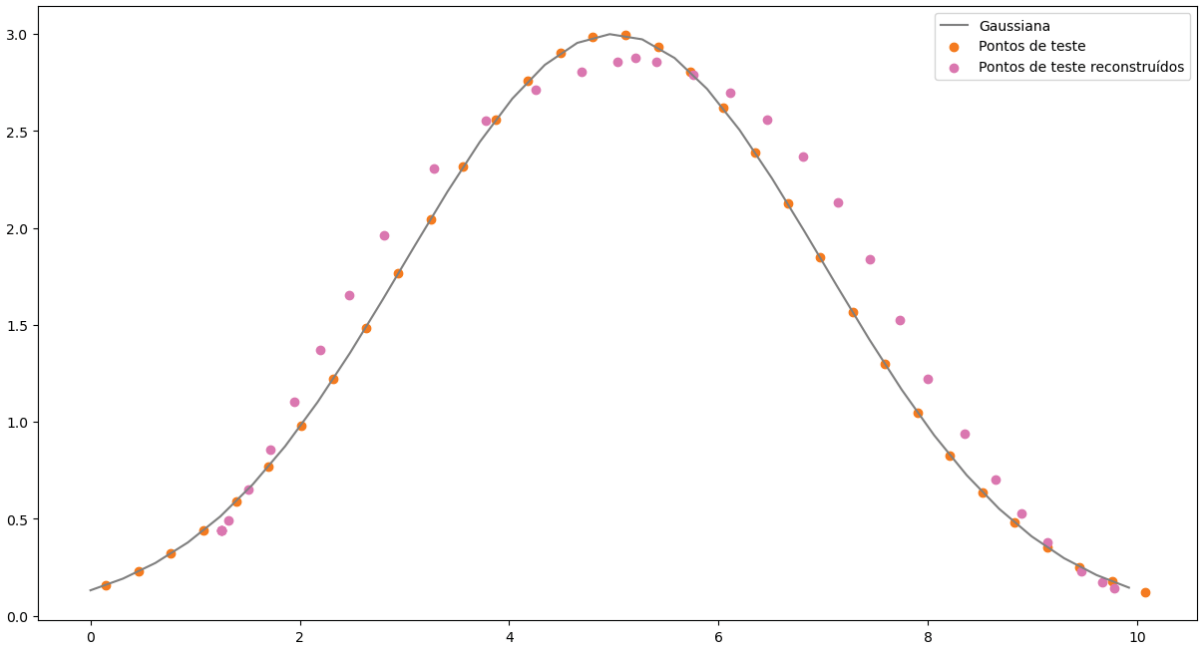


FONTE: O AUTOR (2023)

Embora a reconstrução não tenha sido perfeita, os pontos de treino reconstruídos tenderam a se aproximar dos pontos originais, o que demonstra que o Autoencoder aprendeu uma aproximação das relações existente entre as variáveis. Assuma-se que o erro de reconstrução denotado por "ε_t" (por alguma métrica) resultou em 0,0012. Neste caso, entende-se que 0,0012 é o limite de decisão, ou seja, o maior erro possível para aceitar um caso como normal é "ε_t".

Na Figura 7 está colocado o conjunto de teste, outros 33 pontos não submetidos ao treinamento, e suas respectivas reconstruções. A reconstrução continua não sendo perfeita, mas é possível perceber que os pontos se assemelham aos originais (ver legenda). Assuma-se que um dado ponto (x, y) do conjunto teste ofereça $\varepsilon_{(x,y)} = 0{,}0008$ (pela mesma métrica anterior). Como $\varepsilon_{(x,y)} \leq \varepsilon_t$, tem-se que o ponto de teste em questão não constitui uma anomalia.

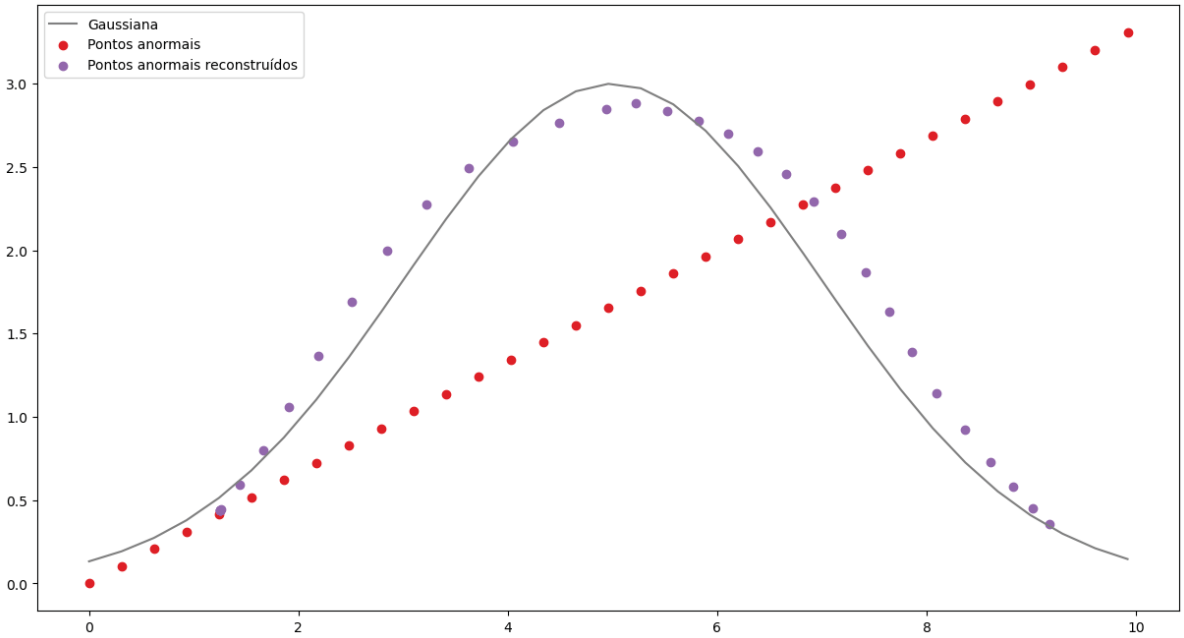
FIGURA 7 - CONJUNTO DE TESTE COM DADOS NORMAIS E SUA RECONSTRUÇÃO APÓS O TREINAMENTO DO AUTOENCODER



FONTE: O AUTOR (2023)

Na Figura 8 está apresentado um exemplo em que pontos não pertencentes à curva gaussiana são submetidos ao mesmo Autoencoder, assim como suas reconstruções.

FIGURA 8 - CONJUNTO DE TESTE COM DADOS ANORMAIS (NÃO EXTRAÍDOS DA GAUSSIANA) E SUBMETIDOS AO AUTOENCODER



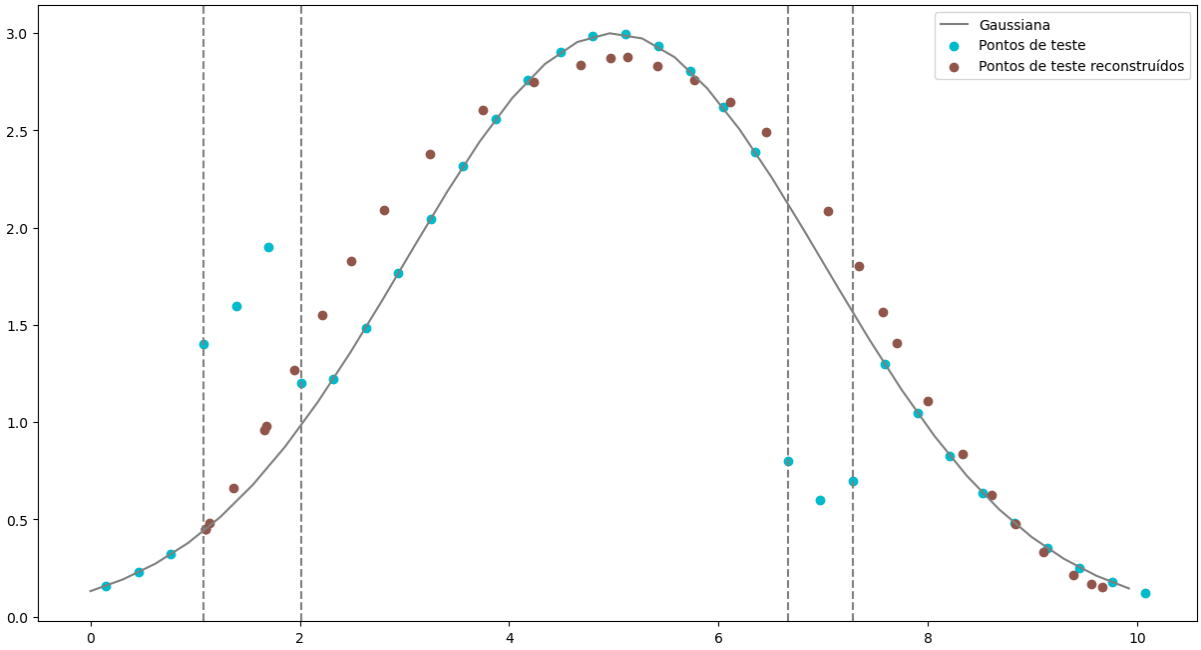
FONTE: O AUTOR (2023)

Neste caso os pontos pertencem a uma reta (com relações matemáticas bem diferentes de uma gaussiana), mas o Autoencoder, ao reconstruí-los, procurou ajustá-

los ao modelo que tinha aprendido (dados normais, ou seja, com distribuição gaussiana). Como consequência, os pontos reconstruídos são muito mais distantes dos seus correspondentes originais. Assuma-se que o cálculo do erro de um dado ponto (x, y) do novo conjunto ofereça $\varepsilon_{(x,y)} = 0{,}07$ (pela mesma métrica anterior). Como $\varepsilon_{(x,y)} > \varepsilon_t$, tem-se que o ponto de teste em questão é entendido como uma anomalia.

No entanto, em problemas de detecção de anomalias, a distribuição é apenas parcialmente destoante da curva aprendida durante o treinamento, como pode ser observado na Figura 9.

FIGURA 9 - CONJUNTO DE DADOS COM DOIS SEGMENTOS ANORMAIS



FONTE: O AUTOR (2023)

Neste caso, apenas duas secções da curva original apresentam pontos significativamente distantes da normalidade. Ainda assim, suas reconstruções foram ajustadas ao formato do que seria a curva normal aprendida, aumentando as distâncias entre os pontos originais e os reconstruídos. Espera-se que estas diferenças, quando medidas, apresentem valores suficientemente grandes para identificar as anomalias.

2.4.12 Erro Quadrático Médio

A forma mais frequente de avaliar o erro de reconstrução de um Autoencoder é por meio do Erro Quadrático Médio (Pratella et al., 2021), aqui referido como MSE (do inglês *Mean Squared Error*) expresso pela equação (2) (Chen et al., 2018).

$$MSE(X,Y) = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n} \quad (2)$$

Na equação precedente, "*n*" identifica a quantidade de valores de entrada, ou seja, a dimensão dos vetores X (entrada) e Y (reconstrução), enquanto "x_i" identifica os valores originais e "y_i" suas reconstruções. O MSE calcula as diferenças entre os valores originais e os reconstruídos pelo Autoencoder e as eleva ao quadrado para evitar cancelamentos de sinal. A média desses quadrados é calculada para representar o erro de toda a inferência.

Para o conjunto de 33 pontos utilizados para treinar o Autoencoder do exemplo anterior (Figura 6) o MSE de treinamento (MSE_t) resultou em 0,0012, enquanto para o conjunto de teste com dados normais (Figura 7) o MSE calculado (MSE_n) foi de 0,0010. O fato de $MSE_n < MSE_t$ valida o pressuposto que dados com características normais são reconstruídos com uma distorção dentro dos parâmetros de treino.

Por outro lado, para o conjunto de teste com dados anormais (Figura 8) o MSE calculado (MSE_a) foi de 0,1047, o que leva a uma classificação como dados anormais, pois o erro neste caso é muito superior ao limite de decisão.

O MSE calcula o erro com base na diferença numérica total entre os dados de entrada e os dados de saída.

2.4.13 Divergência de Kullback-Leibler

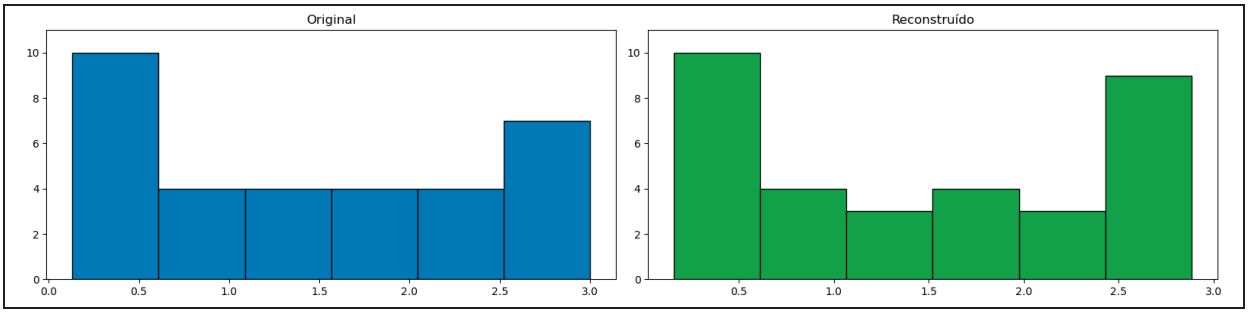
Outra forma de calcular o erro de reconstrução de um Autoencoder é avaliar a distribuição dos valores ao invés da diferença numérica. Para tanto, utiliza-se a Divergência de Kullback-Leibler, ou KLd (*KL Divergence*), uma métrica estatística que avalia numericamente quão divergentes são duas distribuições de probabilidades (Afgani; Sinanović; Haas, 2008). Para distribuições idênticas, a KLd = 0. Valores mais altos indicam maiores divergências. A KLd é dada pela equação (3).

$$KLd(p \,||\, q) = \sum_{x \in X} p(x)\ log_2^{\frac{p(x)}{q(x)}} \quad (3)$$

Na equação anterior, "p" e "q" são as distribuições de probabilidades dos dados originais e reconstruídos, respectivamente, enquanto "X" é o conjunto de todos os valores possíveis dentro do espaço em avaliação.

As distribuições de probabilidades necessárias para o cálculo da KLd podem ser aproximadas por meio de histogramas. Estão exemplificados na Figura 10 os dois histogramas produzidos a partir do conjunto de treinamento de 33 pontos mostrados na Figura 6.

FIGURA 10 - HISTOGRAMAS DOS DADOS ORIGINAIS E RECONSTRUÍDOS A PARTIR DO CONJUNTO DE 33 PONTOS DE TREINAMENTO

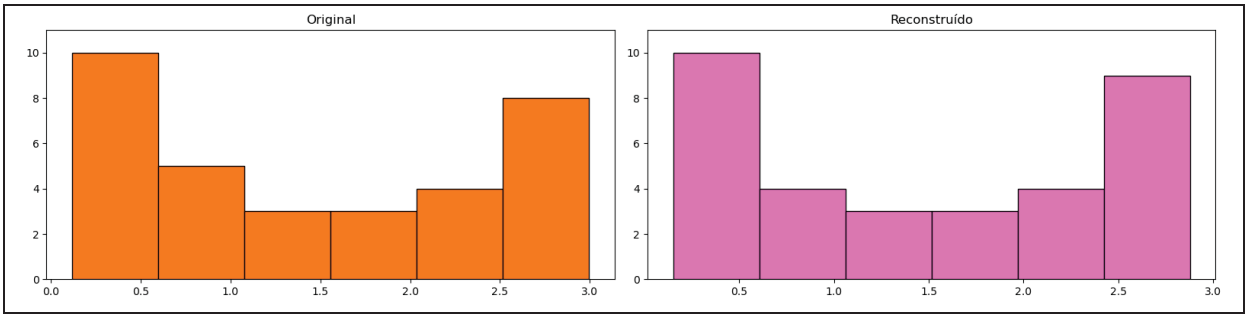


FONTE: O AUTOR (2023)

Os dois histogramas não são idênticos, pois são baseados em dados também não idênticos, mas são similares. Se as contagens representadas em cada uma das colunas forem divididas pela quantidade total de pontos (33 neste caso), o resultado é um conjunto de probabilidades, que podem então ser colocadas na fórmula da KLd. É importante observar que a quantidade de colunas do histograma é crítica, pois afeta as contagem de pontos e, portanto, a aproximação da distribuição de probabilidades (Afgani; Sinanović; Haas, 2008). Como neste caso foram utilizadas seis colunas, a notação a ser utilizada é KLd:6, que foi calculada em 0,0681.

Na Figura 11 estão apresentados os histogramas construídos a partir do conjunto de dados de teste normais (Figura 7).

FIGURA 11 - HISTOGRAMAS DOS DADOS ORIGINAIS E RECONSTRUÍDOS A PARTIR DO CONJUNTO DE 33 PONTOS DE TESTE DE NORMALIDADE



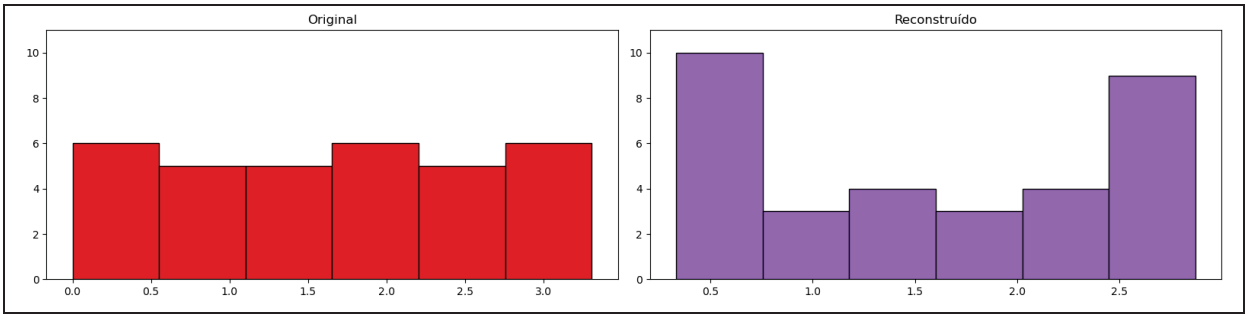
FONTE: O AUTOR (2023)

Neste caso, os dois histogramas também sugerem distribuições semelhantes, com a KLd:6 = 0,0572. Pelo critério do uso do KLd:6 do conjunto de treinamento

(0,0681) como limite de decisão, este conjunto também seria classificado como normal.

Finalmente, na Figura 12 estão dispostos os dois histogramas construídos a partir dos dados de teste anormais (Figura 8).

FIGURA 12 - HISTOGRAMAS DE DADOS ORIGINAIS E CONSTRUÍDOS A PARTIR DO CONJUNTO DE 33 PONTOS DE TESTE DE ANORMALIDADE



FONTE: O AUTOR (2023)

Seguindo a mesma lógica de avaliação, a para este conjunto a KLd:6 = 0,1336, o que implica em classificar o conjunto como uma anomalia, pois supera o limite de decisão calculado com os dados de treinamento (0,0681).

Como já colocado, a escolha da quantidade de colunas tem influência significativa no cálculo da KLd. Um conjunto muito pequeno que colunas poderia fazer com que as anomalias na distribuição não fossem percebidas e uma quantidade muito grande faria com que dados normais extremos pudessem ser considerados como anomalias. A título de comparação, no Quadro 3 estão colocados os cálculos da KLd para três opções de histograma.

QUADRO 3 - COMPARATIVO DOS DIVERSOS VALORES DE KLD PARA DIFERENTES HISTOGRAMAS

Conjunto	KLd:5	KLd:6	KLd:30
Treino	0,0504	0,0681	0,3195
Teste normal	0,0742	0,0572	0,4062
Teste anormal	0,1617	0,1336	0,3629

FONTE: O AUTOR (2023)

O processo para identificar a quantidade ótima de colunas para os histogramas envolve uma busca exaustiva que, embora simples e eficaz, é custoso computacionalmente (Afgani; Sinanović; Haas, 2008).

A divergência de Kullback-Leibler também pode ter uma expressão para o caso gaussiano, quando se considera que a série de valores tem uma distribuição normal. Neste caso a expressão da divergência KL para o caso normal (KLn) é dada pela equação (4) (Belov; Armstrong, 2011):

$$KLn(\mu_p, \mu_q, \sigma_q) = \frac{(\mu_p - \mu_q)^2}{2{\sigma_q}^2} \quad (4)$$

Na equação anterior, "μ_p" é a média da série de valores "p", assim como "μ_q" e "σ_q" são respectivamente a média e o desvio padrão da série "q".

Por esta forma de cálculo presumir uma distribuição próxima da normal, o uso em distribuições outras implicará na presença de um erro. No entanto, seu cálculo não depende da construção de histogramas.

2.5 Mecanismo de Atenção

Um dos domínios em que o Aprendizado de Máquina tem mostrado alto desempenho é no Processamento de Linguagem Natural (PLN). Este problema foi, durante muitos anos, abordado como um problema de série temporal (no inglês, *time series*), para o qual a arquitetura de rede neural mais frequentemente utilizada se chama LSTM (*Long Short-Term Memory*) (Vaswani et al., 2017). No entanto, tudo isso mudou com uma nova arquitetura chamada *Transformer*, que implementa um mecanismo especial de otimização de aprendizado chamado Atenção.

O Transformer é um modelo do tipo encoder-decoder (como os Autoencoders) no qual o mecanismo de Atenção é utilizado para aumentar a capacidade de aprendizado das relações entre diferentes indivíduos do conjunto de entrada, pela maior valoração de alguns elementos de entrada em relação a outros. Após seu sucesso, o Transformer passou a ser um componente chave em outra rede de elevado nível de desempenho em tarefas de PLN, a GPT (*Generative Pretrained Transformer*) (Radford et al., 2018), pavimentando outro campo de estudo, o Entendimento de Linguagem Natural (no inglês *Natural Language Understanding*).

A Atenção foi introduzida inicialmente em um modelo de tradução automática, tendo superado expressivamente soluções baseadas em redes LSTM (Vaswani et al., 2017). O que a Atenção faz é avaliar indivíduos em uma janela no conjunto de treinamento (um subconjunto cujo tamanho é definido como hiperparâmetro) e aprender quais deles contribuem mais para os resultados da inferência, focando especificamente nestes indivíduos. Esta análise é feita por meio de "cabeças de atenção" (no inglês *Attention heads*), que são matrizes cujos elementos são aprendidos ao longo do treinamento.

Existem dois tipos básicos de Atenção: *single-head attention* e *multi-head attention*. No primeiro caso, a Atenção é calculada uma única vez durante cada época

de treinamento, enquanto na segunda é calculada em paralelo diversas vezes, cada uma com uma visão diferente da entrada, que ao final são todas integradas na forma de uma média.

O resultado da Atenção é um conjunto de pesos que identificam quais os indivíduos que mais contribuem para o objetivo da rede. Estes indivíduos recebem um peso de tal forma que a soma de todos é sempre 1. Estes pesos influenciarão a rede para valorizar mais alguns indivíduos em detrimento de outros (“prestar mais atenção” em alguns indivíduos e menos em outros). No caso da atenção *multi-head*, o treinamento poderá identificar vários indivíduos mais importantes que os demais.

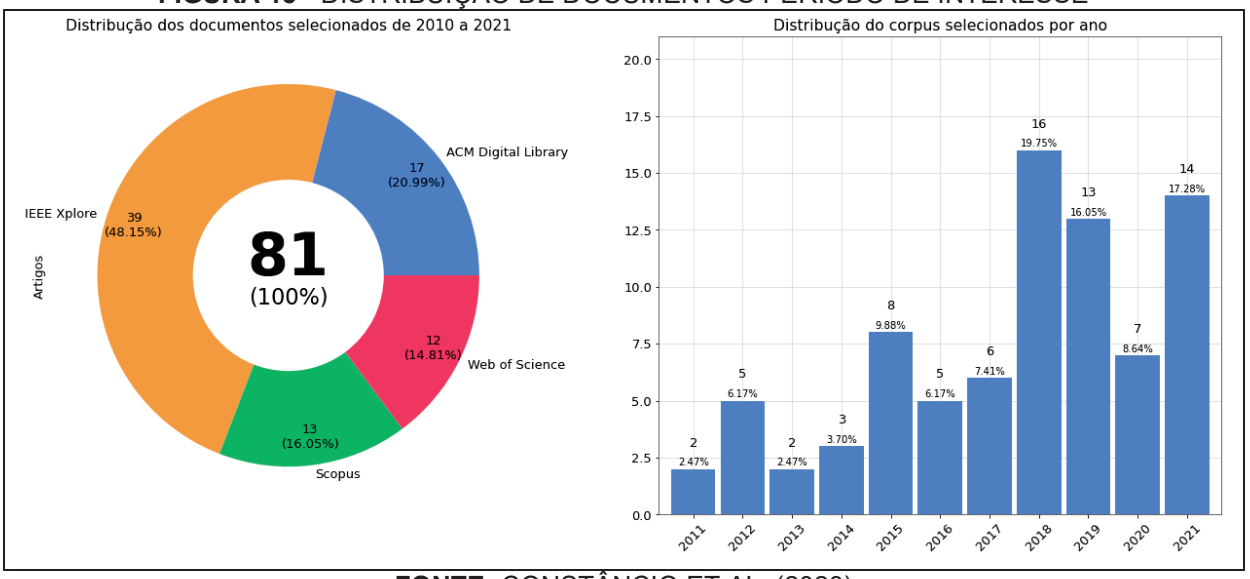
Ao capturar a importância relativa de cada item do conjunto de entrada, valorizando mais alguns e desvalorizando outros, a rede adquire a capacidade de perceber relações não lineares e de longa distância entre os mesmos. Este aprendizado atua para interferir na decisão da rede, que usará a configuração de entrada como um fator adicional durante suas inferências. Em outras palavras, uma rede neural regular é capaz de identificar relações entre as diversas características dos indivíduos do conjunto de entrada, mas não eventuais dependências que as adjacências desses indivíduos possam impor. Por outro lado, uma rede atencional expande sua percepção para identificar relações entre os indivíduos, reconhecendo eventuais padrões de conjunto. Este aprendizado adicional opera para melhorar a percepção da rede em problemas onde estas relações contribuem para o resultado.

Embora o mecanismo de Atenção tenha nascido no domínio do PLN, o mesmo já foi traduzido para outros campos de aplicação, como por exemplo a Atenção visual, utilizada em redes neurais dedicadas à visão computacional (Xu et al., 2015). A capacidade da Atenção de capturar relações de longa distância entre os indivíduos do conjunto de treinamento lhe confere o potencial de aplicação em qualquer problema onde os indivíduos apresentem situações de dependência entre si, sejam estas explícitas ou implícitas.

2.6 Detecção automática de mentiras

Os dados coletados dos documentos revisados pela revisão sistemática de literatura conduzida (Constancio et al., 2023) permitem afirmar que houve um aumento no volume de publicações a respeito de detecção de mentiras apoiada por Aprendizado de Máquina no período de interesse, como pode ser visto na Figura 13.

FIGURA 13 - DISTRIBUIÇÃO DE DOCUMENTOS PERÍODO DE INTERESSE

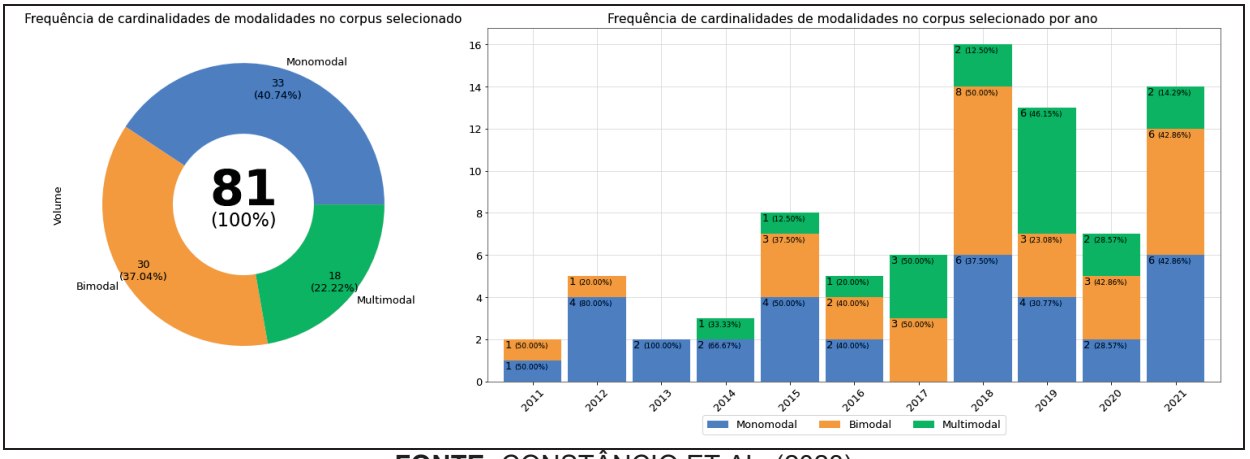


FONTE: CONSTÂNCIO ET AL. (2023)

O gráfico de rosca à esquerda apresenta a distribuição de artigos selecionados nos quatro portais de busca científica. No gráfico de barras à direita está a distribuição anual dos mesmos artigos ao longo do período de 2011 a 2021. É notável o aumento de volume nos últimos quatro anos, que juntos somam 50 (61,73%) dos 81 artigos selecionados.

A análise estatística expôs que a complexidade das abordagens também aumentou, uma vez que diferentes modalidades foram combinadas e exploradas para atingir níveis mais altos de desempenho em diferentes cenários e sob diferentes restrições, demonstrado na Figura 14.

FIGURA 14 - FREQUÊNCIAS DE MODALIDADES NO PERÍODO DE INTERESSE



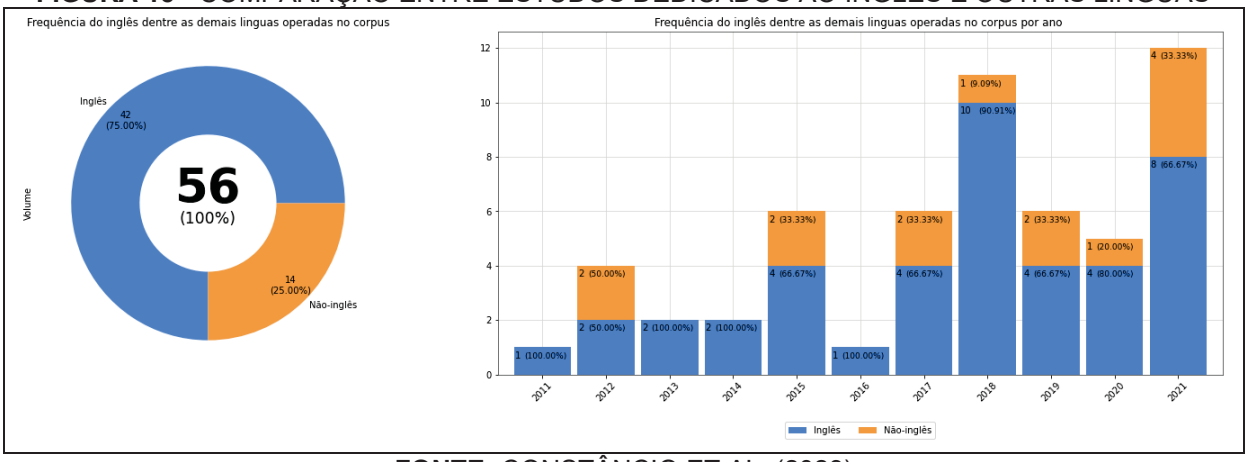
FONTE: CONSTÂNCIO ET AL. (2023)

O gráfico de rosca à esquerda apresenta a distribuição das cardinalidades relatadas nos artigos selecionados. No gráfico de barras empilhadas à direita está a distribuição anual dos mesmos artigos (e as cardinalidades exploradas) ao longo do

período de interesse, evidenciando a mudança de tendência de propostas monomodais (primeiros cinco anos) para bimodais e multimodais (últimos cinco anos).

A falta de investigação a respeito de Detecção de Mentiras em idiomas diferentes do inglês é evidente a partir dos resultados mostrados pela análise estatística, como pode ser percebido na Figura 15.

FIGURA 15 - COMPARAÇÃO ENTRE ESTUDOS DEDICADOS AO INGLÊS E OUTRAS LÍNGUAS



FONTE: CONSTÂNCIO ET AL. (2023)

O gráfico de rosca à esquerda apresenta a distribuição dos estudos dedicados à língua inglesa e outras línguas agrupadas. No gráfico de barras empilhadas à direita está a distribuição anual dos mesmos artigos (e as línguas exploradas) ao longo do período de interesse. Dentre as línguas diferentes do inglês não se encontra o português. Nem todos os estudos do corpus exploraram pistas vocais ou verbais, o que justifica um subconjunto de 56 estudos dos 81 que compreendem a revisão.

A revisão de literatura mostra que a prevalência de dados simulados (65,28%) sobre dados reais (34,72%) desafia alguns dos resultados, uma vez que a influência dos dados simulados sobre os resultados é desconhecida. Mesmo assim, tais resultados não devem ser considerados inválidos.

Graças ao lançamento público do *Real-Life Trial Deception Detection Dataset* (RLTDDD) em 2015 (Pérez-Rosas et al., 2015), o volume de pesquisa com dados reais aumentou, apresentando resultados de alto desempenho. Trata-se de um conjunto de dados multimodal, embora as pesquisas vocais e textuais sofram com o fato de todos os vídeos serem em inglês. Estudos dedicados às indicações vocais de outros idiomas carecem de sua própria versão do conjunto de dados da vida real.

A revisão de literatura também evidenciou que os autores exploraram o arsenal de técnicas de Aprendizado de Máquina na forma de 26 diferentes algoritmos.

Os cinco mais frequentes foram Redes Neurais (34 vezes, 30,09%), Support Vector Machines (SVM) (28 vezes, 24,78%), Random Forest (20 vezes, 17,70%), Árvores de Decisão (21 vezes, 18,58%), e K-Nearest Neighbor (KNN) (10 vezes, 8,85%).

O aprendizado supervisionado foi adotado em 80 estudos visto que nestes o problema foi modelado como uma tarefa de classificação. No entanto, um estudo (Mathur; Mataric, 2021) endereçou o problema de escassez de dados rotulados propondo o uso das Deep Belief Networks (DBN), que é uma rede neural treinada por meio de aprendizado não-supervisionado.

Das 34 ocorrências de Redes Neurais, 12 (35,29%) utilizaram redes Multi-layer Perceptron (MLP), 9 (26,47%) redes Long Short-Term Memory (LSTM), 3 (8,82%) Redes Neurais Convolucionais (CNN) e 2 (5,88%) Autoencoders. Os demais estudos utilizaram variadas arquiteturas. As acurácias relatadas para todos esses modelos variaram de 0,7961 a 0,9674.

2.7 Resumo da fundamentação teórica

Dada a variedade de técnicas e conceitos envolvendo Aprendizado de Máquina, o Quadro 4 foi elaborado para resumir aquelas que foram, ainda que brevemente, discutidas até então.

QUADRO 4 - RESUMO DOS MODELOS E CONCEITOS DE APRENDIZADO DE MÁQUINA

Técnica	Exemplos de Aplicação	Tipo de Aprendizado	Limitações
K-Means	Segmentação de clientes, análise de imagem	Aprendizado não-supervisionado	Sensível à inicialização, requer número de clusters pré-especificado, sensível à forma dos clusters
PCA (Principal Component Analysis)	Redução de dimensionalidade, compressão de imagem	Aprendizado não supervisionado	Assume linearidade, não captura bem relações não-lineares entre características
Clustering Hierárquico	Classificação de espécies, análise de documentos	Aprendizado não-supervisionado	Sensível à escolha de métricas de distância, computacionalmente custoso para grandes conjuntos de dados
Naive Bayes	Filtragem de spam, categorização de texto	Aprendizado supervisionado	Pressupõe independência das características, pode ter desempenho reduzido em dados complexos
Árvores de Decisão	Diagnóstico médico, detecção de fraudes	Aprendizado supervisionado	Tendência a *overfitting* com árvores profundas, sensíveis a pequenas variações nos dados de entrada
SVM	Classificação de imagens, detecção de anomalias	Aprendizado supervisionado	Sensível à escolha de hiperparâmetros, pode ser computacionalmente intensivo em grandes conjuntos de dados

Redes Neurais	Reconhecimento de fala, visão computacional	Aprendizado supervisionado	Requer grande quantidade de dados, ajuste complexo de hiperparâmetros, tendência a *overfitting*
Aprendizado Profundo	Reconhecimento de imagem, processamento de linguagem natural	Aprendizado supervisionado, não-supervisionado, autossupervisionado e semi-supervisionado	Requer grande quantidade de dados, alto custo computacional, tendência a *overfitting*
Autoencoders	Redução de dimensionalidade, detecção de ruídos	Aprendizado autossupervisionado	Dificuldade na escolha do tamanho da camada latente, sensíveis à qualidade dos dados de entrada
Transformers	Tradução automática, processamento de linguagem natural	Aprendizado supervisionado e não supervisionado	Requer grande quantidade de dados de treinamento, computacionalmente intensivo em sua forma completa
GPT	Geração de texto, respostas automáticas	Aprendizado não supervisionado	Limitações similares a outros modelos de linguagem, geração de conteúdo às vezes incoerente

FONTE: O AUTOR (2023)

Tanto a literatura consultada quanto a revisão sistemática de literatura composta pelos **81 artigos** selecionados apresentam um conjunto rico e complexo de conceitos, descobertas e incertezas a respeito da detecção de mentiras. O que todos parecem concordar, no entanto, é na dificuldade intrínseca do problema, dada a variedade de fatores que influenciam tanto a mentira quanto a sua detecção.

No Quadro 5 está listada uma série de conceitos relacionados com a detecção de mentiras e informações adicionais envolvendo fonte de dados, modalidades de pistas, tecnologias relacionadas e alternativas para Aprendizado de Máquina.

QUADRO 5 - QUADRO RESUMO DOS CONCEITOS RELACIONADOS A DETECÇÃO DE MENTIRAS

Tipo	Conceito	Significado	Discussão
Fonte de dados	Vida real (*real-life data*)	Dados coletados em situações não simuladas onde os sujeitos se manifestam espontaneamente	Difíceis de conseguir e de categorizar entre “verdade” e “mentira”, mas com pistas mais representativas dos processos psíquicos, emocionais e fisiológicos dos sujeitos observados
	Simulações (*mock data*)	Dados coletados em situações de laboratório onde os sujeitos respondem a instruções recebidas	Coletados em situações de laboratório, muitas vezes mediante gratificação monetária, fáceis de categorizar entre “verdade” e “mentiras”, representam alguns processos psíquicos, emocionais e fisiológicos que podem não ser os mesmos em situações de vida-real

Complexidade	Monomodal	Apenas uma modalidade de pista é explorada	Mais fácil de aproveitar porque dispensa a combinação e sincronização de diferentes sinais, mas pode deixar de considerar inferências mais ricas
	Multimodal	Pistas de diferentes modalidades são aproveitadas em conjunto	Mais complexa porque requer o entendimento da relação entre sinais de diferentes fontes e sua sincronização, mas pode aumentar a precisão porque observa mais sinais
Modalidade	Visual	Variação nas expressões faciais, na manifestação de gestos, assim como no movimento da cabeça e do corpo	São as mais fáceis para aproveitamento humano, hoje com suporte tecnológico acessível por meio de tecnologias de visão computacional
	Acústica	Variações no tom e na qualidade da voz, a velocidade da fala e a presença de pausas, repetições e erros	Aproveitáveis por humanos apenas para o espectro audível, hoje já com suporte computacional acessível por meio de tecnologias específicas
	Verbal	Variações léxicas e gramaticais no discurso, como a frequência de autorreferências, a frequência de pronomes, a referência a entidades e detalhes sensoriais como cores, texturas, locais e datas	Aproveitáveis por humanos e também por máquina por meio das tecnologias de Processamento de Linguagem Natural, inclusive para o português do Brasil, ainda não explorado na literatura consultada
	Emocional	Variações nas expressões faciais, tom da voz e escolha das palavras motivadas pela mudança no estado emocional do indivíduo	Pouco explorada tecnologicamente, ainda que existam estudos que procuram identificar emoções em fontes visuais, acústicas e verbais
	Fisiológica	Variações em indicadores fisiológicos como temperatura corporal, ritmo cardíaco e respiratório, pressão arterial, dilatação de pupila e ativação de áreas do cérebro	Depende de sensores específicos com contato físico com o sujeito, como medidores fisiológicos, eletrocardiogramas e eletroencefalogramas
Tecnologia	Polígrafo	Dispositivo que registra variações fisiológicas em um indivíduo	Requer uma calibragem preliminar e registra variações fisiológicas que devem ser interpretadas por um operador humano
	Aprendizado de Máquina	Tecnologia computacional que dá ao computador a capacidade de atuar em uma atividade sem ter sido especificamente programado para tal	Pode requerer calibragem ou não, assim como dados rotulados ou não, mas visa identificar padrões intrínsecos nos dados para subsidiar uma inferência automática, sem a necessidade de um operador humano
Aprendizado	Supervisionado	Requer dados rotulados para a construção de modelos preditivos	Requer dados rotulados que são raros e pouco variados, mas oferece a oportunidade de construir modelos preditivos que classificam diretamente um novo exemplar de pistas

	Auto-supervisionado	Dispensa dados supervisionados e permite a construção de modelos reconstrutivos	Dispensa os dados rotulados, mas requer um processo de calibragem para poder criar um modelo de normalidade que servirá de linha de base para a identificação de variações suficientemente salientes para identificar uma anomalia

FONTE: O AUTOR (2023)

Os aspectos evidenciados no quadro precedente foram escolhidos para sintetizar conceitos relevantes para fundamentar as escolhas metodológicas adotadas para validar a hipótese de pesquisa e atingir os objetivos da presente pesquisa.

3 ENCAMINHAMENTOS METODOLÓGICOS

Este capítulo apresenta os procedimentos metodológicos que orientaram a pesquisa para o atingimento dos seus objetivos. A pesquisa foi classificada dentro dos diversos critérios acadêmicos comumente aplicados.

3.1 Caracterização da pesquisa

Os diversos critérios de classificação desta pesquisa estão colocados a seguir (Creswell, 2014).

a) **propósito**: *pesquisa quase-experimental*, com o propósito de avaliar as características necessárias para a identificação de narrativas não sinceras por meio do Aprendizado de Máquina e pistas multimodais, incluindo especificamente pistas verbais para o português do Brasil;
b) **natureza dos dados**: *quantitativa*, visto que, embora a fonte primária dos dados fossem mídias digitais, destas foram extraídas variáveis numéricas específicas, assim como suas relações, capazes de operar como pistas de narrativas mentirosas; os resultados alcançados também foram expressos numericamente;
c) **obtenção dos dados**: coleta de vídeos de acesso público (YouTube) para a construção de um conjunto de dados anotado (narrativas classificadas como sinceras e não sinceras);
d) **natureza**: *aplicada*, pois produziu conhecimento cuja aplicação responde a situações de problemas reais.

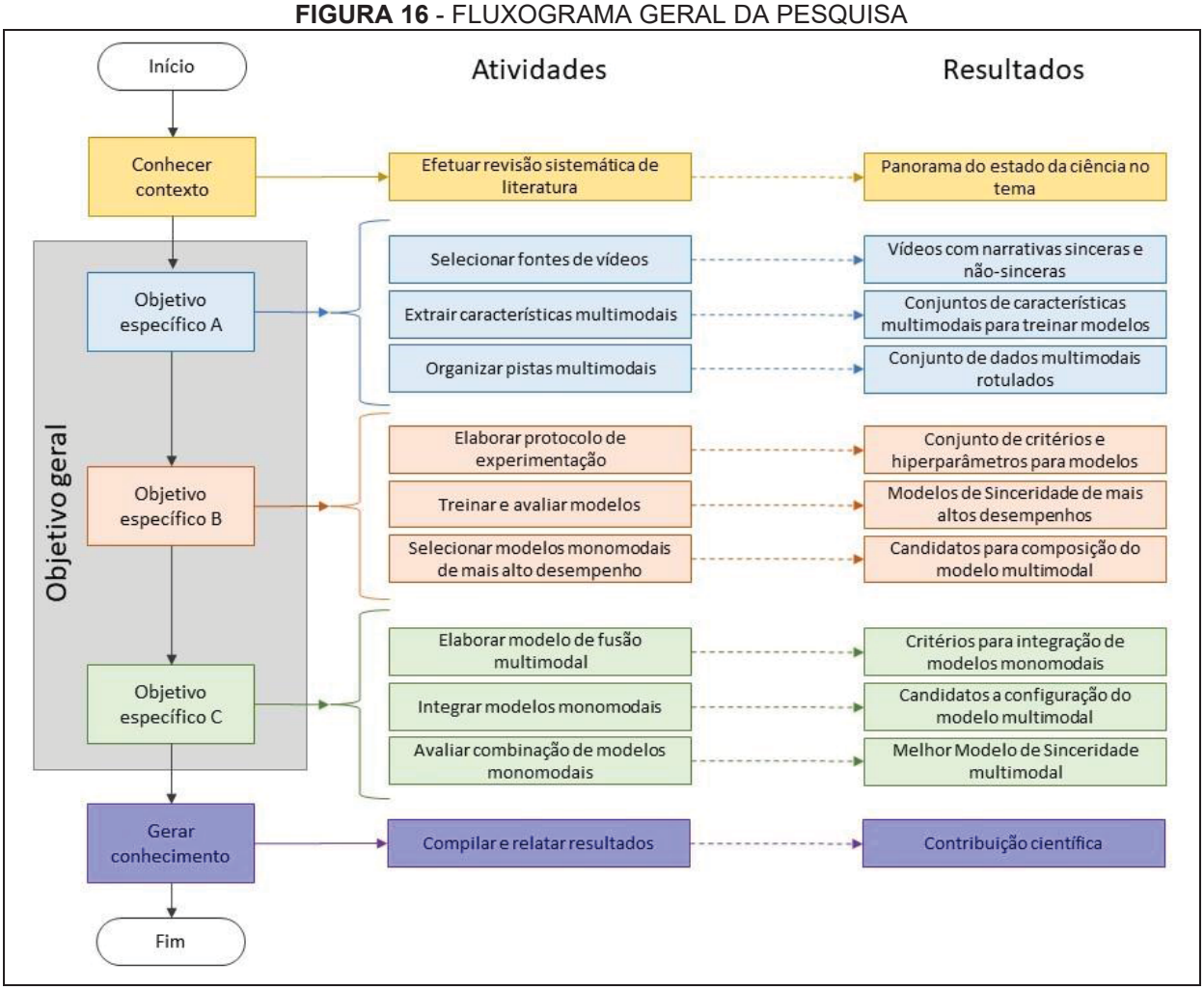
Entendeu-se que estes aspectos são compatíveis com o objetivo geral, a confirmação da hipótese enunciada, a contribuição para a resolução do problema percebido e a resposta à questão de pesquisa.

