



UNIVERSIDADE ESTADUAL PAULISTA

“JÚLIO DE MESQUITA FILHO”

Câmpus de São José do Rio Preto

Jéssica Vicentini

**Comparando Técnicas de Explicabilidade sobre
Modelos de Linguagem: um Estudo de Caso na
Detecção de Notícias Falsas**

São José do Rio Preto - SP

2024

Jéssica Vicentini

**Comparando Técnicas de Explicabilidade sobre Modelos de
Linguagem: um Estudo de Caso na Detecção de Notícias Falsas**

Orientador: Prof. Dr. Arnaldo Candido Junior

Coorientador: Prof. Dr. Ivan Rizzo Guilherme

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

São José do Rio Preto - SP

2024

V633c	Vicentini, Jéssica Comparando Técnicas de Explicabilidade sobre Modelos de Linguagem: um Estudo de Caso na Detecção de Notícias Falsas / Jéssica Vicentini. -- São José do Rio Preto, 2024 147 p. : il., tabs. Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto Orientador: Arnaldo Candido Junior Coorientador: Ivan Rizzo Guilherme 1. Inteligência artificial. 2. Inteligência artificial explicável. 3. Fake news. 4. Processamento de linguagem natural (Computação). 5. Aprendizado de máquinas. I. Título.
-------	--

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Jéssica Vicentini

Comparando Técnicas de Explicabilidade sobre Modelos de Linguagem: um Estudo de Caso na Detecção de Notícias Falsas

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Banca Examinadora

- Prof. Dr. Arnaldo Candido Junior (Orientador)
Universidade Estadual Paulista – Câmpus de São José do Rio Preto
- Prof. Dr. Ivan Rizzo Guilherme (Coorientador)
Universidade Estadual Paulista – Câmpus de Rio Claro,
- Prof. Dr. João Paulo Papa
Universidade Estadual Paulista – Câmpus de Bauru
- Prof. Dr. Frederico Santos de Oliveira
Universidade Federal de Mato Grosso

São José do Rio Preto - SP

29 de Julho de 2024

*Este trabalho é dedicado a todas as mulheres,
que se esforçam diariamente para conquistar o seu lugar.*

Agradecimentos

Agradeço primeiramente ao meu orientador Prof. Dr. Arnaldo Candido Junior e ao meu coorientador Prof. Dr. Ivan Rizzo Guilherme por me apoiarem e me guiarem nessa jornada, secundamente ao Rafael Bezerra de Menezes Rodrigues que desde o começo me ajudou e deu suporte com os experimentos, e por fim agradeço a minha família, amigos e ao meu namorado que sempre me deram suporte e acreditaram que eu seria capaz de atingir os meus objetivos.

Resumo

Modelos de linguagem têm sido amplamente adotados em várias aplicações de processamento de linguagem natural. Seu sucesso em tarefas como tradução automática, classificação e geração de texto impulsionou sua popularidade. No entanto, esses modelos são frequentemente considerados “caixas-opacas” devido à sua complexidade e dificuldade de interpretação. A falta de transparência e de entendimento interno desses modelos levanta questões sobre sua confiabilidade e limita sua adoção em cenários críticos, como a tomada de decisão em áreas sensíveis. Nesse contexto, o presente trabalho visa explorar técnicas de Inteligência Artificial Explicável para interpretar e entender o comportamento dos modelos de linguagem. Especificamente, foca nos métodos Model-Agnostic Explanations (LIME) e Integrated Gradients (IG). O estudo de caso envolve a análise da eficácia de modelos BERTimbau treinados no contexto da classificação de notícias em português brasileiro como reais ou falsas, utilizando os conjuntos de dados FakeRecognize e Fake.Br Corpus, e buscar compreender se esses métodos são eficazes para esse cenário.

Palavras-chave: Inteligência artificial explicável. BERTimbau. Local interpretable model-agnostic explanations. Integrated gradients. Processamento de linguagem natural. Aprendizado de máquina. Deep Learning. Modelos de linguagem. Transformer.

Abstract

Language models have been widely adopted in various natural language processing applications. Their success in tasks such as machine translation, classification, and text generation has driven their popularity. However, these models are often considered “black-boxes” due to their complexity and difficulty of interpretation. The lack of transparency and internal understanding of these models raises questions about their reliability and limits their adoption in critical scenarios, such as decision-making in sensitive areas. In this context, the present work aims to explore techniques of Explainable Artificial Intelligence to interpret and understand the behavior of language models. Specifically, it focuses on the methods Model-Agnostic Explanations (LIME) and Integrated Gradients (IG). The case study involves analyzing the effectiveness of BERTimbau models trained in the context of classifying Brazilian Portuguese news as real or fake, using the FakeRecogna and Fake.Br Corpus datasets, and seeks to understand if these methods are effective for this scenario.

Keywords: Explainable artificial intelligence. BERTimbau. Local interpretable model-agnostic explanations. Integrated gradients. Natural language processing. Machine learning. Deep learning. Language models. Transformer.

Lista de Figuras

Figura 1 – Etapas da análise de PLN - Fonte: Adaptado de Dale (2010)	28
Figura 2 – Comparação das arquiteturas RNN e Auto-atenção - Fonte: Adaptado de Zhang et al. (2021)	32
Figura 3 – Representação da matriz de incorporação da frase “ <i>Hello good morning</i> ” - Fonte: Adaptado de Alammr (2018)	33
Figura 4 – Modelo de arquitetura Transformer - Fonte: Traduzido de Vaswani et al. (2017)	34
Figura 5 – Diagrama da estrutura de um Transformer constituído de uma pilha de codificadores e decodificadores - Fonte: Adaptado de Alammr (2018)	35
Figura 6 – Atenção multi-cabeça e <i>scaled dot-product attention</i> - Fonte: Traduzido de Vaswani et al. (2017)	35
Figura 7 – Geração dos vetores V, Q e K apartir das matrizes de peso - Fonte: Adaptado de Alammr (2018)	36
Figura 8 – Cálculo de atenção da frase “Hello good morning” - Fonte: Adaptado de Alammr (2018)	37
Figura 9 – Geração das matrizes Z pelas 8 cabeças de atenção - Fonte: Adaptado de Alammr (2018)	37
Figura 10 – Concatenação das matrizes Z e multiplicação da matriz resultante pela matriz de pesos W^O - Fonte: Adaptado de Alammr (2018)	38
Figura 11 – Fine Tuning Adaptativo - Fonte: Traduzido de Ruder (2021)	41
Figura 12 – Fine Tuning Comportamental - Fonte: Traduzido de Ruder (2021)	41
Figura 13 – Representação de um entrada do BERT - Fonte: Traduzido de Devlin et al. (2018)	44
Figura 14 – Tipos de técnicas de XAI - Fonte: Adaptado de Gohel, Singh e Mohanty (2021)	48
Figura 15 – Processo de explicação com o LIME - Fonte: Adaptado de Ribeiro, Singh e Guestrin (2016)	53
Figura 16 – Análise individual realizada pelo LIME - Fonte: Adaptado de Ribeiro, Singh e Guestrin (2016)	54

Figura 17 – Exemplo da detecção de componentes importantes para a classificação por meio do método de Integrated Gradients - Fonte: Sundararajan, Taly e Yan (2017)	58
Figura 18 – Comparação quantitativa das pontuações de explicação entre árvores de decisão explicáveis, desvios normalizados de características e LIME - Fonte: Traduzido de Torcianti e Matzka (2021)	62
Figura 19 – Número máximo de unidades de explicação em uma explicação por instância em comparação com os resultados de precisão descritiva - Fonte: Traduzido de Moradi e Samwald (2021)	64
Figura 20 – Etapas seguidas para realização dos experimentos	65
Figura 21 – Distribuição das amostras após a divisão nos conjuntos de treinamento e teste	67
Figura 22 – Acurácia e Perda do Modelo 1 e 2 ao longo das 10 épocas de treinamento . .	69
Figura 23 – Acurácia e Perda do Modelo 3 e 4 ao longo das 10 épocas de treinamento . .	70
Figura 24 – Matrizes de Confusão do Modelo 1 e Modelo 2	72
Figura 25 – Matrizes de Confusão do Modelo 3 e Modelo 4	73
Figura 26 – Explicações geradas pelo LIME do Modelo 1	74
Figura 27 – Explicações geradas pelo LIME do Modelo 2	77
Figura 28 – Explicações geradas pelo LIME do Modelo 3	80
Figura 29 – Explicações geradas pelo LIME do Modelo 4	82
Figura 30 – Escala de cores dos mapas de saliência	84
Figura 31 – Explicações geradas pelo IG do Modelo 1	85
Figura 32 – Explicações geradas pelo IG do Modelo 2	87
Figura 33 – Explicações geradas pelo IG do Modelo 3	89
Figura 34 – Explicações geradas pelo IG do Modelo 4	91
Figura 35 – Métricas <i>Dataset FakeRecogna</i> com <i>stop words</i>	96
Figura 36 – Métricas <i>Dataset FakeRecogna</i> sem <i>stop words</i>	97
Figura 37 – Métricas <i>Dataset Fake.Br Corpus</i> com <i>stop words</i>	98
Figura 38 – Métricas <i>Dataset Fake.Br Corpus</i> sem <i>stop words</i>	99
Figura 39 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.	112
Figura 40 – Modelo 1 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.	112

Figura 41 – Modelo 1 - Explicação da notícias 5 falsa classificada como falsa gerada pelo LIME.	113
Figura 42 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.	113
Figura 43 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.	113
Figura 44 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.	114
Figura 45 – Modelo 1 - Explicação da notícias 5 real classificada como real gerada pelo LIME.	114
Figura 46 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.	114
Figura 47 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.	115
Figura 48 – Modelo 1 - Explicação da notícias 5 real classificada como falsa gerada pelo LIME.	115
Figura 49 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.	116
Figura 50 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.	116
Figura 51 – Modelo 2 - Explicação da notícia 5 falsa classificada como falsa gerada pelo LIME.	117
Figura 52 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.	117
Figura 53 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo LIME.	117
Figura 54 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo LIME.	118
Figura 55 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.	118
Figura 56 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.	118

Figura 57 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo LIME.	119
Figura 58 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.	120
Figura 59 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.	120
Figura 60 – Modelo 2 - Explicação da notícia 5 real classificada como falsa gerada pelo LIME.	120
Figura 61 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.	121
Figura 62 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.	121
Figura 63 – Modelo 3 - Explicação da notícia 5 falsa classificada como falsa gerada pelo LIME.	122
Figura 64 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.	122
Figura 65 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo LIME.	122
Figura 66 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo LIME.	123
Figura 67 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.	123
Figura 68 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.	123
Figura 69 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo LIME.	124
Figura 70 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.	124
Figura 71 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.	124
Figura 72 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.	125

Figura 73 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.	125
Figura 74 – Modelo 4 - Explicação da notícia 5 falsa classificada como falsa gerada pelo LIME.	126
Figura 75 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.	126
Figura 76 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo LIME.	126
Figura 77 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo LIME.	127
Figura 78 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.	127
Figura 79 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.	127
Figura 80 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo LIME.	128
Figura 81 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.	128
Figura 82 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.	128
Figura 83 – Modelo 4 - Explicação da notícia 5 real classificada como falsa gerada pelo LIME.	129
Figura 84 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.	130
Figura 85 – Modelo 1 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.	130
Figura 86 – Modelo 1 - Explicação da notícias 5 falsa classificada como falsa gerada pelo IG.	131
Figura 87 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.	131
Figura 88 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.	131

Figura 89 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.	132
Figura 90 – Modelo 1 - Explicação da notícias 5 real classificada como real gerada pelo IG.	132
Figura 91 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.	132
Figura 92 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.	133
Figura 93 – Modelo 1 - Explicação da notícias 5 real classificada como falsa gerada pelo IG.	133
Figura 94 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.	134
Figura 95 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.	134
Figura 96 – Modelo 2 - Explicação da notícia 5 falsa classificada como falsa gerada pelo IG.	135
Figura 97 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.	135
Figura 98 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo IG.	135
Figura 99 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo IG.	136
Figura 100 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.	136
Figura 101 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.	136
Figura 102 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo IG.	137
Figura 103 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.	137
Figura 104 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.	137
Figura 105 – Modelo 2 - Explicação da notícia 5 real classificada como falsa gerada pelo IG.	138
Figura 106 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.	139
Figura 107 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.	139
Figura 108 – Modelo 3 - Explicação da notícia 5 falsa classificada como falsa gerada pelo IG.	140