3.2 Fluxograma de pesquisa

Como meio de facilitar a compreensão, na Figura 16 está apresentado um fluxograma com as diversas atividades realizadas, incluindo objetivos específicos, e seus resultados.

O fluxograma organiza os três objetivos específicos em série. Dentro de cada um, também em série, estão as atividades desempenhadas para os seus atingimentos. Finalmente, vinculado a cada atividade está o resultado alcançado para cada atividade.

FIGURA 16 - FLUXOGRAMA GERAL DA PESQUISA



FONTE: O AUTOR (2023)

O objetivo específico A contribuiu para a coleta de dados multimodais (acústicos, verbais e visuais) que serviram para a validação dos modelos de Aprendizado de Máquina experimentados.

O objetivo específico B contribuiu para identificar e selecionar os modelos de Aprendizado de Máquina que ofereceram os maiores graus de correção preditiva para cada modalidade. Outra contribuição foi a introdução, no Modelo de Sinceridade, dos aspectos particulares da língua portuguesa e cultura brasileira.

O objetivo específico C contribuiu para a combinação harmônica e sinérgica de todos os modelos parciais monomodais dentro de um modelo integrado que procurou valorizar os aspectos particulares de cada modalidade de informação.

3.3 Principais referências adotadas

O Quadro 6 apresenta um resumo dos principais autores e conceitos que fundamentaram os processos encaminhados.

QUADRO 6 - PRINCIPAIS TEMAS E REFERÊNCIAS DE SUPORTE PARA OS ENCAMINHAMENTOS METODOLÓGICOS

Tema	**Referências**
Aprendizado Profundo	Goodfellow; Yoshua; Courville, (2016)
Autoencoders	Goodfellow; Yoshua; Courville, (2016) Bank; Koenigstein; Giryes, (2021)
Detecção de anomalias	Bank; Koenigstein; Giryes (2021) Pratella et al., (2021)
Detecção acústica de mentiras	Mathur; Matarić (2021)
Detecção verbal de mentiras	Vrij (2008) Papantoniou et al. (2021)
Detecção visual de mentiras	Ekman (1992) Mathur; Matarić (2021)
Divergência de Kullback-Leibler	Afgani; Sinanović; Haas (2008)
Mecanismo de Atenção	Vaswani et al. (2017)

FONTE: O AUTOR (2023)

Os conceitos apresentados operaram como fundamentos científicos e metodológicos para a tomada de decisões estratégicas ao longo da pesquisa.

3.4 Revisão de literatura

Uma revisão de literatura foi realizada para compreender o estado da ciência a respeito da detecção de mentiras assistida por Aprendizado de Máquina, visto que tal revisão não fora, até então, encontrada.

3.4.1 Primeira revisão de literatura

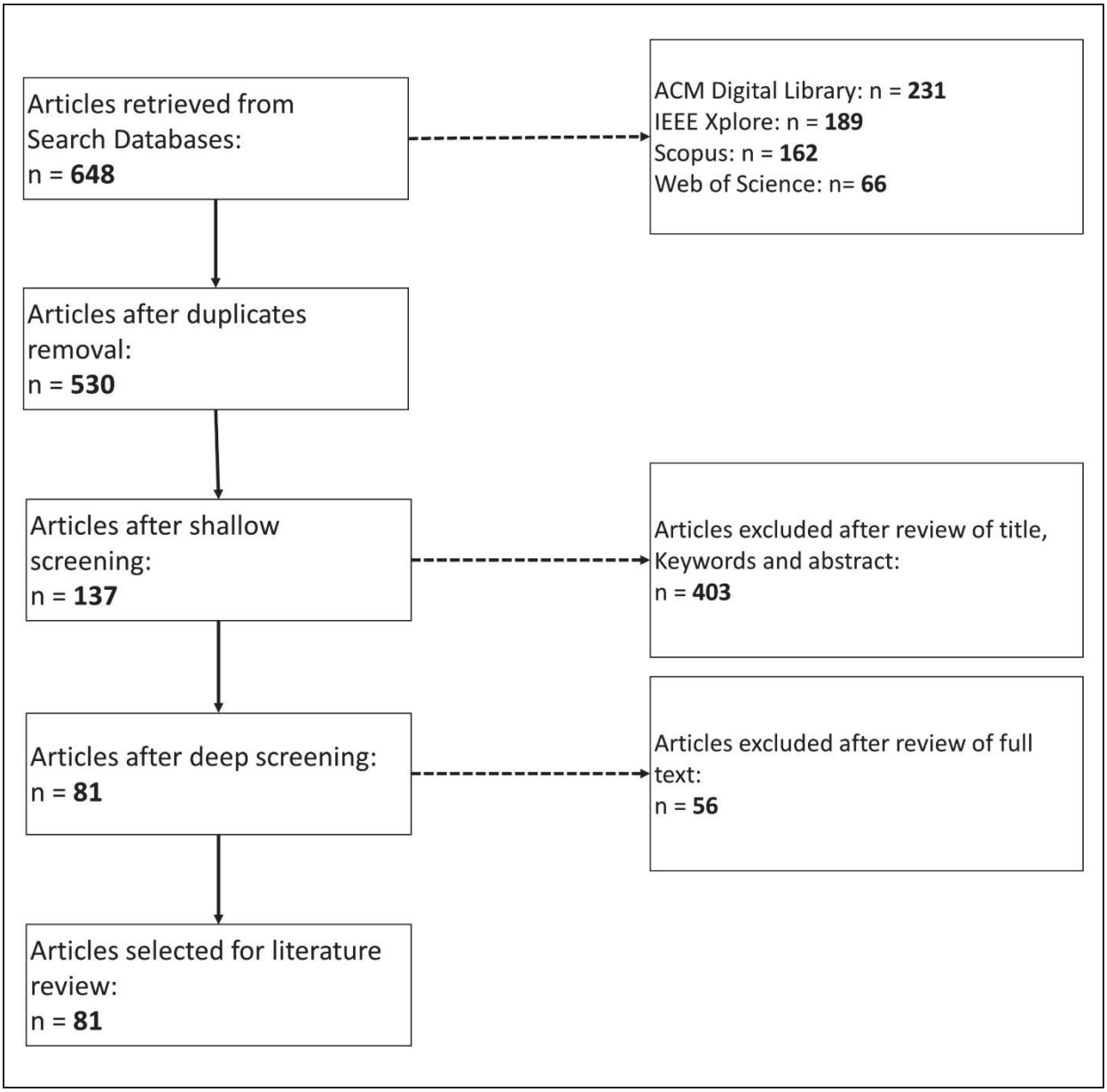
A revisão de literatura teve o propósito de identificar as características, tendências e lacunas de pesquisa no período de 2010 a 2020 e foi constituída pela recuperação sistemática de publicações científicas, lidas na íntegra e resumidas na forma de metadados que vieram a produzir uma análise estatística.

De forma geral, a seleção dos artigos que compuseram a revisão foi dividida em duas etapas. A primeira, chamada de *shallow screening*, consistiu na leitura do título, palavras-chave e resumo de cada um dos artigos recuperados. Somente artigos que parecessem atender aos requisitos de seleção (discutir a respeito de detecção de mentiras por meio de Aprendizado de Máquina)

O corpus resultante foi então submetido ao *deep screening*, ou seja, leitura do texto na íntegra, para validar os critérios de seleção e para proceder com a extração de metadados.

O protocolo PRISMA foi utilizado para nortear a realização e o relato da revisão de literatura. Na Figura 17 é possível visualizar o fluxograma PRISMA que apresenta o processo de seleção das fontes incluídas na revisão.

FIGURA 17 - FLUXOGRAMA PRISMA QUE REPRESENTA AS ETAPAS PARA SELEÇÃO DOS ESTUDOS INCLUÍDOS NA REVISÃO DE LITERATURA



FONTE: CONSTANCIO ET AL. (2023)

A recuperação sistemática da revisão é resultante de consultas conduzidas nas bases de periódicos ACM Digital Library (ACM), IEEE Xplore (IEEE), Scopus e Web of Science (WoS). As consultas foram realizadas em 02/03/2021.

As estratégias de consulta, referências recuperadas (coluna “Rec.”) e referências selecionadas (coluna “Sel.”) estão resumidas no Quadro 7.

QUADRO 7 - CONSULTA E RESULTADOS NA REVISÃO SISTEMÁTICA DE LITERATURA

Base	Estratégia de consulta	Rec.	Sel.
ACM	"deception detection" OR "lie detection"	202	18
IEEE	"deception detection" OR "lie detection"	171	34
Scopus	("deception detection" OR "lie detection") AND ("machine learning" OR "artificial intelligence")	129	10
WoS	("deception detection" OR "lie detection") AND ("machine learning" OR "artificial intelligence")	51	10
	Total	553	72

FONTE: CONSTANCIO ET AL. (2023)

Do *corpus* original de 533 artigos, apenas 72 atenderam aos critérios de seleção estabelecidos:

1. Somente estudos que declaradamente abordaram o problema de detecção de mentiras;
2. Somente estudos que declaradamente identificaram ao menos uma técnica de Aprendizado de Máquina aplicada ao problema;
3. Somente estudos que declaradamente identificaram quais características foram utilizadas como pistas para detecção de mentiras;
4. Somente estudos que apresentaram um nível de desempenho numérico expresso por meio de alguma métrica específica;
5. Somente estudos que usufruíram de dados oriundos de alguma fonte não-invasiva de captação; por não-invasiva, entende-se o uso de dispositivos que absolutamente não toquem nos sujeitos, assim como que tais dispositivos sejam tão portáteis quanto um computador comum.

No entanto, métodos de captação de dados que apresentaram sensores de toque combinados com dispositivos sem contato foram selecionados, com o objetivo de aproveitar os resultados e conhecimentos trazidos por estes útimos.

A revisão em questão forneceu um panorama das técnicas adotadas ao longo de uma década e permitiu identificar tendências, alternativas, dificuldades, descobertas e lacunas pertinentes ao tema.

3.4.2 Atualização da revisão de literatura

Como a revisão de literatura foi conduzida no ano de 2021 e incluiu os anos de 2010 a 2020, publicações posteriores ao período de interesse poderiam conter informações relevantes para a pesquisa. Existiam outras fontes de publicação que não foram considerados na revisão executada. Objetivou-se, então, complementar o corpo de conhecimento com uma atualização da revisão sistemática já realizada.

Ao ser submetido ao periódico PLOS ONE, os revisores solicitaram a inclusão de achados publicados no ano de 2021. Por este motivo, operou-se a atualização da revisão. A atualização foi resultante de consultas realizadas nas bases de periódicos ACM Digital Library (ACM), IEEE Xplore (IEEE), Scopus e Web of Science (WoS) (as mesmas já utilizadas), com os mesmos critérios de busca, mas para o período posterior a 2020. Esta nova consulta foi realizada em 05/05/2022.

O novo *corpus* sofreu o mesmo processo de filtragem (*shallow* e *deep screening*) anteriormente realizado, sob os mesmos critérios de seleção, de onde foram selecionadas outras referências.

Após a atualização, o corpus final passou a ser composto de 684 artigos publicados em revistas científicas e congressos científicos, dos quais 81 atenderam aos critérios de seleção.

3.5 Materiais e métodos

Os resultados atingidos dependeram de experimentos computacionais diversos. Alguns agentes de software foram construídos em linguagem de programação Python versão 3.9, de propósito geral e gratuita. Além de linguagem de fácil uso, uma grande quantidade de pacotes gratuitos estava disponível para a construção de modelos de modelos de Aprendizado de Máquina.

Os experimentos foram compostos por:

1. **conjuntos de dados**: arquivos em formato CSV (*comma-separated values*) constituídos por medições específicas para cada uma das variáveis que descrevem as narrativas multimodais coletadas;
2. **código Python**: código de linguagem que orquestrou tanto os processos pré-experimentais (processos que prepararam os dados para os experimentos propriamente ditos) quanto os experimentais (processos que efetivamente utilizaram os dados para a realização das inferências para detecção de mentiras);
3. **ferramentas de apoio**: diversos agentes de software disponíveis em diversas formas e de diversas origens que foram operados de maneira a produzir resultados intermediários, posteriormente integrados para constituir o modelo final de detecção de mentiras.

Os modelos foram elaborados e experimentados para cada uma das modalidades previstas (acústica, verbal e visual) em separado, para posterior

integração por meio de um critério de fusão. Fusão, no contexto desta pesquisa, é o processo de integrar as diversas detecções parciais para cada modalidade em uma única detecção final, que represente a conclusão multimodal de um caso de teste.

As seções a seguir descrevem e justificam estes componentes.

3.5.1 Hardware

Em todos os experimentos, o hardware utilizado para execução foi um computador munido de:

1. Processador Intel® Core™ i7-10700F, 2.90GHz;
2. 32 GB de memória RAM;
3. Sistema operacional Windows 11 Pro, versão 22H2;
4. GPU Nvidia RTX 3060 com 12GB RAM DDR6.

Tratou-se de equipamento pessoal que teve sua configuração aprimorada especialmente para a execução dos experimentos necessários para a condução da pesquisa.

3.5.2 Vídeos do programa “Acredite em quem quiser”

O programa de TV aberta “Acredite em quem quiser”®, produzido pela Rede Globo de Televisão®[6], era um quadro integrante do “Domingão com Huck”®[7], que ia ao ar aos domingos, durante vários períodos do ano. Vídeos de alta qualidade e resolução deste quadro estavam (durante o período de condução da pesquisa) disponíveis no YouTube®[8] e foram utilizados como fonte primária para a confecção do conjunto de dados para detecção de mentiras.

Nos quadros deste programa, nove convidados tentam convencer interlocutores a respeito de determinado assunto. Enquanto dois dos convidados alternam entre discursos sinceros e não sinceros, um deles necessariamente diz a verdade em resposta a todas as interpelações feitas pelos interlocutores.

Ao final ocorre uma revelação quando, em muitos casos, é possível identificar quais declarações eram sinceras e quais não eram. Por ser de produção nacional, tal quadro forneceu conteúdo em língua portuguesa do Brasil, justificando sua adoção.

Segmentos destes vídeos foram extraídos e representaram a entrada primária de todo o processo. Os critérios para a seleção destes segmentos e os processos

[6] https://redeglobo.globo.com/
[7] https://gshow.globo.com/programas/caldeirao-do-huck/
[8] http://www.youtube.com

específicos para a produção do conjunto de dados a partir dos mesmos estão pormenorizados na seção 3.6.2, página 89.

3.5.3 Jupyter e PyCharm

Tanto o Jupyter[9] quanto o PyCharm Community[10] são ferramentas voltadas ao desenvolvimento de código em linguagem Python. Ambos foram selecionados por causa de sua popularidade e por serem de acesso gratuito.

O Jupyter é um ambiente baseado em interface web (aplicativos que executam dentro de navegadores da Internet) e é voltado à experimentação. Permite a escrita e teste de código Python, com a facilidade de permitir a imediata visualização de resultados, incluindo gráficos e tabelas.

Sua proposta de operação o torna particularmente produtivo para situações experimentais, onde os resultados alimentam a própria atividade investigativa, agilizando o aprendizado, a descoberta e a evolução dos processos em escrutínio.

Já o PyCharm é o que se chama de Ambiente Integrado de Desenvolvimento e oferece um amplo e funcional conjunto de ferramentas próprias para a construção de código Python mais complexo, que foi o caso em alguns momentos.

Dentre os recursos oferecidos pelo PyCharm estão o editor de código multiaparelhado (código colorido, código anotado, indicações de erro e de atenção, dentre outros), *debugger* integrado, facilidades de refatoração (*refactoring*) de código, avaliação de qualidade de código e sugestões de melhoria.

3.5.4 TensorFlow e Keras

O par de ferramentas TensorFlow[11] e Keras[12] operam em conjunto e oferecem recursos prontos para uso de praticantes, profissionais e estudiosos de redes neurais. Estes dois produtos foram utilizados para a construção dos modelos de redes neurais que representaram os Modelos de Sinceridade.

Ambos os produtos são distribuídos gratuitamente, sendo o TensorFlow mantido pela Google enquanto o Keras é mantido por sua própria comunidade de desenvolvedores.

---

[9] https://jupyter.org/
[10] https://www.jetbrains.com/products/compare/?product=pycharm&product=pycharm-ce
[11] https://www.tensorflow.org/
[12] https://keras.io/

O TensorFlow versão 2.9 opera como uma biblioteca de operações matemáticas, tanto genéricas quanto especializadas, para a construção de redes neurais artificiais, tendo como um dos seus grandes diferenciais o aproveitamento de GPUs (*Graphic Processing Units*), o que confere grande aceleração durante os processos de treinamento e inferência das redes neurais onde são empregadas. Já o Keras opera como um facilitador para uso dos recursos do TensorFlow, acelerando grandemente a produtividade do programador.

Na revisão de literatura conduzida (Constancio et al., 2023) foram identificados 31 estudos que exploraram redes neurais em diversas configurações. Embora 17 dos artigos (a maioria) não tenham identificado o *framework* de desenvolvimento utilizado para a construção de seus modelos, o uso combinado do Keras e do TensorFlow foi a preferência dos autores. Sua adoção também se deve à grande quantidade de documentação e exemplos disponíveis na Internet.

3.5.5 OpenFace

O pacote gratuito OpenFace versão 2.2.0[13] foi utilizado para a extração de características faciais dos vídeos selecionados. Este pacote consiste em um modelo de rede neural convolucional especialmente treinado para a extração de diversas características faciais (Baltrusaitis et al., 2018), disponibilizado por meio de um conjunto de aplicativos.

O OpenFace foi escolhido por ser, durante o período de desenvolvimento desta pesquisa, a principal ferramenta gratuita para extração de características faciais, utilizada em diversos estudos científicos, conforme a revisão de literatura (Constancio et al., 2023).

O OpenFace produz diversos arquivos, dentre os quais um arquivo CSV (*Comma-separated values*) consistindo em um conjunto de características para cada um dos quadros de um vídeo em formato MP4 (formato de entrada aceito pelo OpenFace). Cada linha do arquivo corresponde a um quadro do vídeo e cada coluna corresponde a uma característica extraída.

[13] https://github.com/TadasBaltrusaitis/OpenFace

3.5.6 OpenSMILE

O pacote OpenSMILE[14] versão 3.0 foi utilizado para a extração de características acústicas dos vídeos selecionados. É produzido e mantido pela empresa Audeering, que o distribui livremente para uso acadêmico.

Trata-se de uma ferramenta especializada na extração de características de áudio, utilizada largamente para processos de identificação e classificação de emoções vocais e musicais. Uma das grandes vantagens do OpenSMILE 3 é sua integração com a linguagem Python.

O OpenSMILE foi a ferramenta mais adotada pelos estudos de natureza vocal na revisão de literatura (Constancio et al., 2023). Este fator e sua integração com a linguagem Python foram determinantes para sua escolha na condução de experimentos.

3.5.7 MoviePy

O MoviePy[15] versão 1.0.2 é um pacote Python que oferece diversas funcionalidades para a edição de vídeos. O mesmo foi utilizado para recortar os segmentos de vídeo que continham as frases sinceras e não sinceras selecionadas para compor o conjunto de dados para detecção de mentiras.

Dentre as funcionalidades do MoviePy, está a extração de faixas de áudio, que foram submetidas ao OpenSMILE para a extração das características acústicas e também da transcrição das mensagens faladas.

3.5.8 Azure Speech-to-text

O Azure Speech-to-text[16] é um componente do conjunto de ferramentas da Microsoft denominado Azure Cognitive Services. Apesar de ser um produto comercial, o mesmo apresenta uma modalidade de uso gratuita que pôde ser aproveitada para os propósitos desta pesquisa.

Este produto é capaz de receber um arquivo de áudio em língua portuguesa do Brasil e produzir uma transcrição, ou seja, uma versão textual das falas presentes no arquivo. Foi utilizado para extrair o texto das diversas mensagens faladas que foram extraídas de segmentos dos vídeos do programa “Acredite em quem quiser”.

---

[14] https://www.audeering.com/research/opensmile/
[15] https://zulko.github.io/moviepy/
[16]https://azure.microsoft.com/en-us/products/ai-services/speech-to-text

Foi escolhido, frente a seus concorrentes, por gerar transcrições incrementadas com os tempos de cada palavra identificada, recurso que foi aproveitado em um dos processos realizados.

3.5.9 SpaCy

O SpaCy[17] é uma biblioteca Python gratuita para processamento de linguagem natural com suporte para a língua portuguesa do Brasil e foi utilizada para a extração de pistas verbais das transcrições dos segmentos de vídeos selecionados.

O SpaCy oferece recursos como a identificação de entidades, identificação da função sintática de palavras, a extração de grafos de dependência gramatical, lematização (redução do vocábulo ao seu radical), dentre diversos outros, próprios para o campo do Processamento de Linguagem Natural.

3.5.10 SentiWordNet-PT-BR

O SentiWordNet-PT-BR[18] é um léxico voltado à mineração de sentimentos em textos. Segue os mesmos parâmetros de concepção que o SentiWordNet, seu equivalente para o inglês.

O SentiWordNet-PT-BR é distribuído na forma de um arquivo texto estruturado e conta com 77.355 expressões selecionadas a partir do WordNet-PT, dicionário de sinônimos para o português, e para cada uma delas apresenta uma escala de sentimento positivo e negativo.

Os sentimentos anotados para cada palavra constante no SentiWordNet-PT-BR foram utilizados para compor as características verbais, complementando os resultados atingidos com o SpaCy.

3.6 Construção do conjunto de dados MMDDD-PtBr

O objetivo específico A determinou a criação de um conjunto de dados rotulado e multimodal a partir de narrativas em vídeo, pois o conjunto de dados é um fator crítico para qualquer experimento de Aprendizado de Máquina e, até o momento da conclusão desta pesquisa, não existiam conjuntos de dados especificamente dedicados ao problema de detecção de mentiras para o português.

---

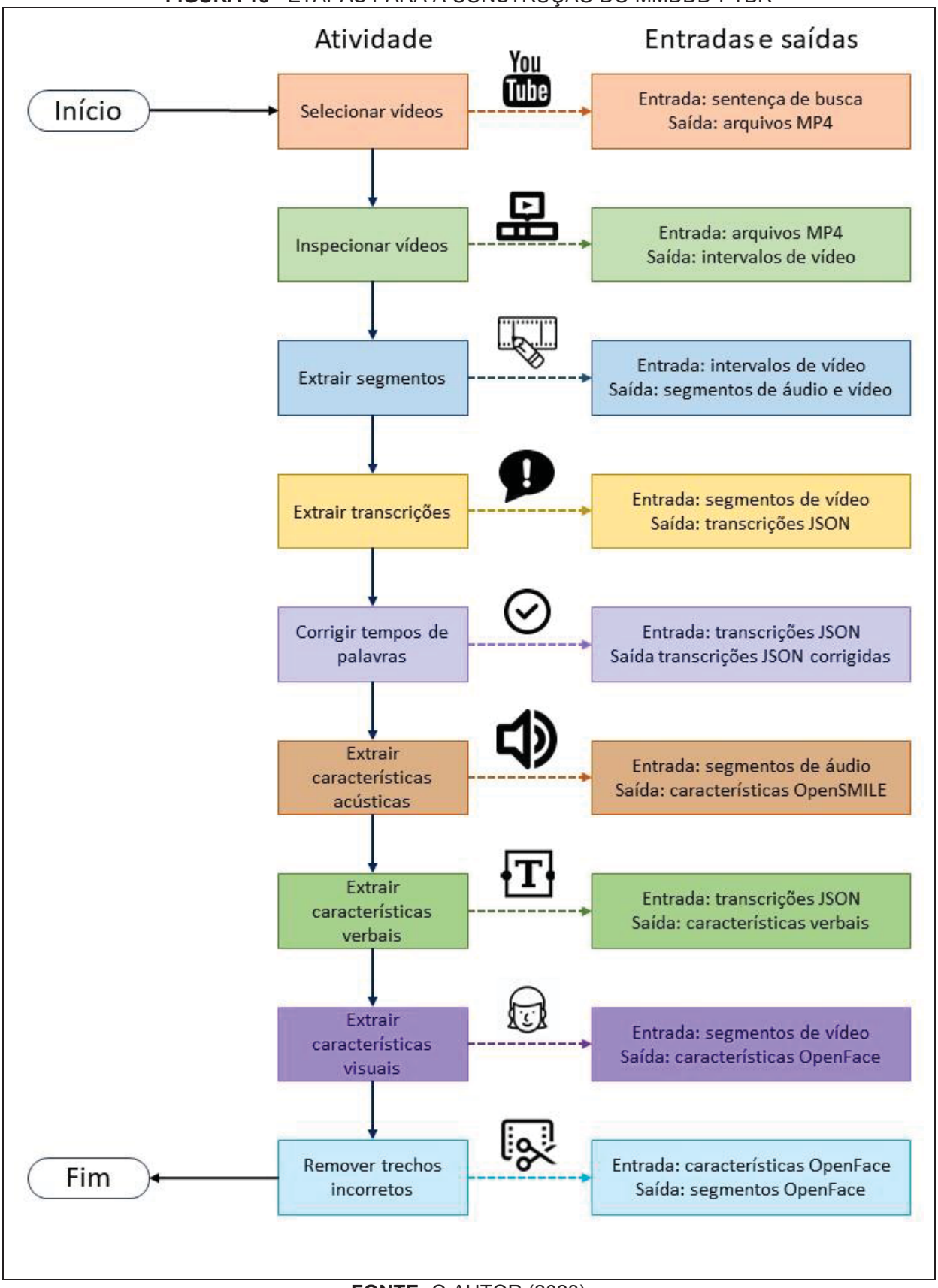
[17] https://spacy.io/
[18] https://github.com/Pedro-Thales/SentiWordNet-PT-BR

O conjunto de dados multimodal para detecção de mentiras em português do Brasil foi denominado de MMDDD-PtBr, sigla para *Multimodal Deception Detection Dataset for Brazilian Portuguese*. Seu objetivo foi o de acumular dados de natureza multimodal (acústica, verbal e visual) extraídos de narrativas em vídeo que pudessem ser rotuladas, ou seja, já classificadas em “sincero” e “não sincero”.

Os dados rotulados permitiram a comparação entre as predições realizadas e a classificação real, de onde métricas de avaliação de qualidade dos modelos foram calculadas.

Na Figura 18 está colocado o processo geral de construção do MMDDD-PtBr. Trata-se de um conjunto de etapas realizadas serialmente, no qual cada uma fez uso do produto da anterior. A primeira etapa recebeu os dados brutos (neste caso, endereços dos vídeos do programa de TV “Acredite em quem quiser”, armazenados no YouTube) e a última produziu o resultado do processo como um todo (neste caso, um conjunto de documentos CSV e JSON, além de trilhas de áudio no formato WAV e segmentos de vídeo no formato MP4).

FIGURA 18 - ETAPAS PARA A CONSTRUÇÃO DO MMDDD-PTBR



FONTE: O AUTOR (2023)

O resultado do processo descrito foi um conjunto de arquivos em formatos diversos que constituem o MMDDD-PtBr. Cada um dos arquivos é responsável por armazenar as características de uma das modalidades, pois diferentes modalidades apresentam diferentes estruturas, diferentes regras de amostragem e diferentes cardinalidades.

Em resumo, o processo:

1. utilizou dois arquivos de vídeo;
2. extraiu 12 indivíduos para construir Modelos de Sinceridade (3 indivíduos do primeiro vídeo e nove do segundo, ao todo sete homens e cinco mulheres);
3. extraiu 37 narrativas sinceras, 24 narrativas não sinceras e 10 narrativas que não puderam ser classificadas (estas não passaram por todas as etapas do processo e não participaram de qualquer experimento);
4. extraiu 61 clipes de vídeo na forma de arquivos MP4;
5. extraiu 61 trilhas de áudio na forma de arquivos WAV;
6. extraiu 61 transcrições de áudio na forma de arquivos JSON;
7. produziu 61 arquivos com características acústicas na forma de arquivos CSV;
8. produziu 61 arquivos com características verbais na forma de arquivos CSV;
9. produziu 61 arquivos com características visuais na forma de arquivos CSV;
10. produziu 61 arquivos com recortes das características visuais na forma de arquivos CSV.

As seções a seguir descrevem detalhadamente cada uma das etapas para a produção do conjunto de dados.

3.6.1 Seleção de vídeos

Os vídeos foram recuperados a partir do YouTube, partindo diretamente dos resultados de uma busca pela frase "acredite em quem quiser domingão huck”. Em 07/08/2023, a busca retornou 26 vídeos que efetivamente correspondiam ao quadro em questão (vários outros também foram retornados), sem qualquer ordem particular.

Os únicos critérios de seleção aplicados foram:

1. O vídeo efetivamente continha o registro do quadro “Acredite em quem quiser”;
2. A resolução mínima do vídeo devia ser de 1280x720 pontos, mas eventuais versões do mesmo vídeo em resolução superior (por exemplo, 1920x1080) foram preferidas.

Uma vez selecionados, os vídeos foram baixados para operação local, pois esta era uma necessidade das ferramentas operadas, que usaram como entrada um arquivo em disco local e não endereços na Internet.

Dois vídeos foram selecionados:

1. Quadro que foi ao ar em 15/07/2021;
2. Quadro que foi ao ar em 23/01/2022.

Para efeitos de codificação, o primeiro vídeo passou a ser chamado de S1 e o segundo de S2. Os vídeos foram salvos localmente no formato MP4, passando a ser submetidos aos processos específicos para extração de dados.

Existiam mais vídeos do programa disponíveis no YouTube e os seus aproveitamentos teriam produzido mais narrativas de mais indivíduos, proporcionando uma percepção mais acurada das virtudes e fragilidades da abordagem baseada em aprendizado autossupervisionado. A limitação foi condicionada pelo tempo necessário para a execução do *pipeline*, notadamente pela existência de etapas manuais, tais como:

1. a correção de diversas das transcrições, total ou parcialmente;
2. a identificação e rotulação de hesitações ao longo dos discursos;
3. a correção dos limites de áudio que separam cada uma das palavras nas narrativas;
4. a identificação de segmentos de vídeo que não mostravam o sujeito durante o seu discurso ou que mostravam sem a necessária qualidade para uso do OpenFace.

Os tempos dessas etapas manuais, quando somados, fizeram com que o processamento de um arquivo de vídeo ocupasse cerca de sete dias para sua conclusão. Aliado a estes custos existia também o tempo de treinamento dos modelos, que é proporcional à quantidade indivíduos (um Modelo de Sinceridade para cada indivíduo) e suas narrativas.

Todos estes tempos foram determinantes para a seleção da quantidade de vídeos em questão.

3.6.2 Segmentação de vídeos

De cada vídeo poderiam ser extraídos até nove sujeitos, pois cada quadro era dividido em três sessões e cada sessão contava com três participantes oferecendo narrativas, sendo que dois poderiam decidir livremente se mentiam ou não, enquanto o remanescente (conhecido como “dono da estória”) necessariamente falava a verdade.

Assim, cada uma das pessoas é numerada de 1 a 9, sendo codificadas como P1, P2, P3... P9. Para efeitos de codificação, cada mensagem extraída foi denominada de forma a identificar o vídeo de origem, a pessoa e um número serial para identificá-lo. Por exemplo, a quarta narrativa do quinto participante do segundo vídeo foi codificada como S2-P5-4.

Os segmentos de vídeo foram identificados por inspeção, ou seja, cada vídeo foi assistido em um *player* de MP4 e os intervalos que delimitavam cada narrativa foram anotados com tempo de início e fim. Esses intervalos posteriormente serviram para realizar recortes do vídeo original, produzindo assim diversos outros arquivos de vídeo com o nome codificado, um para cada narrativa.

Nem todos os participantes puderam efetivamente contribuir para a construção do conjunto de dados. Por exemplo, em uma das sessões os participantes vieram em par, com a alternância de narrativas entre um e outro. Em outros casos, os participantes se manifestavam no mesmo momento em que o público aplaudia ou produzia ruídos, tornando o áudio inutilizável. Em outros casos ainda, o momento final de revelação, efetivamente, identificava quem estava sendo sincero, mas não deixava claro dentre os demais quais narrativas eram ou não sinceras, impedindo sua classificação.

Diante dessas situações, apenas 12 dos 18 potenciais participantes realmente geraram narrativas que puderam ser aproveitadas para compor o conjunto de dados. Especificamente, no caso do vídeo S1, as narrativas dos participantes de 1 a 6 (S1-P1 a S1-P6) não puderam ser utilizadas, como mostrado no Quadro 8.

QUADRO 8 - RESUMO DOS PARTICIPANTES SELECIONADOS A CLASSIFICAÇÃO DE SUAS NARRATIVAS COLETADAS

Sujeito	Narrativas sinceras	Narrativas não sinceras	Total de narrativas
S1-P7	8	5	13
S1-P8	7	0	7
S1-P9	4	5	9
S2-P1	3	0	3

S2-P2	1	3	4
S2-P3	1	4	5
S2-P4	2	0	2
S2-P5	2	1	3
S2-P6	1	2	3
S2-P7	3	3	6
S2-P8	1	1	2
S2-P9	4	0	4
Total	**37**	**24**	**61**

FONTE: O AUTOR (2023)

A partir do momento em que os segmentos de vídeo foram recortados e tiveram seus nomes codificados, esses códigos passaram a identificar todas as outras referências aos mesmos. Os vídeos originais não mais voltaram a ser utilizados.

3.6.3 Clipes, trilhas de áudio e transcrições

Os intervalos de tempo de início e fim de cada segmento permitiram recortar os vídeos originais em clipes contendo unicamente a narrativa que interessou para compor o conjunto de dados. Um código Python utilizando um pacote chamado MoviePy foi utilizado para realizar dois processos importantes: a) recortar efetivamente o arquivo de vídeo para gerar o arquivo de segmento; b) extrair a trilha de áudio correspondente ao segmento.

Por exemplo, o segmento S2-P5-3 (terceira narrativa do quinto participante do segundo vídeo) estava delimitado entre os segundos 1.830,4 e 1.838,3, portanto, um segmento com aproximadamente oito segundos de duração. Com o uso do MoviePy foi possível extrair os arquivos S2-P5-3.MP4 e S2-P5-3-WAV. O primeiro arquivo constitui o recorte de vídeo com imagem e áudio enquanto o segundo apenas a trilha de áudio.

A trilha de áudio foi então enviada para o Azure Speech-to-text (por meio de outro código Python) que produziu o arquivo S2-P5-3.JSON. O serviço do Azure pôde ser configurado para não apenas extrair a transcrição do áudio enviado, mas também separar cada palavra identificada e indicar em que momentos iniciavam e terminavam dentro do tempo da trilha. No Quadro 9 é possível ver um trecho do arquivo JSON gerado a partir do arquivo S2-P5-3.WAV.

QUADRO 9 - SEGMENTO DO ARQUIVO DE TRANSCRIÇÃO GERADO PELO AZURE SPEECH-TO-TEXT


```
{
    "transcript": "eu sou atriz trabalho com eventos eu amo salto uso mesmo já sofri
muito com ele mas eu não durmo eu não tomo banho nada de",
    "offset": 0.09,
    "duration": 7.81,
    "words": [
        {
            "word": "eu",
            "start": 0.09,
            "end": 0.2,
        },
        {
            "word": "sou",
            "start": 0.21,
            "end": 0.3,
...
```


FONTE: O AUTOR (2023)

O arquivo JSON é um documento em texto plano cujo conteúdo é estruturado em pares nome-valor. No caso do exemplo, o nome “transcript” identifica o campo cujo valor é o texto compreendido pelo serviço do Azure, mostrado em azul.

Em alguns casos, o resultado foi perfeito ou muito próximo do perfeito, mas em outros o resultado foi bastante decepcionante, estando muito longe do que foi realmente dito. Esses últimos motivaram uma revisão manual de todas as transcrições, com reescrita do texto. Nessa revisão, vírgulas e pontuação foram incluídas para produzir uma versão da transcrição em melhores condições para o processamento verbal posterior. No Quadro 10 está mostrada a reescrita do texto transcrito, no campo “text”.

QUADRO 10 – SEGMENTO DO ARQUIVO DE TRANSCRIÇÃO APÓS A CORREÇÃO MANUAL DO TEXTO


```
{
   "transcrip": "eu sou atriz trabalho com eventos eu amo salto uso mesmo já sofri
muito com ele mas eu não durmo eu não tomo banho nada d",
   "text": "Eu sou atriz, trabalho com eventos. Eu amo salto, uso mesmo, já sofri
muito com ele, mas eu não durmo, não tomo banho, nada disso.",
   "offset": 0.09,
   "duration": 7.81,
   "words": [
        {
...
```


FONTE: O AUTOR (2023)

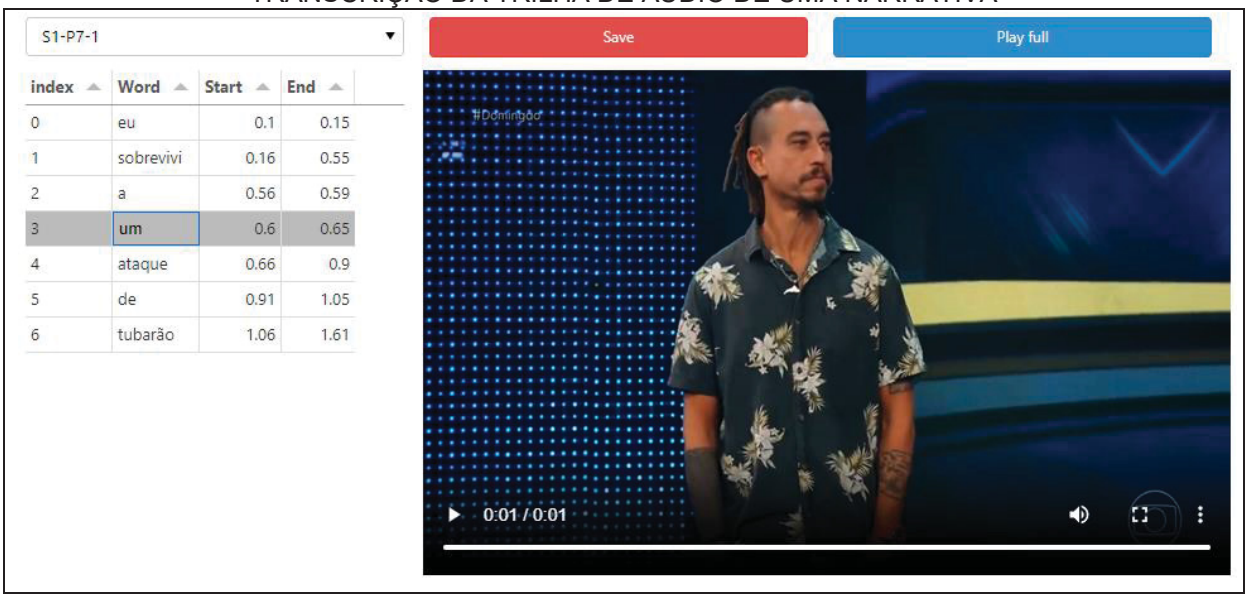
O campo “text” apresenta o texto já revisado (cor vermelha). Foi esse texto, e não a transcrição retornada pelo Azure Speech-to-text, que foi utilizado para a extração de características verbais.

3.6.4 Correção nos tempos das palavras

O Azure Speech-to-text pode ser configurado para identificar os intervalos de tempo em que cada uma das palavras reconhecidas ocorreu, mas assim como nem sempre o serviço reconheceu corretamente as palavras, nem sempre corretamente identificou seus momentos de início e fim, até porque na linguagem falada tais fronteiras são muito tênues.

Visando aumentar ao máximo a precisão dos modelos que viriam a ser treinados, foi construída uma interface específica para editar os arquivos JSON de transcrição para corrigir os intervalos de cada palavra, partindo do trabalho já realizado pelo Azure. Na Figura 19 é possível ver uma imagem desta interface.

FIGURA 19 - INTERFACE DO EDITOR PARA CORREÇÃO DE INTERVALOS DE PALAVRAS DA TRANSCRIÇÃO DA TRILHA DE ÁUDIO DE UMA NARRATIVA



FONTE: O AUTOR (2023)

Da esquerda para a direita e de cima para baixo, a interface foi formada pelos seguintes controles:

1. uma lista retrátil que permitiu selecionar sobre qual das narrativas se deseja operar, no caso da imagem, S1-P7-1;
2. um botão para salvar o trabalho, que atualizava o arquivo da transcrição, no caso da imagem, S1-P7-1.JSON;
3. um botão para tocar todo o vídeo, desde o início até o fim;
4. uma lista com as palavras da transcrição e seus respectivos tempos de início (campo “Start”) e fim (campo “End”); estes são os campos que foram usados para corrigir os tempos de cada palavra;

5. um *player* do segmento de vídeo correspondente à palavra selecionada, no caso da imagem, palavra “um”, de *index*=3.

O processo consistiu em carregar o arquivo de transcrição correspondente à narrativa sobre a qual se desejava operar. Para escutar cada palavra bastava clicar sobre a mesma na lista de palavras. O vídeo do segmento correspondente era tocado, permitindo que o operador avaliasse se os limites da palavra estavam bem estabelecidos. Em caso de necessidade, os campos “Start” e “End” puderam ser editados para redefinir os tempos conforme necessário. A precisão era de centésimos de segundo.

Após a edição de cada valor, o segmento de vídeo era novamente tocado para confirmação. O processo foi concluído quando o operador ficou satisfeito com os tempos de início e fim de cada palavra, quando então clicava no botão “Save” para que o arquivo de transcrição fosse atualizado.

Em maior ou menor grau, todos os 61 arquivos de transcrição foram submetidos ao processo de correção acima descrito, constituindo a etapa mais demorada de todo o *pipeline* para produção do conjunto de dados.

Com os segmentos de vídeo e áudio extraídos, assim como com a transcrição já corrigida, tornou-se possível utilizar as ferramentas selecionadas para extrair as características acústicas, verbais e visuais de cada participante para compor o conjunto de dados.

3.6.5 Extração de características acústicas

Para a extração de características acústicas foi utilizado o OpenSMILE, fazendo uso de sua interface na forma de um pacote Python. O OpenSMILE pôde extrair diversos conjuntos de características, tendo sido escolhido o mais recente deles, o ComParE 2016. Esse conjunto de características podia ser exportado em três diferentes formas: a) Low-level Descriptors (LLD); b) Functionals; c) LLD deltas.

Nas opções Functionals e LLD deltas o arquivo de áudio era fornecido e 65 ou 6.373 características (respectivamente) eram fornecidas na forma de uma tabela com uma única linha. Tratava-se de uma análise de todo o áudio, resumida na forma de todas aquelas características.

Na opção LLD o OpenSMILE segmentou o arquivo de áudio em quadros de 10 milisegundos e extraiu 65 características acústicas para cada uma dessas amostras, produzindo uma tabela de 68 colunas (as três primeiras identificavam o

arquivo e o intervalo de tempo) por tantas linhas quantas necessárias para descrever todos os quadros.

Dado que o objetivo foi construir Autoencoders a partir desses dados, tornou-se forçoso a adoção da opção LLD, até porque desta forma as características extraídas apresentavam as oscilações sonoras ao longo da narrativa que, se esperava, marcariam os pontos de variação na expressão vocal, evidenciando as anomalias que caracterizariam a não sinceridade.

Para exemplificar, quando o arquivo de áudio S1-P7-13.WAV de 17 segundos de duração foi submetido ao OpenSMILE, o arquivo S1-P7-13-opensmile.CSV com 1.710 linhas (a primeira sendo cabeçalho) foi gerado, constituindo as características acústicas daquela narrativa. Este processo foi realizado para todas as 61 narrativas.

As características presentes no arquivo CSV de saída do OpenSMILE correspondem às listadas no Quadro 11.

QUADRO 11 - GRUPOS DE CARACTERÍSTICAS ACÚSTICAS EXPORTADAS PELO OPENSMILE E INCLUÍDAS NO COMPONENTE ACÚSTICO DO MMDDD-PTBR

	Grupo
4 características relacionadas a energia	
Soma do espectro auditivo (intensidade)	Prosódico
Soma do espectro auditivo filtrado no estilo RASTA	Prosódico
Energia RMS; taxa de cruzamento zero	Prosódico
55 características espectrais	
Espectro auditivo no estilo RASTA, bandas 1-26 (0-8 kHz)	Espectral
MFCC 1–14	Cepstral
Energia espectral 250-650 Hz; 1 k-4 kHz	Espectral
Ponto de desativação espectral 0,25; 0,50; 0,75; 0,90	Espectral
Fluxo espectral; centroide; entropia; inclinação	Espectral
Nitidez psicoacústica; harmonicidade	Espectral
Variância espectral; assimetria; curtose	Espectral
6 características relativas a voz	
F0 (SHS e suavização de viterbi)	Prosódica
Probabilidade de som de voz	Qualidade
Logaritmo de HNR; Jitter (local, delta); Shimmer (local)	Qualidade

FONTE: ADAPTADO DE WENINGER ET AL. (2013)

Essas características são a representação numérica dos diversos parâmetros acústicos de uma narrativa e constituem o componente acústico do MMDDD-PtBr.

3.6.6 Extração de características verbais

Para a extração de características verbais foram utilizados o SpaCy, fazendo uso de sua interface na forma de um pacote Python, e o SentiWordNet-PT-BR. A característica “hesitação” foi manualmente extraída, visto que a mesma não foi identificada na transcrição devolvida pelo Azure speech-to-text.

O SpaCy opera recebendo um texto e produzindo diversas estruturas que descrevem inúmeros atributos linguísticos para cada palavra. Já o SentiWordNet-PT-BR é um documento texto estruturado, consultado para retornar as intensidades dos sentimentos positivo e negativo eventualmente vinculados a uma dada palavra.

A primeira etapa foi enriquecer o arquivo JSON de cada narrativa, que originalmente continha a transcrição obtida do Azure a partir da trilha de áudio, a correção manual desta transcrição e os intervalos (também manualmente corrigidos) dos tempos de início e fim de cada palavra proferida. O enriquecimento consiste em adicionar novos atributos a cada uma das palavras, a partir dos recursos do SpaCy.

Após o texto corrigido da transcrição ser fornecido ao SpaCy, os seguintes atributos foram obtidos e utilizados para enriquecer cada palavra:

1. **função sintática (*part-of-speech tag*)**: indica se a palavra opera como substantivo, adjetivo, advérbio, verbo, nome próprio ou pronome;
2. **entidade**: indica se a palavra é ou não uma entidade nomeada (pessoa, empresa, local, ponto de interesse, dentre outros); o tipo específico da entidade não é levado em consideração;
3. **lema**: a versão não flexionada da palavra, quando for o caso, por exemplo “trabalhar” é o lema da palavra “trabalho”;
4. **morfologia**: conjunto de atributos que descrevem aspectos de flexão e adaptação da palavra no seu contexto de uso, como número, gênero, tempo e pessoa verbal.

Adicionalmente, cada palavra foi manualmente marcada como “Hesitação”. Foram consideradas hesitações a repetição de parte das palavras (gagueira) ou a interrupção de uma palavra para a introdução de outra. Também foi classificado como hesitação, expressões de preenchimento como “eh”, “ah”, “hm”.

Posteriormente, a forma lematizada da palavra foi utilizada para procurar os escores de sentimentos no SentiWordNet-PT-BR. Quando localizada, os escores correspondentes presentes foram utilizados (algumas palavras apresentaram tanto o escore positivo quanto o negativo), do contrário foram assumidos como zero.

Durante o processo de identificação do sentimento das palavras no SentiWordNet-PT-BR, foi percebido que algumas palavras ocorriam mais de uma vez e em cada ocorrência podiam apresentar diferentes escores de sentimento. Nestes casos, o escore utilizado foi a média de todas as ocorrências localizadas.

No Quadro 12 apresenta-se um exemplo do arquivo S2-P5-2.JSON.

QUADRO 12 - EXTRATO DO ARQUIVO S2-P5-2.JSON APÓS O ENRIQUECIMENTO DE CADA PALAVRA A PARTIR DO SPACY, DO SENTIWORDNET-PT-BR E MANUALMENTE


```
...
    "text": "Eh... eu sou atriz e trabalho com eventos. Muitas horas em cima de um salto e um belo dia eu cheguei de um evento muito exausta e eu teria que calçar o salto no outro dia cedo e meu pé estava inchado. Foi aí que eu tomei banho com um salto e eu nunca mais tirei.",
    "offset": 0.39,
    "duration": 19.12,
    "words": [
        {
            "word": "eh",
            "start": 0.39,
            "end": 0.73,
            "pos_tag": {
                "tag": "PROPN",
                "explain": "proper noun"
            },
            "ent": {
                "type": "",
                "explain": null
            },
            "lemma": "Eh",
            "hesitation": true,
            "morphology": {
                "Gender": "Masc",
                "Number": "Sing"
            },
            "neg_score": 0.0,
            "pos_score": 0.0
        },
        {
            "word": "eu",
            "start": 0.74,
            "end": 0.83,
            "pos_tag": {
                "tag": "PRON",
                "explain": "pronoun"
            },
            "ent": {
                "type": "",
                "explain": null
            },
            "lemma": "eu",
            "hesitation": false,
            "morphology": {
                "Case": "Nom",
                "Gender": "Masc",
                "Number": "Sing",
                "Person": "1",
                "PronType": "Prs"
            },
            "neg_score": 0.0,
            "pos_score": 0.0
    },
...
```


FONTE: O AUTOR (2023)

Cada arquivo de transcrição foi enriquecido com atributos verbais adicionais e exportado para o formato CSV, mais apropriado para ser aproveitado em experimentos. Para o exemplo do arquivo S2-P5-2.JSON foi gerado o arquivo S2-P5-2-verbal.CSV, contendo 14 colunas, sendo a primeira o texto da palavra e as outras 13 as seguintes características:

1. **duration**: duração da palavra em milissegundos, escolhida para avaliar o efeito da não sinceridade na velocidade com que as palavras são pronunciadas, que pode sugerir alterações na carga cognitiva necessária para forjar uma narrativa (Suchotzki; Gamer, 2019);
2. **is_ent**: com o valor 1 indica que a palavra identifica uma entidade, escolhida porque a frequência do uso de entidades nomeadas pode operar como preditor para narrativas não sinceras (Kleinberg et al., 2018);
3. **is_num**: com o valor 1 indica que a palavra descreve um numeral, escolhido porque espera-se que o uso de numerais identifique uma linguagem mais específica; linguagem vaga e imprecisa pode ser um indicativo de não sinceridade (DePaulo et al., 2003; Porter; Brinke, 2010b; Vrij, 2008);
4. **is_pronoun**: com o valor 1 indica que a palavra opera como pronome, escolhido porque a mudança na frequência de certos tipos de palavras, identificadas por sua função sintática, pode operar como um preditor para a deteção de mentiras (DePaulo et al., 2003; Papantoniou et al., 2021), sendo que pelo menos 11 dos 81 artigos avaliados na revisão de literatura fizeram uso deste tipo de característica;
5. **is_verb**: com o valor 1 indica que a palavra opera como verbo, escolhido pelos mesmos motivos de “is_pronoun”;
6. **is_noun**: com o valor 1 indica que a palavra opera como substantivo, escolhido pelos mesmos motivos de “is_pronoun”;
7. **is_adjective**: com o valor 1 indica que a palavra opera como adjetivo, escolhido pelos mesmos motivos de “is_pronoun” e porque identifica uma linguagem menos vaga (mais específica) (Papantoniou et al., 2021);
8. **is_adverb**: com o valor 1 indica que a palavra opera como advérbio, escolhido pelos menos motivos que “is_adjective”;

9. **hesitation**: com o valor 1 indica que a palavra identifica uma hesitação, escolhido porque pode operar como um preditor de não sinceridade (Vrij, 2008);
10. **pos_score**: escore de sentimento positivo, escolhido porque o sentimento e a emoção são fatores cruciais afetados pelo ato de mentir (Ekman, 1992; Vrij, 2008), além de já ter sido relatado como preditor para não sinceridade (Briscoe; Appling; Hayes, 2014; Jaiswal; Tabibu; Bajpai, 2016; Kamboj et al., 2021; Papantoniou et al., 2021);
11. **neg_score**: escore de sentimento negativo, escolhido pelos mesmos motivos que “pos_score”;
12. **is_1_person**: com o valor 1 indica que o eventual pronome é de primeira pessoa, escolhido por que a alternância no uso de pronomes em primeira e terceira pessoas pode operar como preditor para não sinceridade (Papantoniou et al., 2021);
13. **is_3_person**: com o valor 1 indica que o eventual pronome é de terceira pessoa, escolhido pelo mesmo motivo que “is_1_person”.

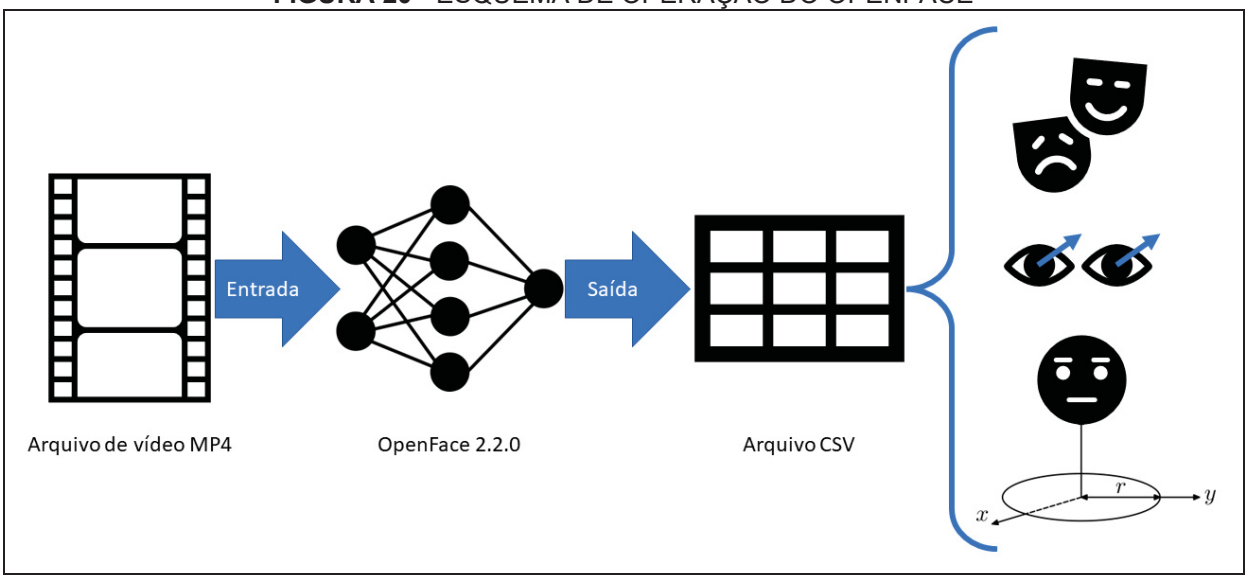
Um fato recorrente em pesquisas de natureza verbal (Constancio et al., 2023) é o uso do LIWC (“*Linguistic Inquiry and Word Count*”), um léxico que suporta múltiplas línguas, inclusive o português. O LIWC classifica as palavras em categorias psicolinguísticas, que supostamente capturam processos psicocognitivos vinculados à expressão verbal. Na revisão de literatura, as categorias oferecidas pelo LIWC figuraram como a característica mais frequentemente experimentada (20 dos 81 estudos). Nesta pesquisa, no entanto, o LIWC não foi adotado por não ser um produto de acesso gratuito.

3.6.7 Extração de características visuais

Para a extração de características visuais foi utilizado o OpenFace. Quando um arquivo de vídeo é fornecido ao OpenFace, diversos arquivos são gerados contendo variadas informações extraídas.

Os aplicativos do OpenFace processam arquivos de imagens ou vídeos e extraem características faciais dos indivíduos presentes, tais como os componentes faciais (*facial landmarks*), expressões faciais baseadas em *Facial Action Units*, a orientação da cabeça (*head pose*) e a estimativa de ângulo de mirada (*gaze estimation*). Na Figura 20 está apresentado um esquema de operação do OpenFace.

FIGURA 20 - ESQUEMA DE OPERAÇÃO DO OPENFACE



FONTE: O AUTOR (2023)

Dentre os arquivos gerados, foi utilizado um arquivo CSV composto por 714 colunas e tantas linhas quantos quadros existirem no vídeo. Por exemplo, o arquivo S2-P4-2-openface.CSV apresentou 436 linhas, sendo a primeira delas o cabeçalho, o que equivale dizer que o arquivo S2-P4-2.MP4, de aproximadamente 14 segundos, é composto por 435 quadros.

A exemplo do arquivo exportado pelo OpenSMILE, o arquivo de características visuais foi mantido na mesma forma em que foi gerado pelo OpenFace.

3.6.8 Remoção de trechos incorretos de vídeos

Como os segmentos de vídeos correspondentes às narrativas selecionadas foram oriundos de um quadro de um programa de TV, em muitos momentos a câmera optou por focalizar o participante sob diferentes ângulos, assim como por mostrar o apresentador ou os demais convidados e até mesmo a plateia. A Figura 21 mostra um quadro do vídeo onde diversas pessoas estão presentes. Para tal quadro o OpenFace, por seus próprios motivos, optou por identificar as características faciais do apresentador e não dos sujeitos.

FIGURA 21 - QUADRO DE UM CLIPE NO QUAL O OPENFACE IDENTIFICOU AS CARACTERÍSTICAS FACIAIS DO APRESENTADOR (CIRCULADO EM VERMELHO) E NÃO DOS PARTICIPANTES

FONTE: O AUTOR (2023)

Por este motivo, existem partes dentro dos segmentos de vídeo que não correspondem ao sujeito que forneceu as trilhas de áudio. Em vários momentos, a imagem não correspondeu à voz registrada.

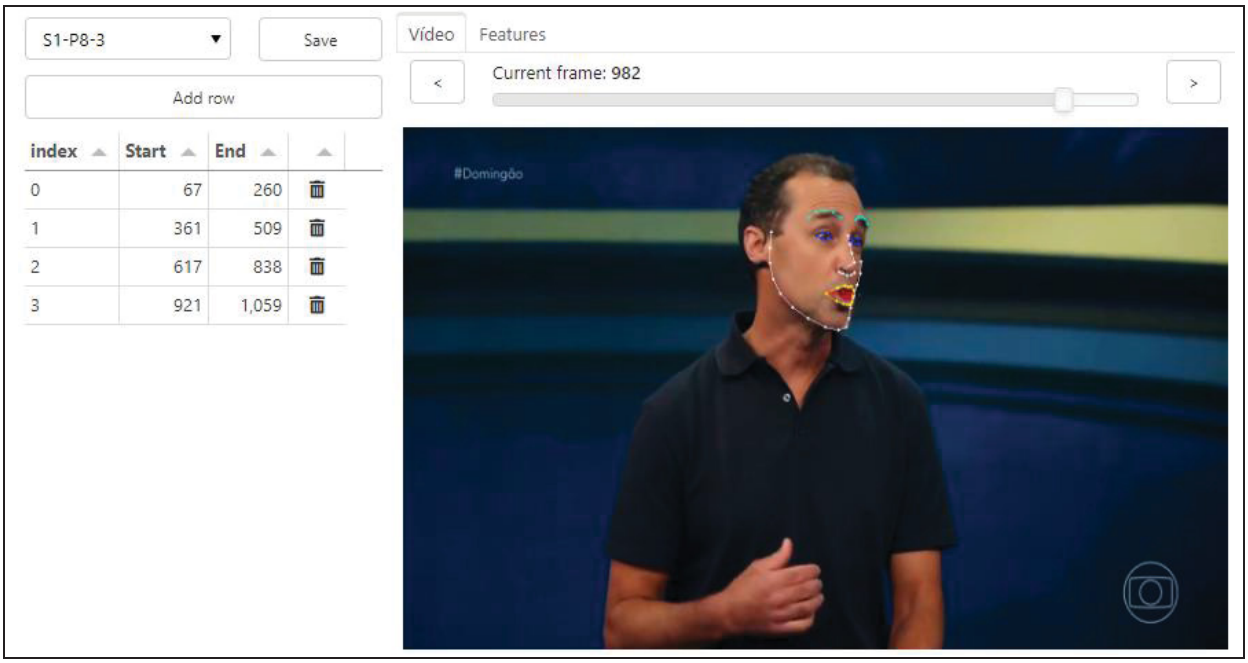
Se as características visuais fossem extraídas indistintamente, sem levar em conta o rosto que está sendo mostrado no vídeo, estas capturariam expressões faciais não correspondentes ao indivíduo que forneceu as características vocais e verbais, produzindo assim um ruído com efeitos danosos para a precisão do processo de detecção.

Fez-se necessário, portanto, realizar recortes dentro dos segmentos de vídeo para que fossem levadas em conta apenas as características faciais do participante cujas narrativas se desejou utilizar nos experimentos.

A estratégia não foi editar os segmentos de vídeo propriamente, mas selecionar das características já geradas pelo OpenFace, os segmentos que correspondessem à exposição do rosto do participante desejado.

Para a realização da tarefa de recorte de características visuais, uma interface foi construída para permitir a seleção criteriosa dos quadros que poderiam participar dos experimentos. Na Figura 22 está contida uma imagem desta interface.

FIGURA 22 - INTERFACE DO EDITOR DE SEGMENTOS DE CARACTERÍSTICAS DO OPENFACE, MOSTRANDO UM DOS PARTICIPANTES E OS PONTOS FACIAIS QUE VÃO GERAR AS CARACTERÍSTICAS VISUAIS



FONTE: O AUTOR (2023)

Da esquerda para a direita e de cima para baixo, a interface foi formada pelos seguintes controles:

1. uma lista retrátil que permitiu selecionar sobre qual das narrativas se deseja operar, no caso da imagem, S1-P8-3;
2. um botão para salvar o trabalho, gerar ou atualizar o arquivo de características visuais recortadas, no caso da imagem, S1-P8-3-openface-cuts.CSV;
3. um *slider* que permitiu selecionar aleatoriamente os quadros do vídeo, além de navegar quadro-a-quadro para frente e para trás;
4. um botão para incluir um novo segmento a ser mantido;
5. uma lista com os intervalos de quadros, definidos por início (campo “Start”) e fim (campo “End”); estes foram os campos usados para identificar porções que foram utilizadas nos experimentos;
6. uma imagem do quadro em observação correspondente ao selecionado no *slider* superior, no caso da imagem, quadro número 982.

O processo consistiu em carregar o arquivo de características visuais correspondente à narrativa sobre o qual se desejou operar e o arquivo de vídeo correspondente. Embora a operação seja feita sobre o arquivo de características, os quadros do vídeo eram mostrados para referência do operador.

Para navegar ao longo dos quadros do vídeo bastava utilizar o *slider*. O quadro selecionado era mostrado com os pontos faciais identificados pelo OpenFace, para facilitar a decisão de manter ou não o quadro para uso posterior. Houve casos em que o OpenFace não foi capaz de extrair as características de um sujeito mostrado no quadro, o que implicou na inexistência daqueles pontos, total ou parcialmente.

Sempre que foi necessário estabelecer um intervalo de quadro a ser mantido, usou-se o botão “Add row”. A lista de segmentos era composta pelos campos “Start” e “End” que puderam ser editados para redefinir os intervalos.

O processo foi concluído quando o operador ficou satisfeito com os intervalos de características a serem mantidas, quando então clicou no botão “Save” para que o arquivo de recortes fosse criado ou atualizado.

Em maior ou menor grau, todos os 61 arquivos de características visuais foram submetidos ao processo de correção acima descrito. Em dois casos (S1-P8-5-openface-cuts.CSV e S2-P9-2-openface-cuts.CSV) todo o arquivo foi excluído. Nesses dois casos, o modelo multimodal não pôde contar com pistas visuais.

3.7 Tipos de Autoencoders

Além do Autoencoder Vanilla (Autoencoder original), foi decidido experimentar algumas variantes na intenção de explorar suas capacidades frente ao problema de capturar caracteres de sinceridade. Especificamente, foram experimentados Autoencoders atencionais, graças aos resultados excepcionais que o mecanismo de Atenção proporcionou no campo de Processamento de Linguagem Natural (Vaswani et al., 2017). As seguintes arquiteturas de Autoencoders foram selecionadas para experimentação:

1. **Autoencoder Vanilla (AE)**: é o modelo de Autoencoder primeiramente proposto, sendo o modelo mais simples e frequentemente aplicado ao problema de detecção de anomalias; tem a vantagem de ter um treinamento mais rápido, mas é incapaz de aprender relações entre diferentes indivíduos de um conjunto de treinamento, o que se mostrou necessário no cenário onde as características variam ao longo da narrativa, como em uma série temporal;
2. **Autoencoder Atencional *Single-head* (AAE)**: é um modelo de Autoencoder acrescido do mecanismo de atenção *single-head*; tem a vantagem de aprender a importância relativa de cada item da entrada do

conjunto de treinamento, mas a desvantagem de requerer mais processamento e memória, elevando o tempo de treinamento;

3. **Autoencoder Atencional *Multi-head* (MAAE)**: é um modelo de Autoencoder acrescido do mecanismo de atenção *multi-head*; tem a vantagem de aprender relações de longa distância que eventualmente existem entre os itens da entrada em um lote de amostras do conjunto de treinamento, mas a desvantagem de envolver cálculos ainda mais caros, elevando ainda mais o tempo de treinamento.

O objetivo de incrementar os Autoencoders com os mecanismos de Atenção foi capturar eventuais relações de longa distância existentes entre diferentes momentos das expressões multimodais, ampliando a percepção da relação entre as características dentro de um intervalo de tempo. Ou seja, enquanto um Autoencoder Vanilla é capaz de identificar as relações entre as características de uma única linha do conjunto de dados, o Autoencoder atencional incrementa esta percepção para identificar também relações entre diferentes indivíduos que compõem um lote de amostras.

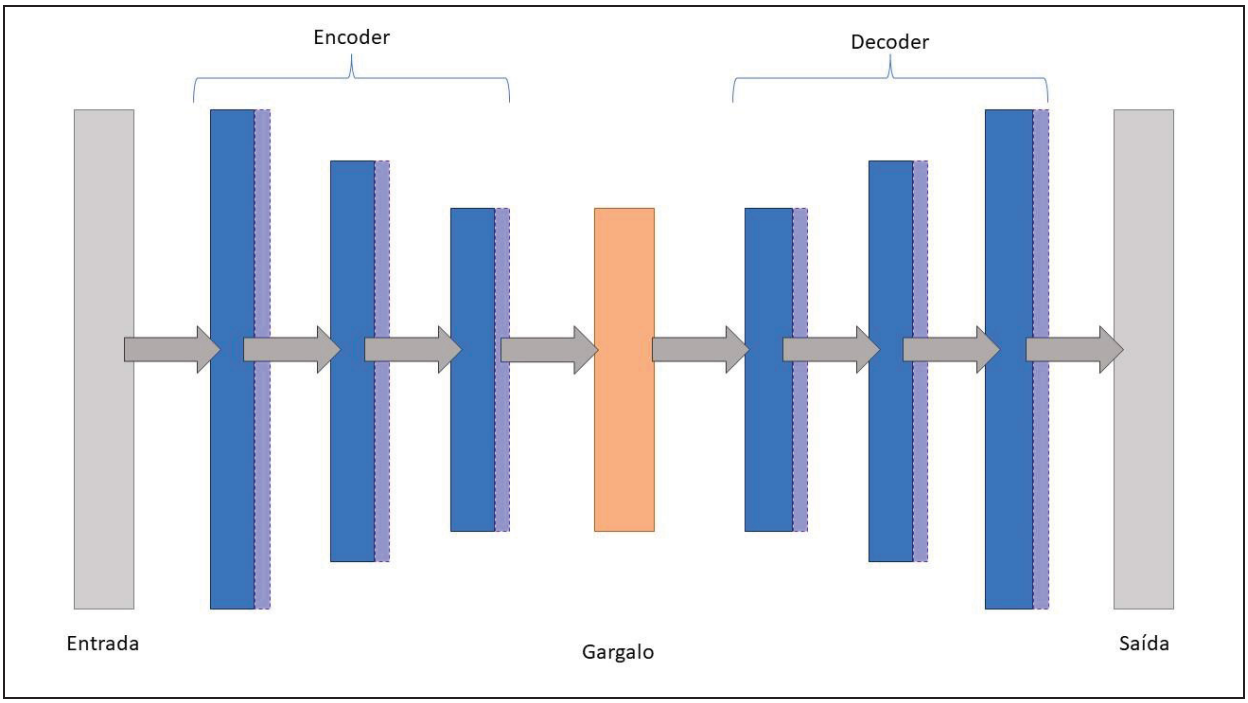
3.8 Arquiteturas de modelos

Os modelos de redes neurais artificiais construídos, operados e avaliados nesta pesquisa foram implementados com Keras/Tensorflow em linguagem de programação Python.

Na Figura 23 é possível vislumbrar um esquema genérico do modelo AE, assim como do AAE e MAAE na Figura 24. Em diversas referências é comum encontrar a camada de gargalo incluída como parte do encoder, neste documento o gargalo será mostrado como um componente isolado dos demais.

Embora a figura apresente três camadas para encoder (e decoder), houve também casos com uma e duas camadas.

FIGURA 23 - MODELO GENÉRICO DE UM AUTOENCODER VANILLA



FONTE: O AUTOR (2023)

As camadas que correspondem ao encoder e decoder (em azul) foram construídas com o que o Keras denomina de camada densa (*dense layer*) e apresentam uma faixa fina (em lilás) à direita para representar a eventual aplicação de *dropout* (no caso do Keras, uma camada de *dropout*, ou *dropout layer*). Quando o recurso esteve ativo, afetou todas as camadas tanto do encoder quanto do decoder.

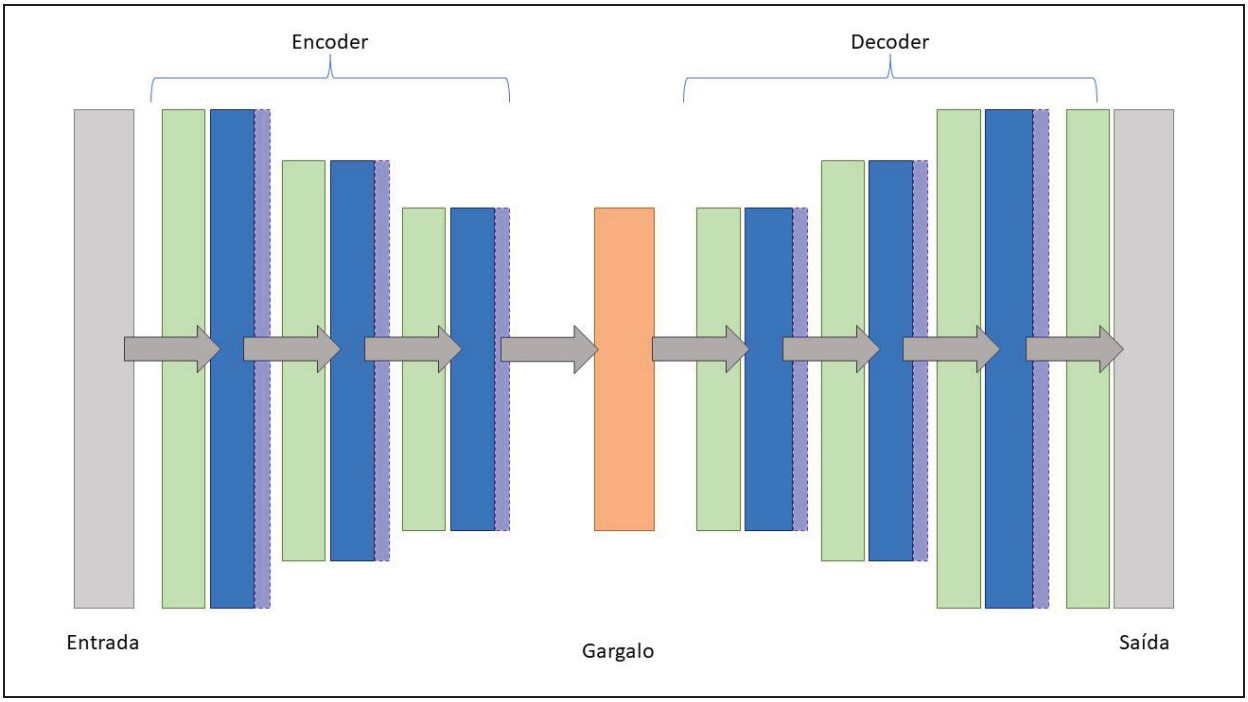
O gargalo é também uma camada densa, sem qualquer diferencial das demais, mas no caso de *dropout* ativo, esta camada não o recebeu, pois atuou como a camada de entrada do decoder. O encoder não utilizou *dropout* após a entrada, portanto, por simetria, o decoder também não utilizou.

A camada de entrada não apresenta função de ativação, pois representa os valores diretamente oriundos do conjunto de entrada (não constitui efetivamente uma camada de neurônios). As camadas escondidas (encoder, decoder e gargalo) utilizaram a função de ativação “relu”. A camada de saída utilizou a função de ativação “sigmoid”.

Na Figura 24 está apresentado o esquema genérico tanto do modelo AAE quanto MAAE. Nos dois modelos atencionais, as camadas do encoder e decoder foram antecedidas por uma camada de Atenção (na terminologia do Keras, *attention layer*). Em um caso foi uma camada de Atenção *single-head* (AAE) e noutro uma camada *multi-head* (MAAE).

Embora, graficamente, os modelos apresentem a mesma estrutura aparente, as camadas *single-head* e *multi-head* implementam formas de cálculo de Atenção diferentes.

FIGURA 24 - MODELO GENÉRICO DE UM AUTOENCODER ATENCIONAL



FONTE: O AUTOR (2023)

No caso particular da Atenção *multi-head*, existiram hiperparâmetros próprios que afetaram seu comportamento, como o tamanho do lote de amostras (quantidade de indivíduos considerados para o cálculo da Atenção), a quantidade de cabeças (quantidade de diferentes variações na ordem dos indivíduos no lote de amostras, que foi sempre configurado como igual à quantidade de características) e o tamanho da chave de consulta (no caso desta pesquisa, mantido sempre igual ao tamanho do vetor de entrada). A aplicação de *dropout* seguiu os mesmos critérios do modelo Vanilla.

As quantidades de neurônios para cada camada de entrada, encoder, gargalo, decoder e saída dependeram dos parâmetros dos experimentos. Por exemplo, para um modelo visual, tanto camada de entrada quanto de saída contavam com 31 neurônios, pois este era o número de características visuais.

As camadas do encoder, decoder e gargalo apresentaram variações em suas quantidades de neurônios, dado que estas quantidades foram exatamente os objetos das experimentações.

3.9 Protocolo de Experimentação

O objetivo específico B determinou que modelos de Aprendizado de Máquina autossupervisionados fossem experimentados na intenção de identificar quais deles ofereceriam os melhores resultados para discernir entre narrativas sinceras e não sinceras presentes no MMDDD-PtBr. Algumas decisões foram tomadas antes de encaminhar os experimentos:

a) **aprendizado profundo**: escolhido por causa dos resultados de excelência atingidos em diversas áreas, tais como visão computacional, compreensão de linguagem natural, veículos autoguiados, reconhecimento e síntese de texto, dentre outros;
b) **aprendizado autossupervisionado**: escolhido porque remete simultaneamente a dois problemas identificados no contexto da detecção de mentiras: pouca disponibilidade de dados rotulados e o tratamento de fatores idiossincráticos e contextuais que interferem na exposição de pistas;
c) **detecção multimodal**: escolhida porque tanto estudos envolvendo Aprendizado de Máquina como os fundamentos teóricos da detecção de mentiras indicaram que o uso sinérgico de múltiplas modalidades eleva o potencial de precisão da detecção

Mesmo restringindo os objetos de estudo ao círculo de modelos autossupervisionados (subconjunto de todos os modelos de Aprendizado de Máquina), existem diversas alternativas possíveis, condicionadas a diversos hiperparâmetros (parâmetros que governam o processo de aprendizado), que poderiam variar grandemente em resultado, tanto para uma mesma modalidade, quanto para diferentes modalidades.

Portanto, a circunstância impôs a necessidade de elaborar e comparar diversos modelos à disposição para a seleção daqueles mais promissores para o problema e para o conjunto de dados coletado.

O protocolo de experimentação elaborado para construir, avaliar e selecionar os modelos mais apropriados para cada modalidade foi estruturado nas seguintes etapas:

1. selecionar narrativas para treinar os Modelos de Sinceridade;
2. selecionar variantes de Autoencoders para participar dos experimentos;

3. treinar diversas arquiteturas de cada variante de Autoencoder sobre cada modalidade (Modelo de Sinceridade monomodal);
4. avaliar a capacidade de detecção de cada Modelo de Sinceridade monomodal;
5. selecionar as arquiteturas de melhor desempenho para compor Modelo de Sinceridade multimodais a partir de suas combinações.
6. avaliar a capacidade de detecção dos Modelos de Sinceridade multimodais para selecionar o de melhor desempenho.

As seções seguintes pormenorizam cada uma destas etapas.

3.9.1 Seleção de narrativas para treinar os Modelos de Sinceridade

Um Modelo de Sinceridade é um modelo de Aprendizado de Máquina que foi treinado para capturar as relações entre características que descrevem a narrativa sincera, entendida como narrativa normal. Para que o modelo pudesse capturar os caracteres de uma narrativa sincera de um sujeito, precisou ser treinado com uma ou mais narrativas sinceras daquele sujeito.

Dada a quantidade restrita de narrativas existentes no MMDDD-PtBr (61 até o momento em que os experimentos foram realizados), foi selecionada exatamente uma narrativa sincera de cada sujeito.

Visando dar ao modelo a maior quantidade possível de dados para que pudesse capturar a maior quantidade possível de padrões de sinceridade por sujeito, o critério de seleção foi o tempo da narrativa. Assim as narrativas sinceras mais longas de cada sujeito foram selecionadas para compor seus respectivos Modelos de Sinceridade monomodais, como mostrado no Quadro 13.

QUADRO 13 - NARRATIVAS SELECIONADAS PARA TREINAR OS MODELOS DE SINCERIDADE DE CADA UM DOS SUJEITOS DO CONJUNTO DE DADOS, COM DURAÇÃO EXPRESSA EM SEGUNDOS

Sujeito	Número de narrativas	Narrativa selecionada	Duração da narrativa (em segundos)
S1-P7	13	S1-P7-13	17
S1-P8	7	S1-P8-3	18
S1-P9	9	S1-P9-11	4
S2-P1	3	S2-P1-1	4
S2-P2	4	S2-P2-5	3
S2-P3	5	S2-P3-5	12
S2-P4	2	S2-P4-2	14
S2-P5	3	S2-P5-3	7
S2-P6	3	S2-P6-5	8
S2-P7	6	S2-P7-8	23

S2-P8	2	S2-P8-5	9
S2-P9	5	S2-P9-4	14

FONTE: O AUTOR (2023)

As demais narrativas dos respectivos sujeitos passaram, então, a servir como casos de teste, sendo submetidos ao Modelo de Sinceridade treinado para avaliar o grau de precisão de detecção que este oferece.

3.9.2 Avaliação de desempenho

Cada modelo treinado foi avaliado primariamente pela métrica acurácia balanceada (Brodersen et al., 2010). Embora a métrica acurácia seja comum em experimentos de classificação binária, percebeu-se em experimentos que a acurácia balanceada alcançou um equilíbrio mais realista na avaliação final das predições, dado que os rótulos de sinceridade apresentavam certo desbalanceamento (37 narrativas sinceras contra 24 narrativas não sinceras).

Em casos de classes com distribuição desbalanceada, é comum o uso da métrica Escore-F1, mas a acurácia balanceada foi preferida por ter uma interpretação mais intuitiva, além do fato do Escore-F1 ser mais indicado quando os dados apresentam maior prevalência de casos negativos (Flach; Kull, 2015), que não foi o caso nestes experimentos. A expressão da acurácia balanceada é dada pela equação (5).

$$Acurácia\ balanceada = \frac{1}{2}\left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP}\right) \quad (5)$$

Onde:

a) VP representa a quantidade de verdadeiros positivos, ou seja, sinceridades classificadas como tal;
b) VN representa a quantidade de verdadeiros negativos, ou seja, não sinceridades classificadas como tal;
c) FP representa a quantidade de falsos positivos, ou seja, não sinceridades classificadas como sinceridades;
d) FN representa a quantidade de falsos negativos, ou seja, não sinceridades classificadas como sinceridade.

Intuitivamente, o que a acurácia mede é a razão entre o total de acertos e o total de casos, não considerando se existem categorias mais numerosas que outras,

o que introduziria um viés na classificação. A acurácia balanceada procura corrigir esta distorção, mantendo a percepção intuitiva da métrica.

Para efeitos tão somente de comparação, os experimentos também foram medidos pela acurácia, que é dada pela equação (6).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (6)$$

Onde VP, VN, FP e FN são as mesmas contagens da equação (5).

Adicionalmente, um outro critério para desempate foi adotado. Em situações nas quais diferentes experimentos ofereceram graus de acurácia e acurácia balanceada idênticos, foi selecionado como modelo vencedor aquele que registrou o menor valor para FN, pois, no contexto deste estudo, um falso negativo significa interpretar erroneamente uma sinceridade com mentira (falsa mentira). Significa acusar injustamente um sujeito de mentir, quando na verdade estava sendo sincero.

Entendeu-se que o efeito ético da presença de falsos negativos (ou seja, falsas não sinceridades) nos resultados do modelo seja grave, de tal forma que se optou por minorá-lo tanto quanto possível.

3.9.3 Treinamento monomodal de modelos

Os componentes multimodais do MMDDD-PtBr apresentam dimensionalidades e cardinalidades diferentes. Por exemplo, para a narrativa S2-P6-3 existem os arquivos apresentados no Quadro 14.

QUADRO 14 - CARDINALIDADES E DIMENSIONALIDADES DOS CONJUNTOS DE DADOS DE CADA MODALIDADE DA NARRATIVA S2-P6-3

Arquivo	Modalidade	Cardinalidade (linhas)	Dimensionalidade (colunas)
S2-P6-3-opensmile.CSV	Acústica	1.130	65/68
S2-P6-3-verbal.CSV	Verbal	32	13/13
S2-P6-3-openface.CSV	Visual	339	31/709/714

FONTE: O AUTOR (2023)

Os arquivos de características acústicas gerados pelo OpenSMILE são compostos por 68 colunas, das quais apenas 65 efetivamente descrevem os aspectos sonoros presentes no arquivo de áudio utilizado. As três primeiras colunas são apenas para controle (nome do arquivo, tempo de início e tempo de fim da amostragem), que não têm valor para o aprendizado e por isso não foram utilizadas. Cada linha corresponde a um quadro de 10 milissegundos de amostragem.

As características verbais foram extraídas especificamente para os experimentos a partir das transcrições e, por isso, todas as 13 características foram utilizadas.

Os arquivos extraídos pelo OpenFace compreendem 714 características visuais, sendo que cinco são de controle (número do quadro, número da face, tempo de quadro, taxa de confiança e indicador de sucesso), que não têm valor para o aprendizado. As demais 709 características descrevem as coordenadas dos diversos pontos de controle que delineiam o rosto, nariz, lábios, olhos e sobrancelhas da face identificada, além de indicadores da presença e ausência de ações faciais, suas intensidades, vetores que descrevem a visada para cada olho e vetores que descrevem a inclinação da cabeça.

O OpenFace extraiu seis características de ângulo da cabeça, oito características de ângulos de visada e 17 características de intensidade das ações faciais, uma delas indicando a piscada, perfazendo 31 características. Essas características são as mesmas exploradas em diversos dos estudos identificados na revisão sistemática de literatura (Constancio et al., 2023), e foram escolhidas porque apresentam o dado de interesse para as análises faciais, ou seja, as ações faciais que constituem as expressões e microexpressões faciais, assim como os ângulos de cabeça e visada.

Enquanto as modalidades acústica e visual apresentaram cardinalidades baseadas em tempo (com diferentes frequências), a cardinalidade da modalidade verbal é estabelecida pela quantidade de palavras presentes na narrativa. Essa incompatibilidade no critério de distribuição das linhas de cada conjunto monomodal de dados impediu que os mesmos fossem alinhados em um único grande conjunto que serviria para alimentar um único modelo, processo conhecido como fusão precoce (*early fusion*, em tradução livre).

Como consequência, os componentes monomodais precisaram ser tratados em separado, utilizados para treinar modelos monomodais que produziram respostas monomodais, posteriormente combinadas para gerar uma única resposta multimodal integrada, processo conhecido como fusão tardia (*late fusion*, em tradução livre).

Enquanto conjuntos de hiperparâmetros que afetaram todos os modelos selecionados, alguns hiperparâmetros eram próprios de certas arquiteturas. Adicionalmente, a quantidade de camadas e a quantidade de neurônios por camada figuraram como dois fatores que interferiram diretamente na capacidade do modelo

de compreender os padrões existentes nos dados. Assim, todos esses hiperparâmetros juntos ofereceram um número elevado de combinações que não puderam ser totalmente experimentados, exigindo uma heurística para guiar a experimentação e avaliação. No Quadro 15 é possível ver a lista de todos os hiperparâmetros que interferiram no treinamento dos modelos e a quais estão vinculados.

QUADRO 15 - HIPERPARÂMETROS PARA CADA TIPO DE AUTOENCODER EXPERIMENTADO

Hiperparâmetro	Autoenconder Vanilla (AE)	Autoencoder atencional *single-head* (AAE)	Autoencoder atencional *multi-head* (MAAE)
Arquitetura	✓	✓	✓
Épocas	✓	✓	✓
Taxa de aprendizado	✓	✓	✓
Tamanho do lote	✓	✓	✓
Dropout	✓	✓	✓
Quantidade de cabeças			✓
Quantidade de amostras			✓
Tamanho da chave de consulta			✓

FONTE: O AUTOR (2023)

O processo de treinamento contou com uma etapa de calibragem da quantidade de épocas que seguiu os seguintes passos:

1. **Treinamento de modelos preliminares com uma arquitetura simples para todas as narrativas**: treinamento com 2000 épocas para avaliar qual a quantidade mínima de épocas necessária para o aprendizado estabilizar; neste estudo, um modelo simples é aquele no qual encoder, decoder e gargalo têm a mesma dimensionalidade que a entrada;
2. **Avaliação da estabilidade do aprendizado em 500 épocas**: quando a evolução das curvas de acurácia e erro apontaram para possibilidades de melhoria no aprendizado com mais épocas, os treinamentos foram repetidos incluindo-se 500 épocas a mais em relação ao experimentado até então;
3. **Registro da quantidade mínima de épocas para a estabilização do aprendizado**: a quantidade de épocas observada até a estabilização foi registrada para uso posterior em todos os demais treinamentos.

Com a calibragem de épocas estabelecida, o treinamento propriamente dito de cada modelo seguiu o seguinte protocolo (especificado apenas o encoder, visto que o decoder é seu simétrico):

1. construção de modelos de uma a três camadas, na mesma dimensionalidade da entrada;
2. construção de modelos com gargalo na mesma dimensionalidade que a entrada, que serve de cardinalidade referencial para posteriores variações;
3. variação dos tamanhos das camadas em duas unidades para cima e duas unidades para baixo em relação à dimensionalidade da camada anterior, a cada camada;
4. variação dos tamanhos do gargalo em duas unidades para cima e duas unidades para baixo em relação ao tamanho da última camada do encoder, com dois passos para cima e dois para baixo;
5. variação no tamanho do lote de treinamento, arbitrariados em 32 e 256, exceto para modelos verbais, pois dada a baixa cardinalidade dos conjuntos de treinamento lotes maiores que 32 não operariam diferença; para modelos verbais, apenas lotes de tamanho 32 foram testados;
6. para o modelo com atenção *multi-head*, a quantidade de cabeças de atenção foi idêntica a dimensionalidade da entrada e a quantidade de amostras foi arbitrariada em 5, 10 e 20; este critério foi experimentado apenas para as modalidades acústica e visual; para a modalidade verbal, que conta com cardinalidades menores, a quantidade de amostras foi de 3, 5, 10 e 20 palavras;
7. *dropout* foi aplicado aos dez modelos de melhor desempenho, até então, em termos de acurácia balanceada de detecção;
8. *dropout* constante de 20% em cada camada e progressivo com 10% de variação por camada, iniciando em 20%; os tamanhos de camada foram mantidos conforme original, assim como acrescidos do percentual de *dropout* aplicado.

Cada modelo foi avaliado logo após seu treinamento, com o resultado em termos de acurácia, acurácia balanceada, escore-F1, precisão, revocação, contagens de falsas sinceridade, verdadeiras sinceridades, falsas não sinceridades e verdadeiras não sinceridades.

3.9.4 Assinatura de modelos

Dada a quantidade de experimentos que foram realizados, tornou-se necessário criar uma codificação para identificar precisamente qual modelo atingiria qual desempenho. Assim, os modelos passaram a ser identificados por meio de uma assinatura, ou seja, uma identificação compacta que o identificasse dentre os demais experimentos.

A assinatura dos modelos foi formada por quatro grandes blocos:

1. código do tipo do modelo;
2. arquitetura do modelo;
3. hiperparâmetros utilizados;
4. sufixo diferencial.

O código do tipo do modelo identificou se o modelo é um Autoencoder Vanilla (AE), um Autoencoder atencional *single-head* (AAE) ou um Autoencoder atencional *multi-head* (MAAE).

A arquitetura do modelo identificou a quantidade e cardinalidade das camadas de neurônios. Por exemplo a assinatura 65_65-63-65_65 identificou um modelo onde o encoder apresentava duas camadas, ambas com 65 neurônios, um gargalo com 63 neurônios e um decoder com duas camadas de 65 neurônios cada.

O bloco com os hiperparâmetros identificou, em ordem, função de perda, o tamanho do *minibatch*, a quantidade de cabeças *multi-head* (quando foi o caso), o tamanho da chave de consulta de atenção (quando foi o caso), a quantidade de amostras de atenção (quando foi o caso).

O bloco eventual de sufixo textual foi utilizado em situações específicas como uma forma alternativa para diferenciar modelos.