Figura 109 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.	140
Figura 110 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo IG.	140
Figura 111 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo IG.	141
Figura 112 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.	141
Figura 113 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.	141
Figura 114 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo IG.	142
Figura 115 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.	142
Figura 116 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.	142
Figura 117 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.	143
Figura 118 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.	143
Figura 119 – Modelo 4 - Explicação da notícia 5 falsa classificada como falsa gerada pelo IG.	144
Figura 120 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.	144
Figura 121 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo IG.	144
Figura 122 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo IG.	145
Figura 123 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.	145
Figura 124 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.	145
Figura 125 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo IG.	146
Figura 126 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.	146
Figura 127 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.	146

Figura 128 – Modelo 4 - Explicação da notícia 5 real classificada como falsa gerada pelo IG. 147

Lista de Tabelas

Tabela 1 – Variações da arquitetura do Transformer utilizadas pelos autores - Fonte: Vaswani et al. (2017)	39
Tabela 2 – Resultados do Transformer em comparação a outros modelos do estado da arte - Fonte: Vaswani et al. (2017)	39
Tabela 3 – Resultados do modelo BERT no GLUE Test - Fonte: Devlin et al. (2018) . .	44
Tabela 4 – BERTimbau resultados STS e RTE - Fonte: Traduzido de Souza, Nogueira e Lotufo (2020)	45
Tabela 5 – BERTimbau resultados NER - Fonte: Traduzido de Souza, Nogueira e Lotufo (2020)	46
Tabela 6 – Modelos treinados	68
Tabela 7 – Taxa de perda e acurácia de teste dos modelos	71
Tabela 8 – Métricas dos modelos	71
Tabela 9 – Tamanho médio das notícias dos <i>datasets</i>	94
Tabela 10 – Sigla das entidades do spaCy	94
Tabela 11 – Siglas das classes gramaticais do spaCy	94

Lista de Abreviaturas

IA	Inteligência Artificial
PLN	Processamento de Linguagem Natural
XAI	Explainable Artificial Intelligence
LIME	Local Interpretable Model-Agnostic Explanations
IG	Integrated Gradients
RNN	Redes neurais recorrentes
RELU	Rectified Linear Unit
GPT	Generative Pre-training
BERT	Bidirectional Encoder Representations from Transformers
MLM	Masked Language Modeling
QA	Question Answering
GLUE	General Language Understanding Evaluation
NSP	Next Sentence Prediction
STS	Semantic Textual Similarity
RTE	Recognizing Textual Entailment
NER	Named Entity Recognition
KNN	K-Nearest Neighbors
CaSE	Contextual and Semantic Explanations
LDA	Latent Dirichlet Allocation
BioCIE	Biomedical Confident Itemsets Explanation
MUSE	Model Understanding through Subspace Explanations

Sumário

1	Introdução	23
1.1	Objetivos	24
1.2	Hipóteses de pesquisa	25
1.3	Organização do texto	25
2	Fundamentação Teórica	26
2.1	Processamento de Linguagem Natural	26
2.1.1	Tipos de Abordagens de PLN	27
2.2	Tipos de Análise em PLN	28
2.2.1	Tokenização	29
2.2.2	Análise Léxica	29
2.2.3	Análise Sintática	29
2.2.4	Análise Semântica	30
2.2.5	Análise Pragmática	30
2.3	Modelos de linguagem modernos	31
2.3.1	Transformer	31
2.3.1.1	Incorporação e codificação posicional	32
2.3.1.2	Codificador e Decodificador	33
2.3.1.3	Atenção multi-cabeça	34
2.3.2	Pré-treinamento e Fine-Tuning	40
2.3.3	Generative Pre-training Transformer GPT	42
2.3.4	BERT	42
2.3.5	BERTimbau	44
3	Explainable Artificial Intelligence	47
3.1	Métodos Transparentes	48
3.1.1	Regressão Linear/Logística	49
3.1.2	Árvore de Decisão	50
3.1.3	K-Nearest Neighbors	50
3.1.4	Aprendizado baseado em regra	50

3.1.5	Modelos Bayesianos	51
3.2	Métodos <i>Post-hoc</i>	51
3.3	Métodos selecionados para o experimento	52
3.3.1	LIME	52
3.3.2	Funcionamento do LIME	54
3.3.3	Mapa de Saliência	56
3.3.4	Integrated Gradients	57
3.4	Trabalhos Correlatos	60
3.4.1	Trabalho de Kiefer (2022)	60
3.4.2	Trabalho de Torcianti e Matzka (2021)	61
3.4.3	Trabalho de Moradi e Samwald (2021)	63
4	Materiais e Métodos	65
4.1	Dataset	66
4.1.1	FakeRecogna	66
4.1.2	Fake.Br Corpus	66
4.2	Divisão do conjunto de dados em treinamento e teste	67
4.3	Treinamento	68
5	Resultados	71
5.1	Teste e Análise do modelo	71
5.2	Amostragem	73
5.3	Explicações geradas pelo LIME	73
5.3.1	Modelo 1: FakeRecogna incluindo <i>stop words</i>	74
5.3.1.1	Notícia falsa classificada como falsa	75
5.3.1.2	Notícia falsa classificada como real	75
5.3.1.3	Notícia real classificada como real	76
5.3.1.4	Notícia real classificada como falsa	76
5.3.2	Modelo 2: FakeRecogna sem <i>stop words</i>	77
5.3.2.1	Notícia falsa classificada como falsa	77
5.3.2.2	Notícias falsa classificada como real	78
5.3.2.3	Notícia real classificada como real	78
5.3.2.4	Notícia real classificada como falsa	79

5.3.3	Modelo 3: Fake.Br incluindo <i>stop words</i>	79
5.3.3.1	Notícia falsa classificada como falsa	79
5.3.3.2	Notícia falsa classificada como real	80
5.3.3.3	Notícia real classificada como real	81
5.3.3.4	Notícia real classificada como falsa	81
5.3.4	Modelo 4: Fake.Br sem <i>stop words</i>	82
5.3.4.1	Notícia falsa classificada como falsa	82
5.3.4.2	Notícia falsa classificada como real	83
5.3.4.3	Notícia real classificada como real	83
5.3.4.4	Notícia real classificada como falsa	83
5.4	Explicações geradas pelo IG	84
5.4.1	Modelo 1: FakeRecogna incluindo <i>stop words</i>	84
5.4.1.1	Notícia falsa classificada como falsa	84
5.4.1.2	Notícia falsa classificada como real	86
5.4.1.3	Notícia real classificada como real	86
5.4.1.4	Notícia real classificada como falsa	87
5.4.2	Modelo 2: FakeRecogna sem <i>stop words</i>	87
5.4.2.1	Notícia falsa classificada como falsa	87
5.4.2.2	Notícia falsa classificada como real	88
5.4.2.3	Notícia real classificada como real	88
5.4.2.4	Notícia real classificada como falsa	88
5.4.3	Modelo 3: Fake.Br incluindo <i>stop words</i>	89
5.4.3.1	Notícia falsa classificada como falsa	89
5.4.3.2	Notícia falsa classificada como real	90
5.4.3.3	Notícia real classificada como real	90
5.4.3.4	Notícia real classificada como falsa	90
5.4.4	Modelo 4: Fake.Br sem <i>stop words</i>	91
5.4.4.1	Notícia falsa classificada como falsa	91
5.4.4.2	Notícia falsa classificada como real	92
5.4.4.3	Notícia real classificada como real	92
5.4.4.4	Notícia real classificada como falsa	92
5.5	Análises dos datasets	93
5.5.1	FakeRecogna incluindo <i>stop words</i>	95

5.5.2	FakeRecogna sem <i>stop words</i>	95
5.5.3	Fake.Br incluindo <i>stop words</i>	95
5.5.4	Fake.Br sem <i>stop words</i>	96
5.6	Considerações Finais	97
5.7	Limitações	101
6	Conclusões	103
	REFERÊNCIAS	106

APÊNDICES 111

APÊNDICE A – Explicações geradas pelo LIME - Modelo 1: Fake- Recogna incluindo *stop words* 112

A.1	Modelo 1 - Explicações do LIME de amostras Falsas classificadas como Falsas	112
A.2	Modelo 1 - Explicações do LIME de amostras Falsas classificadas como Reais	113
A.3	Modelo 1 - Explicações do LIME de amostras Reais classificadas como Reais .	113
A.4	Modelo 1 - Explicações do LIME de amostras Reais classificadas como Falsas	114

APÊNDICE B – Explicações geradas pelo LIME - Modelo 2: Fake- Recogna sem *stop words* 116

B.1	Modelo 2 - Explicações do LIME de amostras Falsas classificadas como Falsas	116
B.2	Modelo 2 - Explicações do LIME de amostras Falsas classificadas como Reais	117
B.3	Modelo 2 - Explicações do LIME de amostras Reais classificadas como Reais .	118
B.4	Modelo 2 - Explicações do LIME de amostras Reais classificadas como Falsas	120

APÊNDICE C – Explicações geradas pelo LIME - Modelo 3: Fake.Br Corpus incluindo *stop words* 121

C.1	Modelo 3 - Explicações do LIME de amostras Falsas classificadas como Falsas	121
C.2	Modelo 3 - Explicações do LIME de amostras Falsas classificadas como Reais	122
C.3	Modelo 3 - Explicações do LIME de amostras Reais classificadas como Reais .	123
C.4	Modelo 3 - Explicações do LIME de amostras Reais classificadas como Falsas	124

APÊNDICE D – Explicações geradas pelo LIME - Modelo 4: Fake.Br

Corpus sem *stop words* 125

D.1	Modelo 4 - Explicações do LIME de amostras Falsas classificadas como Falsas	125
D.2	Modelo 4 - Explicações do LIME de amostras Falsas classificadas como Reais	126
D.3	Modelo 4 - Explicações do LIME de amostras Reais classificadas como Reais	127
D.4	Modelo 4 - Explicações do LIME de amostras Reais classificadas como Falsas	128

APÊNDICE E – Explicações geradas pelo IG - Modelo 1: FakeRe-

cogna incluindo *stop words* 130

E.1	Modelo 1 - Explicações do IG de amostras Falsas classificadas como Falsas	130
E.2	Modelo 1 - Explicações do IG de amostras Falsas classificadas como Reais	131
E.3	Modelo 1 - Explicações do IG de amostras Reais classificadas como Reais	131
E.4	Modelo 1 - Explicações do IG de amostras Reais classificadas como Falsas	132

APÊNDICE F – Explicações geradas pelo IG - Modelo 2: FakeRe-

cogna sem *stop words* 134

F.1	Modelo 2 - Explicações do IG de amostras Falsas classificadas como Falsas	134
F.2	Modelo 2 - Explicações do IG de amostras Falsas classificadas como Reais	135
F.3	Modelo 2 - Explicações do IG de amostras Reais classificadas como Reais	136
F.4	Modelo 2 - Explicações do IG de amostras Reais classificadas como Falsas	137

APÊNDICE G – Explicações geradas pelo IG - Modelo 3: Fake.Br

Corpus incluindo *stop words* 139

G.1	Modelo 3 - Explicações do IG de amostras Falsas classificadas como Falsas	139
G.2	Modelo 3 - Explicações do IG de amostras Falsas classificadas como Reais	140
G.3	Modelo 3 - Explicações do IG de amostras Reais classificadas como Reais	141
G.4	Modelo 3 - Explicações do IG de amostras Reais classificadas como Falsas	142

APÊNDICE H – Explicações geradas pelo IG - Modelo 4: Fake.Br

Corpus sem *stop words* 143

H.1	Modelo 4 - Explicações do IG de amostras Falsas classificadas como Falsas	143
H.2	Modelo 4 - Explicações do IG de amostras Falsas classificadas como Reais	144
H.3	Modelo 4 - Explicações do IG de amostras Reais classificadas como Reais	145
H.4	Modelo 4 - Explicações do IG de amostras Reais classificadas como Falsas	146

1 Introdução

Com o avanço da inteligência artificial (IA) e a incorporação da mesma em aplicativos cada vez mais difundidos pelas pessoas, o uso de modelos de processamento de linguagem natural (PLN) e algoritmos de aprendizado de máquina, vem sendo amplamente adotados, em especial os modelos de linguagem (HAI, 2021).

Mas a transparência desses modelos, muitas vezes chamados de “caixa opaca”, levanta questões sobre sua confiabilidade e compreensão. Pois estes são capazes de fornecer respostas e resolver problemas, mas os resultados obtidos por eles não são fáceis de serem interpretados por humanos, pois os modelos não conseguem elaborar como obtiveram uma solução (SHEVSKAYA, 2021).

Antes de decidir implantar um modelo de linguagem no mundo real, é essencial confiar nas previsões individuais e na capacidade geral de um modelo. Ao contrário dos dados de teste usados para avaliar a precisão de um modelo, os dados do mundo real podem apresentar uma variedade de desafios e nuances, e garantir que o modelo possa lidar com essas diferenças é crucial. Dessa forma, torna-se importante entender os fatores que levaram o modelo a uma dada decisão (AHMED; JEON; PICCIALLI, 2022).