Um exemplo de assinatura de modelo é “AAE-31-33-31-mse-32-AU-Gaze-Head”, usado para identificar um modelo Autoencoder atencional *single-head* com uma camada de encoder com 31 neurônios, um gargalo com 33 neurônios e um decoder com uma camada de 31 neurônios. Este modelo foi treinado com a função de perda MSE e um *minibatch* de 32 épocas. O sufixo “AU-Gaze-Head” indica que as características utilizadas foram, respectivamente, ações faciais, ângulo de mirada e orientação da cabeça.

3.9.5 Avaliação da capacidade de detecção de cada modelo

Em acordo com a literatura, o uso de Autoencoders aplicados à detecção de anomalias utiliza o erro quadrático médio (MSE) de reconstrução como limiar de decisão entre normalidade e anormalidade (Bank; Koenigstein; Giryes, 2021).

No entanto, neste estudo procurou-se adotar formas alternativas de medir o erro de reconstrução pela **diferença na distribuição** dos dados reconstruídos em relação aos dados de treinamento, aproveitando resultados encontrados em outros estudos (Afgani; Sinanović; Haas, 2008). Para tal foi adotada a divergência de Kullback-Leibler (ou divergência KL) nas suas duas formas possíveis, o caso gaussiano (KLn) e o caso genérico (KLd).

A intuição que motivou a experimentação da divergência KL foi a de que dados anormais podem não apenas divergir no cômputo geral das diferenças de reconstrução de cada instância de dado (capturada pela MSE), mas também na forma como os dados se distribuem (capturada pela divergência KL). Os modelos foram avaliados pelas três métricas (MSE, KLd e KLn) para medir o grau de capacidade preditiva de cada uma e assim identificar a melhor delas como métrica definitiva.

Para o caso particular do KLd, o cálculo depende de uma aproximação da distribuição de probabilidades, atingida pela construção de um histograma. Neste caso, a quantidade de colunas do histograma é crítica pois uma quantidade muito pequena de colunas concentra muitos valores, mascarando a presença de valores discrepantes (anomalias) e valores muito altos aumenta demasiadamente a sensibilidade a variações, levando a considerar como anomalias valores normais em situações extremas (Afgani; Sinanović; Haas, 2008).

Assim, foi necessário testar algumas alternativas de quantidade de colunas no momento de realizar o cálculo da KLd. Os valores experimentados foram 5, 10, 15, 20, 30, 40, 50 e 60 colunas. Para efeitos de notação, quando a KLd foi calculada com um histograma de 15 colunas, foi representada como KLd:15.

3.9.6 Escore de Sinceridade

Em geral, o erro de reconstrução com os dados de normalidade opera como um limiar para detecção. Se o erro de reconstrução de um caso de teste for superior ao erro calculado em treinamento, considera-se que os dados são suficientemente discrepantes da normalidade, caracterizando assim uma anomalia. No contexto desta

pesquisa, significa dizer que o caso de teste apresenta discrepância suficiente para ser considerado não sincero.

No entanto, a comparação pura e simples com um limiar de classificação dificulta a percepção do grau de confiança da detecção que o modelo oferece. Por exemplo, se o limiar de sinceridade de um sujeito for 121 e um caso de teste produzir o erro de reconstrução de 121,01, teve-se um caso classificado como não sincero (erro de reconstrução do caso de teste superior ao erro de reconstrução do caso de sinceridade). Em um outro exemplo, poderia-se ter um caso de teste apresentando um erro de reconstrução de 190, também considerado como não sincero.

Embora os dois casos representem a deteção de uma não sinceridade, as margens de diferença entre cada caso são altamente díspares. No primeiro caso apenas 0,01 (uma diferença de 0,008%) e no segundo 69 (uma diferença de 57,025%). O primeiro caso é altamente sensível a pequenas perturbações nos dados, podendo mesmo representar um erro de detecção. Já no segundo caso, apenas uma perturbação mais volumosa poderia reverter a decisão em favor da não sinceridade, oferecendo assim mais confiança na detecção do modelo.

Diante dessa situação, optou-se por realizar um processo de normalização do erro de reconstrução, para que o mesmo passasse a ser uma métrica bipolar, com domínio [-1, 1], de tal forma que a polaridade negativa da métrica indicasse não sinceridade e a polaridade positiva, sinceridade. Dentro desta proposta, a medida de -1 indica 100% de confiança na detecção de não sinceridade, zero indica absoluta incapacidade de decisão e 1 indica 100% de confiança na detecção de sinceridade.

A escolha do significado do sinal da métrica foi apenas intuitividade. Visto que o modelo é de “sinceridade”, pareceu mais intuitivo que a métrica oferecesse um valor positivo para sinceridade e negativo para não sinceridade. Tal métrica passou a ser chamada de **Escore de Sinceridade** (ES).

Por exemplo, se a narrativa S1-P7-5, ao ser submetida a um certo Modelo de Sinceridade treinado com a narrativa S1-P7-12, produziu ES = -0,54, entende-se que aquela foi detectada como não sinceridade, com 54% de confiança.

O cálculo do Escore de Sinceridade segue dois passos. No primeiro é calculada a **distância relativa** entre o erro de reconstrução em sinceridade (erro de reconstrução do Modelo de Sinceridade) e o erro de reconstrução do caso de teste. Tal métrica é representada por $\Delta\varepsilon$ e definida pela equação (**7**):

$$\Delta\varepsilon = \frac{\varepsilon_s - \varepsilon_t}{\lambda} \quad (7)$$

Na equação precedente, "ε_s" significa o erro de reconstrução em sinceridade, ou seja, o **limiar de classificação**, "ε_t" significa o erro de reconstrução do caso de teste (ambos calculados sob qualquer métrica para medir o erro de reconstrução, porém a mesma métrica para os dois cálculos) e "λ" representa a **margem de incerteza**. A margem de incerteza delimita a distância máxima a partir do limiar de classificação para a qual existe um grau de incerteza na detecção. No caso específico desta pesquisa, "λ" corresponde à **média dos desvios padrões dos erros de reconstrução de cada uma das características** do conjunto de entrada usadas para treinar o Modelo de Sinceridade.

Valores maiores de "λ" indicam que houve maior variação dos erros de reconstrução das características, alargando a margem de incerteza, consequentemente exigindo que a diferença dos erros de reconstrução de treinamento e teste sejam também maiores para que se tenha maior confiança no resultado.

A distância relativa (Δε) pode produzir valores menores que -1 e maiores que 1. Para garantir que os mesmos sejam ajustados para o domínio definido, aplica-se a função tangente hiperbólica (que ajusta o valor da distância relativa para o domínio [-1, 1]), como definido pela equação (8):

$$ES_p = \tanh \Delta\varepsilon \quad (8)$$

A métrica ES_p foi chamada de **Escore de Sinceridade parcial**, pois expressou a confiança de detecção de um modelo monomodal apenas, ou seja, de uma parcela do modelo multimodal final.

Um exemplo de cálculo poderá facilitar a compreensão do processo de deteção de mentiras por meio do Escore de Sinceridade. Tome-se a narrativa S1-P7-13 para treinar o Modelo de Sinceridade acústico do sujeito S1-P7. Este sujeito oferece outras 12 narrativas que podem ser sinceras e não.

O modelo acústico tem 65 características de entrada, portanto o Modelo de Sinceridade recebe e reconstrói 65 características ao longo de 1.710 linhas, pois a narrativa S1-P7-13 tem cerca de 17 segundos de duração.

Após o treinamento são calculados o limiar de classificação (“ε_s”) e a margem de incerteza (“λ”) para o modelo. O primeiro parâmetro corresponde à média dos erros de reconstrução de cada uma das 65 características para as 1.710 linhas e o segundo ao desvio padrão da mesma série de valores. Após o cálculo ε_s = 0,271 e λ = 0,348.

Ao submeter a narrativa S1-P7-5 (com 931 linhas de dados) ao modelo de sinceridade, o erro de reconstrução calculado corresponde a ε_t = 0,35. Calcula-se então a distância relativa do erro em relação do limiar de classificação, como mostrado em equação (9).

$$\Delta\varepsilon = \frac{\varepsilon_s - \varepsilon_t}{\lambda} = \frac{0{,}271 - 0{,}35}{0{,}348} = -0{,}227 \quad (9)$$

Esta distância é então normalizada pela função tangente hiperbólica para ser restringida ao domínio [-1, 1], como visto em equação (10).

$$ES_p = \tanh \Delta\varepsilon = \tanh(-0{,}227) = -0{,}223 \quad (10)$$

Com ES_p = 0,223 se tem que a detecção da narrativa S1-P7-5 dada pelo Modelo de Sinceridade monomodal acústico de S1-P7-13 é de “não sinceridade” com 22,3% de confiança.

O processo de cálculo e a interpretação foram os mesmos para as outras modalidades. Em um modelo multimodal composto pelas modalidades acústica, verbal e visual, houve um ES_p para cada modalidade, cada um oferecendo um grau de confiança de classificação diferente.

3.9.7 Seleção dos modelos monomodais

Dadas as três diferentes métricas para o cálculo do erro de reconstrução, os processos de classificação foram realizados à luz de cada uma delas visando identificar a que melhor desempenhou por modalidade. A ideia foi a de selecionar a melhor métrica para cada caso, no intuito de elevar ao máximo o desempenho de cada modelo monomodal para promover o melhor resultado multimodal posterior.

3.10 Fusão de modalidades

O objetivo específico C determinou que um mecanismo de fusão dos Modelos de Sinceridade monomodais fosse elaborado para que uma resposta única e integrada fosse oferecida ao final do processo de detecção. Uma forma comum de

executar a fusão tardia é pelo método da votação, ou seja, conta-se quantos modelos indicaram sinceridade e quantos indicaram não sinceridade e a maioria apontará a classificação final.

Neste caso, no entanto, a proposta do Escore de Sinceridade permite fazer uma espécie de votação ponderada pela confiança, como mostrado na equação (11).

$$ES = \tanh \sum_{p}^{n} ES_p \tag{11}$$

Na equação precedente, "*p*" identifica cada um dos ES_p parciais, para cada um dos "*n*" modelos parciais utilizados. O processo utilizado foi simplesmente o de somar os ES_p de cada modelo monomodal (chamados aqui de escores parciais) e aplicar novamente a função tangente hiperbólica para manter o valor final dentro do domínio [-1, 1], com a mesma interpretação dos valores.

Por exemplo, em um modelo multimodal com modalidades acústica, verbal e visual, os Escores de Sinceridade parciais poderiam ser $ES_{p(acústico)}$ = 0,45; $ES_{p(verbal)}$ = 0,12; $ES_{p(visual)}$ = -0,77. A soma desses escores parciais é de **-0,20**, que após ser submetido função "*tanh*" resultaria em -**0,1974**. Com este resultado, a detecção multimodal da narrativa testada é de **não sinceridade com 19,75% de confiança**. Apesar de haver dois votos em favor da sinceridade (acústico e verbal), suas confianças parciais não foram capazes de superar a confiança do modelo visual, que apontou não sinceridade, apenas a enfraqueceram naquele resultado.

3.11 Estudo de ablação de características verbais

Dado que uma das principais contribuições desta pesquisa é a exploração dos aspectos verbais na detecção de mentiras, foi realizado um estudo para avaliar quais destas contribuem mais para o resultado do modelo. O processo é conhecido como estudo de ablação, que consiste em remover partes de um sistema para avaliar que efeitos tal operação exerceu (Sheikholeslami et al., 2021) no resultado.

No caso particular deste estudo, a ablação consistiu em remover características do conjunto de treinamento original, eventualmente adaptando os modelos para a nova configuração de entradas, e então treinar e avaliar os desempenhos resultantes, para posterior comparação com um modelo referencial.

O protocolo deste estudo foi derivar variantes do modelo verbal individual de mais alto desempenho, que atuou como linha de base (modelo referencial) para as comparações. Estas variantes utilizaram apenas partes das características presentes na entrada original (composta por 13 características). Como a dimensionalidade das entradas mudou, os modelos foram adequados para refletir o novo conjunto de características para treinamento. As características originais foram organizadas em quatro grupos:

1. **Características paralinguísticas**: representadas pela letra “P”, incluem “duration” e “hesitation”;
2. **Características de sentimento**: representadas pela letra “S”, incluem “pos_score” e “neg_score”;
3. **Características de função sintática e entidade**: representadas pela letra “O”, incluem “is_ent”, “is_num”, “is_pronoun”, ‘is_verb”, “is_noun”, “is_adjective” e “is_adverb”;
4. **Características de pessoa verbal**: representadas pela letra “R”, incluem “is_1_person” e “is_3_person”.

As variantes do modelo referencial apresentaram suas assinaturas sufixadas com as letras que identificaram os conjuntos de características presentes. Por exemplo, o modelo que incluiu as características paralinguísticas e de pessoa verbal recebeu o sufixo “P-R”, enquanto o modelo que incluiu as características de sentimento, função sintática e pessoa verbal foi sufixado com “S-O-R”.

Como a ordem das características no conjunto não interfere no resultado ("P-R” é equivalente a “R-P”), as combinações se limitaram a dez, sendo duas dos quatro grupos tomados três a três e oito tomados dois a dois.

Após o treinamento de cada modelo variante, os mesmos foram avaliados e suas acurácias balanceadas comparadas com a do modelo referencial, para assim apreciar os efeitos em termos de desempenho de detecção.

3.12 Experimentos com coletividades

Embora a proposta desta pesquisa tenha sido a de elaborar Modelos de Sinceridade individuais para evidenciar as peculiaridades de cada sujeito durante uma narrativa não sincera, a disponibilidade de um conjunto de dados anotado com diversos sujeitos viabilizou a experimentação de Modelos de Sinceridade coletivos.

O propósito do treinamento de tais modelos foi avaliar a aprendizagem de características de sinceridade de grupo (não individuais) e suas capacidades discriminatórias na detecção de outros sujeitos ainda não conhecidos (generalização). Este tipo de experimento pode ser utilizado para validar a proposição de que existem padrões coletivos de sinceridade.

Assim, conjuntos de dados coletivos foram construídos dentro da política *leave-one-out*, ou seja, para cada um dos 12 sujeitos disponíveis, escolheu-se um para excluir do modelo coletivo, que foi treinado com todas as narrativas sinceras de todos os outros sujeitos. Foram, portanto, criados 12 Modelos de Sinceridade coletivos formados pelas narrativas de 11 sujeitos. Aquele modelo que excluiu dado sujeito seria exatamente usado para a avaliação das suas detecções.

Dado que o Escore de Sinceridade é uma métrica multicomponentes, ou seja, permite que múltiplos resultados de detecção parcial sejam utilizados para compor uma detecção final, a introdução de modelos coletivos de sinceridade não representou desafio técnico.

Como os conjuntos de dados monomodais cresceram significativamente, o protocolo de treinamento com os modelos coletivos utilizou apenas os hiperparâmetros dos cinco melhores modelos individuais já avaliados.

Com os modelos coletivos monomodais parciais treinados, bastou incluí-los no cálculo do Escore de Sinceridade final, que passou a ter seis componentes, ou seja, três componentes monomodais individuais e três componentes monomodais coletivos.

Testes com modelos apenas individuais, apenas coletivos e com todos foram realizados com o intuito de medir o potencial preditivo de cada combinação. No caso dos modelos combinados monomodais, o melhor modelo coletivo foi utilizado para compor um modelo bicomponente com o melhor modelo individual.

3.13 Avaliação de resultados

Como meio de decidir qual dos diversos modelos treinados teve mais alto desempenho em termos de detecção de mentiras, foram inicialmente consideradas duas métricas: a acurácia e a acurácia balanceada. O objetivo foi o de avaliar se o desbalanceamento existente nas rotulações era ou não suficiente para influenciar na qualidade da avaliação.

Os modelos, tanto monomodais quanto multimodais, foram avaliados pelas duas métricas, mas a acurácia balanceada foi a métrica que serviu como critério para avaliar o desempenho. Em casos de empate, uma métrica secundária foi adotada: a contagem de casos de falsas não sinceridades.

Uma falsa não sinceridade ocorre quando o modelo erroneamente classifica uma narrativa sincera como não sincera, o que significa não reconhecer a sinceridade diante do próprio Modelo de Sinceridade. Este erro foi considerado o mais grave porque demonstra que padrões de sinceridade deixaram de fazer parte do modelo. Adicionalmente, em uma situação real, significaria acusar um interrogado de mentir, quando na verdade estava oferecendo um testemunho sincero.

Assim, em casos de empate, o melhor modelo foi aquele que ofereceu uma menor contagem de falsas não sinceridades.

4 RESULTADOS

Os esforços empreendidos na pesquisa consistiram na realização de uma revisão sistemática de literatura, na construção de um conjunto de dados rotulado e multimodal para detecção de mentiras e de experimentos (que foram alimentados com aqueles de dados) para a identificação dos hiperparâmetros que otimizassem a distinção entre casos de teste sinceros e não sinceros para os Modelos de Sinceridade construídos.

A revisão de literatura produziu um panorama amplo e detalhado do estado da ciência no campo da Detecção de Mentiras suportada por Aprendizado de Máquina, o que veio a permitir identificar tendências, lacunas e oportunidades.

O conjunto de dados rotulado, concebido a partir de narrativas expressas em português do Brasil, oportunizou a realização de experimentos com os Modelos de Sinceridade, explorando aspectos linguísticos particulares da língua portuguesa.

Os experimentos permitiram comparar os efeitos dos diversos hiperparâmetros sobre os Modelos de Sinceridade monomodais para então conceber um Modelo de Sinceridade multimodal com os melhores resultados. Também foi possível identificar quais métricas para o cálculo do erro de reconstrução produziram os melhores resultados para cada uma das modalidades.

As seções seguintes apresentam pormenores desses esforços.

4.1 Artigo da Revisão de literatura

A revisão de literatura realizada para compreender o estado da ciência a respeito da Detecção de Mentiras assistida por Aprendizado de Máquina forneceu um panorama amplo e profundo das técnicas e estratégias adotadas ao longo de uma década (período de 2011 a 2021), assim como o grau de sucesso de diversas abordagens propostas como resposta ao problema.

Um artigo científico intitulado “*Deception detection with machine learning: a systematic review and statistical analysis*” (Constancio et al., 2023) foi escrito contendo os achados da revisão e apresenta tendências, alternativas, dificuldades, descobertas e lacunas pertinentes ao tema. O artigo foi submetido ao editorial do periódico PloS One em 2022, portanto a produção científica a partir deste ano não está contida no estudo.

Em 05/01/2024, segundo o site Google Acadêmico[19], tal artigo já contava com nove referências, figurando como o terceiro item de retorno à expressão de busca “deception detection machine learning”. O relatório também destaca as principais técnicas de Aprendizado de Máquina adotadas, os conjuntos de características consumidos e os níveis de desempenho relatados em cada um dos 81 artigos selecionados dentre os 648 recuperados.

O artigo foi publicado no periódico científico PoIS One[20]. Todos os dados encontram-se na forma de diagramas, Jupyter Notebooks e bancos de dados em um repositório do GitHub[21], liberado para acesso público.

Em adição à revisão de literatura também foi construído um conjunto de mapas mentais que atuam como resumo de cada estudo.

4.2 Conjunto de dados para Detecção de Mentiras

Em resposta ao objetivo específico A, o primeiro conjunto de dados multimodal, específico para detecção de mentiras para o português do Brasil foi construído. Trata-se de um Produto Técnico-Tecnológico (PTT) que foi denominado “*Multimodal Deception Detection Dataset for Brazilian Portuguese*”, MMDDD-PtBr, e abrange características de origem acústica, verbal e visual. Apesar de sua direção específica para a língua portuguesa, o mesmo foi denominado em inglês para efeitos de internacionalização.

Uma metodologia foi elaborada para extrair as características multimodais de vídeos selecionados do quadro “Acredite em quem quiser” que integra o programa “Domingão com Huck”, produzido pela Rede Globo de Televisão e que foi ao ar aos domingos durante alguns meses da temporada em que esta pesquisa estava sendo encaminhada. Algumas ferramentas também foram construídas para superar desafios específicos oferecidos pelos vídeos originais selecionados como fonte de dados primária.

Uma vez mais, tanto a metodologia quanto as ferramentas adicionais que foram elaboradas também são PTTs que tanto podem ser diretamente utilizados para a extensão deste conjunto de dados, como para inspirar a construção de outros, para outras línguas ou culturas.

---

[19] https://scholar.google.com/scholar?cites=7800862588371019624&as_sdt=2005&sciodt=0,5&hl=pt-BR

[20] https://journals.plos.org/plosone

[21] https://github.com/gambit4348/deception-detection-review-2022

O MMDDD-PtBr é composto (até o momento da conclusão desta pesquisa) por 61 narrativas (37 sinceras e 24 não sinceras) coletadas de 12 sujeitos (sete homens e cinco mulheres) com sotaques de diversas regiões do país (nordeste, centro-oeste e sul).

As narrativas têm nomes codificados para identificar o vídeo fonte (S1 e S2), o sujeito (P1, P2, P3, P4, P5, P6, P7, P8 e P9, dentro de cada fonte) e narrativa (número serial iniciado em 1 para cada sujeito). Por exemplo, a narrativa S2-P3-3 identifica a terceira narrativa do terceiro sujeito do segundo vídeo fonte. No Apêndice A está disponível um quadro listando cada uma das narrativas e diversos dos seus atributos.

Por tratar-se de um conjunto de dados multimodal, o MMDDD-PtBr não pôde ser construído na forma de um único arquivo. Assim, diversos arquivos, específicos por modalidade, foram produzidos a partir dos vídeos originais. De forma resumida o conjunto de dados é composto de:

1. 61 arquivos MP4 correspondendo às narrativas selecionadas para compor o conjunto de dados;
2. 61 arquivos WAV correspondendo às trilhas de áudio dos arquivos de vídeos das narrativas;
3. 61 documentos CSV com 65 características acústicas extraídas pelo OpenSMILE a partir das trilhas de áudio;
4. 61 documentos JSON com as transcrições corrigidas e palavras enriquecidas pelo uso combinado do SpaCy e do SentiWordNet-PT-BR;
5. 61 documentos CSV com 13 características verbais extraídas a partir dos documentos JSON de transcrições;
6. 61 documentos CSV com 714 características visuais extraídas pelo OpenFace;
7. 61 documentos CSV com recortes de segmentos correspondentes ao sujeito de cada narrativa, sendo que dois deles (S1-P8-5-openface-cuts.CSV e S2-P9-2-openface-cuts.CSV) foram inteiramente rejeitados, não produzindo qualquer característica visual para aquelas narrativas.

Na presente configuração, o MMDDD-PtBr ocupa o espaço aproximado de 387 megabytes.

4.3 Modelos de Sinceridade monomodais

Em resposta ao objetivo específico B, diversos experimentos com Modelos de Sinceridade monomodais foram realizados buscando identificar quais deles produziriam os melhores resultados em termos de acurácia de detecção. Dadas as condições oferecidas pelo MMDDD-PtBr, foi possível construir Modelos de Sinceridade coletivos, que incluiram todas as narrativas sinceras dos 12 indivíduos, tomados 11 a 11.

Foi elaborado o **Escore de Sinceridade parcial** (ES_p), métrica bipolar, com domínio definido em [-1, 1], no qual o lado negativo expressa a confiança na detecção de uma narrativa não sincera, zero indica detecção inconclusiva e o lado positivo a confiança na detecção de uma narrativa sincera.

Nas seções a seguir são apresentados e discutidos resultados específicos de cada modalidade.

4.3.1 Modelos acústicos individuais

Os modelos acústicos foram alimentados com características extraídas pelo OpenSMILE, que recebeu como entrada os arquivos de áudio de narrativas em formato WAV e retornou planilhas compostas por 68 colunas (configuração de dados que o OpenSMILE denomina como “ComParE 2016”), sendo 65 de características acústicas e três de controle com uma linha para cada 10 milissegundos de áudio.

Todas as 65 características acústicas extraídas pelo OpenSMILE foram consideradas, implicando que todos os modelos tiveram 65 neurônios de entrada e 65 de saída, com variadas configurações de camadas escondidas, conforme estabelecido no protocolo de experimentação. Todos os modelos foram treinados com 500 épocas, visto que a calibragem inicial mostrou que tal valor seria mais que suficiente para a estabilidade do aprendizado. A maioria dos modelos atingiu a estabilidade com aproximadamente 200 épocas, alguns em menos de 50.

Modelos Autoencoder Vanilla foram testados em 154 diferentes configurações, sendo 70 sem o uso de *dropout* e 84 com esta técnica de regularização. Os 10 melhores resultados alcançados podem ser vistos tanto no Quadro 16 quanto no Quadro 17 (continuação). Os desempenhos estão expressos em acurácia (“Acc”) e acurácia balanceada (“B Acc”), esta última operando como o critério de avaliação dos modelos. Todas as métricas para cálculo do erro de

reconstrução (KLd:5 a KLd:60, KLn e MSE) estão presentes. A coluna “Modelo” apresenta as assinaturas dos modelos experimentados.

As células são coloridas, a intensidade da cor reflete o valor numérico presente, para mais fácil percepção das magnitudes distribuídas. Nas células mais escuras estão contidos os desempenhos mais elevados.

QUADRO 16 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS ACÚSTICAS DE AUTOENCODERS VANILLA USANDO AS MÉTRICA KL:5, KL:10, KL:15, KL:20 E KL:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AE-65_67_69-67-69_67_65-mse-256	0,490	0,499	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500
AE-65_65-63-65_65-mse-256	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500
AE-65_65_65-69-65_65_65-mse-256	0,510	0,520	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500
AE-65_67-63-67_65-mse-256	0,531	0,540	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500
AE-65_63_61-61-61_63_65-mse-256	0,469	0,479	0,469	0,479	0,469	0,479	0,469	0,479	0,469	0,479
AE-65_02_67_02-63-67_02_65_02-mse-256	0,469	0,474	0,449	0,457	0,571	0,580	0,531	0,540	0,469	0,479
AE-65_02_65_02_65_02-63-65_02_65_02_65_02-mse-256	0,490	0,494	0,429	0,433	0,449	0,453	0,469	0,475	0,469	0,475
AE-65-69-65-mse-256	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500
AE-81_02_93_03_108_04-67-108_04_93_03_81_02-mse-256	0,510	0,513	0,469	0,474	0,510	0,515	0,490	0,496	0,490	0,497
AE-81_02_93_03_108_04-63-108_04_93_03_81_02-mse-256	0,490	0,494	0,490	0,494	0,490	0,495	0,571	0,577	0,571	0,578

FONTE: DADOS DA PESQUISA (2023)

Os resultados mostram que diferentes arquiteturas forneceram diferentes níveis de acurácia balanceada para diferentes métricas de avaliação de erro.

QUADRO 17 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS ACÚSTICAS DE AUTOENCODERS VANILLA USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AE-65_67_69-67-69_67_65-mse-256	0,490	0,500	0,510	0,520	0,490	0,500	0,653	0,658	0,490	0,500
AE-65_65-63-65_65-mse-256	0,490	0,500	0,490	0,500	0,490	0,500	0,633	0,638	0,490	0,500
AE-65_65_65-69-65_65_65-mse-256	0,490	0,500	0,490	0,500	0,490	0,500	0,612	0,619	0,490	0,500
AE-65_67-63-67_65-mse-256	0,490	0,500	0,490	0,500	0,490	0,500	0,592	0,599	0,531	0,540
AE-65_63_61-61-61_63_65-mse-256	0,469	0,479	0,469	0,479	0,469	0,479	0,592	0,598	0,490	0,500
AE-65_02_67_02-63-67_02_65_02-mse-256	0,469	0,479	0,469	0,479	0,469	0,479	0,510	0,518	0,510	0,520
AE-65_02_65_02_65_02-63-65_02_65_02_65_02-mse-256	0,469	0,476	0,469	0,477	0,469	0,477	0,429	0,437	0,571	0,580
AE-65-69-65-mse-256	0,490	0,500	0,490	0,500	0,490	0,500	0,571	0,579	0,490	0,500
AE-81_02_93_03_108_04-67-108_04_93_03_81_02-mse-256	0,469	0,477	0,429	0,437	0,469	0,478	0,571	0,579	0,531	0,540
AE-81_02_93_03_108_04-63-108_04_93_03_81_02-mse-256	0,531	0,538	0,490	0,498	0,490	0,498	0,571	0,578	0,531	0,540

FONTE: DADOS DA PESQUISA (2023)

A aplicação de *dropout* produziu alguns modelos presentes no grupo de 10 melhores, mas os modelos de desempenho mais elevado não necessitaram deste

recurso. A métrica de avaliação teve um efeito decisivo na classificação dos modelos. Os valores elevados de acurácia balanceada somente foram alcançados na métrica KLn. As outras métricas produziram avaliação significativamente mais baixas, notadamente a MSE, que é a métrica mais frequentemente utilizada em problemas de detecção de anomalias.

Chama a atenção que as diversas configurações de histogramas para o calcula das KLd também resultaram em avaliações expressivamente mais baixas. A KLn é uma métrica de medição da diferença de distribuição para conjuntos que se assemelham da distribuição normal. Em sendo a KLd uma métrica mais geral, deveria ter sido capaz de ao menos se aproximar dos resultados medidos pela KLn.

O segundo conjunto de modelos testado foi dos Autoencoders atencionais *single-head* (AAE). Foram ao todo 130 experimentos (70 sem e 60 com *dropout*). Nos Quadro 18 e Quadro 19 apresentam-se os 10 níveis de desempenho mais elevados alcançados com diferentes modelos atencionais *single-head*.

Quando comparados com os modelos AE, tais Autoencoders atingiram níveis mais elevados de acurácia e acurácia balanceada, indicando que a Atenção *single-head* possibilitou maior aprendizado de fatores distintivos nas narrativas sinceras. Diferentemente dos resultados dos AEs, as diversas métricas do erro de reconstrução mostram-se mais homogêneas.

Neste caso, KLd e MSE alcançaram altos níveis de desempenho, significando que as discrepâncias entre os erros de reconstrução em treinamento e em teste tiveram foram capturados tanto no volume quanto na distribuição, sendo esta última aparentemente não fortemente similar à curva normal.

QUADRO 18 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS ACÚSTICAS DE AUTOENCODERS ATENCIONAIS *SINGLE-HEAD* USANDO AS MÉTRICA KL:5, KL:10, KL:15, KL:20 E KL:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AAE_4-65_65-69-65_65-mse-32	0,735	0,733	0,735	0,731	0,694	0,689	0,673	0,668	0,653	0,648
AAE_4-65_65-65-65_65-mse-256	0,592	0,585	0,592	0,585	0,592	0,585	0,592	0,585	0,592	0,585
AAE_4-65_63_61-59-61_63_65-mse-256	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-65_63-65-63_65-mse-256	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-65_63-67-63_65-mse-32	0,633	0,630	0,612	0,609	0,633	0,629	0,633	0,628	0,612	0,608
AAE_4-65_67-63-67_65-mse-256	0,592	0,585	0,592	0,585	0,592	0,585	0,592	0,585	0,592	0,585
AAE_4-65_63-63-63_65-mse-32	0,694	0,689	0,673	0,668	0,673	0,668	0,673	0,668	0,653	0,648
AAE_4-65_67-65-67_65-mse-256	0,490	0,485	0,551	0,545	0,551	0,545	0,551	0,545	0,571	0,565
AAE_4-65-67-65-mse-256	0,388	0,384	0,388	0,383	0,408	0,403	0,490	0,484	0,490	0,484
AAE_4-65_02_65_02-69-65_02_65_02-mse-32	0,673	0,668	0,673	0,668	0,673	0,668	0,653	0,648	0,633	0,627

FONTE: DADOS DA PESQUISA (2023)

Neste caso os piores desempenhos ficaram polarizados na métrica KLn (um efeito reverso em relação aos AEs), enquanto o melhor foi capturado pela KL:5. As diversas granularidades capturadas pela KLd resultaram em variados níveis de acurácia balanceada.

QUADRO 19 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS ACÚSTICAS DE AUTOENCODERS ATENCIONAIS *SINGLE-HEAD* USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AAE_4-65_65-69-65_65-mse-32	0,633	0,627	0,633	0,627	0,633	0,627	0,510	0,516	0,673	0,675
AAE_4-65_65-65-65_65-mse-256	0,592	0,585	0,592	0,585	0,592	0,585	0,490	0,492	0,714	0,713
AAE_4-65_63_61-59-61_63_65-mse-256	0,612	0,606	0,612	0,606	0,612	0,606	0,633	0,635	0,714	0,713
AAE_4-65_63-65-63_65-mse-256	0,612	0,606	0,612	0,606	0,612	0,606	0,449	0,453	0,694	0,697
AAE_4-65_63-67-63_65-mse-32	0,633	0,627	0,612	0,606	0,612	0,606	0,469	0,475	0,694	0,694
AAE_4-65_67-63-67_65-mse-256	0,592	0,585	0,592	0,585	0,592	0,585	0,429	0,431	0,694	0,693
AAE_4-65_63-63-63_65-mse-32	0,633	0,626	0,612	0,606	0,612	0,606	0,469	0,475	0,633	0,633
AAE_4-65_67-65-67_65-mse-256	0,571	0,565	0,592	0,586	0,592	0,586	0,449	0,453	0,673	0,674
AAE_4-65-67-65-mse-256	0,531	0,525	0,531	0,525	0,531	0,525	0,327	0,329	0,673	0,673
AAE_4-65_02_65_02-69-65_02_65_02-mse-32	0,633	0,627	0,633	0,627	0,633	0,627	0,388	0,394	0,673	0,672

FONTE: DADOS DA PESQUISA (2023)

Dentre os melhores modelos atencionais *single-head*, apenas um deles foi melhorado pela aplicação de *dropout*.

O terceiro conjunto de experimentos incluiu os Autoencoders atencionais *multi-head* (MAAE). Foram 826 experimentos ao todo, com 210 sem a aplicação de *dropout* e 616 com a regularização.

QUADRO 20 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS ACÚSTICAS DE AUTOENCODERS ATENCIONAIS *MULTI-HEAD* USANDO AS MÉTRICA KL:5, KL:10, KL:15, KLD:20 E KLD:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-65_67-71-67_65-mse-32-65-65-5	0,633	0,631	0,653	0,649	0,633	0,629	0,612	0,609	0,633	0,629
MAAE-81_02_93_03-63-93_03_81_02-mse-32-65-65-5	0,694	0,691	0,755	0,753	0,735	0,733	0,735	0,733	0,694	0,693
MAAE-81_02_93_03-69-93_03_81_02-mse-32-65-65-10	0,735	0,733	0,755	0,753	0,735	0,733	0,735	0,733	0,694	0,693
MAAE-65_63_61-57-61_63_65-mse-256-65-65-20	0,551	0,548	0,571	0,568	0,571	0,567	0,571	0,566	0,551	0,545
MAAE-65_67-67-67_65-mse-32-65-65-5	0,510	0,512	0,449	0,451	0,429	0,433	0,469	0,474	0,490	0,494
MAAE-65_63-65-63_65-mse-32-65-65-5	0,510	0,513	0,531	0,534	0,510	0,514	0,490	0,494	0,449	0,453
MAAE-65-65-65-mse-256-65-65-10	0,449	0,453	0,388	0,393	0,367	0,373	0,388	0,394	0,408	0,415
MAAE-65_63_61-61-61_63_65-mse-256-65-65-5	0,551	0,545	0,592	0,585	0,612	0,606	0,592	0,585	0,571	0,564
MAAE-65_02_63_02_61_02-61-61_02_63_02_65_02-mse-32-65-65-10	0,633	0,627	0,653	0,648	0,633	0,628	0,612	0,608	0,612	0,606
MAAE-65_65_65-63-65_65_65-mse-256-65-65-10	0,490	0,484	0,571	0,565	0,571	0,565	0,592	0,586	0,612	0,607

FONTE: DADOS DA PESQUISA (2023)

Os melhores desempenhos dos modelos atencionais *multi-head* ficaram distribuídos em diversas configurações de KLd e, em geral, KLn não apresentou bons resultados. Mesmo MSE, em geral, propiciou melhores desempenhos.

QUADRO 21 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS ACÚSTICAS DE AUTOENCODERS ATENCIONAIS *MULTI-HEAD* USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-65_67-71-67_65-mse-32-65-65-5	0,653	0,649	0,633	0,629	0,653	0,649	0,531	0,537	0,755	0,758
MAAE-81_02_93_03-63-93_03_81_02-mse-32-65-65-5	0,673	0,672	0,673	0,672	0,673	0,672	0,592	0,598	0,612	0,611
MAAE-81_02_93_03-69-93_03_81_02-mse-32-65-65-10	0,694	0,693	0,673	0,672	0,612	0,609	0,551	0,554	0,633	0,631
MAAE-65_63_61-57-61_63_65-mse-256-65-65-20	0,571	0,566	0,612	0,607	0,612	0,607	0,653	0,657	0,735	0,737
MAAE-65_67-67-67_65-mse-32-65-65-5	0,469	0,473	0,469	0,473	0,490	0,493	0,510	0,515	0,735	0,736
MAAE-65_63-65-63_65-mse-32-65-65-5	0,449	0,454	0,449	0,453	0,449	0,453	0,469	0,473	0,714	0,718
MAAE-65-65-65-mse-256-65-65-10	0,449	0,456	0,429	0,435	0,429	0,435	0,531	0,537	0,714	0,717
MAAE-65_63_61-61-61_63_65-mse-256-65-65-5	0,551	0,544	0,531	0,524	0,571	0,565	0,653	0,655	0,714	0,717
MAAE-65_02_63_02_61_02-61-61_02_63_02_65_02-mse-32-65-65-10	0,612	0,606	0,592	0,586	0,571	0,565	0,714	0,716	0,551	0,550
MAAE-65_65_65-63-65_65_65-mse-256-65-65-10	0,612	0,607	0,592	0,586	0,592	0,586	0,612	0,614	0,714	0,715

FONTE: DADOS DA PESQUISA (2023)

A Atenção *multi-head* propiciou modelos ainda superiores aos *single-head*, mas por margem menos expressiva. A distribuição de desempenhos apresentou maiores variações em relação às arquiteturas dos modelos, quando comparado com os modelos com Atenção *single-head*.

Os níveis mais elevados de desempenhos foram alcançados com os modelos atencionais. Os resultados dos 1.100 modelos acústicos (AE, AAE e MAAE) foram avaliados em conjunto para produzir os cinco modelos de melhor desempenho dentre todos os experimentos. Estes modelos podem ser vistos no Quadro 22.

QUADRO 22 – CINCO MELHORES MODELOS ACÚSTICOS DENTRE AE, AAE E MAAE

Modelo				Contagens			
#	**Arquitetura**	**Métrica**	**B Acc**	**V S**	**V NS**	**F S**	**F NS**
102	MAAE-65_67-71-67_65-mse-32-65-65-5	MSE	0,758	15	22	2	10
573	MAAE-81_02_93_03-63-93_03_81_02-mse-32-65-65-5	KLd:10	0,753	22	15	9	3
583	MAAE-81_02_93_03-69-93_03_81_02-mse-32-65-65-10	KLd:10	0,753	22	15	9	3
167	MAAE-65_63_61-57-61_63_65-mse-256-65-65-20	MSE	0,737	16	20	4	9
96	MAAE-65_67-67-67_65-mse-32-65-65-5	MSE	0,736	17	19	5	8

FONTE: DADOS DA PESQUISA (2023)

A coluna “Acc.” Informa a acurácia atingida pelo modelo, a coluna “B Acc” a acurácia balanceada, a coluna “V S” informa a contagem de sinceridades verdadeiras detectadas, a coluna “V NS” a contagem de não sinceridades verdadeiras, a coluna “F S” a contagem de sinceridades falsas e a coluna “F NS” a contagem de não sinceridades falsas. A coluna “#” identifica o índice do modelo dentro do conjunto de 1.100 experimentos e a coluna “Arquitetura” apresenta a assinatura do modelo.

O melhor desempenho foi atingido pela métrica MSE, seguido por KLd:10. O baixo desempenho de KLn como métrica de erro de reconstrução sugere que as distribuições não se assemelham a uma distribuição normal, tornando o critério impreciso para este conjunto de características.

No Quadro **23** estão listados os tempos envolvidos nos processos de treinamento de cada tipo de modelo.

QUADRO 23 – TEMPOS CONSUMIDOS PELOS TREINAMENTOS DOS MODELOS ACÚSTICOS INDIVIDUAIS

Modelo	Quantidade	Tempo médio (s)	Tempo total
AE	154	78.303 ± 56.172	2 dias, 6:48:42.802
AAE	130	114.230 ± 795.936	2 dias, 12:55:21.180
MAAE	826	137.181 ± 143.442	11 dias, 10:21:43.366
		Total	16 dias, 6:05:47.348

FONTE: DADOS DA PESQUISA (2023)

Atenção para o fato de que cada configuração de modelo testada foi treinada para todos os 12 sujeitos, visto que são modelos individuais. Assim, uma unidade de um experimento implica no treinamento de 12 diferentes modelos, um para cada sujeito. Os conjuntos de treinamento de cada sujeito apresentavam cardinalidades

diferentes, pois são características extraídas de narrativas. Como consequência, os tempos de treinamento de um mesmo modelo para os dados de diferentes sujeitos sofreram de grande variância, o que justifica os valores elevados para o desvio padrão calculado.

A partir dos resultados globais, é possível perceber que o mecanismo de Atenção aprimorou o desempenho dos modelos, pois a arquitetura com mais alto desempenho foi "MAAE-65_67-71-67_65-mse-32-65-65-5".

Quanto aos diferentes modelos experimentados, é interessante perceber que cada métrica apresentou um desempenho superior em diferentes tipos de Autoencoder. As métricas abordam a questão do erro de reconstrução de forma diferente, pois os modelos com maiores desempenhos medidos a partir de cada métrica são diferentes. Com frequência um modelo com mais alto desempenho a partir de uma métrica (por exemplo KLd:15) tem desempenho significativamente mais baixo pelo cálculo de outra métrica (MSE).

Enquanto a KLd e a KLn avaliam diferenças na distribuição dos dados reconstruídos em relação aos dados de entrada, a MSE avalia a diferença numérica entre os dados. A julgar pelos resultados alcançados, esta forma diferente de encarar os dados reconstruídos promove percepções diferentes o suficiente para alterar de forma bastante crítica (algumas vezes invertendo) o resultado de uma detecção.

O fato de o Autoencoder atencional *multi-head* superar os outros tipos aponta para a existência de relações complexas e de longa distância entre indivíduos do conjunto de dados. Pondo de uma forma diferente, significa que as características de sinceridade do áudio de um sujeito em um quadro (intervalo de 10 milissegundos) guardam algum grau de dependência de outro quadro (não necessariamente adjacente) e que esta relação se modifica em situações de não sinceridade.

É uma evidência de que os parâmetros de voz evoluem de forma diferente em uma narrativa não sincera, quando comparadas com uma narrativa sincera. No caso do modelo de melhor desempenho, o último parâmetro de sua assinatura é o valor "5", que corresponde à quantidade de amostras consideradas (a maior distância considerada para identificar as relações de longa distância). Em outras palavras, para detectar a disparidade na voz do sujeito foi necessária uma janela de 0,05 segundos (50 milissegundos).

4.3.2 Modelos acústicos coletivos

Os modelos acústicos coletivos foram treinados a partir das arquiteturas dos cinco modelos individuais de melhor desempenho (Quadro 22). Nenhum dos hiperparâmetros foi alterado, apenas os dados de entrada, que foram compostos por todas as narrativas sinceras de 11 dos 12 sujeitos, como descrito no protocolo de experimentação. Os resultados podem ser observados no Quadro 24 e no Quadro 25.

QUADRO 24 – DESEMPENHO DAS CINCO MELHORES ARQUITETURAS ACÚSTICAS COLETIVAS DE AUTOENCODERS USANDO AS MÉTRICA KL:5, KL:10, KL:15, KLD:20 E KLD:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-65_67-67-67_65-mse-32-65-65-5	0,551	0,543	0,551	0,543	0,571	0,563	0,571	0,563	0,571	0,563
MAAE-65_67-71-67_65-mse-32-65-65-5	0,551	0,543	0,551	0,542	0,531	0,521	0,531	0,521	0,551	0,542
MAAE-65_63_61-57-61_63_65-mse-256-65-65-20	0,469	0,462	0,531	0,522	0,551	0,543	0,531	0,522	0,531	0,522
MAAE-81_02_93_03-63-93_03_81_02-mse-32-65-65-5	0,347	0,346	0,388	0,383	0,510	0,502	0,531	0,521	0,510	0,500
MAAE-81_02_93_03-63-93_03_81_02-mse-32-65-65-10	0,490	0,480	0,490	0,480	0,490	0,480	0,490	0,480	0,510	0,500

FONTE: DADOS DA PESQUISA (2023)

Os modelos coletivos atingiram níveis de desempenho mais baixos que os modelos individuais. Uma possível justificativa para este efeito é a quantidade de indivíduos que fizeram parte de cada um dos modelos coletivos (11 diferentes sujeitos). Com poucos indivíduos, existe a chance de haver muitos padrões específicos (padrões individuais) e poucos genéricos (padrões de grupo), dificultando o aprendizado e, consequentemente, reduzindo a qualidade da detecção.