Além disso, em setores regulamentados, como saúde e finanças, a explicabilidade é uma exigência para evitar discriminação e práticas injustas. Portanto, a explicabilidade desempenha um papel fundamental no desenvolvimento e na aplicação de sistemas ativados por IA (AHMED; JEON; PICCIALLI, 2022).

Isso pode ser especialmente preocupante nos casos em que é necessária uma explicação, como em aplicações médicas ou jurídicas. Como por exemplo a União Europeia no dia nove de abril de 2023 chegou a um acordo sobre a regulamentação do uso de IA, um acordo do qual tem como objetivo, maximizar os benefícios potenciais da tecnologia, enquanto buscam se proteger dos riscos que ela pode proporcionar, assegurando os direitos fundamentais e possibilitando o acompanhamento do progresso das aplicações baseadas em IA, dentre as obrigações das empresas que produzem essa tecnologia está a elaboração de documento técnico sobre os conteúdos gerados e de resumos detalhados sobre o conteúdo gerado pelos sistemas (REUTERS, 2023).

Outro exemplo seria em áreas como medicina, supondo uma aplicação capaz de analisar

resultados de exames médicos e identificar a enfermidade de um paciente, o médico gostaria de entender como a IA chegou a essa decisão. No estudo de caso deste trabalho, é desejado entender porque o modelo considerou uma notícia como real ou falsa.

Portanto, entender o comportamento de um modelo é crucial para avaliar sua confiabilidade. Além disso, com um melhor entendimento do modelo, é possível obter percepções que podem ser utilizados para melhorar sua capacidade de gerar informações confiáveis.

Em resposta a esse desafio de melhorar a interpretabilidade, o campo emergente de IA explicável, do inglês *Explainable Artificial Intelligence* (XAI) foi desenvolvido para oferecer técnicas que permitam entender como o modelo está tomando suas decisões e explicar essas decisões de forma que possam ser facilmente compreendidas pelos usuários. Com essas ferramentas, é possível obter informações valiosas sobre o modelo, aumentar sua transparência e confiabilidade e permitir a sua melhoria contínua (GOHEL; SINGH; MOHANTY, 2021).

Neste trabalho são explorados os tipos de métodos de XAI existentes, com foco em dois métodos que são amplamente utilizados na literatura, o Local Interpretable Model-Agnostic Explanations (LIME) e o Integrated Gradients (IG). A fim de analisar a eficácia de um modelo BERTimbau ajustado em um conjunto de dados de notícias do português brasileiro, rotuladas como reais ou falsas, ambos os métodos LIME e IG foram aplicados sobre um conjunto selecionado de amostras a fim de compor um estudo de caso.

1.1 Objetivos

O objetivo principal desse trabalho é avaliar modelos de linguagem para classificação de texto em português para entender suas decisões e verificar se possuem limitações.

Este trabalho tem como objetivo específicos:

- Estudo e apresentação dos tipos de métodos de XAI existentes, com ênfase no LIME e Integrated Gradients.
- Estudo de caso no domínio de identificação notícias falsas em português brasileiro.
- Aplicar os métodos LIME e IG para analisar a eficácia de modelos BERTimbau.
- Investigar a importância da presença das *stop words* no domínio apresentado.
- Analisar a eficácia dos métodos LIME e IG no contexto aplicado.

1.2 Hipóteses de pesquisa

Nesta seção, são apresentadas as hipóteses que norteiam a investigação deste trabalho, focadas na classificação de notícias falsas em português e na análise das decisões dos modelos de linguagem.

Primeiramente, acredita-se que os métodos de explicabilidade analisados são eficientes em justificar as escolhas feitas por modelos caixa opaca. Além disso, acredita-se que é possível fazer uma caracterização das notícias falsas e reais dos *datasets* selecionados para este estudo de caso.

Tradicionalmente, *stop words* são removidas do texto em tarefas de processamento de linguagem natural por serem consideradas irrelevantes para a compreensão do conteúdo. No entanto, esta pesquisa propõe a hipótese de que as *stop words* podem desempenhar um papel significativo na classificação de notícias, influenciando o desempenho dos modelos de linguagem.

A última hipótese considera que notícias falsas têm uma maior tendência de mencionar indivíduos, especialmente figuras políticas, em comparação com notícias reais. Este comportamento pode ser motivado pela inclinação política dos autores e pelas fontes de onde as notícias foram originadas. A inserção de nomes de figuras públicas, principalmente políticos, pode ser utilizada para reforçar vieses ideológicos, tornando as notícias falsas mais persuasivas para determinados públicos.

1.3 Organização do texto

Este texto é organizado como segue: O Capítulo 2 apresenta a Fundamentação Teórica do trabalho onde são apresentadas modelos de linguagem clássicos e modernos, No Capítulo 3 *Explainable artificial intelligence* são apresentados os conceitos relacionados a XAI e onde são descritos os trabalhos correlatos, no Capítulo 4 Materiais e Métodos é possível ver como foram realizados os treinamentos dos modelos, no Capítulo 5 Resultados estão os resultados experimentais do estudo de caso junto a uma análise realizada sobre os *datasets*, e por fim o Capítulo 6 Conclusões traz as conclusões obtidas nos experimentos.

2 Fundamentação Teórica

Neste capítulo é apresentada uma visão sobre Processamento de Linguagem Natural e alguns conceitos que o constituem como: os níveis em que ele pode ser aplicado, as diferentes abordagens que pode se enquadrar e os tipos de análise para suas compõem suas etapas.

2.1 Processamento de Linguagem Natural

O PLN é o campo da inteligência artificial responsável pelo desenvolvimento de formas de interpretação da comunicação humana em computadores, seja utilizando áudio ou texto (DSA, 2022).

De acordo com Jurafsky e Martin (2009), o processamento de linguagem e fala é definido de forma abrangente como a capacidade de utilizar técnicas computacionais para processar a linguagem humana falada ou escrita. O que distingue as aplicações de processamento de linguagem de outros sistemas de processamento de dados é o uso do conhecimento da linguagem. Uma aplicação dessa natureza deve ser capaz de realizar tarefas como: entender humanos, o que envolve a compreensão de linguagem natural; comunicar-se com humanos por meio de geração de linguagem natural e síntese de fala; realizar recuperação de informações textuais; extrair informações e fatos importantes desses recursos textuais; e fazer inferências, que é a capacidade de obter conclusões com base nas informações disponíveis.

Os pesquisadores de PLN visam reunir conhecimentos sobre como os seres humanos entendem e usam linguagem para que ferramentas e técnicas apropriadas possam ser desenvolvidas para tornar os sistemas de computador capazes de entender e manipular linguagens naturais para executar as tarefas desejadas (GOBINDA, 2003).

Para compreender uma linguagem, a simples transmissão de palavras não é suficiente, os seres humanos dependem das habilidades no idioma e do conhecimento do discurso, portanto para o desenvolvimento de um programa que seja capaz de entender a linguagem natural, é necessário fazer a apresentação do conhecimento, como o contexto e usos (LIDDY, 2001). O algoritmo precisa ser inserido no domínio da linguagem para poder processar jargões e atribuir sentido a sentenças de palavras. Liddy (2001) define que o processamento da linguagem natural pode se dar em diferentes níveis:

- Fonológico: está relacionado a interpretação dos sons entre as palavras, este nível é composto por três regras, as regras fonéticas que lidam com os sons das palavras, as fonêmicas para variações de pronuncia e as regras prosódicas para entendimento de entonação e tonicidade.
- Morfológico: lida com os componentes (morfemas) que constituem as palavras e sua natureza. Entre esses componentes estão os prefixos, sufixos e regras de formação das palavras.
- Léxico: tem como finalidade a interpretação do sentido de uma palavra individualmente.
- Sintático: estuda a composição da frase, em que por meio de regras palavras podem ser combinadas para a formação de frases e sentenças.
- Semântico: estudo do sentido completo de palavras, frases e sentenças.
- Pragmático: utiliza conceitos extraídos do texto para realizar interpretação, buscando sentido na forma de uso da linguagem.
- Discurso: faz uma análise do texto como um todo.

Um importante aspecto a ser ressaltado é que estes níveis são convenções que são apoiadas por evidências de diferentes áreas de estudos, e cada área tem seu objetivo sendo o da computação desenvolver por meio dos conhecimentos de estruturas de dados e algoritmos, um modelo computacional capaz de trabalhar com essas convenções, gerando questões de como estruturar e modelar, textos e conhecimento.

2.1.1 Tipos de Abordagens de PLN

PLN por meio de aplicações/sistemas, como apontado por Luger (2004), pode ser utilizado para realizar diferentes tarefas, podendo assim se enquadrarem em diferentes tipos de abordagens sendo elas a simbólica, estatística, conexionista e híbrida.

- Simbólica: Essa abordagem utiliza como base as regras linguísticas livres de ambiguidade e bem estruturadas realizando uma análise profunda dos fenômenos aprendidos.
- Estatística: Emprega o uso de modelos matemáticos aproximados de fenômenos linguísticos, utilizando de grandes textos para suas deduções. Diferentemente da simbólica, essa

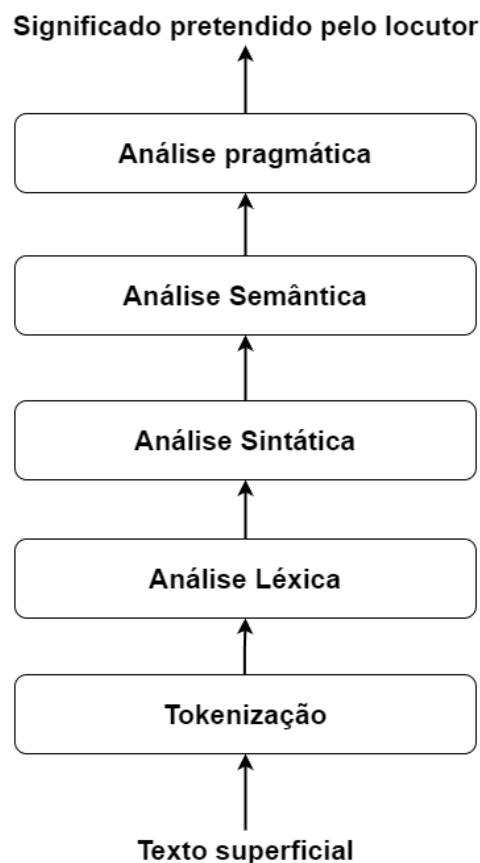
abordagem faz uso de seus próprios textos como fonte de análise, geralmente não recorrendo a conhecimento linguístico ou ao conhecimento do mundo.

- **Conexionista:** Com ela é possível transformar, inferir e manipular textos. Ao contrário da abordagem estatística, que também utiliza modelos baseados em fenômenos linguísticos, o que a diferencia é a combinação de aprendizado estatístico com teorias de representação do conhecimento.
- **Híbrida:** Por fim a abordagem híbrida mescla as abordagens citadas anteriormente, tratando problemas de PLN de forma ampla e flexível.

2.2 Tipos de Análise em PLN

Nesta seção serão abordadas as etapas clássicas de PLN, que como mostrado na Figura 1 são elas: tokenização, análise léxica, sintática, semântica e pragmática (DALE, 2010). No entanto, é importante notar que modelos mais modernos de PLN, frequentemente conexionistas ou estatísticos, tendem a suprimir algumas dessas fases, como análise sintática e semântica.

Figura 1 – Etapas da análise de PLN - Fonte: Adaptado de Dale (2010)



2.2.1 Tokenização

A primeira etapa realizada é a de tokenização ou segmentação. Em línguas como português e inglês, as palavras são separadas por espaços em branco, mas em alguns casos essa forma de separação não é suficiente, por exemplo, no caso da língua Japonesa que não apresenta separação (JURAFSKY; MARTIN, 2009) tornando assim a tarefa de tokenização mais difícil. As palavras adquiridas dessa segmentação são chamadas em linguística computacional de *tokens*.

2.2.2 Análise Léxica

A análise léxica é realizada a nível de palavra sendo assim considerada a unidade básica de análise de sentido em PLN. Uma palavra pode ser pensada de duas maneiras, como uma sequência de caracteres no texto ou como um objeto abstrato. Por exemplo, o verbo “entregar” é o termo principal para um conjunto de sequência de palavras, sendo o verbo “entregar” um objeto abstrato, ou lema, do conjunto: entrega, entregador, entregando, entregue.

A principal tarefa da análise léxica é relacionar as variantes morfológicas aos seus lemas, que nada mais são do que formas canônicas de palavras. Para isso, é necessário utilizar um dicionário de lemas no qual eles são encontrados acompanhados de suas informações semânticas e sintáticas imutáveis (INDURKHYA; DAMERAU, 2010). A análise léxica segundo Indurkhya e Damerau (2010), possui dois pontos importantes, o primeiro é o “*parsing side*” que realiza o mapeamento da palavra para seu lema, e o segundo é a geração morfológica que realiza o contrário.

2.2.3 Análise Sintática

Depois da palavra como unidade básica da análise esta a frase, podendo ser expresso através dela uma ideia, opinião sobre algo no mundo real ou imaginário. Deste modo a extração do significado é uma tarefa importante, sendo utilizado para isso a análise sintática, pois as frases não são apenas sequências de palavras, cada frase precisa ser analisada para ser obtido o seu significado sintático.

Chomsky (2009) define a sintaxe como o estudo das regras da linguagem e “os princípios pelos quais entendemos e produzimos enunciados gramaticais”. Assim, a sintaxe refere-se às partes individuais do discurso, que é a forma em que as palavras são arranjadas. Existem diferentes tipos de notações sintáticas, como ordenação de sequências de palavra, probabilidade para essas

sequências de palavras e o uso de categorias de parte do discurso como uma classe de equivalência gramatical para palavras (JURAFSKY; MARTIN, 2009).

2.2.4 Análise Semântica

Enquanto a análise sintática lida com partes individuais do discurso, a semântica lida com o sentido dessas partes dentro do contexto das frases. Compreender a comunicação humana é um processo complexo que depende dos resultados de análises anteriores (léxicas e sintáticas), bem como de informações contextuais e de senso comum.

Em linguística, a análise semântica refere-se a análise dos sentidos das palavras, expressões fixas, frases inteiras e declarações em contexto. Especificamente, isso significa traduzir as expressões originais em algum tipo de metalinguagem. Em geral, a principal evidência para a linguística semântica vem das interpretações dos falantes nativos sobre o uso de expressões em contexto, padrões de uso, localização e frequência de predicados, detectáveis por meio de técnicas linguísticas em diferentes *corpus* (INDURKHYA; DAMERAU, 2010).

Muitas palavras podem levar a múltiplas interpretações, porque as palavras podem ter mais de um sentido, ou porque elas podem ser utilizadas junto de quantificadores, métodos ou operadores de negação que podem ser aplicados de diferentes maneiras. Por exemplo, trechos do texto podem gerar ambiguidade de escopo, ou por referência intencional de pronomes ou outras expressões referenciais que podem não ser de fácil compreensão (ambiguidade referencial).

2.2.5 Análise Pragmática

A análise pragmática vai além da estrutura de uma única frase, pois para entender uma frase, a análise pragmática busca sentido em outras frases do texto ou fala. Deste modo, frases podem ser utilizadas em contextos diferentes e dependendo dele é obtida uma interpretação diferente. Um ponto a ser levantado é que uma frase pode ser pragmaticamente bem formada, no entanto não ser sintaticamente bem formada, como por exemplo a seguinte frase “Eu vou loja”, é compreensível que a intenção do locutor é ir até a loja, no entanto a frase está sintaticamente errada (ALLEN, 1988). Também pode ocorrer de uma frase ser sintaticamente correta no entanto não apresentar sentido, como por exemplo a frase “Ideias verdes dormem furiosamente”.

2.3 Modelos de linguagem modernos

Os modelos apresentados nesse capítulo se baseiam em algoritmos de aprendizado profundo, Deng, Yu et al. (2014) definem aprendizado profundo como um classe de técnicas de aprendizado de máquina que exploram muitas camadas de processamento de informação não linear. O intuito por trás disso, é realizar a extração e transformação automática de características para análise e classificação de padrões. Essa classe de algoritmos são baseadas em redes neurais, que imitam o processamento feito pelo cérebro humano, por meio de camadas de neurônios artificiais ou unidades de processamento conectadas pelos pesos sinápticos, e assim como o cérebro humano, por meio do processo de aprendizagem elas adquirem conhecimento do ambiente (HAYKIN, 2001).

Os modelos de aprendizado profundo de linguagem natural que são o alvo deste trabalho são capazes de gerar texto, traduzir idiomas, escrever diferentes tipos de conteúdo criativo e responder perguntas de forma informativa. No entanto, eles também têm uma limitação importante: são considerados “caixas opacas”. Caixa opaca é um sistema cujas entradas e saídas são conhecidas, mas cujos mecanismos e processos internos são pouco conhecidos ou desconhecidos (DAS; RAD, 2020). Isso significa que pode ser difícil entender como um modelo de linguagem considerado caixa-opaca funciona e como ele chega em determinados resultados; por isso este trabalho tem como objetivo utilizar de métodos de XAI para analisar o modelo BERTimbau.

2.3.1 Transformer

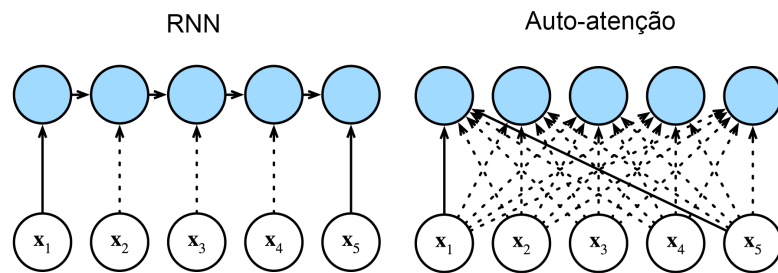
Transformer é uma arquitetura de rede neural artificial introduzida por Vaswani et al. (2017) para processamento de linguagem natural, sendo aplicada no trabalho original na tarefa de tradução automática de texto, em um momento em que as redes neurais recorrentes (RNNs) e suas variações eram as abordagens mais utilizadas.

As RNN's são projetadas para fazer reconhecimento de dados em sequências ou com variação de tempo, elas oferecem em sua estrutura conexões de *feedback* possibilitando a reentrância e a persistência de informações (DSA, 2022). Na Figura 2 está a comparação da arquitetura de um RNN e de um mecanismo de auto-atenção, ao lado esquerdo na arquitetura da RNN, cada nó azul representa um estado oculto h_t em um momento específico t , os nós brancos representam as entradas x_t em diferentes momentos, por meio das setas é possível visualizar como a informação é propagada de um estado oculto para o próximo, indicando que cada estado oculto h_t depende

não só da entrada x_t atual, mas também do estado oculto anterior $h_t - 1$, assim a rede possui uma dimensão temporal que influencia na tomada de decisão.

Vaswani et al. (2017) apontam que a característica de reentrância das RNN's impede que seja feita a paralelização dos exemplos de treinamento tornando a tarefa de trabalhar com sequências de dados de maior comprimento uma tarefa mais difícil de ser executada, devido a limitação de envio de lote pela memória. Na Figura 2 ao lado direito está a representação da arquitetura do mecanismo de auto-atenção do Transformer, que como na representação da RNN o nó azul representa um estado oculto, e pode-se observar pelas setas que diferente da RNN cada entrada x_t pode influenciar diretamente todos os estados ocultos, independente da posição contribuindo assim para que seja evitada a recorrência.

Figura 2 – Comparação das arquiteturas RNN e Auto-atenção - Fonte: Adaptado de Zhang et al. (2021)



Além do mecanismo de auto-atenção os principais componentes e características que permitem que o Transformer evite a recorrência são: camadas de atenção multi-cabeça, codificação posicional e as camadas empilhadas em sua arquitetura. Estas possibilitam mais paralelização, e um menor custo computacional, e elas serão melhor descritas no decorrer do texto.

2.3.1.1 Incorporação e codificação posicional

O primeiro passo do modelo é transformar as palavras de entrada em matrizes chamadas de *embeddings*, esse processo se chama incorporação. No caso do Transformer, a palavra é incorporada a um vetor de tamanho 512. Na Figura 3 há um exemplo de matriz de *embeddings* da frase “Hello good morning”. A incorporação ocorre apenas no codificador mais inferior, os outros codificadores recebem como entrada a saída do codificador anterior a ele.

Na segunda etapa é injetado nas *embeddings* de entrada uma codificação posicional, que contém informação da posição dos tokens na sequência, para que o Transformer possa seguir essa ordem. Nessa codificação a matriz recebe uma perturbação de modo que, para cada índice par é criado um vetor utilizando a função seno representada na Equação 2.1, e para índices ímpares é

Figura 3 – Representação da matriz de incorporação da frase “*Hello good morning*” - Fonte: Adaptado de Alammr (2018)

$$\begin{array}{ccc}
 \text{Hello} & \begin{bmatrix} 1,76 & 2,22 & \dots & 6,66 \end{bmatrix} & X_1 \\
 \text{good} & \begin{bmatrix} 8,45 & 0,45 & \dots & 4,85 \end{bmatrix} & X_2 \\
 \text{morning} & \begin{bmatrix} 11,72 & 10,10 & \dots & 6,97 \end{bmatrix} & X_3 \\
 & & \text{3x512}
 \end{array}$$

utilizada a função cosseno representada na Equação 2.2, em que i representa a coluna da matriz, d o total de colunas, e pos representa a linha da matriz, no caso o vetor de *embedding* de uma palavra.

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \quad (2.1)$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \quad (2.2)$$

Depois de incorporar as palavras em matrizes, elas fluem para as subcamadas que constituem o codificador que estão representadas na Figura 4.

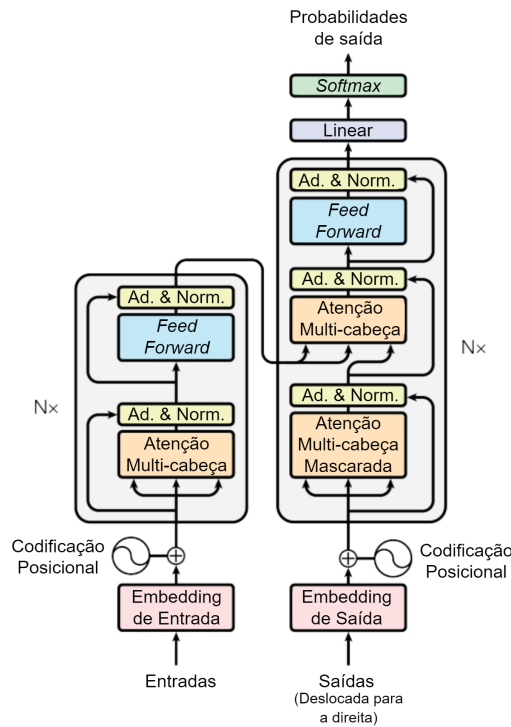
2.3.1.2 Codificador e Decodificador

O modelo Transformer proposto por Vaswani et al. (2017) é composto por duas estruturas principais, o codificador e decodificador como mostrado na Figura 4 na parte esquerda e direita, respectivamente.

Afim de exemplificar, na Figura 5, esta um exemplo do Transformer sendo utilizado para tradução automática de texto, onde temos como entrada uma frase em inglês que passa por esses dois principais componentes o codificador e o decodificador e tem como saída a frase traduzida em português.

Como exibido pela Figura 4 o codificador é composto por uma pilha de N camadas idênticas, com cada camada possuindo 2 subcamadas. Na primeira subcamada se encontra um mecanismo de atenção multi-cabeça e na segunda uma rede *feed-forward* totalmente conectada e aplicada a cada posição, estas subcamadas do modelo e as camadas de *embedding* apresentam saídas com dimensão $d_{model} = 512$. Na Figura 4, é possível visualizar uma seta que representa uma conexão residual em torno de cada uma das subcamadas que é seguida por uma normalização.

Figura 4 – Modelo de arquitetura Transformer - Fonte: Traduzido de Vaswani et al. (2017)



Assim como o codificador, o decodificador também é composto por uma pilha de N camadas idênticas, e além das duas subcamadas que também estão presentes no codificador, o decodificador adiciona uma terceira subcamada de atenção multi-cabeça que opera sobre a saída da pilha de codificador.