QUADRO 25 – DESEMPENHO DAS CINCO MELHORES ARQUITETURAS ACÚSTICAS COLETIVAS DE AUTOENCODERS USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-65_67-67-67_65-mse-32-65-65-5	0,571	0,563	0,571	0,563	0,571	0,563	0,490	0,500	0,490	0,500
MAAE-65_67-71-67_65-mse-32-65-65-5	0,551	0,542	0,551	0,542	0,551	0,542	0,490	0,500	0,490	0,500
MAAE-65_63_61-57-61_63_65-mse-256-65-65-20	0,551	0,542	0,510	0,500	0,510	0,500	0,469	0,478	0,490	0,500
MAAE-81_02_93_03-63-93_03_81_02-mse-32-65-65-5	0,510	0,500	0,510	0,500	0,510	0,500	0,490	0,500	0,490	0,500
MAAE-81_02_93_03-63-93_03_81_02-mse-32-65-65-10	0,510	0,500	0,510	0,500	0,510	0,500	0,490	0,500	0,490	0,500

FONTE: DADOS DA PESQUISA (2023)

Outra possibilidade é que os padrões de sinceridade (padrões de normalidade) sejam muito discrepantes de sujeito para sujeito, além de se confundirem com padrões de anormalidade (não sinceridade) dentro do grupo. Nesta eventualidade, significaria dizer que de fato as expressões de sinceridade variam de

forma expressiva de sujeito para sujeito, corroborando a noção de que os caracteres peculiares de expressão de uma pessoa modificam os tipos de pistas que podem ser observadas.

Visto que os conjuntos de treinamento cresceram, os tempos de treinamento também cresceram de forma expressiva, como pode ser observado no Quadro 26.

QUADRO 26 – TEMPOS CONSUMIDOS PELOS TREINAMENTOS DOS MODELOS ACÚSTICOS COLETIVOS

Modelo	Quantidade	Tempo médio (s)	Tempo total
Coletivo	5	2781.095 ± 722.130	1 dia 22:21:05.724

FONTE: DADOS DA PESQUISA (2023)

Para possibilitar a exploração mais ampla do espaço de hiperparâmetros para modelos coletivos, seria necessário mais tempo ou um hardware (GPU) mais potente.

4.3.3 Modelo acústico bicomponente

Com os modelos individuais e coletivos treinados, foi possível construir um modelo acústico bicomponente. Os componentes foram os modelos individuais de cada sujeito e os modelos coletivos que excluem aquele mesmo sujeito, lembrando que estes últimos modelos foram treinados para avaliar a capacidade de um modelo coletivo distinguir sinceridade de não sinceridade de um sujeito desconhecido (generalização). No Quadro 27 está apresentado o desempenho combinado dos melhores modelos acústicos individual e coletivo.

QUADRO 27 – DESEMPENHO COMBINADO DOS MODELOS ACÚSTICOS INDIVIDUAL E COLETIVO

Arquitetura	Card.	Métrica	Act	Acc	B Acc	V S	V NS	F S	F NS
MAAE-65_67-71-67_65-mse-32-65-65-5	Ind	MSE	37	0,755	0,758	15	22	2	10
MAAE-65_67-67-67_65-mse-32-65-65-5	Col	KLd:15	26	0,531	0,525	20	6	18	5
		Final	27	0,551	0,547	19	8	16	6

FONTE: DADOS DA PESQUISA (2023)

A coluna “Card.” Indica cardinalidade individual (“Ind”) ou coletiva (“Col”), ao passo que a coluna “Act” contém a contagem de acertos. O processo de fusão foi o mesmo determinado para os modelos multimodais, ou seja, pelo uso da soma ajustada dos Escores de Sinceridade parcial de cada um dos componentes. A combinação dos dois modelos não resultou em aprimoramento da capacidade de detecção, ao contrário, a reduziu.

No Quadro 28 é possível observar um detalhamento da combinação dos dois modelos ao nível de sujeito. Os valores rotulados como “S” indicam classificação da narrativa como “sincera”, enquanto “NS” identifica narrativa “não sincera”.

QUADRO 28 – COMPARATIVO DETALHADO DOS DESEMPENHOS DE MODELOS ACÚSTICOS INDIVIDUAL E COLETIVO

Referência	Teste		MAAE-65_67-71-67_65-mse-32-65-65-5 Individual		MAAE-65_67-67-67_65-mse-32-65-65-5 Coletivo		Final		
			MSE		KLd:15				
	Sujeito	Esperado	Escore	Detectado	Escore	Detectado	Soma bruta	Escore	Detectado
S1-P7-13	S1-P7-1	NS	-0,031	NS	0,913	S	0,708	0,609	S
	S1-P7-2	S	0,213	S	0,950	S	0,822	0,676	S
	S1-P7-3	S	0,045	S	0,879	S	0,728	0,622	S
	S1-P7-4	NS	-0,044	NS	0,788	S	0,631	0,559	S
	S1-P7-5	S	-0,137	NS	0,783	S	0,569	0,515	S
	S1-P7-6	NS	-0,046	NS	0,113	S	0,067	0,067	S
	S1-P7-7	S	0,024	S	0,972	S	0,760	0,641	S
	S1-P7-8	NS	-0,041	NS	0,887	S	0,689	0,597	S
	S1-P7-9	NS	-0,035	NS	0,878	S	0,687	0,596	S
	S1-P7-10	S	-0,096	NS	0,879	S	0,654	0,575	S
	S1-P7-11	S	0,331	S	0,715	S	0,780	0,653	S
	S1-P7-12	S	0,201	S	0,671	S	0,702	0,606	S
S1-P8-3	S1-P8-1	S	0,363	S	-0,107	NS	0,251	0,246	S
	S1-P8-2	S	0,082	S	-0,054	NS	0,028	0,028	S
	S1-P8-4	S	0,156	S	0,225	S	0,364	0,348	S
	S1-P8-5	S	0,008	S	-0,204	NS	-0,194	-0,191	NS
	S1-P8-6	S	0,069	S	0,405	S	0,442	0,415	S
	S1-P8-7	S	0,072	S	0,588	S	0,578	0,522	S
S1-P9-11	S1-P9-1	NS	-0,009	NS	0,561	S	0,503	0,464	S
	S1-P9-2	NS	-0,296	NS	0,835	S	0,492	0,456	S
	S1-P9-3	S	-0,040	NS	0,941	S	0,717	0,615	S
	S1-P9-5	NS	-0,026	NS	-0,576	NS	-0,538	-0,492	NS
	S1-P9-6	NS	-0,310	NS	0,671	S	0,346	0,333	S
	S1-P9-8	NS	0,034	S	0,107	S	0,141	0,140	S
	S1-P9-9	S	-0,053	NS	0,405	S	0,339	0,326	S
	S1-P9-10	S	0,088	S	0,328	S	0,393	0,374	S
S2-P1-1	S2-P1-2	S	-0,231	NS	0,323	S	0,092	0,092	S
	S2-P1-3	S	-0,851	NS	0,471	S	-0,363	-0,348	NS
S2-P2-5	S2-P2-1	NS	-0,545	NS	0,126	S	-0,396	-0,377	NS
	S2-P2-2	NS	-0,630	NS	0,633	S	0,003	0,003	S
	S2-P2-3	NS	-0,543	NS	-0,454	NS	-0,760	-0,641	NS
S2-P3-5	S2-P3-1	NS	-0,137	NS	-0,151	NS	-0,280	-0,273	NS
	S2-P3-2	NS	-0,638	NS	-0,353	NS	-0,758	-0,640	NS
	S2-P3-3	NS	-0,119	NS	-0,869	NS	-0,756	-0,639	NS

	S2-P3-4	NS	-0,414	NS	0,556	S	0,141	0,140	S
S2-P4-2	S2-P4-1	S	-0,447	NS	0,182	S	-0,259	-0,254	NS
S2-P5-3	S2-P5-1	NS	-0,313	NS	0,514	S	0,199	0,196	S
	S2-P5-2	S	-0,885	NS	0,222	S	-0,581	-0,523	NS
S2-P6-5	S2-P6-1	NS	-0,015	NS	0,056	S	0,041	0,041	S
	S2-P6-2	NS	-0,181	NS	-0,448	NS	-0,557	-0,506	NS
S2-P7-8	S2-P7-1	NS	0,036	S	0,982	S	0,769	0,646	S
	S2-P7-4	NS	-0,248	NS	0,983	S	0,626	0,555	S
	S2-P7-5	NS	-0,299	NS	0,883	S	0,525	0,482	S
	S2-P7-6	S	-0,035	NS	0,987	S	0,741	0,630	S
	S2-P7-7	S	0,294	S	0,986	S	0,856	0,694	S
S2-P8-5	S2-P8-1	NS	-0,758	NS	0,576	S	-0,180	-0,178	NS
S2-P9-4	S2-P9-1	S	0,072	S	-0,984	NS	-0,722	-0,618	NS
	S2-P9-2	S	0,013	S	0,683	S	0,601	0,538	S
	S2-P9-3	S	-0,014	NS	-0,963	NS	-0,752	-0,636	NS

FONTE: DADOS DA PESQUISA (2023)

As células foram coloridas para facilitar a identificação das detecções frente ao resultado esperado. As células em vermelho identificam “não sinceridade”, enquanto as verdes, “sinceridade”.

O modelo coletivo parece ter uma tendência de classificar mais casos como sinceros, com confiança elevada, enquanto o modelo individual classificou corretamente alguns casos de não sinceridade, mas com confiança mais baixa, o que veio a inverter incorretamente a classificação. Tais resultados sugerem que os modelos coletivos de sinceridade são compostos por padrões que se assemelham bastante aos padrões de não sinceridade do sujeito testado. Esses padrões possivelmente abrangem uma região grande do espaço de características onde também se encontram os padrões de não sinceridade do sujeito não incluído.

É observável também que alguns casos de sinceridade esperada foram classificados como não sinceros pelo modelo coletivo, levando a ainda mais erros na classificação final.

Considerando os resultados, foi possível divisar a hipótese de que o conjunto de narrativas sinceras que constituíram os modelos de sinceridade coletivos eram muito diversas e em número muito pequeno para a produção de padrões significativos de grupo.

Outra hipótese foi de que os padrões de sinceridade individuais efetivamente eram mais discriminantes que os padrões coletivos, o que está em concordância com o pressuposto da pesquisa (alto potencial discriminante das características

idiossincráticas). As duas hipóteses não são mutuamente excludentes, havendo a possibilidade da ocorrência conjugada de ambas na produção do resultado.

4.3.4 Modelos verbais individuais

Os modelos verbais foram alimentados com características extraídas a partir das transcrições das narrativas e de seus pós-processamentos para incluir 9 características linguísticas geradas pelo SpaCy (função sintática, pessoa verbal e detecção de entidade), duas características paralinguísticas (hesitação e duração, incluídas manualmente) e duas características de sentimento (intensidades positiva e negativa, derivados do SentiWordNet-PT-BR).

Todas as 13 características foram consideradas, implicando que todos os modelos tiveram 13 neurônios de entrada e 13 de saída, com variadas configurações de camadas escondidas, conforme estabelecido no protocolo de experimentação. A calibragem inicial mostrou que 500 épocas seriam mais que suficientes para que os modelos alcançassem a estabilidade no aprendizado. A maioria dos modelos atingiu a estabilidade com aproximadamente 200 épocas.

Modelos Autoencoder Vanilla foram testados em 90 diferentes configurações, sendo 35 sem e 65 com a aplicação de *dropout*. Os 10 melhores resultados alcançados podem ser vistos tanto no Quadro 29 quanto no Quadro 30 (continuação).

QUADRO 29 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VERBAIS DE AUTOENCODERS VANILLA USANDO AS MÉTRICA KL:5, KL:10, KL:15, KL:20 E KL:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AE-13_15-17-15_13-mse-32	0,449	0,455	0,408	0,413	0,388	0,392	0,367	0,372	0,408	0,413
AE-13_11-11-11_13-mse-32	0,429	0,434	0,429	0,435	0,490	0,497	0,429	0,435	0,469	0,475
AE-13_02_13_02_13_02-15-13_02_13_02_13_02-mse-32	0,408	0,413	0,469	0,474	0,367	0,371	0,388	0,392	0,388	0,392
AE-16_02_16_03_15_04-9-15_04_16_03_16_02-mse-32	0,449	0,453	0,367	0,369	0,408	0,411	0,449	0,450	0,469	0,470
AE-16_02_16_03_15_04-5-15_04_16_03_16_02-mse-32	0,429	0,433	0,388	0,391	0,367	0,371	0,306	0,309	0,388	0,391
AE-13-13-13-mse-32	0,429	0,434	0,449	0,455	0,449	0,456	0,408	0,413	0,449	0,455
AE-16_02_19_03_22_04-17-22_04_19_03_16_02-mse-32	0,408	0,413	0,388	0,393	0,388	0,391	0,429	0,432	0,449	0,453
AE-13_11-13-11_13-mse-32	0,367	0,372	0,327	0,330	0,367	0,372	0,388	0,393	0,449	0,455
AE-13_13-13-13_13-mse-32	0,449	0,455	0,449	0,454	0,429	0,433	0,429	0,433	0,388	0,393
AE-16_02_19_03_22_04-9-22_04_19_03_16_02-mse-32	0,490	0,493	0,388	0,389	0,388	0,386	0,347	0,345	0,327	0,324

FONTE: DADOS DA PESQUISA (2023)

Os desempenhos estão expressos em acurácia (“Acc”) e acurácia balanceada (“B Acc”) e a coluna “Modelo” apresenta as assinaturas dos modelos experimentados.

As células mais escuras apresentam os desempenhos mais elevados.

QUADRO 30 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VERBAIS DE AUTOENCODERS VANILLA USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AE-13_15-17-15_13-mse-32	0,449	0,453	0,510	0,515	0,490	0,494	0,653	0,656	0,531	0,538
AE-13_11-11-11_13-mse-32	0,429	0,435	0,429	0,435	0,408	0,413	0,612	0,618	0,490	0,498
AE-13_02_13_02_13_02-15-13_02_13_02_13_02-mse-32	0,388	0,392	0,367	0,371	0,367	0,371	0,490	0,496	0,612	0,618
AE-16_02_16_03_15_04-9-15_04_16_03_16_02-mse-32	0,449	0,449	0,469	0,470	0,469	0,470	0,571	0,575	0,612	0,618
AE-16_02_16_03_15_04-5-15_04_16_03_16_02-mse-32	0,347	0,351	0,367	0,371	0,347	0,350	0,612	0,616	0,592	0,596
AE-13-13-13-mse-32	0,367	0,373	0,408	0,413	0,388	0,393	0,592	0,598	0,490	0,498
AE-16_02_19_03_22_04-17-22_04_19_03_16_02-mse-32	0,449	0,453	0,449	0,453	0,469	0,473	0,510	0,516	0,592	0,598
AE-13_11-13-11_13-mse-32	0,490	0,496	0,449	0,455	0,429	0,435	0,592	0,597	0,490	0,499
AE-13_13-13-13_13-mse-32	0,408	0,413	0,408	0,413	0,408	0,413	0,592	0,596	0,510	0,519
AE-16_02_19_03_22_04-9-22_04_19_03_16_02-mse-32	0,408	0,405	0,388	0,385	0,429	0,425	0,469	0,477	0,592	0,595

FONTE: DADOS DA PESQUISA (2023)

Os resultados mostram que os níveis mais altos de acurácia balanceada ficaram polarizadas nas métricas KLn e MSE. Embora vários modelos com *dropout* tenham produzidos bons resultados, os melhores não o utilizaram.

O segundo conjunto de modelos testado foi dos Autoencoders atencionais *single-head* (AAE). Foram ao todo 70 experimentos (35 sem e 35 com *dropout*). Tanto no Quadro 31 quanto no Quadro 32 apresentam-se os 10 níveis de desempenho mais elevados alcançados com diferentes modelos atencionais *single-head*.

QUADRO 31 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VERBAIS DE AUTOENCODERS ATENCIONAIS *SINGLE-HEAD* USANDO AS MÉTRICA KL:5, KL:10, KL:15, KL:20 E KL:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AAE_4-13_02_13_02-9-13_02_13_02-mse-32	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-13_02-9-13_02-mse-32	0,633	0,630	0,592	0,588	0,571	0,568	0,571	0,568	0,571	0,568
AAE_4-13_15_17-19-17_15_13-mse-32	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-16_02_19_03-11-19_03_16_02-mse-32	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-13_13-13-13_13-mse-32	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-13-9-13-mse-32	0,551	0,549	0,571	0,569	0,571	0,569	0,571	0,569	0,571	0,569
AAE_4-13-13-13-mse-32	0,633	0,629	0,571	0,568	0,612	0,608	0,612	0,608	0,592	0,588
AAE_4-13_15-13-15_13-mse-32	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-13_02_15_02_17_02-13-17_02_15_02_13_02-mse-32	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606
AAE_4-13_02_11_02-7-11_02_13_02-mse-32	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606	0,612	0,606

FONTE: DADOS DA PESQUISA (2023)

Neste caso os melhores desempenhos ficaram polarizados na métrica KLn, com variações diversas e próximas nas outras métricas. As diversas granularidades capturadas pela KLd resultaram em variados níveis de acurácia balanceada, mas inferiores.

A exemplo que observado nos modelos acústicos, a Atenção *single-head* aprimorou a capacidade dos modelos de aprender a distinção entre sinceridade e não sinceridade, mas de maneira menos expressiva, embora as avaliações mostrem-se bem mais homogêneas.

QUADRO 32 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VERBAIS DE AUTOENCODERS ATENCIONAIS *SINGLE-HEAD* USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AAE_4-13_02_13_02-9-13_02_13_02-mse-32	0,612	0,606	0,612	0,606	0,592	0,585	0,673	0,679	0,510	0,518
AAE_4-13_02-9-13_02-mse-32	0,571	0,568	0,571	0,568	0,571	0,568	0,653	0,652	0,531	0,537
AAE_4-13_15_17-19-17_15_13-mse-32	0,612	0,606	0,612	0,606	0,592	0,585	0,653	0,651	0,612	0,617
AAE_4-16_02_19_03-11-19_03_16_02-mse-32	0,612	0,606	0,612	0,606	0,592	0,585	0,633	0,639	0,510	0,517
AAE_4-13_13-13-13_13-mse-32	0,612	0,606	0,612	0,606	0,592	0,585	0,633	0,638	0,531	0,534
AAE_4-13-9-13-mse-32	0,571	0,569	0,571	0,569	0,551	0,548	0,612	0,612	0,633	0,637
AAE_4-13-13-13-mse-32	0,612	0,608	0,612	0,608	0,612	0,608	0,551	0,556	0,408	0,413
AAE_4-13_15-13-15_13-mse-32	0,612	0,606	0,612	0,606	0,592	0,585	0,469	0,472	0,612	0,619
AAE_4-13_02_15_02_17_02-13-17_02_15_02_13_02-mse-32	0,612	0,606	0,612	0,606	0,592	0,585	0,449	0,453	0,612	0,618
AAE_4-13_02_11_02-7-11_02_13_02-mse-32	0,612	0,606	0,612	0,606	0,592	0,585	0,612	0,618	0,531	0,538

FONTE: DADOS DA PESQUISA (2023)

Para o modelo atencional *single-head*, nem todos os modelos foram melhorados pela aplicação de *dropout*. De forma geral estes modelos resultaram em acurácias balanceadas mais elevadas que as dos modelos Vanilla, o que parece indicar o efeito benéfico do mecanismo de Atenção no aprendizado das características de sinceridade e na sua capacidade de distinção de narrativas não sinceras.

O terceiro conjunto de experimentos incluiu os Autoencoders atencionais *multi-head* (MAAE). Foram 300 experimentos ao todo, com 140 sem a aplicação de *dropout* e 160 com. Os dez melhores resultados foram alcançados pelos modelos apresentados tanto no Quadro 33 quanto no Quadro 34.

QUADRO 33 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VERBAIS DE AUTOENCODERS ATENCIONAIS *MULTI-HEAD* USANDO AS MÉTRICA KL:5, KL:10, KL:15, KLD:20 E KLD:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10	0,367	0,369	0,449	0,449	0,490	0,489	0,510	0,509	0,531	0,529
MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-15	0,694	0,698	0,592	0,595	0,592	0,595	0,592	0,595	0,592	0,595
MAAE-13_15_17-19-17_15_13-mse-32-13-13-5	0,490	0,494	0,551	0,553	0,551	0,552	0,510	0,510	0,469	0,468
MAAE-13_02_11_02-15-11_02_13_02-mse-32-13-13-15	0,551	0,558	0,571	0,578	0,592	0,598	0,633	0,638	0,633	0,638
MAAE-13_11-9-11_13-mse-32-13-13-15	0,510	0,516	0,571	0,576	0,551	0,556	0,510	0,514	0,531	0,534
MAAE-13_11_9-7-9_11_13-mse-32-13-13-15	0,551	0,558	0,592	0,597	0,612	0,618	0,612	0,617	0,612	0,617
MAAE-13_02_15_02_17_02-17-17_02_15_02_13_02-mse-32-13-13-15	0,531	0,536	0,592	0,597	0,551	0,555	0,531	0,534	0,531	0,534
MAAE-13_13_13-15-13_13_13-mse-32-13-13-15	0,490	0,494	0,490	0,494	0,571	0,574	0,551	0,554	0,571	0,574
MAAE-13_02_11_02_9_02-5-9_02_11_02_13_02-mse-32-13-13-15	0,510	0,516	0,551	0,555	0,510	0,513	0,510	0,513	0,551	0,554
MAAE-13_15-17-15_13-mse-32-13-13-15	0,633	0,639	0,592	0,596	0,592	0,596	0,612	0,617	0,612	0,617

FONTE: DADOS DA PESQUISA (2023)

Os melhores desempenhos ficaram distribuídos preponderantemente na métrica MSE, com alguns resultados notáveis em diversas configurações de KLd. Com uma exceção, KLn não apresentou bons resultados.

QUADRO 34 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VERBAIS DE AUTOENCODERS ATENCIONAIS *MULTI-HEAD* USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10	0,531	0,529	0,551	0,550	0,551	0,550	0,510	0,518	0,714	0,717
MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-15	0,592	0,595	0,592	0,595	0,592	0,595	0,469	0,476	0,592	0,596
MAAE-13_15_17-19-17_15_13-mse-32-13-13-5	0,449	0,447	0,429	0,425	0,469	0,465	0,510	0,520	0,653	0,659
MAAE-13_02_11_02-15-11_02_13_02-mse-32-13-13-15	0,653	0,658	0,653	0,658	0,653	0,658	0,469	0,476	0,551	0,555
MAAE-13_11-9-11_13-mse-32-13-13-15	0,531	0,534	0,531	0,534	0,551	0,554	0,449	0,456	0,653	0,657
MAAE-13_11_9-7-9_11_13-mse-32-13-13-15	0,592	0,595	0,653	0,657	0,612	0,616	0,469	0,477	0,571	0,575
MAAE-13_02_15_02_17_02-17-17_02_15_02_13_02-mse-32-13-13-15	0,531	0,534	0,531	0,534	0,531	0,534	0,449	0,454	0,653	0,657
MAAE-13_13_13-15-13_13_13-mse-32-13-13-15	0,571	0,574	0,571	0,574	0,571	0,574	0,510	0,518	0,653	0,656
MAAE-13_02_11_02_9_02-5-9_02_11_02_13_02-mse-32-13-13-15	0,531	0,534	0,531	0,534	0,531	0,534	0,592	0,596	0,653	0,655
MAAE-13_15-17-15_13-mse-32-13-13-15	0,612	0,617	0,612	0,617	0,612	0,617	0,429	0,436	0,510	0,513

FONTE: DADOS DA PESQUISA (2023)

Os níveis mais elevados de desempenho foram alcançados com os modelos atencionais *multi-head*. Os resultados dos modelos AE, AAE e MAAE combinados foram avaliados (somando 470 experimentos) para produzir os cinco modelos de mais elevada acurácia balanceada. Estes modelos podem ser vistos no Quadro 35.

QUADRO 35 – CINCO MELHORES MODELOS VERBAIS DENTRE AS ARQUITETURAS AE, AAE E MAAE

Modelo		Melhores			KLd:5		KLd:40		KLn		MSE	
#	Assinatura	Métrica	B Acc	Falsa NS	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
282	MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10	MSE	0,717	10	0,367	0,369	0,531	0,529	0,510	0,518	0,714	0,717
283	MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-15	KLd:5	0,698	13	0,694	0,698	0,592	0,595	0,469	0,476	0,592	0,596
50	AAE-13_02_13_02-9-13_02_13_02-mse-32	KLn	0,679	15	0,612	0,606	0,612	0,606	0,673	0,679	0,510	0,518
133	MAAE-13_15_17-19-17_15_13-mse-32-13-13-5	MSE	0,659	16	0,490	0,494	0,449	0,447	0,510	0,520	0,653	0,659
199	MAAE-13_02_11_02-15-11_02_13_02-mse-32-13-13-15	KLd:40	0,658	15	0,551	0,558	0,653	0,658	0,469	0,476	0,551	0,555

FONTE: DADOS DA PESQUISA (2023)

A coluna “#” informa o número do experimento dentro de todo o conjunto. Sob o título “Melhores”, a coluna “Métrica” identifica qual métrica de avaliação do erro de reconstrução adotada para o cálculo da acurácia balanceada (coluna “B Acc”) produziu aquele resultado, enquanto a coluna “Falsa NS” informa a quantidade de verdades incorretamente identificadas como mentiras (falsas não sinceridades).

O uso de *dropout* também foi um fator de elevação da precisão dos modelos. Todas as métricas participaram da seleção dos melhores modelos, mas a métrica de avaliação do erro de reconstrução que identificou o modelo de maior acurácia balanceada foi o erro médio quadrático (MSE).

No caso do modelo de melhor desempenho, o componente final de sua assinatura é “10”, significando que o contexto para reconhecer discrepâncias verbais entre sinceridade e não sinceridade pela Atenção *multi-head* é de 10 palavras.

No Quadro 36 estão listados os tempos envolvidos nos processos de treinamento de cada tipo de modelo.

QUADRO 36 – TEMPOS CONSUMIDOS PELOS TREINAMENTOS DOS MODELOS VERBAIS INDIVIDUAIS

Modelo	Quantidade	Tempo médio (s)	Tempo total
AE	100	20,215 ± 2,342	6:44:17,866
AAE	70	22,027 ± 3,154	5:08:22,353
MAAE	300	40,776 ± 5,373	1 dia, 16:46:33,336
		Total	2 dias, 4:39:13,555

FONTE: DADOS DA PESQUISA (2023)

Cada configuração de modelo testada foi treinada para todos os 12 sujeitos, visto que são modelos individuais. Assim, uma unidade de um experimento implica no treinamento de 12 diferentes modelos, um para cada sujeito.

Assim como ocorreu com os modelos acústicos, os resultados globais evidenciam que o mecanismo de Atenção também aprimorou o desempenho, pois a arquitetura com mais alto desempenho foi "MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10".

Novamente o Autoencoder atencional *multi-head* superou os outros tipos, uma vez mais sugerindo a existência de relações complexas e de longa distância entre indivíduos do conjunto de dados. Neste caso, a janela de 10 palavras foi a que forneceu melhores resultados.

O protocolo de experimentação estabeleceu um estudo de ablação para avaliar a contribuição relativa das características verbais no desempenho das detecções. Modelos foram construídos a partir daquele que atingiu a maior acurácia balanceada. Diversas configurações de conjuntos de características foram testadas, o que resultou em 10 diferentes combinações. No Quadro 37 pode-se ver os resultados numéricos resultante daqueles modelos.

QUADRO 37 – DESEMPENHOS DO MELHOR MODELO VERBAL E SUAS VARIANTES NO ESTUDO DE ABLAÇÃO

#	Modelo	Métrica	B Acc	%	V S	V NS	F S	F NS
1	MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10	MSE	0,717	100,0%	15	20	4	10
2	MAAE-8_02_11_03_17_04-6-17_04_11_03_8_02-mse-32-6-6-10-P-S-R	MSE	0,617	86,0%	10	20	4	15
3	MAAE-11_02_16_03_22_04-9-22_04_16_03_11_02-mse-32-9-9-10-R-O	MSE	0,593	82,8%	13	16	8	12
4	MAAE-14_02_19_03_25_04-11-25_04_19_03_14_02-mse-32-11-11-10-P-S-O	KLd:60	0,592	82,6%	15	14	10	10
5	MAAE-14_02_19_03_25_04-11-25_04_19_03_14_02-mse-32-11-11-10-S-R-O	MSE	0,577	80,5%	8	20	4	17
6	MAAE-11_02_16_03_22_04-9-22_04_16_03_11_02-mse-32-9-9-10-P-O	MSE	0,576	80,3%	9	19	5	16
7	MAAE-14_02_19_03_25_04-11-25_04_19_03_14_02-mse-32-11-11-10-P-R-O	KLd:20	0,573	80,0%	12	16	8	13
8	MAAE-5_02_9_03_13_04-4-13_04_9_03_5_02-mse-32-4-4-10-S-R	KLn	0,557	77,7%	7	20	4	18
9	MAAE-11_02_16_03_22_04-9-22_04_16_03_11_02-mse-32-9-9-10-S-O	MSE	0,552	77,0%	13	14	10	12
10	MAAE-5_02_9_03_13_04-4-13_04_9_03_5_02-mse-32-4-4-10-P-R	MSE	0,535	74,7%	8	18	6	17
11	MAAE-5_02_9_03_13_04-4-13_04_9_03_5_02-mse-32-4-4-10-P-S	KLn	0,533	74,3%	11	15	9	14

FONTE: DADOS DA PESQUISA (2023)

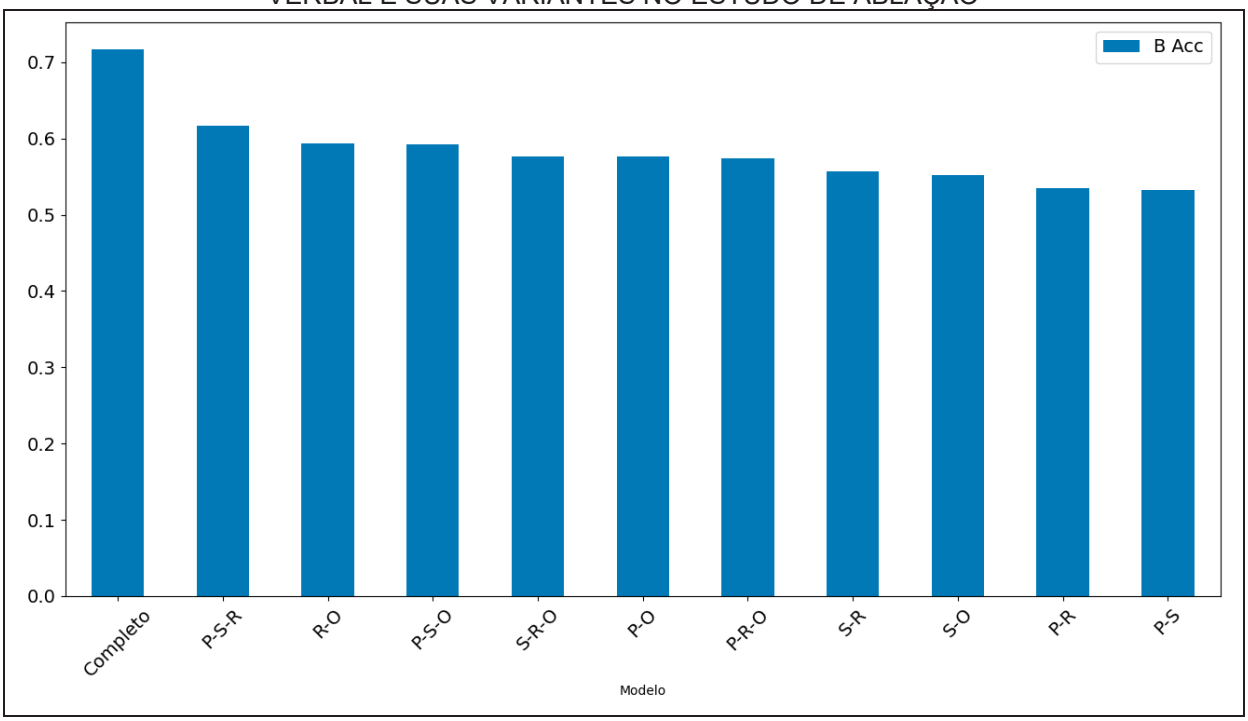
A primeira linha mostra o modelo verbal individual completo que atingiu o mais alto desempenho medido em acurácia balanceada. Este serviu como linha de base (referencial) para comparação com os demais. A coluna “%” contém o desempenho relativo de cada modelo em relação ao modelo referencial.

As perdas variam de 14 a 25,7%, evidenciando que todas as características participaram para elevar o desempenho do modelo. De forma geral, as variantes sofreram com a elevação de falsas não sinceridades (“F NS”), com exceção do modelo 4. Este, no entanto, sofreu da elevação das falsas sinceridades (“F S”).

O modelo que assumiu a segunda colocação não fez uso das características de função sintática e entidade (‘is_ent’, ‘is_num’, ‘is_pronoun’, ‘is_verb’, ‘is_noun’, ‘is_adjective’, ‘is_adverb’). No entanto, o modelo que assumiu a terceira colocação fez uso apenas das características que identificam a pessoa verbal (‘is_1_person’, ‘is_3_person’) e exatamente as características de função sintática e entidade.

Uma visão gráfica pode mais facilmente transmitir a noção da degradação de acurácia de detecção que a remoção de características imprimiu no modelo de linha de base. O gráfico de barras apresentado na Figura 25 compara os níveis de acurácia balanceada resultantes.

FIGURA 25 – COMPARATIVO DE ACURÁCIA BALANCEADA ENTRE O MELHOR MODELO VERBAL E SUAS VARIANTES NO ESTUDO DE ABLAÇÃO



FONTE: DADOS DA PESQUISA (2023)

Os resultados sugerem que todos os grupos de características influenciaram na detecção, pois a diferença em relação ao segundo colocado (que conta com 11 características) para o modelo completo (que conta com 13 características) foi mais expressiva (14 pontos percentuais) do que quando comparando os níveis de desempenho de todas as variantes entre si (11,7 pontos percentuais).

Cabe colocar que as características paraverbais (“duration” e “hesitation”) foram responsáveis por uma queda de 19,5 pontos percentuais em relação ao modelo completo. Estas características foram extraídas manualmente das transcrições obtidas do Azure speech-to-text e ocuparam a quase totalidade do tempo de extração das características verbais. É uma deficiência na tecnologia do produto que tem um impacto significativo em matéria de esforço e acurácia do modelo.

Uma percepção mais detalhada de como os grupos de características influenciaram os resultados de casos concretos pode ser alcançada pela observação do Quadro 38.

QUADRO 38 – COMPARATIVO DETALHADO DE DETEÇÃO DO MODELO REFERENCIAL E DUAS VARIANTES

#	Referência	Teste		MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10		MAAE-8_02_11_03_17_04-6-17_04_11_03_8_02-mse-32-6-6-10-P-S-R		MAAE-5_02_9_03_13_04-4-13_04_9_03_5_02-mse-32-4-4-10-P-S	
				MSE		MSE		KLn	
		Sujeito	Esperado	Escore	Detectado	Escore	Detectado	Escore	Detectado
1	S1-P7-13	S1-P7-1	NS	0,027	S	-0,040	NS	-1,000	NS
2		S1-P7-2	S	-0,127	NS	0,201	S	0,567	S
3		S1-P7-3	S	0,049	S	-0,041	NS	-1,000	NS
4		S1-P7-4	NS	0,169	S	-0,103	NS	0,614	S
5		S1-P7-5	S	0,083	S	-0,137	NS	-1,000	NS
6		S1-P7-6	NS	-0,226	NS	-0,079	NS	0,691	S
7		S1-P7-7	S	-0,138	NS	-0,137	NS	-1,000	NS
8		S1-P7-8	NS	0,159	S	-0,171	NS	0,693	S
9		S1-P7-9	NS	-0,238	NS	-0,132	NS	-1,000	NS
10		S1-P7-10	S	0,074	S	-0,185	NS	-1,000	NS
11		S1-P7-11	S	0,087	S	0,056	S	-1,000	NS
12		S1-P7-12	S	0,252	S	-0,019	NS	0,669	S
13	S1-P8-3	S1-P8-1	S	0,202	S	0,059	S	-1,000	NS
14		S1-P8-2	S	0,000	S	0,018	S	0,585	S
15		S1-P8-4	S	0,013	S	0,064	S	-0,755	NS
16		S1-P8-5	S	0,052	S	0,000	S	-1,000	NS
17		S1-P8-6	S	-0,025	NS	0,066	S	-1,000	NS
18		S1-P8-7	S	-0,104	NS	-0,062	NS	-1,000	NS
19	S1-P9-11	S1-P9-1	NS	-0,270	NS	-0,556	NS	-1,000	NS

20		S1-P9-2	NS	-0,392	NS	-0,758	NS	0,494	S
21		S1-P9-3	S	0,041	S	-0,147	NS	0,524	S
22		S1-P9-5	NS	-0,398	NS	-0,776	NS	0,523	S
23		S1-P9-6	NS	-0,057	NS	-0,508	NS	-1,000	NS
24		S1-P9-8	NS	-0,438	NS	-0,614	NS	-0,549	NS
25		S1-P9-9	S	-0,086	NS	-0,640	NS	-1,000	NS
26		S1-P9-10	S	0,101	S	0,313	S	-1,000	NS
27	S2-P1-1	S2-P1-2	S	-0,532	NS	-0,475	NS	0,728	S
28		S2-P1-3	S	0,194	S	0,050	S	-1,000	NS
29	S2-P2-5	S2-P2-1	NS	-0,293	NS	0,065	S	-0,320	NS
30		S2-P2-2	NS	-0,229	NS	-0,800	NS	0,591	S
31		S2-P2-3	NS	-0,347	NS	-0,508	NS	0,587	S
32		S2-P3-1	NS	-0,388	NS	0,038	S	-1,000	NS
33		S2-P3-2	NS	-0,561	NS	-0,737	NS	-0,830	NS
34		S2-P3-3	NS	-0,423	NS	-0,285	NS	-0,985	NS
35		S2-P3-4	NS	-0,486	NS	-0,253	NS	-1,000	NS
36	S2-P4-2	S2-P4-1	S	0,031	S	-0,085	NS	-0,549	NS
37	S2-P5-3	S2-P5-1	NS	-0,038	NS	0,260	S	-1,000	NS
38		S2-P5-2	S	-0,083	NS	0,172	S	0,615	S
39		S2-P6-1	NS	-0,132	NS	-0,021	NS	-1,000	NS
40		S2-P6-2	NS	-0,245	NS	-0,624	NS	-1,000	NS
41	S2-P7-8	S2-P7-1	NS	-0,018	NS	-0,067	NS	-1,000	NS
42		S2-P7-4	NS	0,071	S	0,040	S	-1,000	NS
43		S2-P7-5	NS	-0,049	NS	-0,259	NS	-1,000	NS
44		S2-P7-6	S	0,165	S	-0,079	NS	-0,914	NS
45		S2-P7-7	S	0,010	S	-0,093	NS	-0,997	NS
46	S2-P8-5	S2-P8-1	NS	-0,180	NS	-0,318	NS	0,672	S
47	S2-P9-4	S2-P9-1	S	-0,178	NS	-0,325	NS	0,362	S
48		S2-P9-2	S	-0,542	NS	-0,806	NS	0,531	S
49		S2-P9-3	S	-0,251	NS	-0,867	NS	0,528	S

FONTE: DADOS DA PESQUISA (2023)

O modelo mais à esquerda é o modelo completo, com todos os quatro tipos de características, o modelo central é o segundo colocado, alimentado com características paralinguísticas (“P”), de sentimento (“S”) e de pessoa verbal (“R”), e o modelo mais à direita corresponde à variante de pior desempenho, que fez uso de características paralinguísticas e de sentimento.

O pior modelo parece ter uma forte tendência a classificar as narrativas como não sinceras e com alto grau de confiança, o que significa que pequenas variações entre as narrativas produziram grandes variações na distribuição das reconstruções.

4.3.5 Modelos verbais coletivos

As versões coletivas dos cinco melhores modelos individuais foram treinadas e avaliadas. Os resultados podem ser encontrados no Quadro 39.

O melhor modelo coletivo não teve um desempenho tão elevado quando os melhores modelos atencionais individuais, *single-head* ou *multi-head*, mas superou o melhor modelo Vanilla individual.

QUADRO 39 – CINCO MELHORES MODELOS VERBAIS COLETIVOS

#	Modelo	Melhores			KLd:15		KLn		MSE	
		Métrica	B Acc	Falsa NS	Acc	B Acc	Acc	B Acc	Acc	B Acc
4	MAAE-13_02_11_02-15-11_02_13_02-mse-32-13-13-15	MSE	0,612	9	0,531	0,525	0,531	0,540	0,612	0,612
3	MAAE-13_15_17-19-17_15_13-mse-32-13-13-5	KLd:15	0,598	17	0,592	0,598	0,449	0,455	0,490	0,490
1	MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-15	MSE	0,591	9	0,408	0,402	0,531	0,540	0,592	0,591
2	AAE_4-13_02_13_02-9-13_02_13_02-mse-32	MSE	0,590	8	0,510	0,500	0,571	0,579	0,592	0,590
0	MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10	KLn	0,579	20	0,388	0,384	0,571	0,579	0,571	0,570

FONTE: DADOS DA PESQUISA (2023)

Visto que os conjuntos de treinamento cresceram, os tempos de treinamento também cresceram de forma expressiva, como pode ser observado no Quadro 40.

QUADRO 40 – TEMPOS CONSUMIDOS PELOS TREINAMENTOS DOS MODELOS VERBAIS COLETIVOS

Modelo	Quantidade	Tempo médio (s)	Tempo total
Coletivo	5	86,935 ± 30,859	1:26:56,128

FONTE: DADOS DA PESQUISA (2023)

Os modelos coletivos atingiram níveis de desempenho mais baixos que os modelos individuais, mas superaram o desempenho dos seus pares no caso acústico. Para o caso verbal, os padrões coletivos foram mais discriminantes que os acústicos, mas ainda inferiores aos padrões individuais.

Uma possível justificativa é a menor quantidade de características (13 para modelos verbais, 65 para modelos acústicos), em conjunção com a natureza predominantemente categórica dos dados (apenas as características “duration”, “pos_score” e “neg_score” podem assumir valores com ponto flutuante), que deve oferecer menor variância nos dados, e que pode facilitar a identificação de padrões.

4.3.6 Modelo verbal bicomponente

Os componentes do modelo verbal bicomponente estão apresentados no Quadro 41.

QUADRO 41 – DESEMPENHO COMBINADO DOS MODELOS VERBAIS INDIVIDUAL E COLETIVO

Arquitetura	Card.	Métrica	Act	Acc	B Acc	V S	V NS	F S	F NS
MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10	Ind	MSE	35	0,714	0,717	15	20	4	10
MAAE-13_02_11_02-15-11_02_13_02-mse-32-13-13-15	Col	MSE	27	0,551	0,553	11	16	8	14
		Final	28	0,571	0,575	10	18	6	15

FONTE: DADOS DA PESQUISA (2023)

A coluna “Card.” Indica cardinalidade individual (“Ind”) ou coletiva (“Col”), ao passo que a coluna “Act” contém a contagem de acertos. O processo de fusão foi o mesmo determinado para os modelos multimodais, ou seja, pelo uso da soma ajustada dos Escores de Sinceridade de cada um dos componentes. A combinação dos dois modelos não resultou em aprimoramento da capacidade de detecção, vindo a reduzi-la da mesma forma que ocorreu no modelo acústico.

No Quadro 42 é possível observar o detalhamento da combinação dos dois modelos ao nível de sujeito. Os valores rotulados como “S” indicam classificação da narrativa como “sincera”, enquanto “NS” identifica narrativa “não sincera”.