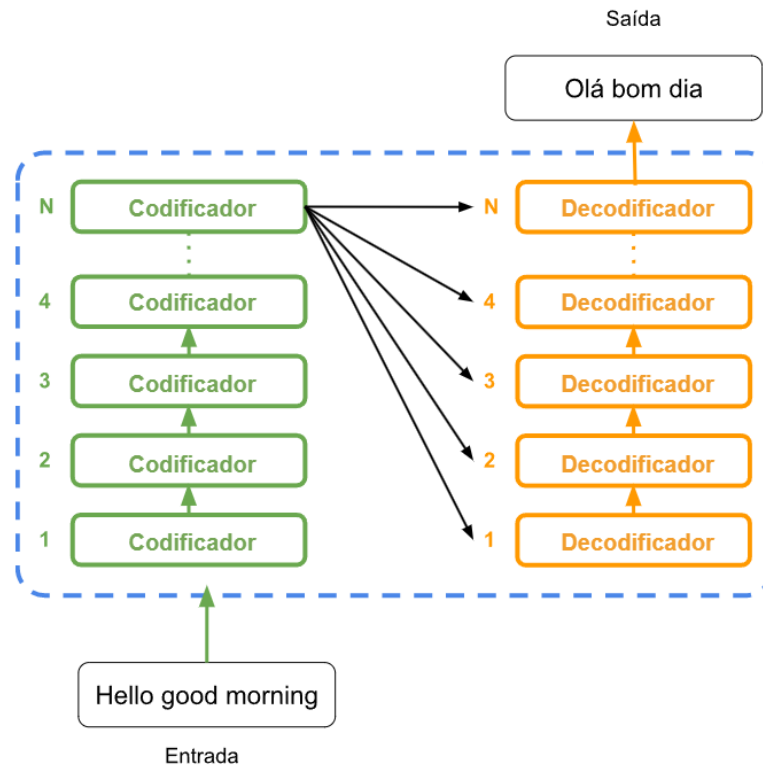
E também no decodificador a sub-camada de auto-atenção é modificada para que cada posição só considere posições anteriores e nunca futuras, sendo representada na Figura 4 como a subcamada de atenção multi-cabeça mascarada. Esse mascaramento, junto com o deslocamento das *embeddings* de saída em uma posição, permite que as previsões feitas para a posição i dependam apenas das saídas conhecidas nas posições anteriores a i .

2.3.1.3 Atenção multi-cabeça

Depois de incorporar as palavras em matrizes, elas fluem para as subcamadas que constituem o codificador que estão representadas na Figura 4. A camada de atenção multi-cabeça, permite que o modelo relacione cada palavra de entrada as outras palavras da frase. A saída gerada nesta camada é utilizada para alimentar a segunda sub-camada do codificador, uma rede neural *feed-forward*.

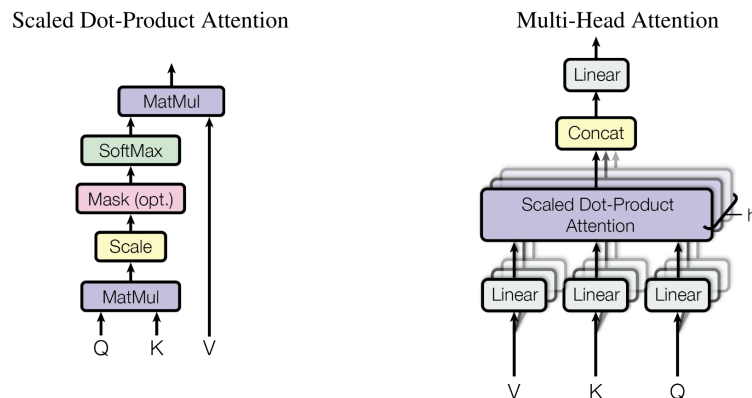
Na camada de atenção multi-cabeça, o primeiro passo para realizar o cálculo da atenção

Figura 5 – Diagrama da estrutura de um Transformer constituído de uma pilha de codificadores e decodificadores - Fonte: Adaptado de Alammr (2018)



é criar três vetores para cada uma das *embeddings* de entrada, sendo denominados pesquisa (Q – *Query*), chave (K – *Key*) e (V – *Value*). Esses vetores são criados na camada linear representada na Figura 6, multiplicando a *embedding* X por três matrizes de peso W_q , W_k e W_v como exibido na Figura 7 que são inicializadas aleatoriamente e os valores melhores são obtidos durante o processo de treinamento.

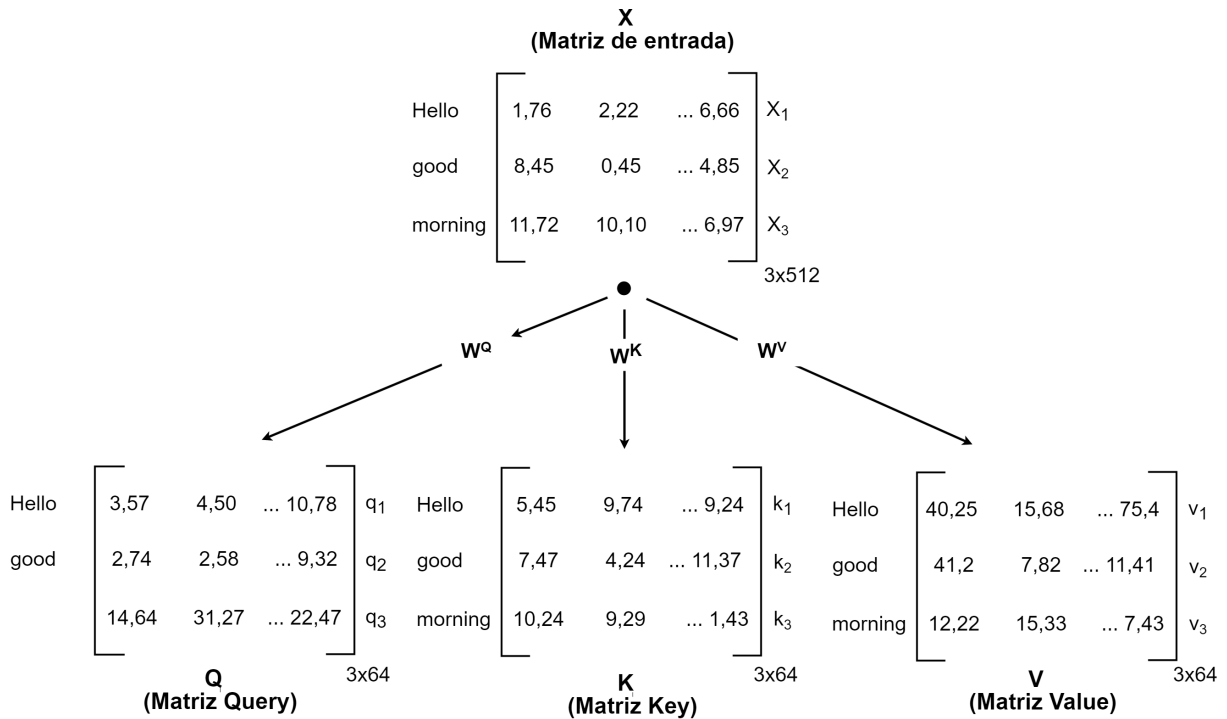
Figura 6 – Atenção multi-cabeça e *scaled dot-product attention* - Fonte: Traduzido de Vaswani et al. (2017)



A camada de atenção multi-cabeça recebe esse nome pois ao em vez de realizar a função de atenção uma única vez com os vetores Q , K e V de dimensão d_{model} , eles são projetados

linearmente h vezes, com diferentes projeções lineares aprendidas para dimensões d_q , d_k e d_v , no caso do modelo Transformer base, são projetadas 8 vezes pois são utilizadas 8 cabeças de atenção paralelas, sendo assim para da cabeça de atenção $d_q = d_k = d_v = d_{model}/h = 64$.

Figura 7 – Geração dos vetores V, Q e K a partir das matrizes de peso - Fonte: Adaptado de Alammar (2018)



Na Figura 7, é possível observar o resultado da operação descrita anteriormente, em que as novas matrizes são menores em dimensão do que a matriz de entrada. As matrizes novas tem um dimensionalidade de 64, enquanto as de entrada/saída de incorporação e codificador têm dimensionalidade de 512 quando utilizada 8 cabeças de atenção.

O segundo passo no cálculo da auto-atenção, como mostrado ao lado esquerdo da Figura 6 é o calculo do *scaled dot-product attention* que determina quanto foco deve ser colocado em uma dada palavra de uma frase de entrada. Para isso é preciso pontuar cada uma das outras palavras que compõem a sentença de entrada contra essa palavra. No exemplo da Figura 8, está sendo calculada a atenção da frase “Hello good morning”.

A pontuação é calculada por meio do produto escalar da matriz Q com a matriz K . Como demonstrado pela Figura 8, a matriz Q é multiplicada pela matriz K transposta representado por K^T . Em seguida para diminuir a variância nas ativações dos neurônios, as pontuações da matriz são divididas por $\sqrt{d_k}$ em que d_k é a dimensão dos vetores K . No caso do trabalho de Vaswani

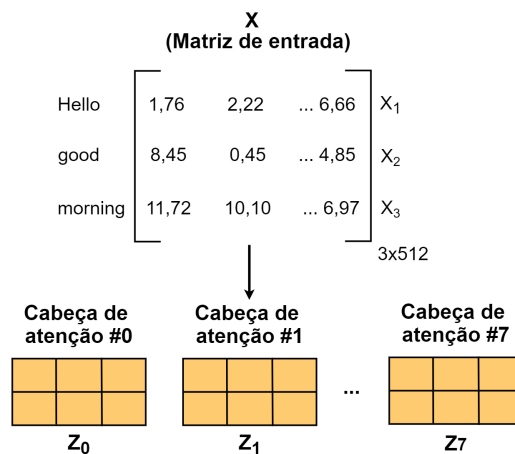
Figura 8 – Cálculo de atenção da frase “Hello good morning” - Fonte: Adaptado de Alammar (2018)

$$\begin{array}{c}
 \begin{array}{c} \text{Hello} \\ \text{good} \\ \text{morning} \end{array} \begin{bmatrix} 3,57 & 4,50 & \dots & 10,78 \\ 2,74 & 2,58 & \dots & 9,32 \\ 14,64 & 31,27 & \dots & 22,47 \end{bmatrix} \begin{array}{c} q_1 \\ q_2 \\ q_3 \end{array} \\
 \text{3x64} \\
 \mathbf{Q} \\
 \text{(Matriz Query)}
 \end{array}
 \bullet
 \begin{array}{c}
 \begin{array}{c} \text{Hello} & \text{good} & \text{morning} \end{array} \\
 \begin{bmatrix} 5,45 & 7,47 & 10,24 \\ 9,74 & 4,24 & 9,29 \\ \vdots & \vdots & \vdots \\ 9,24 & 11,37 & 1,43 \end{bmatrix} \\
 \begin{array}{c} k_1 & k_2 & k_3 \end{array} \\
 \text{3x64} \\
 \mathbf{K}^T \\
 \text{(Matriz Key)}
 \end{array}
 =
 \begin{array}{c}
 \begin{array}{c} \text{Hello} & \text{good} & \text{morning} \end{array} \\
 \begin{bmatrix} q_1.k_1 & q_1.k_2 & q_1.k_3 \\ q_2.k_1 & q_2.k_2 & q_2.k_3 \\ q_3.k_1 & q_3.k_2 & q_3.k_3 \end{bmatrix} \\
 \mathbf{QK}^T
 \end{array}
 \end{array}$$

et al. (2017) a versão base que apresenta dimensão 64, as pontuações são divididas por 8. Essa operação leva a gradientes mais estáveis.

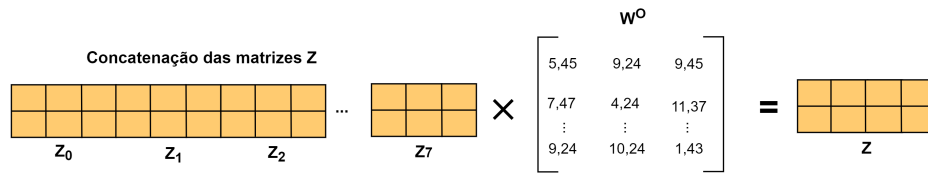
Em seguida, pode ser passada uma máscara, sendo esse um passo opcional, e após esse passo, as pontuações são normalizadas por uma operação de softmax para que sejam todas positivas e somem 1. Por fim, a matriz resultante é multiplicada pela matriz V com o intuito de manter intactos os valores de palavras que devem ser focadas e abafar palavras irrelevantes.

Levando em conta que são utilizadas h cabeças de atenção diferentes, como representado na Figura 6, isso significa que o cálculo do *scaled dot-product* será realizado h vezes, gerando h matrizes diferentes denominadas Z , como mostrado na Figura 9.

Figura 9 – Geração das matrizes Z pelas 8 cabeças de atenção - Fonte: Adaptado de Alammar (2018)

Como a camada *Feed-Forward* espera uma única entrada, é preciso concatenar as matrizes Z como exibido na Figura 6 e na Figura 10, em seguida o resultado dessa concatenação passa por uma camada linear, onde é multiplicado por uma matriz de pesos adicional W^O .

Figura 10 – Concatenação das matrizes Z e multiplicação da matriz resultante pela matriz de pesos W^O -
 Fonte: Adaptado de Alammr (2018)



Como mostrado na Figura 4, os resultados vindos da camada de atenção multi-cabeça são adicionados à uma conexão residual vinda da camada de entrada. Essa conexão ajuda os gradientes a seguirem pela rede, após isso é feita uma normalização, que evita de que os valores em cada camada sejam fortemente alterados, possibilitando um treinamento mais rápido.

Em seguida os dados seguem para a camada da rede *feed-forward*, que é composta por duas camadas densas que fazem uso da função de ativação Rectified Linear Unit (ReLU) (JARRETT et al., 2009). Enfim a matriz gerada pelo codificador é utilizada para alimentar o decodificador, que se assemelha ao codificador, possuindo uma pilha de N camadas idênticas, como mostrado pela Figura 4, desse modo o decodificador recebe duas entradas, uma representação gerada pelo codificador e uma do decodificador da etapa de tempo anterior.

Além das subcamadas que estão presentes no codificador, o decodificador adiciona uma subcamada que aplica o cálculo de atenção sobre a saída do codificador. Os decodificadores geram como saída um vetor de números decimais que passa pela camada Linear que é uma rede neural simples totalmente conectada e em seguida por uma camada que aplica a função de softmax que transforma as pontuações em probabilidades.

O modelo proposto por Vaswani et al. (2017) foi avaliado utilizando o *dataset* WMT 2014 English-German padrão (BOJAR et al., 2014), que contém aproximadamente 4,5 milhões de pares de frases, e também foi utilizado o *dataset* WMT 2014 English-French com 36 milhões de frases. A máquina utilizada pelos autores contém 8 NVIDIA P100 GPUs, cada etapa do treinamento levou cerca de 0,4 segundos, o modelo básico implementado foi treinado durante 100.000 etapas o que corresponde a 12 horas de treinamento e para os modelos grandes o tempo gasto por etapa foi de 1,0 segundo e estes foram treinados por 300.000 etapas cerca de 3,5 dias.

Os autores treinaram o modelo com diferentes variações de arquitetura estando todas listadas na Tabela 1, foram testados diferentes números de cabeças de atenção representado na Tabela por h . Na tabela é possível ver também a pontuação alcançada na avaliação feita pelo algoritmo BLEU (PAPINENI et al., 2002).

Tabela 1 – Variações da arquitetura do Transformer utilizadas pelos autores - Fonte: Vaswani et al. (2017)

	N	dmodelo	dff	h	dk	dv	Pdrop	Els	train steps	PPL (dev)	BLEU (dev)	parâmetros 6 ×10
base	6	512	2048	8	64	64	0,1	0,1	100K	4,92	25,8	65
(A)				1	512	512				5,29	24,9	
				4	128	128				5,00	25,5	
				16	32	32				4,91	25,8	
				32	16	16				5,01	25,4	
(B)					16					5,16	25,1	58
					32					5,01	25,4	60
(C)	2									6,11	23,7	36
	4									5,19	25,3	50
	8									4,88	25,5	80
		256			32	32				5,75	24,5	28
		1024			128	128				4,66	26,0	168
			1024							5,12	25,4	53
			4096							4,75	26,2	90
(D)							0,0			5,77	24,6	
							0,2	0,0		4,95	25,5	
								0,2		4,67	25,3	
										5,47	25,7	
(E)		incorporação posicional em vez de sinusóides								4,92	25,7	
big	6	1024	4096	16			0,3		300K	4,33	26,4	213

Na Tabela 2 estão os resultados levantados pelo Transformer em comparação à outros modelos do estado da arte, é possível ver que na tarefa de tradução de inglês para alemão do WMT 2014, a variante *big* do modelo Transformer, atingiu uma melhor pontuação BLEU em relação ao estado da arte, uma melhora de mais de 2,0 pontos na pontuação BLEU, atingindo 28,4 pontos, quanto a tarefa de tradução de inglês para francês o mesmo modelo atingiu uma pontuação de 41,8 consumindo $\frac{1}{4}$ do custo do modelo anterior do estado da arte.

Tabela 2 – Resultados do Transformer em comparação a outros modelos do estado da arte - Fonte: Vaswani et al. (2017)

Modelo	BLEU		Custo Treinamento (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet	23,75			
Deep-Att + PosUnk		39,2		$1,0 * 10^{20}$
GNMT + RL	24,6	39,92	$2,3 * 10^{19}$	$1,4 * 10^{20}$
ConvS2S	25,16	40,46	$9,6 * 10^{18}$	$1,5 * 10^{20}$
MoE	26,03	40,56	$2,0 * 10^{19}$	$1,2 * 10^{20}$
Deep-Att + PosUnk Ensemble		40,4		$8,0 * 10^{20}$
GNMT + RL Ensemble	26,30	41,16	$1,8 * 10^{20}$	$1,1 * 10^{21}$
ConvS2S Ensemble	26,36	41,29	$7,7 * 10^{19}$	$1,2 * 10^{21}$
Transformer (base model)	27,3	38,1	$3,3 * 10^{18}$	
Transformer (big)	28,4	41,8	$2,3 * 10^{19}$	

2.3.2 Pré-treinamento e Fine-Tuning

Treinar um modelo de linguagem, é uma tarefa laboriosa. Requer a preparação de uma base de dados e o treinamento demanda muito tempo, além disso, realizar o treinamento de um modelo apresenta um alto custo computacional, em vista disso, surgem os modelos pré-treinados.

Estes modelos comumente são pré-treinados em um grande conjunto de dados não rotulados. Eles aprendem uma representação de linguagem e podem posteriormente, serem reutilizados para construir soluções específicas capazes de atingir bons resultados, de forma mais fácil, sem a necessidade de elaboração de um conjunto de treinamento muito robusto.

Tendo sido apresentado ao conceito de pré-treinamento, para que um modelo pré-treinado, seja aplicado a um escopo específico diferente da base de dados do pré-treinamento se faz necessária a realização do *fine-tuning* (DEVLIN et al., 2018).

Enquanto o pré-treinamento é uma atividade mais custosa em termos de recursos computacionais, o *fine-tuning* é uma tarefa mais simples. O pré-treinamento, mesmo realizado em dados de escopos diferentes da atividade do *fine-tuning*, é capaz de apresentar resultados satisfatórios. Desse modo, o *fine-tuning* pode ser visto como um mecanismo para adequar os dados de treinamento com a população de interesse, essa adequação é feita pelo *fine-tuning* utilizando dados rotulados para a realização de uma tarefa específica.

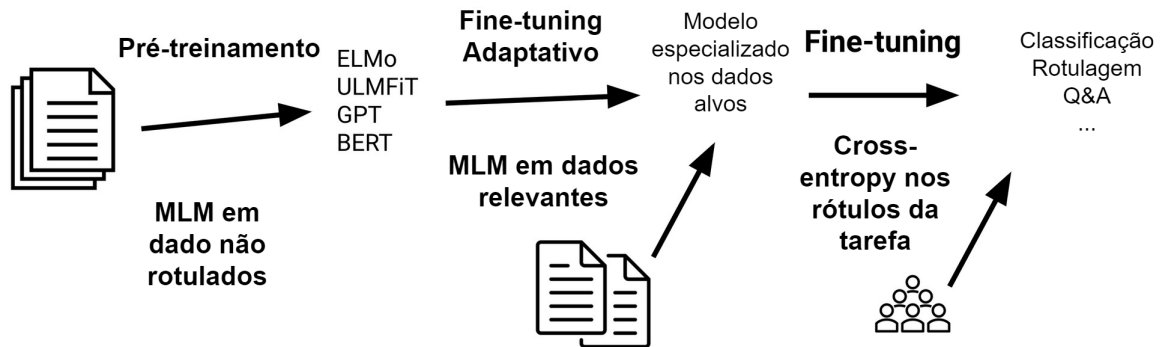
Ruder (2021) apresentou uma detalhada separação dos métodos de *fine-tuning* em diferentes categorias, apresentadas à seguir:

- **Fine-Tuning Adaptativo**

O *fine-tuning* adaptativo é uma metodologia para lidar com o problema que os modelos pré-treinados enfrentam ao trabalhar com dados que pertençam a um contexto que seja muito diferente aos que ele foi apresentado na etapa de pré-treinamento. Com esse intuito, antes de realizar o *fine-tuning* específico da tarefa, é realizado o *fine-tuning* em dados adicionais como representado na Figura 11, que podem se concentrar em realizar o ajuste adaptativo ao domínio ou a tarefa. No primeiro são utilizados dados não rotulados que se aproximam dos dados de treinamento da tarefa de *downstream*, enquanto no segundo, os dados também não rotulados utilizados são uma versão dos dados utilizados no treinamento da tarefa que tiveram seus rótulos removidos.

Ao realizar este tipo de *fine-tuning*, o modelo é capaz de atingir resultados relevantes para

Figura 11 – Fine Tuning Adaptativo - Fonte: Traduzido de Ruder (2021)

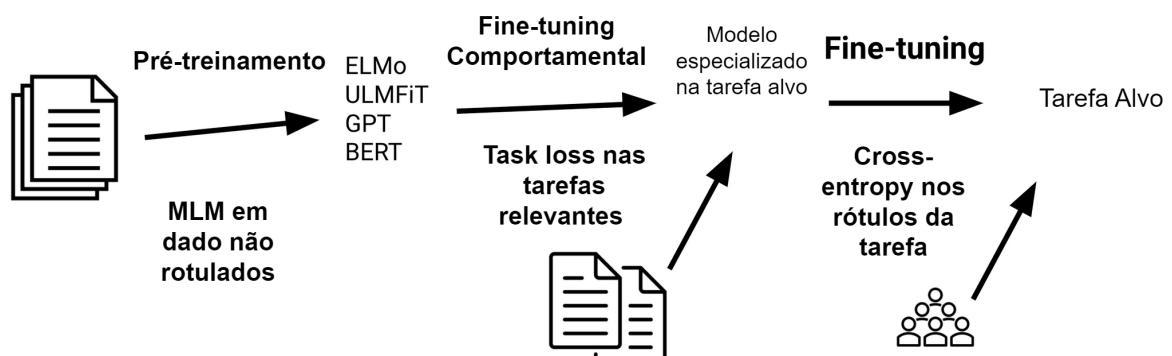


uma distribuição de dados em que se especializou. Contudo, essa característica pode ser custosa para a sua capacidade como um modelo geral de linguagem. Portanto, este tipo de *fine-tuning* é uma boa solução para quando o objetivo é alcançar um desempenho maior em tarefas de domínio único. Este tipo de ajuste foi utilizado em trabalhos como: Logeswaran et al. (2019); Han e Eisenstein (2019) e Mehri et al. (2019).

• Fine-tuning Comportamental

O *fine-tuning* comportamental se concentra em ensinar ao modelo comportamentos que sejam úteis para a realização da tarefa alvo, utilizando dados rotulados relevantes para a tarefa. Uma das maneiras de se fazer isso é realizando o treinamento de tarefas intermediárias, como exibido na Figura 12, em que o modelo é ajustado utilizando dados rotulados que sejam relevantes para estas tarefas antes de realizar o *fine-tuning* na tarefa em específico como feito por Phang, Févry e Bowman (2018). Esse tipo de abordagem é preferível para tarefas que exijam alto nível de inferência. Outra maneira de realizar esse tipo de *fine-tuning* é pelo enquadramento da tarefa de objetivo final, como um tipo de modelagem mascarada como realizado por por Sellam, Das e Parikh (2020).

Figura 12 – Fine Tuning Comportamental - Fonte: Traduzido de Ruder (2021)



- **Fine-tuning com eficiência de parâmetros**

Este tipo de *fine-tuning* permite salvar pequenas alterações para atender tarefas em específico, este *fine-tuning* é feito de modo em que a maioria dos parâmetros são mantidos fixos no modelo sendo possível alterar uma pequena quantidade de parâmetros para cada tarefa, mantendo assim uma única cópia do modelo e as modificações para cada tarefa em diferentes pequenos arquivos. Esta metodologia contribui em situações em que um modelo precisa ser ajustado em diferentes configurações, pois fazer o armazenamento de diversas cópias do modelo com os ajustes para cada tarefa é custoso.

- **Fine-tuning texto para texto**

Este método de *fine-tuning* é utilizado para desenvolvimento de modelos de transferência de aprendizado em PLN, transicionando modelos de linguagem mascarada para modelos auto-regressivos. O *fine-tuning* é importante para adaptar esses modelos a novas tarefas, e a forma mais eficaz de usá-los é ajustar alguns ou todos os parâmetros. Embora o aprendizado sem supervisão seja possível, o *fine-tuning* é crucial para a capacidade de adaptação.

2.3.3 Generative Pre-training Transformer GPT

O modelo Generative Pre-training Transformer (GPT) (RADFORD et al., 2018) foi o primeiro modelo de linguagem pré-treinado desenvolvido baseado em Transformer, sendo capaz de manipular a semântica das palavras dentro de contexto. O modelo foi proposto com a necessidade de aliviar a dependência presente no processamento de linguagem natural com aprendizado supervisionado, pois a maioria dos métodos de aprendizado profundo do estado da arte anteriores necessitavam de grandes quantidades de dados já rotulados manualmente.