QUADRO 42 – COMPARATIVO DETALHADO DOS DESEMPENHOS DE MODELOS VERBAIS INDIVIDUAL E COLETIVO

#	Referência	Teste		MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10 Individual		MAAE-13_02_11_02-15-11_02_13_02-mse-32-13-13-15 Coletivo		Final		
				MSE		MSE				
		Sujeito	Esperado	Escore	Detectado	Escore	Detectado	Soma bruta	Escore	Detectado
1	S1-P7-13	S1-P7-1	NS	0,027	S	0,144	S	0,170	0,169	S
2		S1-P7-2	S	-0,127	NS	-0,067	NS	-0,195	-0,192	NS
3		S1-P7-3	S	0,049	S	0,102	S	0,152	0,150	S
4		S1-P7-4	NS	0,169	S	0,196	S	0,366	0,350	S
5		S1-P7-5	S	0,083	S	0,072	S	0,155	0,154	S
6		S1-P7-6	NS	-0,226	NS	-0,134	NS	-0,360	-0,345	NS
7		S1-P7-7	S	-0,138	NS	-0,146	NS	-0,284	-0,277	NS
8		S1-P7-8	NS	0,159	S	0,228	S	0,387	0,369	S
9		S1-P7-9	NS	-0,238	NS	-0,185	NS	-0,423	-0,399	NS
10		S1-P7-10	S	0,074	S	0,074	S	0,149	0,148	S
11		S1-P7-11	S	0,087	S	0,034	S	0,121	0,120	S
12		S1-P7-12	S	0,252	S	0,239	S	0,492	0,455	S
13	S1-P8-3	S1-P8-1	S	0,202	S	0,230	S	0,433	0,408	S
14		S1-P8-2	S	0,000	S	-0,004	NS	-0,003	-0,003	NS
15		S1-P8-4	S	0,013	S	-0,143	NS	-0,130	-0,129	NS
16		S1-P8-5	S	0,052	S	-0,062	NS	-0,009	-0,009	NS
17		S1-P8-6	S	-0,025	NS	-0,161	NS	-0,185	-0,183	NS
18		S1-P8-7	S	-0,104	NS	-0,289	NS	-0,393	-0,374	NS

19	S1-P9-11	S1-P9-1	NS	-0,270	NS	0,144	S	-0,126	-0,125	NS
20		S1-P9-2	NS	-0,392	NS	-0,234	NS	-0,625	-0,555	NS
21		S1-P9-3	S	0,041	S	0,044	S	0,085	0,085	S
22		S1-P9-5	NS	-0,398	NS	-0,137	NS	-0,534	-0,489	NS
23		S1-P9-6	NS	-0,057	NS	0,275	S	0,218	0,214	S
24		S1-P9-8	NS	-0,438	NS	-0,233	NS	-0,670	-0,585	NS
25		S1-P9-9	S	-0,086	NS	0,051	S	-0,035	-0,035	NS
26		S1-P9-10	S	0,101	S	0,087	S	0,188	0,186	S
27	S2-P1-1	S2-P1-2	S	-0,532	NS	-0,327	NS	-0,859	-0,696	NS
28		S2-P1-3	S	0,194	S	0,108	S	0,302	0,293	S
29	S2-P2-5	S2-P2-1	NS	-0,293	NS	-0,060	NS	-0,353	-0,339	NS
30		S2-P2-2	NS	-0,229	NS	0,104	S	-0,125	-0,124	NS
31		S2-P2-3	NS	-0,347	NS	-0,039	NS	-0,386	-0,368	NS
32		S2-P3-1	NS	-0,388	NS	-0,235	NS	-0,623	-0,553	NS
33		S2-P3-2	NS	-0,561	NS	-0,418	NS	-0,979	-0,753	NS
34		S2-P3-3	NS	-0,423	NS	-0,274	NS	-0,697	-0,602	NS
35		S2-P3-4	NS	-0,486	NS	-0,301	NS	-0,788	-0,657	NS
36	S2-P4-2	S2-P4-1	S	0,031	S	-0,214	NS	-0,183	-0,181	NS
37	S2-P5-3	S2-P5-1	NS	-0,038	NS	0,232	S	0,194	0,192	S
38		S2-P5-2	S	-0,083	NS	0,347	S	0,264	0,258	S
39	S2-P6-5	S2-P6-1	NS	-0,132	NS	0,041	S	-0,091	-0,091	NS
40		S2-P6-2	NS	-0,245	NS	-0,110	NS	-0,355	-0,341	NS
41	S2-P7-8	S2-P7-1	NS	-0,018	NS	-0,291	NS	-0,309	-0,299	NS
42		S2-P7-4	NS	0,071	S	-0,005	NS	0,066	0,066	S
43		S2-P7-5	NS	-0,049	NS	-0,268	NS	-0,317	-0,307	NS
44		S2-P7-6	S	0,165	S	-0,167	NS	-0,002	-0,002	NS
45		S2-P7-7	S	0,010	S	-0,121	NS	-0,112	-0,111	NS
46	S2-P8-5	S2-P8-1	NS	-0,180	NS	-0,211	NS	-0,391	-0,372	NS
47	S2-P9-4	S2-P9-1	S	-0,178	NS	-0,146	NS	-0,324	-0,313	NS
48		S2-P9-2	S	-0,542	NS	-0,505	NS	-1,047	-0,781	NS
49		S2-P9-3	S	-0,251	NS	-0,156	NS	-0,408	-0,387	NS

FONTE: DADOS DA PESQUISA (2023)

As células foram coloridas para facilitar a identificação das detecções frente ao resultado esperado. As células em vermelho identificam “não sinceridade”, enquanto as verdes, “sinceridade”.

O modelo coletivo parece ter uma tendência de classificar mais casos como não sinceros, com confiança mais elevada, enquanto o modelo individual classificou corretamente alguns casos de não sinceridade, mas com confiança mais baixa, o que veio a inverter incorretamente a classificação. Tais resultados sugerem que os modelos coletivos de sinceridade são compostos por padrões que se assemelham bastante aos padrões de não sinceridade do sujeito testado. Esses padrões

possivelmente abrangem uma região grande do espaço de características onde também se encontram os padrões de não sinceridade do sujeito não incluído. É observável também que alguns casos de sinceridade esperada foram classificados como não sinceros pelo modelo coletivo, levando a mais erros na classificação final.

Dados os resultados, as duas hipóteses aventadas para o modelo coletivo acústico (insuficiência de variedade de dados e padrões individuais mais discriminantes) são aplicáveis também a este caso. Analogamente, existe a possibilidade da ocorrência conjugada de ambas na produção do resultado.

4.3.7 Modelos visuais individuais

Os modelos visuais foram alimentados com características extraídas pelo OpenFace, que processou arquivos de vídeo de narrativas em formato MP4 e retornou planilhas compostas por 714 colunas, sendo 709 de características visuais e seis de controle.

Das 709 características visuais extraídas pelo OpenFace, apenas 31 foram consideradas, implicando que todos os modelos tiveram 31 neurônios de entrada e 31 de saída, com variadas configurações de camadas escondidas, conforme estabelecido no protocolo de experimentação. Todos os modelos foram treinados com 1000 épocas, a partir da calibragem inicial. A maioria dos modelos atingiu a estabilidade com aproximadamente de 200 a 600 épocas, mas alguns precisaram de quantidades próximas a 1000, especialmente quando do uso de *dropout*.

Modelos Autoencoder Vanilla foram testados em 130 diferentes configurações, sendo 70 com e 60 sem a aplicação de *dropout*. Os 10 melhores resultados podem ser vistos tanto no Quadro 43 quanto no Quadro 44 (continuação).

QUADRO 43 - DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VISUAIS DE AUTOENCODERS VANILLA USANDO AS MÉTRICA KL:5, KL:10, KL:15, KL:20 E KL:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-32	0,571	0,577	0,571	0,578	0,592	0,598	0,571	0,578	0,571	0,578
AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-32	0,571	0,577	0,571	0,578	0,592	0,598	0,571	0,578	0,571	0,578
AE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32	0,531	0,538	0,551	0,558	0,551	0,558	0,551	0,559	0,551	0,558
AE-31_29_27-31-27_29_31-mse-32	0,571	0,579	0,531	0,540	0,531	0,540	0,510	0,520	0,510	0,520
AE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256	0,571	0,578	0,551	0,558	0,551	0,558	0,551	0,558	0,510	0,518
AE-39_02_41_03_45_04-27-45_04_41_03_39_02-mse-256	0,551	0,558	0,551	0,558	0,571	0,578	0,531	0,538	0,571	0,578
AE-39_02_41_03_45_04-27-45_04_41_03_39_02-mse-256	0,551	0,558	0,551	0,558	0,571	0,578	0,531	0,538	0,571	0,578

AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-256	0,571	0,578	0,531	0,538	0,551	0,558	0,551	0,559	0,551	0,559
AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-256	0,571	0,578	0,531	0,538	0,551	0,558	0,551	0,559	0,551	0,559
AE-31_29_27-27-27_29_31-mse-256	0,551	0,560	0,510	0,520	0,531	0,540	0,510	0,520	0,531	0,540

FONTE: DADOS DA PESQUISA (2023)

Os desempenhos estão expressos em acurácia (“Acc”) e acurácia balanceada (“B Acc”), esta última atuando como o critério de avaliação dos modelos. Todas as métricas para cálculo do erro de reconstrução (KLd:5 a KLd:60, KLn e MSE) estão presentes. A coluna “Modelo” apresenta as assinaturas dos modelos experimentados.

QUADRO 44 - DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VISUAIS DE AUTOENCODERS VANILLA USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-32	0,551	0,558	0,510	0,518	0,510	0,518	0,490	0,500	0,490	0,500
AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-32	0,551	0,558	0,510	0,518	0,510	0,518	0,490	0,500	0,490	0,500
AE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32	0,551	0,559	0,571	0,580	0,571	0,580	0,469	0,479	0,490	0,500
AE-31_29_27-31-27_29_31-mse-32	0,510	0,520	0,490	0,500	0,490	0,500	0,490	0,500	0,490	0,500
AE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256	0,510	0,518	0,551	0,559	0,571	0,579	0,490	0,500	0,490	0,500
AE-39_02_41_03_45_04-27-45_04_41_03_39_02-mse-256	0,571	0,579	0,551	0,559	0,531	0,538	0,490	0,500	0,490	0,500
AE-39_02_41_03_45_04-27-45_04_41_03_39_02-mse-256	0,571	0,579	0,551	0,559	0,531	0,538	0,490	0,500	0,490	0,500
AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-256	0,531	0,538	0,510	0,518	0,510	0,518	0,490	0,500	0,490	0,500
AE-39_02_41_03_45_04-29-45_04_41_03_39_02-mse-256	0,531	0,538	0,510	0,518	0,510	0,518	0,490	0,500	0,490	0,500
AE-31_29_27-27-27_29_31-mse-256	0,531	0,540	0,531	0,540	0,531	0,540	0,449	0,458	0,490	0,500

FONTE: DADOS DA PESQUISA (2023)

A aplicação de *dropout* aprimorou a qualidade da maioria dos modelos, com apenas dois não tendo feito uso deste recurso. Tanto KLn quanto MSE não produziram boas avaliações, sugerindo que os volumes dos erros não variaram significativamente, mas as distribuições sim.

O segundo conjunto de modelos testado foi dos Autoencoders atencionais *single-head* (AAE). Foram ao todo 180 experimentos (70 sem e 110 com *dropout*). Nos Quadro 45 e Quadro 46 apresentam-se os 10 níveis de desempenho mais elevados alcançados com diferentes modelos atencionais *single-head*.

QUADRO 45 - DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VISUAIS DE AUTOENCODERS ATENCIONAIS *SINGLE-HEAD* USANDO AS MÉTRICA KL:5, KL:10, KL:15, KL:20 E KL:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AAE-31-35-31-mse-32	0,510	0,513	0,551	0,553	0,531	0,533	0,551	0,553	0,612	0,613
AAE-31_33-29-33_31-mse-256	0,429	0,423	0,429	0,423	0,449	0,443	0,449	0,443	0,429	0,423
AAE-31_02_33_02-29-33_02_31_02-mse-32	0,510	0,506	0,510	0,506	0,510	0,506	0,510	0,506	0,490	0,485
AAE-39_02_47_03-31-47_03_39_02-mse-32	0,510	0,506	0,510	0,506	0,510	0,506	0,510	0,506	0,490	0,485
AAE-39_02_47_03-35-47_03_39_02-mse-32	0,510	0,506	0,510	0,506	0,510	0,506	0,510	0,506	0,490	0,485
AAE-31_29_27-25-27_29_31-mse-256	0,429	0,423	0,429	0,423	0,449	0,443	0,449	0,443	0,429	0,423
AAE-31_33-37-33_31-mse-256	0,429	0,423	0,429	0,423	0,449	0,443	0,449	0,443	0,429	0,423
AAE-31_31_31-31-31_31_31-mse-256	0,429	0,423	0,429	0,423	0,449	0,443	0,449	0,443	0,429	0,423
AAE-31_02_33_02-31-33_02_31_02-mse-32	0,510	0,506	0,510	0,506	0,510	0,506	0,510	0,506	0,490	0,485
AAE-31_02_31_02-31-31_02_31_02-mse-32	0,510	0,506	0,510	0,506	0,510	0,506	0,510	0,506	0,490	0,485

FONTE: DADOS DA PESQUISA (2023)

O uso de *dropout* aprimorou cinco dos dez modelos, mas os restantes, incluindo o modelo de mais alto desempenho, não precisaram do recurso.

QUADRO 46 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VISUAIS DE AUTOENCODERS ATENCIONAIS *SINGLE-HEAD* USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
AAE-31-35-31-mse-32	0,653	0,654	0,633	0,634	0,633	0,634	0,490	0,495	0,449	0,458
AAE-31_33-29-33_31-mse-256	0,429	0,423	0,429	0,423	0,429	0,423	0,429	0,431	0,633	0,636
AAE-31_02_33_02-29-33_02_31_02-mse-32	0,490	0,485	0,490	0,485	0,490	0,485	0,490	0,493	0,612	0,618
AAE-39_02_47_03-31-47_03_39_02-mse-32	0,490	0,485	0,490	0,485	0,490	0,485	0,408	0,415	0,612	0,618
AAE-39_02_47_03-35-47_03_39_02-mse-32	0,490	0,485	0,490	0,485	0,490	0,485	0,531	0,533	0,612	0,618
AAE-31_29_27-25-27_29_31-mse-256	0,429	0,423	0,429	0,423	0,429	0,423	0,469	0,471	0,612	0,616
AAE-31_33-37-33_31-mse-256	0,429	0,423	0,429	0,423	0,429	0,423	0,224	0,225	0,612	0,615
AAE-31_31_31-31-31_31_31-mse-256	0,429	0,423	0,429	0,423	0,429	0,423	0,367	0,369	0,612	0,615
AAE-31_02_33_02-31-33_02_31_02-mse-32	0,490	0,485	0,490	0,485	0,490	0,485	0,490	0,494	0,592	0,597
AAE-31_02_31_02-31-31_02_31_02-mse-32	0,490	0,485	0,490	0,485	0,490	0,485	0,429	0,431	0,592	0,597

FONTE: DADOS DA PESQUISA (2023)

A Atenção *single-head* trouxe um aumento de desempenho em relação aos modelos Vanilla, mas a margem não foi expressiva, de apenas quatro pontos percentuais.

O terceiro conjunto de experimentos incluiu os Autoencoders atencionais *multi-head* (MAAE). Foram 510 experimentos ao todo, com 210 sem a aplicação de

dropout e 310 com. Os dez melhores resultados foram alcançados pelos modelos apresentados no Quadro 47 e Quadro 48 (continuação).

Divergindo dos modelos verbais e acústicos, os níveis mais elevados de desempenho foram capturados por variações da métrica KLd para avaliação do erro de reconstrução, sem correspondente desempenho das métricas KLn e MSE, com exceção de um caso.

QUADRO 47 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VISUAIS DE AUTOENCODERS ATENCIONAIS *MULTI-HEAD* USANDO AS MÉTRICA KL:5, KL:10, KL:15, KLD:20 E KLD:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	0,653	0,653	0,673	0,674	0,714	0,715	0,653	0,653	0,633	0,633
MAAE-39_02_44_03-35-44_03_39_02-mse-256-31-31-5	0,633	0,637	0,653	0,657	0,653	0,657	0,612	0,617	0,612	0,617
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256-31-31-5	0,510	0,513	0,551	0,554	0,571	0,575	0,551	0,554	0,571	0,574
MAAE-31_02_33_02_35_02-35-35_02_33_02_31_02-mse-32-31-31-10	0,694	0,695	0,633	0,634	0,551	0,553	0,551	0,553	0,551	0,552
MAAE-31_02_31_02_31_02-33-31_02_31_02_31_02-mse-32-31-31-10	0,531	0,529	0,490	0,488	0,510	0,508	0,531	0,528	0,551	0,546
MAAE-31_02_29_02_27_02-31-27_02_29_02_31_02-mse-32-31-31-20	0,653	0,651	0,673	0,673	0,592	0,589	0,571	0,568	0,510	0,505
MAAE-39_02_44_03_52_04-27-52_04_44_03_39_02-mse-256-31-31-5	0,633	0,632	0,653	0,653	0,673	0,673	0,673	0,673	0,673	0,673
MAAE-31_02_29_02_27_02-27-27_02_29_02_31_02-mse-256-31-31-10	0,612	0,612	0,612	0,611	0,673	0,672	0,653	0,651	0,653	0,651
MAAE-39_02_41_03_45_04-31-45_04_41_03_39_02-mse-32-31-31-10	0,571	0,571	0,653	0,653	0,612	0,613	0,592	0,592	0,673	0,671
MAAE-31_02_29_02_27_02-27-27_02_29_02_31_02-mse-32-31-31-10	0,571	0,570	0,531	0,530	0,510	0,511	0,551	0,550	0,571	0,570

FONTE: DADOS DA PESQUISA (2023)

Todos os dez melhores casos apresentaram o uso de *dropout*, o que é outro diferencial em relação aos modelos acústico e verbal. Aqueles modelos apresentaram algumas configurações de alto desempenho sem a necessidade deste recurso. No entanto, em convergência com os experimentos das outras duas modalidades, os modelos com Atenção *multi-head* foram os que atingiram níveis mais elevados de acurácia balanceada.

QUADRO 48 – DEZ MELHORES DESEMPENHOS PARA ARQUITETURAS VISUAIS DE AUTOENCODERS ATENCIONAIS *MULTI-HEAD* USANDO AS MÉTRICA KL:40, KL:50, KL:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	0,694	0,693	0,653	0,652	0,673	0,673	0,531	0,536	0,510	0,515
MAAE-39_02_44_03-35-44_03_39_02-mse-256-31-31-5	0,694	0,697	0,653	0,657	0,653	0,657	0,612	0,617	0,551	0,558
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256-31-31-5	0,633	0,634	0,694	0,695	0,653	0,653	0,551	0,553	0,592	0,595
MAAE-31_02_33_02_35_02-35-35_02_33_02_31_02-mse-32-31-31-10	0,531	0,529	0,571	0,568	0,571	0,568	0,551	0,550	0,592	0,596

MAAE-31_02_31_02_31_02-33-31_02_31_02_31_02-mse-32-31-31-10	0,531	0,524	0,510	0,503	0,510	0,503	0,551	0,553	0,673	0,678
MAAE-31_02_29_02_27_02-31-27_02_29_02_31_02-mse-32-31-31-20	0,490	0,484	0,490	0,483	0,510	0,503	0,429	0,432	0,551	0,555
MAAE-39_02_44_03_52_04-27-52_04_44_03_39_02-mse-256-31-31-5	0,653	0,652	0,653	0,651	0,612	0,610	0,510	0,510	0,571	0,578
MAAE-31_02_29_02_27_02-27-27_02_29_02_31_02-mse-256-31-31-10	0,612	0,609	0,633	0,629	0,612	0,608	0,592	0,596	0,551	0,557
MAAE-39_02_41_03_45_04-31-45_04_41_03_39_02-mse-32-31-31-10	0,592	0,588	0,571	0,567	0,571	0,567	0,490	0,488	0,571	0,575
MAAE-31_02_29_02_27_02-27-27_02_29_02_31_02-mse-32-31-31-10	0,612	0,609	0,571	0,568	0,673	0,670	0,592	0,592	0,551	0,556

FONTE: DADOS DA PESQUISA (2023)

Os resultados combinados de todos os modelos AE, AAE e MAAE (somando 820 experimentos) estão apresentados no Quadro 49.

QUADRO 49 - CINCO MELHORES MODELOS VISUAIS DENTRE AS ARQUITETURAS AE, AAE E MAAE

Modelo		Melhores			KLd:5		KLd:15		KLd:40		KLd:50		MSE	
#	Assinatura	Métrica	B Acc	Falsa NS	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
219	MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	KLd:15	0,715	8	0,653	0,653	0,714	0,715	0,694	0,693	0,653	0,652	0,510	0,515
327	MAAE-39_02_44_03-35-44_03_39_02-mse-256-31-31-5	KLd:40	0,697	11	0,633	0,637	0,653	0,657	0,694	0,697	0,653	0,657	0,551	0,558
234	MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256-31-31-5	KLd:50	0,695	9	0,510	0,513	0,571	0,575	0,633	0,634	0,694	0,695	0,592	0,595
337	MAAE-31_02_33_02_35_02-35-35_02_33_02_31_02-mse-32-31-31-10	KLd:5	0,695	9	0,694	0,695	0,551	0,553	0,531	0,529	0,571	0,568	0,592	0,596
460	MAAE-31_02_31_02_31_02-33-31_02_31_02_31_02-mse-32-31-31-10	MSE	0,678	13	0,531	0,529	0,510	0,508	0,531	0,524	0,510	0,503	0,673	0,678

FONTE: DADOS DA PESQUISA (2023)

A coluna “#” informa o número do experimento dentro de todo o conjunto. Sob o título “Melhores”, a coluna “Métrica” identifica qual métrica de avaliação do erro de reconstrução adotada para o cálculo da acurácia balanceada (coluna “B Acc”) produziu aquele resultado, enquanto a coluna “Falsa NS” informa a quantidade de verdades que foram incorretamente identificadas como mentiras (falsas não sinceridades).

Para a modalidade visual, há uma predominância de melhores resultados que foram atingidos pela métrica KLd, com apenas um resultado atingido com MSE e nenhum com KLn.

O melhor modelo apresenta tamanho do lote igual a 5 (último parâmetro da assinatura), indicando que foram necessários cinco quadros para avaliar as relações de longa distância entre as características.

No Quadro 50 estão listados os tempos envolvidos nos processos de treinamento de cada tipo de modelo. Assim como ocorreu com os modelos acústicos e verbais, os resultados globais evidenciam que o mecanismo de atenção também aprimorou o desempenho, pois a arquitetura com mais alto desempenho foi "MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5".

QUADRO 50 – TEMPOS CONSUMIDOS PELOS TREINAMENTOS DOS MODELOS VISUAIS INDIVIDUAIS

Modelo	Quantidade	Tempo médio (s)	Tempo total
AE	130	66,413 ± 41,893	1 dia, 4:46:44,248
AAE	180	79,464 ± 55,502	1 dia, 23:40:41,189
MAAE	510	146,789 ± 73,885	10 dias, 9:32:31,266
		Total	13 dias, 13:59:56,703

FONTE: DADOS DA PESQUISA (2023)

Novamente o Autoencoder atencional *multi-head* superou os outros tipos, uma vez mais sugerindo a existência de relações complexas e de longa distância entre as características dos quadros do conjunto de dados. Neste caso, o lote de amostras de 5 quadros foi a que forneceu melhores resultados.

4.3.8 Modelos visuais coletivos

A exemplo das modalidades acústica e verbal, modelos coletivos correspondentes aos cinco melhores modelos individuais foram treinados e avaliados, cujos resultados podem ser observados tanto no Quadro 51 quanto noAssim como os demais modelos coletivos, o desempenho geral expresso pela acurácia balanceada permaneceu abaixo dos correspondentes individuais. Um fato que chama a atenção é que, enquanto os modelos individuais apresentaram os melhores desempenhos medidos por variações de KLd, as versões coletivas apresentaram as avaliações mais elevadas quando medidos por KLn e MSE.

Quadro 52.

QUADRO 51 - DESEMPENHO DAS CINCO MELHORES ARQUITETURAS VISUAIS COLETIVAS DE AUTOENCODERS USANDO AS MÉTRICA KLD:5, KLD:10, KLD:15, KLD:20 E KLD:30

Modelo	KLd:5		KLd:10		KLd:15		KLd:20		KLd:30	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-31_02_33_02_35_02-35-35_02_33_02_31_02-mse-32-31-31-10	0,469	0,472	0,449	0,448	0,469	0,468	0,408	0,406	0,449	0,446
MAAE-31_02_31_02_31_02-33-31_02_31_02_31_02-mse-32-31-31-10	0,531	0,533	0,490	0,492	0,469	0,470	0,469	0,468	0,490	0,488

MAAE-39_02_44_02-35-44_02_39_02-mse-256-31-31-5	0,449	0,444	0,449	0,444	0,449	0,444	0,449	0,444	0,449	0,444
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256-31-31-5	0,469	0,469	0,449	0,448	0,469	0,468	0,490	0,488	0,449	0,446
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	0,367	0,366	0,388	0,383	0,408	0,403	0,429	0,423	0,408	0,403

FONTE: DADOS DA PESQUISA (2023)

Assim como os demais modelos coletivos, o desempenho geral expresso pela acurácia balanceada permaneceu abaixo dos correspondentes individuais. Um fato que chama a atenção é que, enquanto os modelos individuais apresentaram os melhores desempenhos medidos por variações de KLd, as versões coletivas apresentaram as avaliações mais elevadas quando medidos por KLn e MSE.

QUADRO 52 - DESEMPENHO DAS CINCO MELHORES ARQUITETURAS VISUAIS COLETIVAS DE AUTOENCODERS USANDO AS MÉTRICA KLD:40, KLD:50, KLD:60, KLN E MSE

Modelo	KLd:40		KLd:50		KLd:60		KLn		MSE	
	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc	Acc	B Acc
MAAE-31_02_33_02_35_02-35-35_02_33_02_31_02-mse-32-31-31-10	0,388	0,383	0,429	0,423	0,367	0,361	0,571	0,580	0,490	0,500
MAAE-31_02_31_02_31_02-33-31_02_31_02_31_02-mse-32-31-31-10	0,429	0,424	0,388	0,383	0,408	0,402	0,551	0,553	0,490	0,500
MAAE-39_02_44_02-35-44_02_39_02-mse-256-31-31-5	0,449	0,444	0,449	0,444	0,449	0,444	0,551	0,549	0,490	0,500
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256-31-31-5	0,429	0,425	0,429	0,425	0,510	0,505	0,510	0,515	0,490	0,500
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	0,429	0,423	0,449	0,443	0,490	0,483	0,408	0,414	0,490	0,500

FONTE: DADOS DA PESQUISA (2023)

De forma análoga, os tempos para o treinamento destes modelos foram superiores em função do volume de dados envolvido, como pode ser confirmado no Quadro 53.

QUADRO 53 - TEMPOS CONSUMIDOS PELOS TREINAMENTOS DOS MODELOS VISUAIS COLETIVOS

Modelo	Quantidade	Tempo médio (s)	Tempo total
Coletivo	5	1.265,959 ± 533,001	21:05:57,553

FONTE: DADOS DA PESQUISA (2023)

Adicionalmente, os resultados deste modelo coletivo seguiram a tendência observada nos modelos acústico e verbal, com desempenho inferior os correspondentes individuais.

4.3.9 Modelo visual bicomponente

Um modelo visual bicomponente também foi construído e avaliado, cujo desempenho se apresenta no Quadro 54.

QUADRO 54 - DESEMPENHO COMBINADO DOS MODELOS VISUAIS INDIVIDUAL E COLETIVO

Model	Card.	Métrica	Act	Acc	B Acc	V S	V NS	F S	F NS

MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	Ind	KLd:15	35	0,714	0,715	17	18	6	8
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256-31-31-5	Col	KLn	29	0,592	0,591	16	13	11	9
		Final	30	0,612	0,614	13	17	7	12

FONTE: DADOS DA PESQUISA (2023)

A coluna “Card.” Indica cardinalidade individual (“Ind”) ou coletiva (“Col”), ao passo que a coluna “Act” contém a contagem de acertos, “Acc” a acurácia, “B Acc” a acurácia balanceada, “V S” a contagem de verdadeiras sinceridades, “V NS” verdadeiras não sinceridades, “F S” falsas sinceridades e “F NS” falsas não sinceridades. O processo de fusão foi o mesmo determinado para os modelos multimodais, ou seja, pelo uso da soma ajustada dos Escores de Sinceridade de cada um dos componentes. A combinação dos dois modelos não resultou em aprimoramento da capacidade de detecção, vindo a reduzi-la da mesma forma que ocorreu nos modelos acústico e verbal.

No Quadro 55 é possível observar o detalhamento da combinação dos dois modelos ao nível de sujeito. Os valores rotulados como “S” indicam classificação da narrativa como “sincera”, enquanto “NS” identifica narrativa “não sincera”.

QUADRO 55 - COMPARATIVO DETALHADO DOS DESEMPENHOS DE MODELOS VISUAIS INDIVIDUAL E COLETIVO

#	Referência	Teste		MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5 Individual		MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-256-31-31-5 Coletivo		Final		
				KLd:15		KLn				
		Sujeito	Esperado	Escore	Detectado	Escore	Detectado	Soma bruta	Escore	Detectado
0	S1-P7-13	S1-P7-1	NS	0,696	S	0,267	S	0,964	0,746	S
1	S1-P7-13	S1-P7-2	S	0,581	S	0,446	S	1,027	0,773	S
2	S1-P7-13	S1-P7-3	S	0,402	S	-1,000	NS	-0,598	-0,536	NS
3	S1-P7-13	S1-P7-4	NS	0,426	S	0,473	S	0,899	0,716	S
4	S1-P7-13	S1-P7-5	S	0,380	S	0,451	S	0,831	0,681	S
5	S1-P7-13	S1-P7-6	NS	-0,744	NS	0,233	S	-0,511	-0,470	NS
6	S1-P7-13	S1-P7-7	S	0,407	S	0,490	S	0,897	0,715	S
7	S1-P7-13	S1-P7-8	NS	0,286	S	-1,000	NS	-0,714	-0,613	NS
8	S1-P7-13	S1-P7-9	NS	0,390	S	0,374	S	0,764	0,643	S
9	S1-P7-13	S1-P7-10	S	0,633	S	0,334	S	0,968	0,748	S
10	S1-P7-13	S1-P7-11	S	0,539	S	-1,000	NS	-0,461	-0,431	NS
11	S1-P7-13	S1-P7-12	S	0,445	S	-1,000	NS	-0,555	-0,504	NS
12	S1-P8-3	S1-P8-1	S	0,874	S	0,379	S	1,252	0,849	S
13	S1-P8-3	S1-P8-2	S	-0,579	NS	0,018	S	-0,561	-0,509	NS
14	S1-P8-3	S1-P8-4	S	0,498	S	-0,890	NS	-0,392	-0,373	NS

15	S1-P8-3	S1-P8-5	S	0,000	***	0,000	***	0,000	0,000	***
16	S1-P8-3	S1-P8-6	S	0,820	S	0,390	S	1,210	0,837	S
17	S1-P8-3	S1-P8-7	S	0,273	S	0,211	S	0,484	0,450	S
18	S1-P9-11	S1-P9-1	NS	-0,576	NS	-0,047	NS	-0,623	-0,553	NS
19	S1-P9-11	S1-P9-2	NS	-0,268	NS	0,562	S	0,293	0,285	S
20	S1-P9-11	S1-P9-3	S	-0,691	NS	0,565	S	-0,126	-0,125	NS
21	S1-P9-11	S1-P9-5	NS	-0,394	NS	0,541	S	0,147	0,146	S
22	S1-P9-11	S1-P9-6	NS	-0,568	NS	0,541	S	-0,027	-0,027	NS
23	S1-P9-11	S1-P9-8	NS	-0,819	NS	0,513	S	-0,306	-0,297	NS
24	S1-P9-11	S1-P9-9	S	-0,499	NS	0,366	S	-0,133	-0,133	NS
25	S1-P9-11	S1-P9-10	S	-0,386	NS	-0,197	NS	-0,583	-0,525	NS
26	S2-P1-1	S2-P1-2	S	0,032	S	-0,148	NS	-0,116	-0,116	NS
27	S2-P1-1	S2-P1-3	S	-0,061	NS	-1,000	NS	-1,061	-0,786	NS
28	S2-P2-5	S2-P2-1	NS	-0,025	NS	-0,040	NS	-0,065	-0,065	NS
29	S2-P2-5	S2-P2-2	NS	-0,126	NS	0,229	S	0,102	0,102	S
30	S2-P2-5	S2-P2-3	NS	-0,510	NS	0,380	S	-0,130	-0,129	NS
31	S2-P3-5	S2-P3-1	NS	0,668	S	-0,734	NS	-0,066	-0,066	NS
32	S2-P3-5	S2-P3-2	NS	-0,150	NS	-1,000	NS	-1,150	-0,818	NS
33	S2-P3-5	S2-P3-3	NS	-0,011	NS	-0,156	NS	-0,167	-0,165	NS
34	S2-P3-5	S2-P3-4	NS	-0,317	NS	0,323	S	0,006	0,006	S
35	S2-P4-2	S2-P4-1	S	0,993	S	0,221	S	1,213	0,838	S
36	S2-P5-3	S2-P5-1	NS	-0,102	NS	-0,722	NS	-0,823	-0,677	NS
37	S2-P5-3	S2-P5-2	S	-0,319	NS	0,383	S	0,063	0,063	S
38	S2-P6-5	S2-P6-1	NS	0,038	S	-1,000	NS	-0,962	-0,745	NS
39	S2-P6-5	S2-P6-2	NS	-0,286	NS	-0,842	NS	-1,129	-0,811	NS
40	S2-P7-8	S2-P7-1	NS	-0,107	NS	-1,000	NS	-1,107	-0,803	NS
41	S2-P7-8	S2-P7-4	NS	-0,039	NS	-1,000	NS	-1,039	-0,778	NS
42	S2-P7-8	S2-P7-5	NS	-0,059	NS	-1,000	NS	-1,059	-0,785	NS
43	S2-P7-8	S2-P7-6	S	0,150	S	0,147	S	0,297	0,289	S
44	S2-P7-8	S2-P7-7	S	0,429	S	0,132	S	0,561	0,509	S
45	S2-P8-5	S2-P8-1	NS	-0,667	NS	-1,000	NS	-1,667	-0,931	NS
46	S2-P9-4	S2-P9-1	S	0,470	S	0,398	S	0,868	0,700	S
47	S2-P9-4	S2-P9-2	S	0,000	***	0,000	***	0,000	0,000	***
48	S2-P9-4	S2-P9-3	S	0,836	S	0,130	S	0,966	0,747	S

FONTE: DADOS DA PESQUISA (2023)

As células foram coloridas para facilitar a identificação das detecções frente ao resultado esperado. As células em vermelho identificam “não sinceridade”, enquanto as verdes, “sinceridade”.

Os dois casos (15 e 47) foram os que tiveram seus dados visuais totalmente removidos em função da insuficiência de condições para extração de características. São dois casos em que a detecção visual é impossível e foram contabilizados como erros, o que reduziu a acurácia balanceada dos dois componentes e do modelo final.

Dados os resultados, as duas hipóteses aventadas para os modelos coletivos acústico e verbal (insuficiência de variedade de dados e padrões individuais mais discriminantes) são aplicáveis a este caso também. Analogamente, existe a possibilidade da ocorrência conjugada de ambas na produção do resultado.

4.4 Modelo de Sinceridade Multimodal

Com os modelos individuais monomodais treinados, o próximo passo foi de avaliar a fusão destes na forma de um modelo multimodal. O Escore de Sinceridade parcial de cada um dos modelos monomodais foi fundido na forma de um Escore de Sinceridade final conforme estabelecido nos encaminhamentos metodológicos.

Os melhores exemplares de cada conjunto de modelos monomodais treinados foram selecionados para compor o modelo multimodal, cujo desempenho está apresentado resumidamente no Quadro 56.

QUADRO 56 - DESEMPENHO RESUMIDO DO MODELO MULTIMODAL

Modelo	Métrica	Modalid.	Acerto	Acc	B Acc	V S	V NS	F S	F NS
MAAE-65_67-71-67_65-mse-32-65-65-5	MSE	Acústica	37	0,755	0,758	15	22	2	10
MAAE-16_02_21_03_28_04-13-28_04_21_03_16_02-mse-32-13-13-10	MSE	Verbal	35	0,714	0,717	15	20	4	10
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	KLd:15	Visual	35	0,714	0,715	17	18	6	8
		Multimodal	35	0,714	0,716	16	19	5	9

FONTE: DADOS DA PESQUISA (2023)

Percebe-se que a fusão dos modelos não produziu o efeito sinérgico de aumentar a acurácia balanceada final. O modelo parcial de mais alta acurácia foi o acústico, mas os modelos verbais e visuais não contribuíram para reverter os erros de classificação. Percepção mais granular de como os modelos parciais interagiram para produzir o resultado final pode ser alcançado pela observação do Quadro 57.

QUADRO 57 - DESEMPENHO DETALHADO DO MODELO MULTIMODAL

#	Ref	Teste		MAAE-65_67-71-67_65-mse-32-65-65-5 Acústico		MAAE-13_02_15_02_17_02-13-17_02_15_02_13_02-mse-32-13-13-10 Verbal		MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5 Visual		Final		
				MSE		MSE		KLd:15				
		Sujeito	Esp	Esc	Det	Esc	Det	Esc	Det	Soma bruta	Esc	Det
0	S1-P7-13	S1-P7-1	NS	-0,031	NS	0,136	S	0,696	S	0,801	0,665	S
1		S1-P7-2	S	0,213	S	0,104	S	0,581	S	0,899	0,716	S
2		S1-P7-3	S	0,045	S	0,056	S	0,402	S	0,503	0,465	S
3		S1-P7-4	NS	-0,044	NS	0,167	S	0,426	S	0,548	0,499	S

4		S1-P7-5	S	-0,137	NS	0,047	S	0,380	S	0,291	0,283	S
5		S1-P7-6	NS	-0,046	NS	-0,224	NS	-0,744	NS	-1,014	-0,767	NS
6		S1-P7-7	S	0,024	S	-0,138	NS	0,407	S	0,293	0,284	S
7		S1-P7-8	NS	-0,041	NS	0,135	S	0,286	S	0,380	0,362	S
8		S1-P7-9	NS	-0,035	NS	-0,226	NS	0,390	S	0,129	0,129	S
9		S1-P7-10	S	-0,096	NS	0,109	S	0,633	S	0,646	0,569	S
10		S1-P7-11	S	0,331	S	0,076	S	0,539	S	0,946	0,738	S
11		S1-P7-12	S	0,201	S	0,285	S	0,445	S	0,931	0,731	S
12	S1-P8-3	S1-P8-1	S	0,363	S	0,174	S	0,874	S	1,411	0,888	S
13		S1-P8-2	S	0,082	S	-0,003	NS	-0,579	NS	-0,499	-0,462	NS
14		S1-P8-4	S	0,156	S	-0,155	NS	0,498	S	0,499	0,461	S
15		S1-P8-5	S	0,008	S	0,004	S	0,000	***	0,012	0,012	S
16		S1-P8-6	S	0,069	S	-0,145	NS	0,820	S	0,744	0,632	S
17		S1-P8-7	S	0,072	S	-0,237	NS	0,273	S	0,108	0,107	S
18	S1-P9-11	S1-P9-1	NS	-0,009	NS	-0,254	NS	-0,576	NS	-0,839	-0,685	NS
19		S1-P9-2	NS	-0,296	NS	-0,464	NS	-0,268	NS	-1,029	-0,773	NS
20		S1-P9-3	S	-0,040	NS	-0,082	NS	-0,691	NS	-0,813	-0,671	NS
21		S1-P9-5	NS	-0,026	NS	-0,436	NS	-0,394	NS	-0,856	-0,694	NS
22		S1-P9-6	NS	-0,310	NS	-0,008	NS	-0,568	NS	-0,887	-0,710	NS
23		S1-P9-8	NS	0,034	S	-0,465	NS	-0,819	NS	-1,249	-0,848	NS
24		S1-P9-9	S	-0,053	NS	-0,277	NS	-0,499	NS	-0,829	-0,680	NS
25		S1-P9-10	S	0,088	S	-0,208	NS	-0,386	NS	-0,506	-0,467	NS
26	S2-P1-1	S2-P1-2	S	-0,231	NS	-0,248	NS	0,032	S	-0,448	-0,420	NS
27		S2-P1-3	S	-0,851	NS	0,003	S	-0,061	NS	-0,910	-0,721	NS
28	S2-P2-5	S2-P2-1	NS	-0,545	NS	-0,262	NS	-0,025	NS	-0,832	-0,681	NS
29		S2-P2-2	NS	-0,630	NS	-0,209	NS	-0,126	NS	-0,965	-0,747	NS
30		S2-P2-3	NS	-0,543	NS	-0,321	NS	-0,510	NS	-1,374	-0,880	NS
31	S2-P3-5	S2-P3-1	NS	-0,137	NS	-0,440	NS	0,668	S	0,091	0,091	S
32		S2-P3-2	NS	-0,638	NS	-0,604	NS	-0,150	NS	-1,392	-0,884	NS
33		S2-P3-3	NS	-0,119	NS	-0,446	NS	-0,011	NS	-0,576	-0,520	NS
34		S2-P3-4	NS	-0,414	NS	-0,496	NS	-0,317	NS	-1,227	-0,842	NS
35	S2-P4-2	S2-P4-1	S	-0,447	NS	-0,065	NS	0,993	S	0,481	0,447	S
36	S2-P5-3	S2-P5-1	NS	-0,313	NS	0,108	S	-0,102	NS	-0,306	-0,297	NS
37		S2-P5-2	S	-0,885	NS	0,150	S	-0,319	NS	-1,055	-0,784	NS
38	S2-P6-5	S2-P6-1	NS	-0,015	NS	-0,082	NS	0,038	S	-0,059	-0,059	NS
39		S2-P6-2	NS	-0,181	NS	-0,264	NS	-0,286	NS	-0,731	-0,624	NS
40	S2-P7-8	S2-P7-1	NS	0,036	S	-0,343	NS	-0,107	NS	-0,413	-0,391	NS
41		S2-P7-4	NS	-0,248	NS	0,024	S	-0,039	NS	-0,264	-0,258	NS
42		S2-P7-5	NS	-0,299	NS	-0,285	NS	-0,059	NS	-0,644	-0,568	NS
43		S2-P7-6	S	-0,035	NS	-0,244	NS	0,150	S	-0,128	-0,127	NS
44		S2-P7-7	S	0,294	S	-0,126	NS	0,429	S	0,597	0,535	S
45	S2-P8-5	S2-P8-1	NS	-0,758	NS	-0,230	NS	-0,667	NS	-1,655	-0,930	NS
46	S2-P9-4	S2-P9-1	S	0,072	S	-0,194	NS	0,470	S	0,348	0,335	S
47		S2-P9-2	S	0,013	S	-0,552	NS	0,000	***	-0,539	-0,492	NS
48		S2-P9-3	S	-0,014	NS	-0,239	NS	0,836	S	0,583	0,525	S

FONTE: DADOS DA PESQUISA (2023)

A coluna “Esp” informa a classificação esperada, enquanto as colunas “Esc” informam o Escore de Sinceridade calculado e as colunas “Det” as classificações detectadas. Os casos 15 e 47 foram os que não produziram características visuais em função das condições de vídeo que impediram a extração por parte do OpenFace.

A redução da acurácia parece demonstrar que os modelos monomodais tiveram critérios distintos de classificação. As perturbações na forma do erro de reconstrução tiveram dissidências significativas em cada modalidade, evidenciando que cada modelo parece ter sofrido de diferentes efeitos de confusão.

O que veio a intensificar os erros de classificação final foram altos graus de confiança em detecções errôneas e não tão altos graus de confiança em detecções corretas. Nos casos em que este fenômeno ocorreu, o processo de fusão pelo Escore de Sinceridade privilegiou os erros.

No entanto, os casos 23, 35, 36, 40, 41 e 48 demonstram que o uso do Escore de Sinceridade contribuiu para maior grau de acerto se comparado com a votação. Em tais casos, a votação teria privilegiado duas classificações parciais de baixa confiança e desprivilegiado uma classificação de alta confiança, produzindo uma classificação final errônea que reduziria ainda mais a acurácia modelo.

Os casos 8 e 31 representam o efeito inverso, onde o uso da Escore de Sinceridade ocasionou um erro de julgamento que não teria ocorrido pela votação.

4.5 Modelo de Sinceridade Multicomponente

Atendendo ao disposto nos encaminhamentos metodológicos e com os modelos individuais e coletivos treinados para as três modalidades, foi possível construir um modelo multicomponente que integrasse a todos. O desempenho resumido deste modelo final está apresentado no Quadro 58.

QUADRO 58 - DESEMPENHO DO MODELO MULTICOMPONENTE COMPLETO

Modelo	Card.	Métrica	Modal.	Act	Acc	B Acc	V S	V NS	F S	F NS
MAAE-65_67-71-67_65-mse-32-65-65-5	Ind	MSE	Acústico	37	0,755	0,758	15	22	2	10
MAAE-65_67-67-67_65-mse-32-65-65-5	Col	KLd:15	Acústico	26	0,531	0,525	20	6	18	5
MAAE-13_02_15_02_17_02-13-17_02_15_02_13_02-mse-32-13-13-10	Ind	MSE	Verbal	29	0,592	0,596	10	19	5	15
AAE_4-13_02_13_02-9-13_02_13_02-mse-32	Col	MSE	Verbal	28	0,571	0,573	13	15	9	12
MAAE-31_02_29_02_27_02-29-27_02_29_02_31_02-mse-32-31-31-5	Ind	KLd:15	Visual	35	0,714	0,715	17	18	6	8
MAAE-31_02_33_02_35_02-35-35_02_33_02_31_02-mse-32-31-31-10	Col	MSE	Visual	26	0,531	0,528	16	10	14	9

Multicomponente	29	0,592	0,591	16	13	11	9

FONTE: DADOS DA PESQUISA (2023)

O modelo multicomponente teve seu desempenho afetado pela presença dos componentes coletivos, que para a atual configuração dos dados, não foi capaz de alcançar um nível de distinção de sinceridade de narrativas que acompanhasse os respectivos modelos individuais.

4.6 Discussão

Os experimentos realizados ao longo de percurso metodológico deste estudo diferem de maneira fundamental dos identificados na maciça maioria dos artigos recuperados na revisão sistemática de literatura que culminou com a publicação artigo no periódico PloS One (Constancio et al., 2023).

Enquanto aqueles artigos propunham a identificação de padrões coletivos para as expressões de verdade e mentira, para em seguida buscar enquadrar um novo indivíduo nestes padrões e assim detectar uma eventual expressão acústica, verbal, visual ou multimodal de não sinceridade, a proposta aqui apresentada procura por padrões individuais de sinceridade e a eventual identificação de antipadrões, que são interpretados como não sinceridade.

Os modelos de Aprendizado de Máquina das abordagens coletivas entendem o problema de detectar mentiras como uma tarefa de classificação e portanto requerem o uso de dados rotulados (Goodfellow; Yoshua; Courville, 2016). No cenário destas abordagens, o modelo é treinado uma única vez e as sessões de detecção são realizadas pela submissão de novas narrativas àquele, o qual terá como saída a probabilidade de as mesmas serem sinceras ou não.

Por outro lado, a abordagem individual entende o problema de detectar mentiras como a detecção de uma anomalia, dispensando os dados rotulados, mas requerendo um processo de calibração inicial para cada indivíduo, de tal forma que o modelo aprenda quais as suas expressões de sinceridade (suas expressões de normalidade). Posteriormente, as demais narrativas são submetidas ao modelo treinado que retornará a probabilidade de as mesmas se conformarem ou não aos padrões de sinceridade previamente aprendidos. Neste cenário, não existe a necessidade de dados rotulados, mas há a necessidade de narrativas reconhecidamente sinceras para a calibragem, ou seja, o aprendizado das expressões de sinceridade.

A proposta de abordar o problema de detectar mentiras como um problema de detectar anomalias está fundamentada na noção de que as pistas expressas para a identificação de uma eventual não sinceridade são altamente dependentes das circunstâncias e de caracteres próprios do interrogado (Porter; Brinke, 2010b). Em outras palavras, a abordagem coletiva é sensível a aspectos idiossincráticos da mentira. Essa noção é sugerida pela própria literatura que discute a detecção de mentiras, visto que é frequente encontrar descrições de pistas que ocorrem “com frequência” ou que “foram observadas” em alguns experimentos. Nota-se que essas pistas não se manifestam sempre e quando se manifestam, não se dão necessariamente da mesma forma.

No entanto, a revisão de literatura identificou variados níveis de desempenho, em uma diversidade de configurações e abordagens. É conveniente comparar esses níveis de desempenho e abordagem com a proposta substanciada por esta pesquisa. Tais comparações podem auxiliar a identificar limitações, virtudes e oportunidades para a continuidade de estudos dentro desta estratégia.

Os resultados trazidos pelos experimentos ofereceram diversos níveis de desempenhos monomodais e multimodal, medidos pela acurácia balanceada e acurácia regular. A comparação destes níveis de desempenho permite avaliar quão promissora é a proposta da detecção de mentiras pelo Modelo de Sinceridade.

O Quadro 59 é uma adaptação de um quadro existente na revisão de literatura, onde os estudos monomodais acústicos são listados. Incluído naquele quadro está o resultado do melhor Modelo de Sinceridade alcançado a partir dos experimentos realizados.

QUADRO 59 - COMPARATIVO DE DESEMPENHOS DE ESTUDOS MONOMODAIS ACÚSTICOS

#	Referência	Acr.	Técnica	Características
1	Fernandes; Ullah, 2021a	1,000	Rede Neural	Parâmetros cepstrais, Energia
2	Fernandes; Ullah, 2021b	0,917	Rede Neural	Parâmetros cepstrais, Energia
3	Nasri; Ouarda; Alimi, 2016	0,864	SVM	MFCC, Tom vocal
4	Sanaullah; Gopalan, 2013	0,833	Rede Neural	*Bark*, Energia
5	Tao et al., 2019	0,825	SVM	MFCC, Energia, Tom vocal, *Zero-crossing*
6	**Esta tese, 2023**	**0,755**	**Autoencoder Atencional *Multihead***	**ComParE 2016**
7	Xie et al., 2018	0,749	Rede Neural	Parâmetros espectrais, Tom vocal
8	Velichko et al., 2018	0,696	Árvore de decisão	INTERSPEECH 2013
9	Fu et al., 2019	0,628	Rede Neural	INTERSPEECH 2009

FONTE: ADAPTADO DE CONSTÂNCIO (2023)

Para efeitos de comparação, a coluna “Acr.” apresenta a acurácia relatada nos respectivos artigos, assim como a acurácia resultante do experimento que atingiu o mais alto nível de desempenho.

Os dois estudos de mais alto desempenho envolvem os mesmos autores e são abordagens diferentes sobre o mesmo cenário de experimentação, o registro em áudio de três sessões de interrogatório policial de um mesmo suspeito que foi posteriormente considerado culpado, num conjunto total de 12 respostas. Tais respostas consistiam unicamente na palavra “no” (“não” em inglês), que puderam ser posteriormente classificadas com sinceras e não sinceras a partir da investigação policial conduzida.

Em ambos os casos foram utilizadas duas diferentes arquiteturas de redes neurais artificiais (Levenberg-Marquardt e LSTM) e as características foram parâmetros cepstrais e energia. A principal diferença entre eles está no conjunto de características acústicas utilizadas, visto que no segundo colocado experimentos foram realizados com a redução da dimensionalidade da entrada por meio do algoritmo PCA.

Em comparação com o estudo encaminhado nesta pesquisa, o conjunto de dados MMDDD-PtBr é mais diversificado (12 indivíduos com 61 narrativas ao todo), sendo que cada narrativa é constituída por uma frase completa e não apenas uma única palavra. O conjunto características extraídas pelo OpenSMILE, conhecido como “ComParE 2016” (conjunto de características utilizados na confecção do MMDDD-PtBr), inclui aquelas características.

O terceiro colocado é um estudo que apresenta um conjunto de dados chamado ReliDDBI, composto por narrativas gravadas de 40 voluntários masculinos e femininos. O processo está baseado no treinamento de 137.640 fragmentos de áudio de 30 milissegundos cada e a classificação foi realizada pelo algoritmo SVM.

Neste caso, chama a atenção o volume de vetores utilizados para treinamento, mas não são frases completas e sim partículas de 30 milissegundos que, segundo os autores, já são suficientes para a detecção da mentira por humanos. Os autores relatam o poder discriminante das frequências MFCC e do tom vocal, sinais que se encontram no ComParE 2016. Aqui, novamente, existe a diferença na composição dos itens de dados, visto que o conjunto de dados proposto nesta tese é baseado em sentenças completas e não fragmentos.

O quarto colocado é um estudo baseado em um cenário similar ao primeiro, seis respostas foram coletadas de três sessões de interrogatório, onde a respeito era unicamente “no”. O teste do polígrafo foi utilizado para classificar as respostas como sinceras ou não, e o modelo de classificação foi uma rede neural artificial do tipo Levenberg-Marquandt. As características acústicas foram as faixas da escala acústica Bark e na energia significante presente nelas. Tais características não estão presentes no conjunto extraído pelo OpenSMILE.

O quinto colocado foi um estudo que explorou um conjunto de dados chamado KWOLF, formado por 388 frases extraídas dos registros em vídeo de um jogo conhecido como “Werewolf Kill”. Os autores não são específicos ao afirmar que os registros de áudio são frases, mas informam que têm duração de dois a seis segundos. As características vocais foram as frequências MFCC, tom e energia vocais e o *zero-crossing*, todas características pertencentes ao ComParE 2016.

Os estudos nas posições de sete a nove não serão comentados.

Analogamente, no Quadro 60 se apresentam os estudos de natureza monomodal e modalidade verbal nos mesmos moldes do quadro anterior.

QUADRO 60 - COMPARATIVO DE DESEMPENHOS DE ESTUDOS MONOMODAIS VERBAIS

#	Referência	Acr.	Técnica	Características
1	Pak; Zhou, 2015	0,980	Árvore de decisão	Categorias LIWC, Complexidade de sintaxe, Unigramas
2	Barsever; Singh; Neftci, 2020	0,936	Rede Neural	BERT *embeddings*
3	Feng; Banerjee; Choi, 2012	0,912	SVM	Bigramas, Categorias sintáticas, Complexidade de sintaxe, Unigramas
4	Briscoe; Appling; Hayes, 2014	0,910	Gradient Boosting	Emoticons, Informalidade, Sentimentos, Complexidade de sintaxe
5	Kleinberg et al., 2018	0,774	SVM	Categorias LIWC, Entidades, Processos psicológicos
6	Mihalcea; Pérez-Rosas; Burzo, 2013	0,737	SVM	Unigramas
7	**Esta tese, 2023**	**0,717**	**Autoencoder Atencional *Multihead***	**Categorias sintáticas, duração, hesitação, sentimentos, pessoa verbal**
8	Kleinberg; Verschuere, 2021	0,690	Random Forest	Categorias LIWC, categorias sintáticas
9	Fornaciari; Poesio, 2012	0,660	SVM	Categorias LIWC, Medidas léxicas, N-gramas, categorias sintáticas
10	Rubin; Conroy, 2012	0,650	Árvore de decisão	Categorias LIWC, Medidas léxicas
11	Rubin; Conroy, 2011	0,650	SMO	Categorias LIWC, Medidas léxicas

12	Mbaziira; Murphy, 2018	0,633	Rede Neural	Complexidade de sintaxe

FONTE: ADAPTADO DE CONSTÂNCIO (2023)

O estudo verbal de mais alto desempenho procurou combinar características estruturais (grafos que relacionam pessoas em uma conversação) e linguísticas (Unigramas, Categorias LIWC e complexidade de sintaxe) presentes em narrativas para atingir um critério discriminante entre sinceridade e não sinceridade. A pesquisa construiu um conjunto de dados composto por 142 narrativas extraídas dos registros de um jogo online chamado de Mafia Game. A classificação foi feita por meio do modelo de Árvore de decisão.

O modelo verbal proposto na pesquisa desta tese não inclui diretamente palavras ou suas representações no conjunto de dados, enquanto no artigo em questão existem os unigramas. Analogamente, os autores exploraram as categorias LIWC, prática repetida em muitos outros estudos, mas que não ocorreu nesta pesquisa por ser um recurso que incorre em custos. No entanto, foram utilizadas categorias sintáticas, que operaram para identificar as categorias de palavras utilizadas em uma narrativa sincera para formar o padrão de expressão verbal.

A complexidade de sintaxe é referida em alguns estudos como um fator de potencial capacidade distintiva. No modelo apresentado por este estudo, a complexidade de sintaxe pode ser potencialmente aprendida pelos Autoencoders profundos, pois as diversas camadas codificam padrões formados pelos atributos que descrevem cada palavra, mas não foi possível averiguar se esse fato efetivamente se deu e se participou do processo de distinção entre sinceridade e não sinceridade implementado pelo Modelo de Sinceridade.

O estudo em segunda colocação experimentou um modelo de classificação baseado em vetores gerados pelo BERT (BERT *embeddings*). Estes vetores são representações numéricas multidimensional de palavras geradas por uma rede pré-treinada chamada BERT (*Bidirectional Encoder Representations from Transformers*).

Os autores testaram seu classificador sobre um conjunto de 1600 narrativas extraídas do “Ott Deceptive Opinion Spam Corpus” e relatam ter percebido um padrão formulaico em sentenças não sinceras.

Após a conversão das sentenças em vetores BERT, estes foram utilizados para treinar uma outra rede neural, especializada na tarefa de detecção de mentiras.

Em comparação com a proposta desta tese, o estudo baseado em BERT se assemelha por operar sobre as narrativas transcritas na forma dos vetores BERT, que

preservam a estrutura das frases, apenas substituindo cada palavra por um vetor. Os experimentos verbais aqui realizados operam de forma similar, mas os vetores que alimentaram os Autoencoders são formados pelas 13 características escolhidas para compor o MMDDD-PtBr.

O estudo que assumiu a terceira colocação combinou a complexidade de sintaxe com unigramas, bigramas e categorias sintáticas para treinar um classificador baseado em SVM a partir de um conjunto de dados de 2.692 avaliações de restaurantes italianos. A complexidade da sintaxe é obtida a partir das árvores geradas pelo PCFG (*Probabilistic Context-Free Grammar*), que são codificadas para operar como características juntamente com as demais.

Neste caso, assim como no caso do artigo em primeira colocação, a estrutura do texto não é preservada no conjunto de dados, passando a ser representada pelas diversas características selecionadas pelos autores.

O quarto colocado explora um classificador baseado em Gradient Boosting treinado sobre um conjunto de 254 declarações fornecidas por voluntários em um cenário de conversação por computador. Cada declaração é codificada na forma de emoticons complexidade de sintaxe, informalidade e sentimento.

Assim, como nesta tese, os sentimentos são derivados de um dicionário de palavras associadas a emoções chamado AFINN. A informalidade é medida pela contagem de palavras que precisam de correção de escrita e os emoticons operam como rótulos diretos de emoções expressas nas sentenças. Este é outro caso em que a estrutura do texto é perdida em favor de diversas características descritivas que são então utilizadas para alimentar o classificador.

O quinto colocado treinou um classificador baseado em SVM com 142 sentenças de planos para atividades em um fim de semana. A hipótese era de que planos não sinceros (sem intenção de realização) seriam linguisticamente menos detalhados. As características exploradas foram categorias LIWC, a presença de entidades e os processos psicológicos operantes durante a formulação das sentenças. Infelizmente, não foi possível identificar as características que caracterizam aqueles processos ou como as mesmas foram extraídas.

O sexto colocado experimentou um classificador SVM treinado a partir de unigramas extraídos de 140 transcrições de vídeos gravados a partir de narrativas de voluntários. Os demais colocados não serão comentados.

O Quadro 61 registra o comparativo dos desempenhos dos estudos monomodais visuais, seguindo a mesma lógica dos dois quadros anteriores.

QUADRO 61 - COMPARATIVO DE DESEMPENHOS DE ESTUDOS MONOMODAIS VISUAIS

#	Referência	Acr.	Técnica	Características
1	Ding et al., 2019	0,970	Rede Neural	Facial expressions, Head motion
2	Labibah; Nasrun; Setianingsih, 2018	0,950	Árvore de decisão	Ângulo de visão, dilatação de pupila
3	Avola et al., 2019	0,768	SVM	Ângulo de visão, expressões faciais, ângulo da cabeça
4	**Esta tese, 2023**	**0,714**	**Autoencoder Atencional *Multihead***	**Ângulo de visão, expressões faciais, ângulo da cabeça**
4	Bailey et al., 2015	0,703	Regressão logística	Expressões faciais
5	Rybar; Bielikova, 2016	0,620	SVM	Sacada ocular, dilatação de pupila, tempo de resposta
6	Islam et al., 2021	0,615	SVM	Expressões faciais
7	Takabatake; Shimada; Saitoh, 2018	0,552	SVM	Microexpressões faciais

FONTE: ADAPTADO DE CONSTÂNCIO (2023)

O estudo em primeira colocação empregou um complexo modelo para explorar a relação entre expressões faciais e movimentos da cabeça como um fator de distinção entre discursos sinceros e não sinceros. As características foram extraídas da parte visual do RLTDDD e foram utilizadas para treinar redes neurais artificiais com o objetivo de aprender a relação temporal entre expressões e movimentos da cabeça.

Quando comparado com os experimentos com Modelos de Sinceridade, a proposta do artigo se baseia em dois classificadores (um para expressões e outro para os movimentos) que são combinados por um módulo denominado *Cross-stream fusion*. Já com os Modelos de Sinceridade, a eventual relação temporal que possa existir entre as características é aprendida pelo mecanismo de Atenção dos modelos atencionais *multihead*, processo mais simples por estar inteiramente contido no Autoencoder na forma de uma camada de Atenção.

O segundo estudo explorou um classificador baseado em Árvore de decisão treinado com 40 registros de voluntários quando estes respondiam a um questionário. As características extraídas foram a dilatação da pupila e o ângulo de visada. Os autores apontam alto grau de discriminância do ângulo de visada, declarando que o

lado para o qual uma pessoa olha indica o lado do cérebro que está em atividade, que por sua vez identificas os centros de memória (sinceridade) ou criatividade (mentira). No entanto, tais sinais já foram desafiados por outros estudos (Burgoon; Guerrero; Floyd, 2016; Vrij, 2008), que argumentos em favor da aversão de visada (*gaze aversion*) e não especificamente em ângulos particulares de visada, o que poderia ser considerado como uma instanciação do Perigo de Brokaw (Ekman, 1992).

Embora os estudos encaminhados nesta tese explorem o ângulo de visada, tal característica não é interpretada como um sinal inequívoco de não sinceridade, especialmente no que diz respeito a ângulos específicos. Essa característica figura como um indicador a mais que é aprendido durante o treinamento do Modelo de Sinceridade, mas seu significado é específico para cada sujeito.

Por limitações do OpenFace e dos próprios vídeos utilizados para conceber o MMDDD-PtBr, a dilatação de pupila não pôde ser aproveitada para característica, ainda que estudos relacionem este fenômeno ao aumento de atividade cognitiva (*cognitive load*), e, esta, com o ato de mentir (Burgoon; Guerrero; Floyd, 2016; DePaulo et al., 2003; Ekman, 1992; Vrij, 2008).

O terceiro colocado treinou um classificador SVN com kernel RBF, utilizando as características dos 121 vídeos presentes no RLTDDD (Pérez-Rosas et al., 2015). As características foram as mesmas utilizadas por esta tese, o ângulo de visada, as expressões faciais e o ângulo de cabeça, já que os autores utilizaram o OpenFace para extração de características.

O kernel de função RBF utilizado nos experimentos daquele estudo reforça a noção de que as relações entre as características discriminantes entre sinceridade e não sinceridade são não lineares, também capturadas por Autoencoders profundos como os utilizados para construir os Modelos de Sinceridade.

Finalmente, no Quadro 62 estão listados os desempenhos dos diversos estudos multimodais identificados na revisão de literatura, incluindo também o melhor resultado alcançado pelos experimentos com Modelos de Sinceridade multimodais.

Apenas os estudos multimodais que exploraram as modalidades acústica, verbal e visual foram incluídos naquele quadro, embora a revisão de literatura tenha identificado outras combinações.

QUADRO 62 – COMPARATIVO DE DESEMPENHOS DE ESTUDOS MULTIMODAIS

#	Referência	Acr.	Técnica	Características

1	Venkatesh; Ramachandra; Bours, 2019	0,970	Diversas	Body motion, Facial micro-expressions, MFCC, N-grams
2	Gogate; Adeel; Hussain, 2018	0,964	Rede Neural	Facial expressions, GloVe embeddings, Hand motion, INTERSPEECH 2013
3	Mathur; Matarić, 2020	0,840	AdaBoost	Eye gaze, Facial affect, Facial expressions, Head motion, LIWC categories, MFCC, Spectral parameters, Voice pitch, Voice quality
4	Jaiswal; Tabibu; Bajpai, 2016	0,790	SVM	Facial expressions, MFCC, POS tags, Prosody, Sentiment, Unigrams, Voice energy
5	**Esta tese, 2023**	**0,714**	**Autoencoder Atencional *Multihead***	**ComParE 2016, Categorias sintáticas, duração, hesitação, sentimentos, pessoa verbal, Ângulo de visão, expressões faciais, ângulo da cabeça**
6	Kamboj et al., 2021	0,700	Árvore de decisão	Eye gaze, Facial emotion, Facial expressions, GloVe embeddings, Head pose, INTERSPEECH 2009, INTERSPEECH 2013, LIWC categories, POS tags, Sentiment, Unigrams

FONTE: ADAPTADO DE CONSTÂNCIO (2023)

O artigo que relatou o mais alto nível de acurácia utilizou os 121 vídeos coletados pelo RLTDDD para construir uma proposta multimodal baseada em diferentes modelos combinados. As características acústicas alimentaram um classificador baseado no Spectral Regression Kernel Discriminant Analysis (SRKDA), as características verbais foram classificadas pelo SVM e as características visuais por um classificador AdaBoost. Estes três classificadores foram combinados por meio de um processo de votação simples.

As características acústicas foram os coeficientes 13 cepstrais (frequências MFCC) e mais um coeficiente de energia. Tais características estão incluídas nos 65 atributos extraídos pelo OpenSMILE e utilizados nos experimentos com Modelos de Sinceridade.

As características verbais foram contagens de n-gramas (unigramas, bigramas e trigramas), que diferem grandemente das características presentes no MDDD-PtBr, que está baseado em características descritivas das palavras das narrativas, mas não as palavras propriamente.

As características visuais, segundo os autores, são 39 características relacionadas com microexpressões e micromovimentos, mas o texto não é específico

em informar suas origens. Os Modelos de Sinceridade visuais experimentados nesta tese foram alimentados com 31 características extraídas pelo OpenFace.

O modelo de fusão dos três classificadores monomodais foi a votação simples, enquanto nos experimentos com os Modelos de Sinceridade foi uma votação ponderada pela confiança, implementada por meio do Escore de Sinceridade multimodal.

O artigo em segunda colocação também utilizou o RLTDDD como fonte de dados para elaborar um modelo baseado em redes neurais profundas que explorou características acústicas extraídas do conjunto INTERSPEECH 2013 do OpenSMILE, GloVe *embeddings* como características verbais, e expressões faciais com movimentos das mãos para características visuais.

Os GloVe *embeddings*, a exemplo dos BERT *embeddings*, são representações vetoriais 300D de palavras extraídas por meio de um modelo de linguagem chamado GloVe. Tais representações são utilizadas para treinar uma rede neural convolucional em uma atividade similar ao que ocorre com os Modelos de Sinceridade verbais.

Os autores experimentaram modelos com fusão precoce e tardia das características e relatam que a primeira ofereceu melhores resultados porque os modelos aprenderam as correlações existentes entre as três modalidades. Nesta pesquisa, os Modelos de Sinceridade monomodais foram fundidos no que seria equivalente à fusão tardia, representando um contraste com o artigo em questão.

O artigo terceiro colocado é mais um a utilizar o RLTDDD como fonte de características multimodais e adiciona a estas o conceito de afeto facial (*facial affect*) como mais um conjunto de características de modalidade visual.

As características acústicas são extraídas pelo OpenSMILE e são as memas utilizadas nos Modelos de Sinceridade acústicos. As características verbais foram as categorias LIWC. Finalmente, as características de afeto são extraídas por um modelo especialmente treinado chamado AffWildNet, um conjunto de dados visuais especializado em afeto visual. As demais características visuais ações visuais, ângulo de visada e ângulo de cabeça (as mesmas utilizadas nos Modelos de Sinceridade).

Os autores relatam o uso de fusão precoce e um classificador baseado em SVM.

O quarto estudo multimodal fez uso de um subconjunto de 100 dos 121 vídeos existentes no RLTDDD, submetidos a um classificador SVM. As características

acústicas foram extraídas pelo OpenSMILE e compreendem 28 das 65 possíveis. As características verbais se limitaram a unigramas e as características visuais incluíram 18 ações faciais. As características foram combinadas de diversas formas, sendo que a que ofereceu os maiores níveis de desempenho foi a fusão precoce.

Ao avaliar comparativamente os desempenhos dos Modelos de Sinceridade em relação os estudos identificados na revisão sistemática de literatura, percebe-se que alguns daqueles relataram acurácias mais elevadas, algumas vezes operando de forma bem similar em matéria de características e algumas vezes bem diferentes.

Alguns estudos relataram que o uso de fusão precoce de modalidades oferece o benefício de dar aos modelos experimentados a oportunidade de aprender correlações entre elas, algo que efetivamente não foi aproveitado do processo de fusão pelo Escore de Sinceridade multimodal.

5 CONSIDERAÇÕES FINAIS

Entende-se que a comunicação desprovida de sinceridade pode representar risco pessoal e social em determinadas circunstâncias. Nestes casos, tais mensagens são entendidas como mentiras sérias, cuja descoberta precoce pode até mesmo salvar vidas. Tal fato consolida o problema que esta pesquisa procurou mitigar (“Distinguir a sinceridade da não sinceridade em uma narrativa de um sujeito durante uma comunicação”).

Em resposta, os encaminhamentos postos em movimento ao longo desta pesquisa propuseram a elaboração de um modelo de Aprendizado de Máquina para detecção de mentiras expressas em língua portuguesa. A pesquisa objetivou, por estes meios, contribuir para a ciência ao desenvolver uma abordagem alternativa para o problema de detecção de mentiras.

A abordagem desenvolvida está fundamentada na aplicação de Autoencoders, redes neurais artificiais treinadas por meio de aprendizado autossupervisionado, utilizadas em diversos cenários, dentre eles a detecção de anomalias. Neste sentido, a abordagem operada consistiu em entender a detecção de mentiras como a detecção de anomalias.

Visto que os Autoencoders foram treinados com dados extraídos de narrativas sinceras, os mesmos foram denominados como **Modelos de Sinceridade**. Diante deste modelo, as narrativas não sinceras figuram como anomalias, descritas por características que as fazem discrepar das narrativas sinceras.

Os Modelos de Sinceridade foram treinados a partir de narrativas presentes em um conjunto de dados anotados, especialmente construído para aplicação em experimentos de detecção de mentiras, denominado *“Multimodal Deception Detection Dataset for Brazilian Portuguese”*, MMDDD-PtBr. Os dados foram extraídos de sequências de vídeos, que permitiram a obtenção de características acústicas (descrevem aspectos sonoros de uma comunicação), verbais (descrevem aspectos linguísticos e paralinguísticos) e visuais (especificamente, descrevem aspectos faciais). No caso particular das características verbais, o conjunto de dados é o primeiro no mundo dedicado à língua portuguesa.

Uma metodologia foi delineada e aplicada, consistindo na cooperação de ferramentas existentes (OpenFace, OpenSMILE, SentiWordNet-PT-Br, SpaCy, MoviePy, Azure Speech-to-text, linguagem Python) com ferramentas de apoio

desenvolvidas (Sincronizador de palavras e recortador de vídeos) para viabilizar a construção do conjunto de dados em seu conteúdo atual, assim como sua eventual extensão, com a inclusão de novas narrativas.

No tocante aos modelos de redes neurais, ao todo 2.390 diferentes arquiteturas foram construídas em TensorFlow/Keras (1.100 acústicas, 470 verbais e 820 visuais). Estes modelos exploraram a presença e ausência do mecanismo de Atenção nas modalidades *single-head* e *multi-head*, além da aplicação de *dropout* e diferentes configurações de hiperparâmetros, com o objetivo de alcançar os mais altos níveis de desempenho de deteção monomodal, para em seguida serem combinados em um único modelo multimodal por meio de uma métrica proposta chamada de **Escore de Sinceridade**.

Os modelos construídos focaram principalmente em narrativas individuais, visto que a proposta original da pesquisa é a de explorar os efeitos situacionais e idiossincráticos atuantes no processo de detectar mentiras, ao invés de buscar a elaboração de um modelo coletivo que procurasse encontrar os padrões distintivos gerais. Ainda assim, a existência de um conjunto de dados anotado (rotulado) permitiu a experimentação de modelos coletivos, que foram harmonicamente operados com os modelos individuais, também graças ao Escore de Sinceridade.

Com a combinação dos modelos individuais, o modelo resultante multimodal alcançou a acurácia balanceada de 0,714. Adicionalmente, 15 modelos coletivos (cinco para cada modalidade) foram treinados e avaliados, tendo alcançado como o melhor desempenho as acurácias balanceadas de 0,571 para o modelo acústico, 0,612 para o modelo verbal e 0,571 para o modelo visual. Modelos multicomponentes (utilizando modelos individuais e coletivos) foram também avaliados, tendo atingido a acurácia balanceada máxima de 0,591.

Considerando os objetivos específicos desta pesquisa, tem-se que:

1. **o objetivo específico A** (“Coletar narrativas e organizar um conjunto de dados rotulado voltado à detecção de mentiras para a língua portuguesa do Brasil”) foi atingido pela construção do **MMDDD-PtBr**, que é um Produto Técnico-Tecnológico de acesso público e gratuito, assim como a metodologia para a sua concepção;
2. **o objetivo específico B** (“Elaborar, avaliar e identificar os modelos autossupervisionados de melhor desempenho para cada modalidade”) foi atingido pela realização de **2.390 experimentos** e decorrente identificação

dos modelos de mais elevado desempenho frente aos dados e medidos pelo **Escore de Sinceridade parcial**;

3. **o objetivo específico C** (“Elaborar um modelo de fusão das modalidades individuais em um modelo integrado multimodal”) foi atingido pela proposta do **Escore de Sinceridade**, que permite combinar quaisquer modalidades de forma harmônica, desde que avaliadas por um Escore de Sinceridade parcial.

Diante do exposto, o **objetivo geral** da pesquisa (“Elaborar modelos de Aprendizado de Máquina autossupervisionados para a detecção individual e multimodal de mentiras para a língua portuguesa”) foi considerado atingido, visto que modelos monomodais e multimodais foram elaborados e testados com diversos graus de desempenho que superam a taxa de 54% de precisão, frequentemente utilizada em pesquisas no campo como linha de base para o ser humano não treinado.

Tendo o objetivo geral atingido, esta pesquisa alcançou a seguinte resposta à **questão de pesquisa** (“Qual modelo de Aprendizado de Máquina autossupervisionado é capaz de utilizar pistas multimodais de um indivíduo específico para distinguir uma narrativa sincera de uma não sincera expressa em língua portuguesa?”): o modelo de Autoencoder atencional *multi-head*, em diferentes configurações a depender da modalidade de fonte de informação utilizada.

Com a resposta à questão de pesquisa, é possível concluir que a **hipótese de pesquisa** (“O Aprendizado de Máquina autossupervisionado é capaz de viabilizar um modelo apto a distinguir uma narrativa sincera de uma não sincera proferida por um indivíduo específico”) pode ser considerada confirmada, pois os níveis de desempenho atingidos ultrapassam a probabilidade do acaso e da linha de base de 54% do ser humano não treinado.

5.1 Conclusões

Algumas conclusões puderam ser atingidas a partir dos resultados alcançados pelos experimentos realizados:

1. o mecanismo de Atenção contribui para aumento do aprendizado dos dados de normalidade (sinceridade), conferindo maior acurácia aos modelos;
2. o fato de o mecanismo de Atenção, especialmente *multi-head*, ter efeito positivo sobre o aprendizado sugere que existem relações de longa

distância entre partes das narrativas sinceras que não evoluem da mesma forma em uma narrativa não sincera; isso significa dizer que o modelo de Aprendizado de Máquina confirmou o pressuposto teórico (Hipótese de Undeutsch) da alteração de expressão de um indivíduo quando este está mentindo;

3. a Divergência de Kullback-Leibler ofereceu, em algumas circunstâncias (especialmente nos modelos visuais), melhores resultados na avaliação do erro de reconstrução, quando comparado com o Erro Médio Quadrático, mas ainda não foi possível estabelecer quando uma métrica oferece mais vantagens que outra e se é possível estabelecer critérios baseados nos dados para esta decisão;
4. a modelagem do problema de detecção de mentiras como uma detecção de anomalias mostrou resultados positivos, mas também alta sensibilidade às narrativas, possivelmente porque algumas variações detectadas não tinham gênese na não sinceridade, mas em outros processos psicológicos e emocionais operantes (por exemplo, o nervosismo de estar participando de um programa de TV ao vivo); tal justificativa não é mais que uma hipótese, que suscita novas pesquisas;
5. embora a acurácia balanceada final do Modelo de Sinceridade Multimodal tenha atingido o patamar de 0,714, superando a linha de base de 54% apontada pela literatura como a expectativa de desempenho de um humano não treinado, os experimentos mostraram que ainda existem fatores de confusão suficientes para produzir erros de julgamento que vieram a prejudicar o resultado final; há necessidade de compreender quais das etapas do processo contribuíram para esses desvios de percepção dos modelos e consecutiva correção;
6. a análise detalhada da resposta de cada componente do modelo multimodal mostrou diferenças de percepção da sinceridade e não sinceridade em cada modalidade; aparentemente, os fatores de confusão em cada modalidade são distintos e precisam ser estudados e mitigados individualmente, dado que os modelos utilizados em cada modalidade apresentam variadas arquiteturas; especificamente os modelos verbais demonstraram mais baixa acurácia de detecção, com erros de julgamento

apresentando alto nível de confiança, o que estimula o estudo dos fatores de confusão particularmente existentes neste componente;

7. o Escore de Sinceridade introduziu a noção de confiança para o problema da detecção de anomalias por meio de Autoencoders; foi observado que na maioria das vezes o uso desta métrica contínua (ao invés da estratégia discreta da votação) promoveu seis casos de acerto, que de outra forma seriam incorretamente classificados; ao mesmo tempo, foram observados apenas dois casos de erro motivados pela aplicação do Escore de Sinceridade, o que sugere que esta pode ser uma estratégia a contribuir para maiores níveis de acurácia em quaisquer problemas de detecção de anomalias;
8. dado que as narrativas utilizadas nos experimentos foram extraídas de um programa de TV, no qual perguntas foram enunciadas por pessoas interessadas em descobrir a verdade, mas sem treinamento específico para tal; em alguns casos, as narrativas de sinceridade coletadas foram muito curtas o que ofereceu pouco material para que os Modelos de Sinceridade pudessem identificar padrões informativos suficientes, fator que pode ter interferido no desempenho de detecção;
9. a estratégia baseada em Aprendizado Autossupervisionado dispensa a necessidade de dados rotulados, mas requer um conjunto de narrativas seguramente sinceras para a construção do Modelo de Sinceridade; ainda é necessário mais estudo para estabelecer um perfil de qualidade dessas narrativas para que o modelo treinado possa ser utilizado confiavelmente;
10. a abordagem individual, ou seja, a construção de um **Modelo de Sinceridade** que seja específico para um dado indivíduo, aproxima o processo computadorizado do seu equivalente humano, ao menos sob a ótica de Ekman, que defende a importância de valorizar os aspectos circunstanciais e idiossincráticos de um testemunho; no entanto, os relatos dos magos da detecção de mentiras dão conta de uma experiência adquirida ao longo de várias observações; assim, aparentemente há dois conjuntos de critérios operantes em um caso de alta taxa de acurácia na detecção, os parâmetros individuais próprios do caso em avaliação e também parâmetros de coletividade que formaram a capacidade do detector; os critérios coletivos não foram bem aprendidos pelos

experimentos realizados nesta pesquisa, evidenciando a necessidade de mais aprofundamento;

11. os modelos coletivos apresentaram desempenho significativamente pior que os modelos individuais (mais de dez pontos percentuais); a suposição que se faz é que o conjunto de dados em sua configuração atual oferece pouca condição para a formação de padrões de grupo, dada a grande variedade de regionalidades associadas aos sujeitos e suas narrativas; uma forma de procurar validar esta hipótese e superar tal situação é pela adição de mais sujeitos e narrativas; igualmente, modelos de classificação a exemplo do realizado em outros estudos também pode auxiliar na compreensão dos resultados; uma justificativa alternativa é que os padrões de sinceridade e não sinceridade sejam muito similares quando avaliados em grupo, o que reforça o pressuposto de que os caracteres particulares de um indivíduo podem ser tão discrepantes de outro a ponto de requererem análises pessoa-a-pessoa; finalmente, é possível que as duas hipóteses ocorram simultaneamente;
12. a precisão oferecida pelo OpenFace pode ter sido um fator de redução de acurácia na detecção visual, pois houve casos em que impediu a extração de características faciais em diversos segmentos, efetivamente inviabilizando a utilização de dois dos vídeos que compõem o MMDDD-Pt-Br.

Estas conclusões, longe de encerrar a discussão a respeito do tema, propiciam a concepção de novas linhas de pesquisa para aprofundamento em cada um dos temas operados.

5.2 Continuidade da pesquisa

Dadas as escolhas realizadas, consolidadas pelos encaminhamentos metodológicos estabelecidos, foi possível identificar certas limitações na pesquisa que talvez possam ser superadas com os seguintes estudos complementares:

1. **diversificação do conjunto de dados**: o conjunto de dados construído foi elaborado com base em apenas uma situação (programa de TV ao vivo), o que necessariamente insere um viés nos dados, podendo mascarar tanto deficiências quanto virtudes da abordagem explorada; a introdução de novas narrativas obtidas em outras circunstâncias (por

exemplo, interrogatórios judiciais, entrevistas em sessões de terapias, atuações de atores profissionais, dentre outras) poderá reduzir esse viés e sugerir aprimoramentos a partir da observação das respostas oferecidas pelos modelos;

2. **pistas visuais corporais**: nesta pesquisa, apenas pistas visuais faciais foram exploradas, mas as variações de expressão podem ocorrer em todo o corpo (movimentação do tronco e braços, além de gestos com as mãos); nesse sentido, a inclusão de pistas visuais corporais pode oferecer mais insumos para o enriquecimento do modelo de sinceridade visual e potencial elevação da acurácia de detecção;
3. **estudo de ablação para todas as modalidades**: embora existam na literatura diversos estudos de ablação para as modalidades acústica e visual, nenhum deles foi realizado para o MMDDD-PtBr, o que significa dizer que nenhum deles capturou eventuais particularidades das mesmas modalidades para a língua portuguesa do Brasil; um estudo de ablação para as modalidades acústica e visual poderia reforçar a noção universalidade das descobertas já relatadas ou desafiá-las, introduzindo a noção de que tais dimensões são regionais;
4. **estudo de falsas não sinceridades**: em todos os experimentos foi percebi a ocorrência de falsas não sinceridades, que carregam um efeito ético importante e podem ser consideradas como o mais grave erro cometido por uma solução de detecção de mentiras; a gênese desse tipo de resultado, os fatores que os influenciam e formas de evitar o problema podem estabelecer novos critérios tanto para a concepção dos Modelos de Sinceridade quanto do processo de coleta de narrativas;
5. **exploração da emoção em todas as modalidades**: apenas a modalidade verbal fez uso de informação emocional (duas características que descrevem sentimento), mas o fator emocional é referenciado por alguns autores da área de detecção de mentiras (notadamente Ekman) como uma pista fundamental; neste sentido, os dados acústicos e visuais poderiam ser complementados com características que de alguma forma estimassem as emoções presentes na narrativa, para enriquecer ainda mais o Modelo de Sinceridade;

6. **substituição do SentiWordNet-PT-Br**: o estudo de ablação demonstrou que as características de sentimento participaram para elevar o nível de acurácia balanceada dos modelos verbais, mas o SentiWordNet-PT-Br é um dicionário de palavras que oferece as polaridades positiva e negativa do sentimento sem levar em conta o contexto do uso de cada vocábulo (em alguns casos, a mesma palavra tem as duas polaridades), o que possivelmente reduz a percepção da verdadeira emoção participante da narrativa; ao invés de utilizar um dicionário estático, as emoções presentes em cada palavra poderiam ser obtidas de um modelo de linguagem que considerasse o contexto, para maior precisão (por exemplo, o BERTimbau[22]);
7. **aprimoramento dos modelos coletivos**: os modelos coletivos experimentados e avaliados tiveram suas arquiteturas baseadas diretamente nos modelos individuais, mas dada a complexidade das relações potencialmente existentes, assim como a variedade de dados, é possível que aqueles modelos não sejam os mais apropriados; experimentos e estudos específicos para os modelos coletivos poderiam levar a ganhos em acurácia para assim promover um modelo multicomponentes mais preciso;
8. **sincronização de dados de entrada**: os modelos multimodais propostos não buscaram vincular partículas de uma modalidade em outra; por exemplo, quais expressões faciais e características acústicas mais frequentemente acompanham certas palavras; este tipo de análise poderia ser utilizado como um fator a mais para detectar discrepâncias na expressão e, assim, identificar variações de conduta que podem ser sinais de não sinceridade;
9. **combinação de MSE e KLd/KLn**: as métricas MSE e KLd/KLn operam de forma distinta para avaliar o erro de reconstrução, capturando aspectos diferentes do erro; um estudo que acomodasse os melhores aspectos das duas métricas poderia resultar em uma outra, mais granular e perceptiva, conferindo maior precisão a modelos para detecção de anomalias em geral e também para o problema de detecção de mentiras;

[22] https://github.com/neuralmind-ai/portuguese-bert

10. **novas formas de calcular o Escore de Sinceridade**: o Escore de Sinceridade foi calculado como a distância relativa que o erro de reconstrução de uma narrativa apresenta em relação a um erro de reconstrução esperado (referencial) para sinceridade; o limiar que modula o erro de reconstrução (zona de incerteza) é o desvio padrão médio dos erros de reconstrução de todas as características; formas alternativas de calcular o Escore de Sinceridade têm o potencial de aprimorar a sensibilidade da métrica de forma a elevar a precisão da detecção.

Adicionalmente, considerando especificamente as descobertas alcançadas pelos resultados dos experimentos, foi possível divisar novos temas para continuação desta pesquisa:

1. **modelos generativos**: os modelos generativos são modelos de aprendizado profundo capazes de produzir novos dados a partir dos dados originais de treinamento; neste sentido, estes tipos de modelo poderiam ser utilizados para tentar prever qual seria a consecução de uma narrativa a partir de seu início; uma avaliação das diferenças entre o que foi gerado e o que foi efetivamente dito pode evidenciar uma expressão de não sinceridade;
2. **análise palavra-a-palavra**: após a eventual detecção de uma não sinceridade, uma das informações mais importantes é localizar o ponto da narrativa onde se encontra a mentira; estudos que conjuguem detecções multimodais poderiam ser utilizados para localizar os pontos mais prováveis da presença da não sinceridade ao longo da frase, identificando assim as palavras que a exprimiram;
3. **modelos de linguagem para dados verbais**: dado o grande avanço que os modelos de linguagem têm apresentado na extração de caracteres linguísticos (um exemplo é o ChatGPT[23]), é possível que a introdução daqueles modelos aprimore a extração de informações relevantes na narrativa, produzindo novas características mais distintivas no componente verbal do MMDDD-PtBr;
4. **identificação de pontos-chave da narrativa**: o mecanismo de Atenção foi aplicado nos modelos experimentados unicamente visando a elevação

[23] https://chat.openai.com/

da acurácia de detecção; embora os resultados tenham demonstrado que tal efetivamente se deu, a Atenção oferece também a oportunidade não explorada de identificar quais partes da narrativa mais contribuíram para a detecção, o que pode participar para alguma forma de explicabilidade da inferência;

5. **estratégias de explicabilidade**: de forma a aprimorar a compreensão da detecção, aumentar seu grau de confiabilidade e até mesmo auxiliar no ajuste dos modelos, estratégias de explicabilidade das inferências e conclusões dos modelos podem ser aplicadas;
6. **critério de qualidade dos dados**: a proposição da detecção por aprendizado autossupervisionado carrega em seu bojo a ideia de independer de dados rotulados, pois estes são raros; no entanto, os experimentos realizados nesta pesquisa necessitaram de dados rotulados para efeitos de comparação; os experimentos mostraram que ainda existe uma carência de critérios que apontem quando esta estratégia pode ser aplicada ou não; assim, estudos que identifiquem o grau de qualidade e confiabilidade dos dados de entrada poderão preencher esta lacuna;
7. **comparação com aprendizado supervisionado**: como o MMDDD-PtBr é um conjunto de dados rotulado, é possível fazer um estudo com aprendizado supervisionado para comparação;
8. **operação paralela com *frameworks* existentes**: a psicologia já produziu diversas ferramentas para a detecção de mentiras e avaliação de credibilidade, como o *Reality Monitoring* (RM), o *Criteria Based Content Analisys* (CBCA) o *Scientific Content Analysis* (SCAN), o *Behavior Analysis Interview* (BAI), a *Ekman's Deception Theory* (EDT) e o *Statement Validity Assessment* (SVA); esses *frameworks* poderiam ser aplicados em situações de mentiras sérias ou em ambiente de laboratório em paralelo com um Modelo de Sinceridade para avaliar o efeito sinérgico que a assistência tecnológica pode oferecer e assim propiciar ajustes finos em ambas as estratégias de detecção;
9. **modelo de entrevista e coleta de dados**: uma vez que o Modelo de Sinceridade é sensível à qualidade dos dados de treinamento (expressões de sinceridade), percebe-se que um roteiro para entrevista pode ser benéfico para identificar quais narrativas oferecem maiores oportunidades

de captura de padrões de sinceridade; o mesmo pode ser dito a respeito dos aspectos técnicos de coleta do áudio e do vídeo das entrevistas; um estudo de quais os melhores critérios para a realização dessas entrevistas possivelmente elevaria a qualidade dos modelos treinados e suas acurácias.

As linhas de pesquisa divisadas, se encaminhadas, poderiam incrementar grandemente o estado do conhecimento a respeito de detecção de mentiras e Aprendizado de Máquina, tanto isoladamente quanto em cooperação.

5.3 Contribuições

A jornada percorrida ao longo desta pesquisa trouxe resultados entendidos como contribuições para o campo, em particular, por ser esta a primeira pesquisa dentro do tema de detecção de mentiras por Aprendizado de Máquina realizada por um pesquisador brasileiro, abordando especificamente a língua portuguesa do Brasil, assim como a abordagem baseada em aprendizado autossupervisionado que, em evoluindo, dispensaria a necessidade de conjuntos de dados rotulados.

Dada ser esta a primeira pesquisa de origem brasileira a endereçar os temas Aprendizado de Máquina e detecção de mentiras, acredita-se que esteja contribuindo para estimular novos estudos que abordem as especificidades da língua e da cultura do Brasil. No caso particular da psicologia, o modelo de trabalho aqui apresentado pode cooperar com estudos para auxiliar na avaliação de pacientes, especialmente quando há suspeitas de distúrbios de personalidade, transtornos de mentira patológica ou outros problemas psicológicos que envolvem comportamento enganoso. Terapeutas podem utilizar o Modelo de Sinceridade para ajudar os pacientes no desenvolvimento de habilidades de comunicação mais honestas e eficazes, sendo particularmente útil em terapias de casal ou familiar, onde as questões de confiança são centrais.

Psicólogos que estudam o comportamento enganoso podem usar os resultados desta pesquisa para entender melhor os padrões de não sinceridade em diferentes contextos, o que pode levar a uma compreensão mais profunda dos motivos por trás do comportamento enganoso e das estratégias empregadas para ocultar a verdade. Tal esforço, ao mesmo tempo, retroalimentaria as conclusões até então alcançadas, incrementando o conhecimento nos dois campos.

Outra contribuição desta pesquisa é o destaque para as condições de coleta de testemunhos, em quaisquer circunstâncias e domínios. Foi demonstrado que a qualidade de áudio e vídeo têm efeito significativo na eficácia das características reconhecidas. No entanto, por falta de critérios técnicos específicos, muitas vezes se percebe o uso de equipamentos para captura de vídeo e áudio de qualidade inferior[24], ou em condições desfavoráveis de configuração[25], que previnem quase que completamente o uso de métodos automatizados.

Esta pesquisa pode ser utilizada para justificar a concepção de um conjunto de diretrizes, até mesmo um modelo de maturidade, para a aquisição de narrativas de vida real em condições úteis para novas pesquisas e mesmo a operacionalização de sistemas automatizados de detecção de mentiras em diversos setores.

No campo da Ciência da Informação, a existência de um modelo para detecção de mentiras para o português pode auxiliar na análise de declarações públicas de figuras políticas, identificando possíveis discursos enganosos, o que contribui para um melhor entendimento de como o viés e a manipulação podem afetar a disseminação da informação.