O GPT usa o decodificador do Transformer localizado ao lado direito da Figura 4 para poder modelar a linguagem, pois é um modelo auto-regressivo em que o modelo prevê a próxima palavra de acordo com seu contexto anterior, ou seja, é um modelo unidirecional, treinado de modo que apenas possa prever o futuro contexto da esquerda para a direita.

2.3.4 BERT

O modelo de representação de linguagem Bidirectional Encoder Representations from Transformers (BERT) introduzido por Devlin et al. (2018), foi projetado para pré-treinar representações bidirecionais profundas, enquanto as técnicas apresentadas até momento em que foi

publicado são modelos unidirecionais, que restringem o poder de pré-processamento dos modelos e abordagens de realização de *fine-tuning*. O modelo é capaz de lidar com diferentes tarefas, como por exemplo classificação e Question-Answering (QA) sem modificações substanciais em sua arquitetura.

O BERT é um codificador de Transformer (lado esquerdo da Figura da Figura 4) bidirecional multi-camada, baseado no Transformer original de Vaswani et al. (2017). Esse modelo passa por dois processos principais, o pré-treinamento e o fine-tuning, no primeiro o modelo é treinado com dados não rotulados em duas diferentes tarefas Masked-Language Modeling (MLM) e Next Sentence Prediction (NSP). E no fine-tuning o modelo é inicializado com os parâmetros gerados na etapa de pré-treinamento, que são ajustados utilizando um conjunto dados rotulados.

Os autores descreveram os resultados de dois diferentes tamanhos de modelos denominados BERTbase e BERTlarge, em que variam as seguintes parâmetros, L que representa o número de camadas ou blocos de Transformer que compõem o modelo, H a quantidade de neurônios na camada oculta na rede *feed forward* e A o número de cabeças de auto-atenção. O BERTbase possui as seguintes configurações $L = 12$, $H = 768$, $A = 12$ e este modelo foi treinado com 220M de parâmetros. Enquanto o BERTlarge foi treinado com um total de 340M parâmetros, com $L = 24$, $H = 1024$ e $A = 16$.

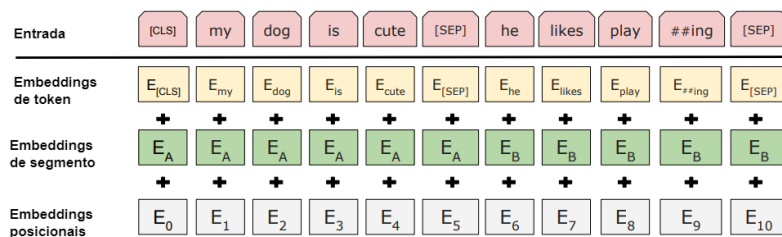
Como dito anteriormente o modelo BERT é pré-treinado em duas tarefas, MLM e NSP. Na primeira tarefa, para realizar a treinamento de uma representação bidirecional profunda, uma porcentagem dos tokens de entrada são mascarados de modo aleatório, enquanto na segunda tarefa o treinamento é utilizada para resolver problemas como Question Answering (QA) e Inferência de Linguagem Natural, que se baseiam na relação entre duas sentenças que não é obtida pela modelagem de linguagem tradicional, para isso o modelo é pre treinado em uma tarefa de previsão de próxima sentença.

O *fine-tuning* no BERT tem como diferencial a simplicidade ao trabalhar com pares de texto, pois diferentes de outros modelos que normalmente codificam individualmente esses pares, para depois aplicar atenção bidirecional cruzada como Parikh et al. (2016); Seo et al. (2017), o BERT unifica essas duas etapas por meio do mecanismo de atenção, isso é feito conectando as entradas e saídas do modelo e ajustando seus parâmetros.

A representação de uma única sentença ou de duas sentenças (como no caso da tarefa de QA) é feita numa única sequência de tokens, quando há duas sentenças elas são agrupadas

em um única. Estes tokens são gerados utilizando as Embedding WordPiece de Wu et al. (2016) que contam com um vocabulário de 30.000 tokens, na qual sempre o primeiro token adicionado à esquerda dessa entrada é um símbolo especial [CLS] e no caso de duas sentenças elas são separadas por um token especial denominado [SEP], como mostrado na Figura 13 em que esta representada duas frases, a primeira “my dog is cute” tendo no início da frase o token [CLS] e a segunda frase “he likes playing”, que é separada e finalizada com o token [SEP].

Figura 13 – Representação de um entrada do BERT - Fonte: Traduzido de Devlin et al. (2018)



Os autores utilizaram General Language Understanding Evaluation (GLUE) (WANG et al., 2018) *benchmark* para realizar testes. Na Tabela 3, estão os resultados obtidos, é possível observar que ambos BERTbase e BERTlarge, superaram os resultados dos outros algoritmos, em todas as tarefas, tendo um aumento na média de acurácia em relação aos modelos anteriores do estado da arte de 4,5 e 7,0 respectivamente.

Tabela 3 – Resultados do modelo BERT no GLUE Test - Fonte: Devlin et al. (2018)

Modelo	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80,6/80,1	66,1	82,3	93,2	35,0	81,0	86,0	61,7	74,0
BiLSTM+ELMo+Attn	80,6/80,2	64,8	79,8	90,4	36,0	73,3	84,9	56,8	71,0
OpenAI GPT	80,6/80,3	70,3	87,4	91,3	45,4	80,0	82,3	56,0	75,1
BERTBASE	80,6/80,4	71,2	90,5	93,5	52,1	85,8	88,9	66,4	79,6
BERTLARGE	80,6/80,5	72,1	92,7	94,9	60,5	86,5	89,3	70,1	82,1

Esses modelos foram avaliados na tarefa de classificação de similaridade de linguagem cruzada, utilizando o *dataset* XNLI (CONNEAU et al., 2018), sendo ele uma versão do MultiNLI (WILLIAMS; NANGIA; BOWMAN, 2018) em que seus conjuntos de dados foram traduzidos de forma manual em 15 idiomas sendo o conjunto de treinamento traduzido de forma automática.

2.3.5 BERTimbau

Como apresentado anteriormente, os robustos modelos pré-treinados são recursos que auxiliam muito nas atividades de PLN. Em vista disso Souza, Nogueira e Lotufo (2020) desenvolveram um modelo baseado em BERT com dados em português brasileiro, chamado BERTimbau,

sendo este o modelo aplicado no estudo de caso deste trabalho. O *corpus* utilizado para o pré-treinamento do modelo é o brWaC (FILHO et al., 2018), que contém dados que foram retirados de páginas da web, sendo um total de 2,68 bilhões de tokens de 3,53 milhões de documentos. A metodologia de criação desse *corpus* é repensável por fornecer um conjunto de dados com uma grande diversidade de domínios.

O BERTimbau utiliza da mesma arquitetura e procedimentos de pré-treinamento do BERT original tendo poucas mudanças, sendo treinados dois modelos em tamanhos diferentes o Base (12 camadas, 768 neurônios na camada oculta, 12 cabeças de atenção e 110M parâmetros) e o Large (24 camadas, 1024 neurônios ocultos, 16 cabeças atenção e 330M parâmetros). Em ambos os modelos o comprimento máximo de sentença é definido como 512 tokens.

Assim como no BERT, os modelos BERTimbau foram pré-treinados nas tarefas de MLM e NSP, e para realizar o uso deles em tarefas de *downstream* diferentes, são removidas as cabeças de classificação MLM e NSP e são anexadas cabeças relevantes para cada tarefa.

Os autores avaliaram os modelos em três tarefas diferentes de PLN, Sentence Textual Similarity (STS), Recognizing Textual Entailment (RTE) e Named Entity Recognition (NER).

Tabela 4 – BERTimbau resultados STS e RTE - Fonte: Traduzido de Souza, Nogueira e Lotufo (2020)

Índice	Modelo	STS		RTE	
		Pearson(*)	MSE	F1(*)	Acurácia
1	mBERT+Features[8]	0,817	0,47	86,6	86,6
2	mBERT(ours)	0,809	0,58	86,8	86,8
3	BERTimbauBase	0,836	0,58	89,2	89,2
4	BERTimbauLarge	0,852	0,50	90,0	90,0

Para a avaliação das tarefas STS e RTE foi utilizado o conjunto de dados ASSIN2 (RODRIGUES et al., 2020). Os resultados apurados foram comparados com os de outros modelos como o mBERT e podem ser visualizados na Tabela 4. Por meio da tabela, é possível notar que o modelo BERTimbauLarge, foi o que obteve as melhores métricas.

Para a avaliação na tarefa de NER, foi utilizado o *dataset* MiniHAREM (SANTOS et al., 2006), na Tabela 5 podem ser observados os resultados. E assim como nas tarefas anteriores, o modelo BERTimbauLarge juntamente do modelo BERTimbauLarge+CRF alcançaram o melhor desempenho.

Tabela 5 – BERTimbau resultados NER - Fonte: Traduzido de Souza, Nogueira e Lotufo (2020)

Índice	Arquitetura	Cenário Total			Cenário seletivo		
		Prec.	Rec.	F1(*)	Prec.	Rec.	F1(*)
1	mBERT	71,6	72,7	72,2	77,0	78,8	77,9
2	mBERT+CRF	74,1	72,2	73,1	80,1	78,3	79,2
3	BERTimbauBase	76,8	77,1	77,2	81,9	82,7	82,2
4	BERTimbauBase+CRF	78,5	76,8	77,6	84,6	81,6	83,1
5	BERTimbauLarge	77,9	78,0	77,9	81,3	82,2	81,7
6	BERTimbauLarge+CRF	79,6	77,4	78,5	84,9	82,5	83,7

3 Explainable Artificial Intelligence

A *Explainable Artificial Intelligence*(XAI) é uma área emergente de estudo dentro do campo da IA, tendo como propósito esclarecer, o processo pelo qual um modelo de IA chegou a uma determinada solução em atividades como classificação de texto e reconhecimento de objetos (GOHEL; SINGH; MOHANTY, 2021). Esse fator de explicabilidade é comumente ausente nos algoritmos de IA de abordagem preditiva, que como já dito anteriormente são considerados como “caixas opacas” (SHEVSKAYA, 2021).

Esses modelos são capazes de trazer resultados em tarefas mas não de elaborar como essa resposta foi obtida. Não é possível decompor os processos realizados por eles, em partes ou sub-tarefas simples e compreensíveis para os humanos. Em vista disso, o principal desafio para a XAI se torna desenvolver abordagens que possam mostrar ao ser humano de forma compreensível porque um modelo alcança determinado resultado.

Os algoritmos modernos de explicabilidade são locais, deste modo eles revelam quais características ou aspectos da entrada foram mais influentes na decisão específica do modelo para um exemplo particular. Eles se concentram na interpretação detalhada das decisões individuais, em vez de buscar entender padrões gerais ou regras aplicáveis a todo o modelo ou a uma ampla classe de dados.

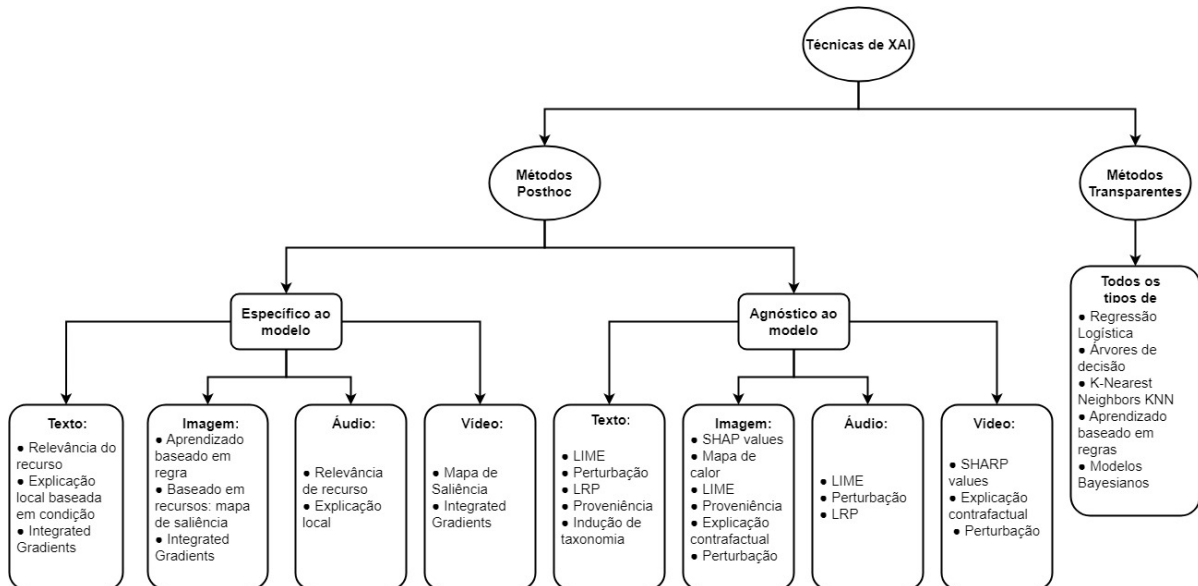
XAI portanto, tem atraído um grande interesse em áreas de aplicação críticas, como aplicações nas áreas de medicina, direito, entre outras. Como já mencionado, em 9 de abril de 2023, a União Europeia estabeleceu um acordo para regulamentar o uso de IA. Entre as obrigações impostas às empresas que desenvolvem essa tecnologia está a criação de documentos técnicos e resumos detalhados sobre o conteúdo gerado pelos sistemas (REUTERS, 2023).