Outra aplicação especialmente crítica é a elaboração de sistemas de alerta automatizado para uso em mídias sociais, visando identificar tentativas de aliciamento e abuso de crianças e adolescentes. Narrativas pré-categorizadas de aliciadores e abusadores poderiam ser utilizadas para a construção de um “Modelo de Risco”, que seria utilizado para avaliar as chances de uma conversação incluir os caracteres próprios daquele tipo expressão, nos mesmos moldes do Modelo de Sinceridade.

Na pesquisa científica, modelos de detecção de mentiras podem ajudar a verificar a autenticidade das respostas em enquetes e coletas de dados, elevando a validade e confiabilidade dos resultados obtidos. Pesquisas dentro deste escopo poderiam ser encaminhadas a partir do Modelo de Sinceridade.

Os resultados até então alcançados apenas oferecem os primeiros passos dentro da abordagem apresentada. Novos estudos que expandam os horizontes do conhecimento podem e devem ser levados a efeito. Tais estudos poderão contribuir para todos os domínios de pesquisa envolvidos, ou seja, Inteligência Artificial, Gestão e Ciência da Informação, assim como da Psicologia.

---

[24] https://www.youtube.com/watch?v=4gXZJjcLkDs
[25] https://www.youtube.com/watch?v=G7bU_lw9qeE

REFERÊNCIAS

AAMONDT, G. MIchael; CUSTER, Heather. Who can best catch a liar? A meta-analysis of individual differences in detecting deception. **The Forensic Examiner**, *[S. l.]*, v. 15, n. 611, 2006.

AFGANI, Mostafa; SINANOVIĆ, Sinan; HAAS, Harald. Anomaly detection using the Kullback-Leibler divergence metric. **2008 1st International Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL 2008**, *[S. l.]*, 2008. DOI: 10.1109/ISABEL.2008.4712573.

ALOM, Md Zahangir; TAHA, Tarek M.; YAKOPCIC, Christopher; WESTBERG, Stefan; SIDIKE, Paheding; NASRIN, Mst Shamima; VAN ESESN, Brian C.; AWWAL, Abdul A. S.; ASARI, Vijayan K. The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. *[S. l.]*, 2018.

AMADO, Bárbara G.; ARCE, Ramón; FARIÑA, Francisca. Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. **European Journal of Psychology Applied to Legal Context**, *[S. l.]*, v. 7, n. 1, p. 3–12, 2015. DOI: 10.1016/j.ejpal.2014.11.002.

AVOLA, Danilo; FORESTI, Gian Luca; CINQUE, Luigi; PANNONE, Daniele. Automatic deception detection in RGB videos using facial action units. **ACM International Conference Proceeding Series**, *[S. l.]*, 2019. DOI: 10.1145/3349801.3349806.

BAILEY, James; DEMYANOV, Sergey; RAMAMOHANARAO, Kotagiri; LECKIE, Christopher. Detection of deception in the Mafia party game. **ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction**, *[S. l.]*, p. 335–342, 2015. DOI: 10.1145/2818346.2820745.

BALL, Terry J. **The Polygraph Museum**. [s.d.]. Disponível em: http://www.lie2me.net/thepolygraphmuseum/id16.html. Acesso em: 17 mar. 2022.

BALTRUSAITIS, Tadas; ZADEH, Amir; LIM, Yao Chong; MORENCY, Louis Philippe. OpenFace 2.0: Facial behavior analysis toolkit. **Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018**, *[S. l.]*, p. 59–66, 2018. DOI: 10.1109/FG.2018.00019.

BANK, Dor; KOENIGSTEIN, Noam; GIRYES, Raja. Autoencoders. **arXiv preprint arXiv:2003.05991**, *[S. l.]*, 2021. DOI: 10.1016/B978-0-12-815739-8.00011-0.

BARSEVER, Dan; SINGH, Sameer; NEFTCI, Emre. Building a Better Lie Detector with BERT: The Difference between Truth and Lies. **Proceedings of the International Joint Conference on Neural Networks**, *[S. l.]*, 2020. DOI: 10.1109/IJCNN48605.2020.9206937.

BELL, Jason. **Machine Learning**. Indianapolis: Wiley, 2015.

BELOV, Dmitry I.; ARMSTRONG, Ronald D. Distributions of the Kullback-Leibler divergence with applications. **British Journal of Mathematical and Statistical Psychology**, *[S. l.]*, v. 64, n. 2, p. 291–309, 2011. DOI: 10.1348/000711010X522227.

BINI, Stefano A. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health

Care? **Journal of Arthroplasty**, *[S. l.]*, v. 33, n. 8, p. 2358–2361, 2018. DOI: 10.1016/j.arth.2018.02.067.

BORGHESI, Andrea; MOLAN, Martin; MILANO, Michela; BARTOLINI, Andrea. Anomaly Detection and Anticipation in High Performance Computing Systems. **IEEE Transactions on Parallel and Distributed Systems**, *[S. l.]*, v. 33, n. 4, p. 739–750, 2022. DOI: 10.1109/TPDS.2021.3082802.

BRISCOE, Erica J.; APPLING, D. Scott; HAYES, Heather. Cues to deception in social media communications. **Proceedings of the Annual Hawaii International Conference on System Sciences**, *[S. l.]*, p. 1435–1443, 2014. DOI: 10.1109/HICSS.2014.186.

BRODERSEN, Kay H.; ONG, Cheng Soon; STEPHAN, Klaas E.; BUHMANN, Joachim M. The balanced accuracy and its posterior distribution. **Proceedings - International Conference on Pattern Recognition**, *[S. l.]*, p. 3121–3124, 2010. DOI: 10.1109/ICPR.2010.764.

BURGOON, Judee K.; GUERRERO, Laura K.; FLOYD, Kory. **Nonverbal communication**. 2nd. ed. New York, NY: Routledgr, 2016. DOI: 10.4324/9781315663425.

CHALAPATHY, Raghavendra; CHAWLA, Sanjay. Deep Learning for Anomaly Detection: A Survey. *[S. l.]*, p. 1–50, 2019. Disponível em: http://arxiv.org/abs/1901.03407.

CHEN, Zhaomin; YEO, Chai Kiat; LEE, Bu Sung; LAU, Chiew Tong. Autoencoder-based network anomaly detection. **Wireless Telecommunications Symposium**, *[S. l.]*, v. 2018- April, p. 1–5, 2018. DOI: 10.1109/WTS.2018.8363930.

CONSTANCIO, Alex Sebastião; TSUNODA, Denise Fukumi; DE FÁTIMA NUNES SILVA, Helena; DA SILVEIRA, Jocelaine Martins; CARVALHO, Deborah Ribeiro. Deception detection with machine learning: A systematic review and statistical analysis. **PLoS ONE**, *[S. l.]*, v. 18, n. 2 February, p. 1–31, 2023. DOI: 10.1371/journal.pone.0281323.

CRESWELL, John W. **Research Design Qualitative, Quantitativa, and Mixed Methods Approaches**. 4th. ed. Thousand Oaks, California: SAGE Publications, Ltd, 2014.

DENAULT, Vincent; TALWAR, Victoria; PLUSQUELLEC, Pierrich; LARIVIÈRE, Vincent. On deception and lying: An overview of over 100 years of social science research. **Applied Cognitive Psychology**, *[S. l.]*, v. 36, n. 4, p. 805–819, 2022. DOI: 10.1002/acp.3971.

DEPAULO, Bella M.; MALONE, Brian E.; LINDSAY, James J.; MUHLENBRUCK, Laura; CHARLTON, Kelly; COOPER, Harris. Cues to deception. **Psychological Bulletin**, *[S. l.]*, v. 129, n. 1, p. 74–118, 2003. DOI: 10.1037/0033-2909.129.1.74.

DING, Mingyu; ZHAO, An; LU, Zhiwu; XIANG, Tao; WEN, Ji Rong. Face-focused cross-stream network for deception detection in videos. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, *[S. l.]*, v. 2019-June, n. 2, p. 7794–7803, 2019. DOI: 10.1109/CVPR.2019.00799.

EKMAN, Paul. **Telling Lies**. New York, NY: W. W. Norton & Company, Inc, 1992. DOI: 10.11647/obp.0125.13.

EKMAN, Paul; FRIESEN, Wallace V.; HAGER, Joseph C. **Facial Action Coding System - The manual**. CD-ROM Man ed. Salt Lake City: Research Nexus division of Network Information Research Corporation, 2002.

FENG, Song; BANERJEE, Ritwik; CHOI, Yejin. Syntactic stylometry for deception detection. **50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference**, *[S. l.]*, v. 2, n. July, p. 171–175, 2012.

FERNANDES, Sinead V.; ULLAH, Muhammad S. Use of Machine Learning for Deception Detection from Spectral and Cepstral Features of Speech Signals. **IEEE Access**, *[S. l.]*, v. 9, p. 78925–78935, 2021. a. DOI: 10.1109/ACCESS.2021.3084200.

FERNANDES, Sinead V.; ULLAH, Muhammad S. Development of Spectral Speech Features for Deception Detection Using Neural Networks. **2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2021**, *[S. l.]*, p. 198–203, 2021. b. DOI: 10.1109/IEMCON53756.2021.9623077.

FLACH, Peter A.; KULL, Meelis. Precision-Recall-Gain curves: PR analysis done right. **Advances in Neural Information Processing Systems**, *[S. l.]*, v. 2015- Janua, p. 838–846, 2015.

FORNACIARI, Tommaso; POESIO, Massimo. On the use of homogenous sets of subjects in deceptive language analysis. **Computational Linguistics, Proceedings of the Workshop on Computational Approaches to Deception Detection**, *[S. l.]*, p. 39–47, 2012.

FU, Hongliang; LEI, Peizhi; TAO, Huawei; ZHAO, Li; YANG, Jing. Improved semi-supervised autoencoder for deception detection. **PLoS ONE**, *[S. l.]*, v. 14, n. 10, p. 1–13, 2019. DOI: 10.1371/journal.pone.0223361. Disponível em: http://dx.doi.org/10.1371/journal.pone.0223361.

GOGATE, Mandar; ADEEL, Ahsan; HUSSAIN, Amir. Deep learning driven multimodal fusion for automated deception detection. **2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings**, *[S. l.]*, v. 2018- Janua, p. 1–6, 2018. DOI: 10.1109/SSCI.2017.8285382.

GONG, Dong; LIU, Lingqiao; LE, Vuong; SAHA, Budhaditya; MANSOUR, Moussa Reda; VENKATESH, Svetha; VAN DEN HENGEL, Anton. Memorizing normality to detect anomaly. **Proceedings of the IEEE International Conference on Computer Vision**, *[S. l.]*, v. 2019- Octob, p. 1705–1714, 2019.

GOODFELLOW, Ian; YOSHUA, Bengio; COURVILLE, Aaron. **Deep learning**. Cambridge, MA, USA: MIT Press, 2016.

IACONO, William G.; BEN-SHAKHAR, Gershon. Current status of forensic lie detection with the comparison question technique: An update of the 2003 National Academy of Sciences report on polygraph testing. **Law and Human Behavior**, *[S. l.]*, v. 43, n. 1, p. 86–98, 2019. DOI: 10.1037/lhb0000307.

ISLAM, Siam; SAHA, Popin; CHOWDHURY, Touhidul; SOROWAR, Asif; RAB,

Raqeebir. Non-invasive Deception Detection in Videos Using Machine Learning Techniques. **2021 5th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2021**, *[S. l.]*, 2021. DOI: 10.1109/ICEEICT53905.2021.9667928.

JAISWAL, Mimansa; TABIBU, Sairam; BAJPAI, Rajiv. The Truth and Nothing but the Truth: Multimodal Analysis for Deception Detection. **IEEE International Conference on Data Mining Workshops, ICDMW**, *[S. l.]*, v. 0, p. 938–943, 2016. DOI: 10.1109/ICDMW.2016.0137.

JAKHAR, D.; KAUR, I. Artificial intelligence, machine learning and deep learning: definitions and differences. **Clinical and Experimental Dermatology**, *[S. l.]*, v. 45, n. 1, p. 131–132, 2020. DOI: 10.1111/ced.14029.

KAMBOJ, Manvi; HESSLER, Christian; ASNANI, Priyanka; RIANI, Kais; ABOUELENIEN, Mohamed. Multimodal Political Deception Detection. **IEEE Multimedia**, *[S. l.]*, v. 28, n. 1, p. 94–102, 2021. DOI: 10.1109/MMUL.2020.3048044.

KLEINBERG, Bennett; VAN DER TOOLEN, Yaloe; VRIJ, Aldert; ARNTZ, Arnoud; VERSCHUERE, Bruno. Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. **Applied Cognitive Psychology**, *[S. l.]*, v. 32, n. 3, p. 354–366, 2018. DOI: 10.1002/acp.3407.

KLEINBERG, Bennett; VERSCHUERE, Bruno. How humans impair automated deception detection performance. **Acta Psychologica**, *[S. l.]*, v. 213, n. March 2020, p. 103250, 2021. DOI: 10.1016/j.actpsy.2020.103250. Disponível em: https://doi.org/10.1016/j.actpsy.2020.103250.

KOPEV, Daniel; ALI, Ahmed; KOYCHEV, Ivan; NAKOV, Preslav. Detecting Deception in Political Debates Using Acoustic and Textual Features. **2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings**, *[S. l.]*, p. 652–659, 2019. DOI: 10.1109/ASRU46091.2019.9003892.

LABIBAH, Zuhrah; NASRUN, Muhammad; SETIANINGSIH, Casi. Lie Detector With The Analysis Of The Change Of Diameter Pupil and The. *[S. l.]*, p. 214–220, 2018.

LEVINE, Timothy R. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. **Journal of Language and Social Psychology**, *[S. l.]*, v. 33, n. 4, p. 378–392, 2014. DOI: 10.1177/0261927X14535916.

LU, Yuchen; XU, Peng. Anomaly Detection for Skin Disease Images Using Variational Autoencoder. *[S. l.]*, 2018. Disponível em: http://arxiv.org/abs/1807.01349.

MAHESH, Batta. Machine Learning Algorithms - A Review. **International Journal of Computer Science and Information Technologies**, *[S. l.]*, v. 9, n. 1, p. 381–386, 2018. DOI: 10.21275/ART20203995.

MATHUR, Leena; MATARIC, Maja J. Affect-Aware Deep Belief Network Representations for Multimodal Unsupervised Deception Detection. **Proceedings - 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021**, *[S. l.]*, 2021. DOI: 10.1109/FG52635.2021.9667050.

MATHUR, Leena; MATARIĆ, Maja J. Introducing Representations of Facial Affect in Automated Multimodal Deception Detection. **ICMI 2020 - Proceedings of the 2020**

International Conference on Multimodal Interaction, *[S. l.]*, p. 305–314, 2020. DOI: 10.1145/3382507.3418864.

MATHUR, Leena; MATARIĆ, Maja J. Unsupervised Audio-Visual Subspace Alignment for High-Stakes Deception Detection. **Proceedings - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021**, *[S. l.]*, p. 2255–2259, 2021. DOI: 10.1109/ICASSP39728.2021.9413550.

MBAZIIRA, Alex V.; MURPHY, Diane R. An empirical study on detecting deception and cybercrime using artificial neural networks. **ACM International Conference Proceeding Series**, *[S. l.]*, p. 42–46, 2018. DOI: 10.1145/3193077.3193080.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, *[S. l.]*, v. 5, n. 4, p. 115–133, 1943. DOI: 10.1007/BF02478259.

MIHALCEA, Rada; PÉREZ-ROSAS, Verónica; BURZO, Mihai. Automatic detection of deceit in verbal communication. **ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction**, *[S. l.]*, p. 131–134, 2013. DOI: 10.1145/2522848.2522888.

MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. **Foundations of Machine learning**. London: The MIT Press, 2012.

NASRI, Hanen; OUARDA, Wael; ALIMI, Adel M. ReLiDSS: Novel lie detection system from speech signal. **Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA**, *[S. l.]*, v. 0, 2016. DOI: 10.1109/AICCSA.2016.7945789.

O'SULLIVAN, Maureen; EKMAN, Paul. **The wizards of deception detection**. **The Detection of Deception in Forensic Contexts**, 2004. DOI: 10.1017/CBO9780511490071.012.

PAK, Jinie; ZHOU, Lina. A comparison of features for automatic deception detection in synchronous computer-mediated communication. **2015 IEEE International Conference on Intelligence and Security Informatics: Securing the World through an Alignment of Technology, Intelligence, Humans and Organizations, ISI 2015**, *[S. l.]*, p. 141–143, 2015. DOI: 10.1109/ISI.2015.7165955.

PAPANTONIOU, Katerina; PAPADAKOS, Panagiotis; PATKOS, Theodore; FLOURIS, George; ANDROUTSOPOULOS, Ion; PLEXOUSAKIS, Dimitris. Deception detection in text and its relation to the cultural dimension of individualism/collectivism. **Natural Language Engineering**, *[S. l.]*, p. 1–62, 2021. DOI: 10.1017/S1351324921000152.

PATTERSON, Josh; GIBSON, Adam. **Deep Learning: A practitioner's approach**. 1st Ed ed. Sebastopol: O'Reilly Media, 2017.

PÉREZ-ROSAS, Verónica; ABOUELENIEN, Mohamed; MIHALCEA, Rada; BURZO, Mihai. Deception detection using real-life trial data. **ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction**, *[S. l.]*, p. 59–66, 2015. DOI: 10.1145/2818346.2820758.

PORTER, Stephen (University of British Columbia); BRINKE, Leanne ten (University of British Columnbia). The truth about lies: What works in detecting high-stakes

deception.pdf. **Legal and Criminological Psychology**, *[S. l.]*, n. 15, p. 57–75, 2010. a. DOI: 10.1348/135532509X433151.

PORTER, Stephen (University of British Columbia); BRINKE, Leanne ten (University of British Columnbia). The truth about lies - What works in detecting high-stakes deception. **Legal and Criminological Psychology**, *[S. l.]*, v. 15, n. 1, p. 57–75, 2010. b. DOI: 10.1348/135532509X433151. Disponível em: www.bpsjournals.co.uk.

PRATELLA, David; SAADI, Samira Ait El Mkadem; BANNWARTH, Sylvie; PAQUIS-FLUCKINGER, Véronique; BOTTINI, Silvia. A survey of autoencoder algorithms to pave the diagnosis of rare diseases. **International Journal of Molecular Sciences**, *[S. l.]*, v. 22, n. 19, 2021. DOI: 10.3390/ijms221910891.

RADFORD, Alec; NARASIMHAN, Karthik; SALIMANS, Tim; SUTSKEVER, Ilya. Improving Language Understanding by Generative Pre-Training. **OpenAI.com**, *[S. l.]*, p. 1–12, 2018. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, US, v. 65, p. 386–408, 1958. DOI: 10.1037/h0042519.

RUBIN, Victoria L.; CONROY, Niall. Discerning truth from deception: Human judgments and automation efforts. **First Monday**, *[S. l.]*, v. 17, n. 3, 2012. DOI: 10.5210/fm.v17i3.3933.

RUBIN, Victoria L.; CONROY, Niall J. Challenges in automated deception detection in computer-mediated communication. **Proceedings of the ASIST Annual Meeting**, *[S. l.]*, v. 48, 2011. DOI: 10.1002/meet.2011.14504801098.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning Internal Representations by Error Propagation. **Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence**, *[S. l.]*, v. 1, n. ICS Report 8506, 1985. DOI: 10.1016/B978-1-4832-1446-7.50035-2.

RYBAR, Metod; BIELIKOVA, Maria. Automated detection of user deception in on-line questionnaires with focus on eye tracking use. **Proceedings - 11th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2016**, *[S. l.]*, n. i, p. 24–28, 2016. DOI: 10.1109/SMAP.2016.7753379.

SALLES, Bruno. Do Lie Detection Tools Really Catch Liars? A Guide for Forensic Professionals. **Brazilian Journal of Forensic Sciences, Medical Law and Bioethics**, *[S. l.]*, v. 9, n. 3, p. 373–393, 2020. DOI: 10.17063/bjfs9(3)y2020373-393.

SANAULLAH, Muhammad; GOPALAN, Kaliappan. Deception detection in speech using bark band and perceptually significant energy features. **Midwest Symposium on Circuits and Systems**, *[S. l.]*, p. 1212–1215, 2013. DOI: 10.1109/MWSCAS.2013.6674872.

SHEIKHOLESLAMI, Sina; MEISTER, Moritz; WANG, Tianze; PAYBERAH, Amir H.; VLASSOV, Vladimir; DOWLING, Jim. AutoAblation: Automated Parallel Ablation Studies for Deep Learning. **Proceedings of the 1st Workshop on Machine Learning and Systems, EuroMLSys 2021**, *[S. l.]*, p. 55–61, 2021. DOI:

10.1145/3437984.3458834.

SRIVASTAVA, Nitish; HINTON, Geoffrey; KRIZHEVSKY, Alex; SUTSKEVER, Ilya; SALAKHUTDINOV, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, *[S. l.]*, v. 15, p. 1929–1958, 2014.

STELLER, Max. Recent Developments in Statement Analysis. *In*: YUILLE, John C. (org.). **Credibility Assessment**. Dordrecht: Springer Netherlands, 1989. p. 135–154. DOI: 10.1007/978-94-015-7856-1_8. Disponível em: https://doi.org/10.1007/978-94-015-7856-1_8.

SUCHOTZKI, Kristina; GAMER, Matthias. Effect of negative motivation on the behavioral and autonomic correlates of deception. **Psychophysiology**, *[S. l.]*, v. 56, n. 1, p. 1–11, 2019. DOI: 10.1111/psyp.13284.

TAKABATAKE, Shohei; SHIMADA, Kazutaka; SAITOH, Takeshi. Construction of a liar corpus and detection of lying situations. **Proceedings - 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2018**, *[S. l.]*, p. 971–976, 2018. DOI: 10.1109/SCIS-ISIS.2018.00161.

TAO, Huawei; LEI, Peizhi; WANG, Mengzhe; WANG, Jie; FU, Hongliang. Speech Deception Detection Algorithm Based on SVM and Acoustic Features. **Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019**, *[S. l.]*, p. 31–33, 2019. DOI: 10.1109/ICCSNT47585.2019.8962491.

TAYLOR, Paul J.; LARNER, Samuel; CONCHIE, Stacey M.; VAN DER ZEE, Sophie. Cross-Cultural Deception Detection. **Detecting Deception: Current Challenges and Cognitive Approaches**, *[S. l.]*, n. January, p. 175–201, 2015. DOI: 10.1002/9781118510001.ch8.

TEN BRINKE, Leanne; PORTER, Stephen. Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception. **Law and Human Behavior**, *[S. l.]*, v. 36, n. 6, p. 469–477, 2012. DOI: 10.1037/h0093929.

TURNER, Ronny E.; EDGLEY, Charles; OLMSTEAD, Glen. Information control in conversatrions: Honesty is not always the baset policy. **The Kansas Journal of Sociology**, *[S. l.]*, v. 11, n. 1, p. 69–89, 1975. Disponível em: http://www.jstor.org/stable/23255229.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Łukasz; POLOSUKHIN, Illia. Attention is all you need. **Advances in Neural Information Processing Systems**, *[S. l.]*, v. 2017-Decem, n. Nips, p. 5999–6009, 2017.

VELICHKO, Alena; BUDKOV, Viktor; KAGIROV, Ildar; KARPOV, Alexey. Comparative Analysis of Classification Methods for Automatic Deception Detection in Speech. *In*: (Alexey Karpov, Oliver Jokisch, Rodmonga Potapova, Org.)SPEECH AND COMPUTER 2018, Cham. **Anais** [...]. Cham: Springer International Publishing, 2018. p. 737–746.

VENKATESH, Sushma; RAMACHANDRA, Raghavendra; BOURS, Patrick. Robust

Algorithm for Multimodal Deception Detection. **Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019**, *[S. l.]*, p. 534–537, 2019. DOI: 10.1109/MIPR.2019.00108.

VRIJ, Aldert. **Detecting Lies and Deceit: Pitfalls and Opportunities**. 2nd. ed. Chichester: John Wiley & Sons, Ltd, 2008.

WALCZYK, Jeffrey J.; HARRIS, Laura L.; DUCK, Terri K.; MULAY, Devyani. A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. **New Ideas in Psychology**, *[S. l.]*, v. 34, n. 1, p. 22–36, 2014. DOI: 10.1016/j.newideapsych.2014.03.001. Disponível em: http://dx.doi.org/10.1016/j.newideapsych.2014.03.001.

WANG, Hongzhi; BAH, Mohamed Jaward; HAMMAD, Mohamed. Progress in Outlier Detection Techniques: A Survey. **IEEE Access**, *[S. l.]*, v. 7, p. 107964–108000, 2019. DOI: 10.1109/ACCESS.2019.2932769.

WANI, M. Arif; BHAT, Farooq Ahmad; AFZAL, Saduf; KHAN, Asif Iqbal. **Advances in Deep Learning**. Warsaw: Springer International Publishing, 2019. v. 57 DOI: 10.1007/978-981-13-6794-6.

WENINGER, Felix; EYBEN, Florian; SCHULLER, Björn W.; MORTILLARO, Marcello; SCHERER, Klaus R. On the acoustics of emotion in audio: What speech, music, and sound have in common. **Frontiers in Psychology**, *[S. l.]*, v. 4, n. MAY, p. 1–12, 2013. DOI: 10.3389/fpsyg.2013.00292.

XIE, Yue; LIANG, Ruiyu; TAO, Huawei; ZHU, Yue; ZHAO, Li. Convolutional bidirectional long short-term memory for deception detection with acoustic features. **IEEE Access**, *[S. l.]*, v. 6, p. 76527–76534, 2018. DOI: 10.1109/ACCESS.2018.2882917.

XU, Kelvin; BAY, Jimmy Lei; KIROSY, Ryan; CHO, Kyunghyun; COURVILLE, Aaron; SALAKHUTDINOVY, Ruslan; S. ZEMELY, Richard; BENGIO, Yoshua. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *In*: (Francis Bach, David Blei, Org.)PROCEEDINGS OF THE 32ND INTERNATIONAL CONFERENCE ON MACHINE LEARNING 2015, lille. **Anais** [...]. lille: PMLR, 2015. p. 2048–2057. DOI: 10.1016/j.scitotenv.2016.07.196. Disponível em: https://proceedings.mlr.press/v37/xuc15.html.

YONG, Bang Xiang; BRINTRUP, Alexandra. Do Autoencoders Need a Bottleneck for Anomaly Detection? **IEEE Access**, *[S. l.]*, v. 10, n. June, p. 78455–78471, 2022. DOI: 10.1109/ACCESS.2022.3192134.

ZHAI, Junhai; ZHANG, Sufang; CHEN, Junfen; HE, Qiang. Autoencoder and Its Various Variants. **Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018**, *[S. l.]*, p. 415–419, 2019. DOI: 10.1109/SMC.2018.00080.

ZUCKERMAN, Miron; DEPAULO, Bella M.; ROSENTHAL, Robert. Verbal and nonverbal communication of deception. *In*: ADVANCES IN EXPERIMENTAL SOCIAL PSYCHOLOGY 1981, **Anais** [...]. : Academic Press Inc., 1981. DOI: 10.1002/aic.690020228.

APÊNDICE A

O Quadro 63 apresenta um resumo das principais características do MMDDD-PtBr.

QUADRO 63 - RESUMO DAS PRINCIPAIS CARACTERÍSTICAS DO MMDDD-PTBR

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
0	S1-P7-1	Não sincero	1909.8	1911.4	1.6	S1-P7-1-opensmile.csv	161	S1-P7-1-verbal.csv	7	Eu sobrevivi a um ataque de tubarão.	S1-P7-1-openface.csv	97		0	97	100.0
1	S1-P7-2	Sincero	2112.5	2113.5	1.0	S1-P7-2-opensmile.csv	101	S1-P7-2-verbal.csv	2	Sou biólogo.	S1-P7-2-openface.csv	60		0	60	100.0
2	S1-P7-3	Sincero	2114.0	2116.5	2.5	S1-P7-3-opensmile.csv	248	S1-P7-3-verbal.csv	11	Eu trabalho com baleias e golfinhos há	S1-P7-3-openface.csv	150		0	150	100.0
3	S1-P7-4	Não sincero	2116.7	2126.5	9.8	S1-P7-4-opensmile.csv	981	S1-P7-4-verbal.csv	35	Nesse momento eu tava na verdade ajudando	S1-P7-4-openface.csv	589	S1-P7-4-openface-cuts.csv	3	398	67.6
4	S1-P7-5	Sincero	2126.5	2135.8	9.3	S1-P7-5-opensmile.csv	931	S1-P7-5-verbal.csv	30	Eh... a pesca de espinhel, pra quem	S1-P7-5-openface.csv	559	S1-P7-5-openface-cuts.csv	1	385	68.9
5	S1-P7-6	Não sincero	2135.8	2171.8	36.0	S1-P7-6-opensmile.csv	3696	S1-P7-6-verbal.csv	121	Então nesse caso era um tubarão azul,	S1-P7-6-openface.csv	2160	S1-P7-6-openface-cuts.csv	3	1886	87.3

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
6	S1-P7-7	Sincero	2177.2	2186.0	8.8	S1-P7-7-opensmile.csv	879	S1-P7-7-verbal.csv	30	Na verdade, assim... a segurança não era	S1-P7-7-openface.csv	529	S1-P7-7-openface-cuts.csv	2	399	75.4
7	S1-P7-8	Não sincero	2328.0	2333.0	5.0	S1-P7-8-opensmile.csv	500	S1-P7-8-verbal.csv	19	Meu parceiro de mergulho foi o primeiro	S1-P7-8-openface.csv	300	S1-P7-8-openface-cuts.csv	2	208	69.3
8	S1-P7-9	Não sincero	2355.4	2358.1	2.7	S1-P7-9-opensmile.csv	270	S1-P7-9-verbal.csv	10	Eh... for déi... eh... ao todo dez	S1-P7-9-openface.csv	162	S1-P7-9-openface-cuts.csv	1	97	59.9
9	S1-P7-10	Sincero	2542.5	2545.0	2.5	S1-P7-10-opensmile.csv	248	S1-P7-10-verbal.csv	9	Meu nome é Clarêncio, sou biólogo de...	S1-P7-10-openface.csv	150	S1-P7-10-openface-cuts.csv	1	117	78.0
10	S1-P7-11	Sincero	2546.0	2550.0	4.0	S1-P7-11-opensmile.csv	401	S1-P7-11-verbal.csv	13	Mergulhador também, trabalho com mamíferos marinhos, baleias	S1-P7-11-openface.csv	240	S1-P7-11-openface-cuts.csv	1	149	62.1
11	S1-P7-12	Sincero	2551.7	2555.7	4.0	S1-P7-12-opensmile.csv	401	S1-P7-12-verbal.csv	17	Na costa do Brasil todo, principalmente na	S1-P7-12-openface.csv	240		0	240	100.0
12	S1-P7-13	Sincero	2557.8	2574.8	17.1	S1-P7-13-opensmile.csv	1710	S1-P7-13-verbal.csv	55	Atualmente eu... eh... trabalho... me especializei em	S1-P7-13-openface.csv	1026	S1-P7-13-openface-cuts.csv	1	994	96.9

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
13	S1-P8-1	Sincero	1918.4	1920.4	2.0	S1-P8-1-opensmile.csv	200	S1-P8-1-verbal.csv	7	eu sobrevivi a um ataque de tubarão	S1-P8-1-openface.csv	120		0	120	100.0
14	S1-P8-2	Sincero	2051.5	2068.8	17.3	S1-P8-2-opensmile.csv	1731	S1-P8-2-verbal.csv	65	Eu consegui velejar pra praia, voltar pra	S1-P8-2-openface.csv	1039	S1-P8-2-openface-cuts.csv	2	747	71.9
15	S1-P8-3	Sincero	2073.7	2092.2	18.5	S1-P8-3-opensmile.csv	1851	S1-P8-3-verbal.csv	67	Eu fiz uma manobra e entrei no	S1-P8-3-openface.csv	1110	S1-P8-3-openface-cuts.csv	4	704	63.4
16	S1-P8-4	Sincero	2094.8	2101.8	7.0	S1-P8-4-opensmile.csv	699	S1-P8-4-verbal.csv	28	Eu tava velejando na praia de Manguinhos,	S1-P8-4-openface.csv	420	S1-P8-4-openface-cuts.csv	2	285	67.9
17	S1-P8-5	Sincero	2105.2	2107.2	2.0	S1-P8-5-opensmile.csv	200	S1-P8-5-verbal.csv	5	Sei, foi um tubarão branco.	S1-P8-5-openface.csv	120	S1-P8-5-openface-cuts.csv	1	0	0.0
18	S1-P8-6	Sincero	2334.5	2337.2	2.7	S1-P8-6-opensmile.csv	270	S1-P8-6-verbal.csv	13	Esse casal que tava... que... que eu	S1-P8-6-openface.csv	162	S1-P8-6-openface-cuts.csv	1	101	62.3
19	S1-P8-7	Sincero	2360.0	2365.6	5.6	S1-P8-7-opensmile.csv	560	S1-P8-7-verbal.csv	15	Durou dez meses também, eu tomei duzentos	S1-P8-7-openface.csv	336	S1-P8-7-openface-cuts.csv	1	253	75.3
20	S1-P9-1	Não sincero	1922.4	1924.6	2.2	S1-P9-1-opensmile.csv	221	S1-P9-1-verbal.csv	7	Eu sobrevivi a um ataque de tubarão.	S1-P9-1-openface.csv	132		0	132	100.0

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
21	S1-P9-2	Não sincero	2040.5	2048.0	7.5	S1-P9-2-opensmile.csv	751	S1-P9-2-verbal.csv	23	Pior momento foi na hora que ele	S1-P9-2-openface.csv	450	S1-P9-2-openface-cuts.csv	2	297	66.0
22	S1-P9-3	Sincero	2241.8	2248.8	7.0	S1-P9-3-opensmile.csv	704	S1-P9-3-verbal.csv	25	Então, eu fui criado... nascido e criado	S1-P9-3-openface.csv	422	S1-P9-3-openface-cuts.csv	1	375	88.9
23	S1-P9-5	Não sincero	2303.5	2324.5	21.0	S1-P9-5-opensmile.csv	2099	S1-P9-5-verbal.csv	66	Eh... eu até... eu não tinha visto	S1-P9-5-openface.csv	1260	S1-P9-5-openface-cuts.csv	4	779	61.8
24	S1-P9-6	Não sincero	2339.5	2342.0	2.5	S1-P9-6-opensmile.csv	248	S1-P9-6-verbal.csv	10	Quando eu cheguei na praia foram os	S1-P9-6-openface.csv	150	S1-P9-6-openface-cuts.csv	1	107	71.3
25	S1-P9-8	Não sincero	2381.9	2387.0	5.1	S1-P9-8-opensmile.csv	511	S1-P9-8-verbal.csv	13	Eu ainda puxo um pouco na perna,	S1-P9-8-openface.csv	306	S1-P9-8-openface-cuts.csv	2	301	98.4
26	S1-P9-9	Sincero	2587.0	2590.0	3.0	S1-P9-9-opensmile.csv	299	S1-P9-9-verbal.csv	8	Sou Paulo Carvalho, sou empresário, chefe de	S1-P9-9-openface.csv	180	S1-P9-9-openface-cuts.csv	1	135	75.0
27	S1-P9-10	Sincero	2598.6	2602.3	3.7	S1-P9-10-opensmile.csv	370	S1-P9-10-verbal.csv	13	Em dois mil e quatro fui campeão	S1-P9-10-openface.csv	223		0	223	100.0

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
28	S1-P9-11	Sincero	2603.7	2619.1	15.4	S1-P9-11-opensmile.csv	1540	S1-P9-11-verbal.csv	56	Eh... velejo de kite surf também lá	S1-P9-11-openface.csv	925	S1-P9-11-openface-cuts.csv	3	574	62.1
29	S2-P1-1	Sincero	244.0	248.0	4.0	S2-P1-1-opensmile.csv	401	S2-P1-1-verbal.csv	11	Comprei um porco pra comer, mas ele	S2-P1-1-openface.csv	120	S2-P1-1-openface-cuts.csv	1	96	80.0
30	S2-P1-2	Sincero	349.4	352.0	2.6	S2-P1-2-opensmile.csv	265	S2-P1-2-verbal.csv	6	Pra engordar ele. Ficar bem gordinho.	S2-P1-2-openface.csv	80		0	80	100.0
31	S2-P1-3	Sincero	355.0	357.5	2.5	S2-P1-3-opensmile.csv	248	S2-P1-3-verbal.csv	7	Ah, mas... numsei, foi coisa de momento.	S2-P1-3-openface.csv	75	S2-P1-3-openface-cuts.csv	1	68	90.7
32	S2-P2-1	Não sincero	251.4	255.0	3.6	S2-P2-1-opensmile.csv	359	S2-P2-1-verbal.csv	11	Comprei um porco pra comer, mas ele	S2-P2-1-openface.csv	108	S2-P2-1-openface-cuts.csv	1	101	93.5
33	S2-P2-2	Não sincero	442.4	448.5	6.1	S2-P2-2-opensmile.csv	610	S2-P2-2-verbal.csv	22	Ele dorme pelo quintal da casa. Aí,	S2-P2-2-openface.csv	183	S2-P2-2-openface-cuts.csv	2	120	65.6
34	S2-P2-3	Não sincero	559.0	571.4	12.4	S2-P2-3-opensmile.csv	1240	S2-P2-3-verbal.csv	44	Eh, eu só dou ração pra ele.	S2-P2-3-openface.csv	372	S2-P2-3-openface-cuts.csv	2	314	84.4

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
35	S2-P2-5	Sincero	1158.5	1162.0	3.5	S2-P2-5-opensmile.csv	349	S2-P2-5-verbal.csv	11	Eu sou Patrícia Salia, sou chefe de	S2-P2-5-openface.csv	105	S2-P2-5-openface-cuts.csv	1	84	80.0
36	S2-P3-1	Não sincero	256.0	260.0	4.0	S2-P3-1-opensmile.csv	401	S2-P3-1-verbal.csv	11	Comprei um porco pra comer, mas ele	S2-P3-1-openface.csv	120	S2-P3-1-openface-cuts.csv	1	73	60.8
37	S2-P3-2	Não sincero	329.7	342.5	12.8	S2-P3-2-opensmile.csv	1279	S2-P3-2-verbal.csv	25	Sim, eu... a minha ideia era fazer	S2-P3-2-openface.csv	384	S2-P3-2-openface-cuts.csv	2	237	61.7
38	S2-P3-3	Não sincero	408.3	423.0	14.7	S2-P3-3-opensmile.csv	1469	S2-P3-3-verbal.csv	38	A convivência é excelente. Vi... ele vive	S2-P3-3-openface.csv	441	S2-P3-3-openface-cuts.csv	2	291	66.0
39	S2-P3-4	Não sincero	580.0	587.0	7.0	S2-P3-4-opensmile.csv	704	S2-P3-4-verbal.csv	16	Sim, ele entra. Ele entra uma vez,	S2-P3-4-openface.csv	212	S2-P3-4-openface-cuts.csv	1	149	70.3
40	S2-P3-5	Sincero	1168.5	1180.5	12.0	S2-P3-5-opensmile.csv	1200	S2-P3-5-verbal.csv	23	Meu nome é Teodósio Mitzko, sou de	S2-P3-5-openface.csv	360	S2-P3-5-openface-cuts.csv	2	190	52.8
41	S2-P4-1	Sincero	1309.0	1313.5	4.5	S2-P4-1-opensmile.csv	451	S2-P4-1-verbal.csv	10	Eu nunca deço do salto alto, nem	S2-P4-1-openface.csv	135	S2-P4-1-openface-cuts.csv	1	119	88.1

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
42	S2-P4-2	Sincero	1428.0	1442.5	14.5	S2-P4-2-opensmile.csv	1451	S2-P4-2-verbal.csv	28	Fácil! Eu encosto minhas mãos na parede,	S2-P4-2-openface.csv	435	S2-P4-2-openface-cuts.csv	2	260	59.8
43	S2-P5-1	Não sincero	1316.0	1319.5	3.5	S2-P5-1-opensmile.csv	349	S2-P5-1-verbal.csv	10	Eu nunca deço do salto alto, nem	S2-P5-1-openface.csv	105	S2-P5-1-openface-cuts.csv	1	75	71.4
44	S2-P5-2	Sincero	1525.1	1545.0	19.9	S2-P5-2-opensmile.csv	1989	S2-P5-2-verbal.csv	56	Eh... eu sou atriz e trabalho com	S2-P5-2-openface.csv	596	S2-P5-2-openface-cuts.csv	3	425	71.3
45	S2-P5-3	Sincero	1830.4	1838.3	7.9	S2-P5-3-opensmile.csv	790	S2-P5-3-verbal.csv	25	Eu sou atriz, trabalho com eventos. Eu	S2-P5-3-openface.csv	237	S2-P5-3-openface-cuts.csv	2	169	71.3
46	S2-P6-1	Não sincero	1320.5	1323.5	3.0	S2-P6-1-opensmile.csv	299	S2-P6-1-verbal.csv	10	Eu nunca deço do salto alto, nem	S2-P6-1-openface.csv	90	S2-P6-1-openface-cuts.csv	1	59	65.6
47	S2-P6-2	Não sincero	1499.4	1503.0	3.6	S2-P6-2-opensmile.csv	359	S2-P6-2-verbal.csv	13	Eu fico de salto porque é uma	S2-P6-2-openface.csv	108	S2-P6-2-openface-cuts.csv	1	83	76.9
48	S2-P6-5	Sincero	1840.0	1848.5	8.5	S2-P6-5-opensmile.csv	849	S2-P6-5-verbal.csv	26	Eu sou a Vera Loyola, sou terapeuta	S2-P6-5-openface.csv	255	S2-P6-5-openface-cuts.csv	1	107	42.0

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
49	S2-P7-1	Não sincero	1912.0	1918.0	6.0	S2-P7-1-opensmile.csv	600	S2-P7-1-verbal.csv	15	Meu nome é Darko Hunter e eu	S2-P7-1-openface.csv	180	S2-P7-1-openface-cuts.csv	1	131	72.8
50	S2-P7-4	Não sincero	2295.7	2301.1	5.4	S2-P7-4-opensmile.csv	539	S2-P7-4-verbal.csv	9	Minha profissão é coordenador de busca de	S2-P7-4-openface.csv	162	S2-P7-4-openface-cuts.csv	1	142	87.7
51	S2-P7-5	Não sincero	2342.0	2355.0	13.0	S2-P7-5-opensmile.csv	1300	S2-P7-5-verbal.csv	34	Padre, todos são emocionantes. São seis mil	S2-P7-5-openface.csv	390	S2-P7-5-openface-cuts.csv	3	232	59.5
52	S2-P7-6	Sincero	2466.0	2472.7	6.7	S2-P7-6-opensmile.csv	670	S2-P7-6-verbal.csv	24	Sou coordenador de cuidados com o passageiro	S2-P7-6-openface.csv	201	S2-P7-6-openface-cuts.csv	1	167	83.1
53	S2-P7-7	Sincero	2483.2	2489.7	6.5	S2-P7-7-opensmile.csv	649	S2-P7-7-verbal.csv	18	Muita coisa, eh... a gente encontra muito	S2-P7-7-openface.csv	195	S2-P7-7-openface-cuts.csv	1	184	94.4
54	S2-P7-8	Sincero	2495.2	2518.2	23.0	S2-P7-8-opensmile.csv	2300	S2-P7-8-verbal.csv	75	A pessoa se envolveu em alguma... em	S2-P7-8-openface.csv	689	S2-P7-8-openface-cuts.csv	5	387	56.2
55	S2-P8-1	Não sincero	1920.0	1926.0	6.0	S2-P8-1-opensmile.csv	600	S2-P8-1-verbal.csv	16	Meu nome é Darko Hunter e já	S2-P8-1-openface.csv	180	S2-P8-1-openface-cuts.csv	1	146	81.1

#	Sujeito		Segmento			Acústico		Verbal			Visual		Recortes			
	Narrativa	Categoria	Início	Fim	Duração	Arquivo	Linhas x 68	Arquivo	Palavras	Primeiras palavras	Arquivo	Linhas x 714	Arquivo	Contagem	Linhas	Linhas %
56	S2-P8-5	Sincero	2537.4	2547.0	9.6	S2-P8-5-opensmile.csv	955	S2-P8-5-verbal.csv	32	Sim, sou detetive de verdade há trinta	S2-P8-5-openface.csv	286	S2-P8-5-openface-cuts.csv	1	252	88.1
57	S2-P9-1	Sincero	1926.6	1931.5	4.9	S2-P9-1-opensmile.csv	490	S2-P9-1-verbal.csv	14	Meu nome é darko hunter e já	S2-P9-1-openface.csv	147	S2-P9-1-openface-cuts.csv	1	96	65.3
58	S2-P9-2	Sincero	1982.0	1984.5	2.5	S2-P9-2-opensmile.csv	248	S2-P9-2-verbal.csv	5	Eu trabalho procurando pessoas desaparecidas.	S2-P9-2-openface.csv	75	S2-P9-2-openface-cuts.csv	1	0	0.0
59	S2-P9-3	Sincero	2046.6	2106.4	59.8	S2-P9-3-opensmile.csv	5981	S2-P9-3-verbal.csv	178	Foi um caso... que era um senhor	S2-P9-3-openface.csv	1791	S2-P9-3-openface-cuts.csv	10	963	53.8
60	S2-P9-4	Sincero	2562.4	2576.8	14.4	S2-P9-4-opensmile.csv	1441	S2-P9-4-verbal.csv	46	Exatamente, desde dois mil e sete eu	S2-P9-4-openface.csv	432	S2-P9-4-openface-cuts.csv	2	334	77.3

FONTE: DADOS DA PESQUISA (2023)