O ato de explicar como as respostas são obtidas é tão relevante quanto obter as próprias respostas. Em ambos os cenários acadêmico e industrial, a pesquisa em XAI tem se tornado uma prioridade (GOHEL; SINGH; MOHANTY, 2021).

As técnicas de XAI, como exibido na Figura 14, podem ser classificadas em duas categorias, métodos transparentes e *post-hoc* (GOHEL; SINGH; MOHANTY, 2021). Os métodos transparentes são aqueles em que o funcionamento interno e o processo de tomada de decisão do modelo são fáceis de entender e representar, sendo métodos aplicáveis em situações em que as

relações internas entre os recursos não são muito complexas ou são lineares. Alguns exemplos de modelos transparentes incluem modelos bayesianos, árvores de decisão e modelos de regressão linear.

Figura 14 – Tipos de técnicas de XAI - Fonte: Adaptado de Gohel, Singh e Mohanty (2021)



Enquanto os métodos *post-hoc* estabelecem modelos substitutos menos intrincados para simular técnicas complicadas de aprendizado de máquina de modelos considerados caixas opacas. São úteis para lidar com uma relação não linear ou com uma maior complexidade de dados, permitindo compreender a complexidade do modelo explicando o que “aprendeu”, quando o mesmo não está seguindo uma relação simples entre dados e características (CHAMOLA et al., 2023).

3.1 Métodos Transparentes

Esses modelos satisfazem três propriedades chamadas de transparência algorítmica, decomponibilidade e simulabilidade (CHAMOLA et al., 2023).

- **Simulabilidade:** é a capacidade de um modelo ser simulado/executado por um ser humano. Um fator que influencia a simulabilidade é a complexidade do modelo, que pode dificultar ou facilitar a sua interpretação e visualização.
- **Decomponibilidade:** um modelo é decomponível quando cada aspecto dele pode ser dissecado e explicado, desde os dados de entrada, hiperparâmetros que ele usa e as operações

que ele realiza.

- **A transparência algorítmica:** refere-se a interpretabilidade do nível do algoritmo desde a entrada de dados fornecidos até a decisão ou classificação final. Isso permite que os usuários compreendam o processo de tomada de decisão de maneira transparente. Por exemplo, o modelo linear é considerado transparente porque seu gráfico de erro é fácil de visualizar e interpretar. Através da visualização, os usuários podem entender como o modelo se comporta em diferentes situações.

A seguir estão descritos alguns modelos, pertencentes a esta categoria de modelos transparentes.

3.1.1 Regressão Linear/Logística

Modelos de regressão linear e logística são dois tipos de algoritmos de aprendizado de máquina que podem ser usados para prever ou explicar variáveis dependentes a partir de variáveis independentes. A principal diferença entre eles é que a regressão linear assume que a variável dependente é contínua e tem uma distribuição normal, enquanto a regressão logística assume que a variável dependente é categórica e tem uma distribuição binomial.

A regressão linear busca encontrar uma relação linear entre as variáveis, por exemplo, uma equação de regressão linear univariada da forma $y = a + bx$, onde y é a variável dependente, x é a variável independente, e a e b são os coeficientes estimados pelo método dos mínimos quadrados. A regressão linear pode ser usada para estimar valores numéricos, como o salário de uma pessoa ou o preço de uma casa, com base em outras características.

A regressão logística busca encontrar uma relação logística entre as variáveis, ou seja, uma equação da forma $\text{logit}(p) = a + bx$, onde p é a probabilidade de ocorrência de um evento (como pagar uma dívida ou comprar um produto), e $\text{logit}(p)$ é o logaritmo natural das chances desse evento. A regressão logística pode ser usada para estimar probabilidades ou classificar observações em duas categorias (como sim ou não, sucesso ou fracasso, spam ou não spam), com base em outras características (UFRGS, 2022).

3.1.2 Árvore de Decisão

Modelos de árvore de decisão são uma forma de aprendizado de máquina supervisionado que pode ser usado para classificação ou regressão. Uma árvore de decisão é uma estrutura que divide os dados em subconjuntos com base em regras simples derivadas dos atributos dos dados (FAN et al., 2018).

Cada nó da árvore representa um teste em um atributo, cada ramo representa o resultado do teste e cada folha representa uma classe ou um valor. O objetivo é criar uma árvore que minimize o erro de predição nos dados de treinamento e generalize bem para os dados de teste. As árvores de decisão de menor escala são facilmente simuláveis, e o incremento no número de níveis nas árvores as torna menos algoritmicamente transparentes (CHAMOLA et al., 2023).

3.1.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para classificação ou regressão. O princípio básico do KNN é que objetos semelhantes tendem a ter rótulos semelhantes. Portanto, para prever o rótulo de um novo objeto, o algoritmo procura os k objetos mais próximos ao novo objeto no conjunto de dados de treinamento e atribui o rótulo mais frequente entre eles.

Existem diferentes formas de definir a proximidade entre os objetos, como a distância euclidiana, a distância de Manhattan ou a similaridade do cosseno. A escolha da medida de proximidade depende do tipo e da escala dos dados. Além disso, o valor de k também afeta o desempenho do algoritmo. Um valor muito pequeno de k pode levar a um modelo sensível ao ruído, enquanto um valor muito grande de k pode levar a um modelo que ignora as diferenças entre as classes.

O KNN é um modelo simples e intuitivo, que não requer treinamento explícito e pode lidar com dados multi-classe. No entanto, ele também tem algumas desvantagens, como o alto custo computacional na fase de teste, a sensibilidade à escala e à distribuição dos dados e a dificuldade de escolher o valor ótimo de k (MLADENOVA; VALOVA, 2022).

3.1.4 Aprendizado baseado em regra

Aprendizado baseado em regra é um método que usa regras lógicas para representar e inferir conhecimento a partir de dados. Esse tipo de modelo é composto por um conjunto de

regras que descrevem as características e relações entre as variáveis de entrada e saída.

Essas regras são geralmente da forma “se condição então ação”, em que a condição é uma combinação de atributos e valores dos dados, e a ação é uma decisão ou uma classificação. Por exemplo, uma regra simples para identificar se um animal é um mamífero pode ser: “se tem pelos e produz leite então é mamífero”.

As regras podem ser geradas manualmente por especialistas do domínio ou automaticamente por algoritmos de indução de regras. Os modelos de aprendizado baseado em regras têm algumas vantagens, como a interpretabilidade, a modularidade e a facilidade de incorporar conhecimento prévio. No entanto, eles também apresentam alguns desafios, como a complexidade, a redundância e a inconsistência das regras (PRÖLLOCHS; FEUERRIEGEL; NEUMANN, 2019).

3.1.5 Modelos Bayesianos

Os modelos bayesianos são uma forma de análise estatística baseada no teorema de Bayes, que permite atualizar o conhecimento sobre os parâmetros de um modelo estatístico com a informação dos dados observados. Os modelos bayesianos requerem a especificação de distribuições *a priori* para os parâmetros desconhecidos, que expressam o conhecimento prévio ou as crenças sobre os mesmos (MURPHY, 2022).

Essas distribuições *a priori* são combinadas com os dados observados através de uma função de verossimilhança, que mede a probabilidade dos dados. O resultado é uma distribuição *a posteriori*, que reflete o conhecimento atualizado sobre os parâmetros, equilibrando o conhecimento prévio com os dados observados. A distribuição *a posteriori* pode ser usada para fazer inferências sobre os parâmetros e para fazer previsões sobre eventos futuros. Os modelos Bayesianos têm muitas aplicações em diversas áreas de pesquisa, como ciências sociais, ecologia, genética, medicina e outras (MURPHY, 2022).

3.2 Métodos *Post-hoc*

Os métodos *post-hoc* recebem como entrada um modelo de IA treinado, e geram formas de representações que sejam mais compreensíveis, como pontuações de importância de recursos, conjunto de regras, mapas de calor ou por meio de linguagem natural. Para isso esse tipo de método cria aproximações do funcionamento interno do modelo e da lógica de decisão. Os métodos do tipo *post-hoc* são locais, tentam esclarecer as relações entre os valores dos recursos e

as saídas de um modelo de previsão, independentemente de seus detalhes internos em relação a um exemplo particular não revelando um padrões e regras gerais (GOHEL; SINGH; MOHANTY, 2021). As técnicas de explicabilidade *post-hoc*, podem ser classificadas em duas categorias: específico ao modelo e agnóstico ao modelo, como exibido na Figura 14.

Neste trabalho serão aplicadas duas técnicas de XAI diferentes, ambas pertencentes a categoria *post-hoc*, o LIME (RIBEIRO; SINGH; GUESTRIN, 2016) e o Integrated Gradients Sundararajan, Taly e Yan (2017), métodos que são amplamente utilizadas na literatura e serão melhor explorados no decorrer deste trabalho.

As técnicas específicas ao modelo permitem restrições de explicabilidade em relação ao algoritmo de aprendizado e à estrutura interna de determinado modelo de aprendizado profundo, ou seja, levam em conta as características e os parâmetros do modelo e requerem um conhecimento mais aprofundado do mesmo, o algoritmo Integrated Gradients se encaixa nessa categoria.

As técnicas agnósticas ao modelo aplicam análise de pares de entradas e previsões do modelo para entender o mecanismo de aprendizado e gerar explicações. Elas podem ser aplicadas a qualquer modelo, sem considerar o seu funcionamento interno. Tratam o modelo como uma caixa-opaca e analisam os seus resultados, como por exemplo o LIME.

3.3 Métodos selecionados para o experimento

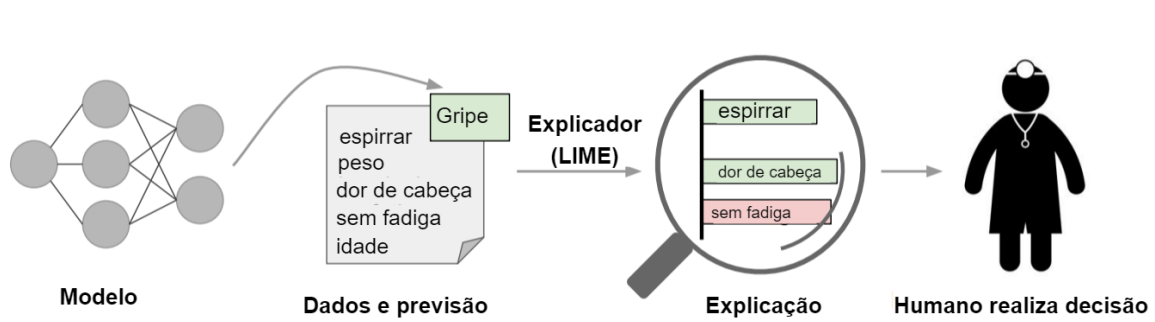
3.3.1 LIME

Como citado anteriormente, um dos métodos usados nos experimentos é o LIME (RIBEIRO; SINGH; GUESTRIN, 2016), uma técnica que explica as previsões de modelos classificadores ou regressores. Em vez de tentar entender o modelo inteiro, o LIME foca em explicar como o modelo chegou a uma previsão específica para uma instância de dados (como uma imagem ou texto). O LIME gera um modelo simples, chamado de modelo de interpretação, que é treinado localmente (ou seja, apenas para a instância de dados em questão), que pode ser facilmente interpretado.

Para explicar uma previsão, o LIME apresenta artefatos textuais e visuais que fornecem uma compreensão qualitativa da relação entre os componentes da instância, como palavras em um texto ou manchas em uma imagem, e a própria previsão do modelo. Explicar as previsões é um aspecto importante para fazer com que os humanos confiem e usem efetivamente o aprendizado de

máquina. Os humanos geralmente têm conhecimento prévio do domínio (campo de conhecimento específico) do aplicativo, que podem usar para aceitar (confiar) ou rejeitar uma previsão se entenderem o raciocínio por trás dela.

Figura 15 – Processo de explicação com o LIME - Fonte: Adaptado de Ribeiro, Singh e Guestrin (2016)



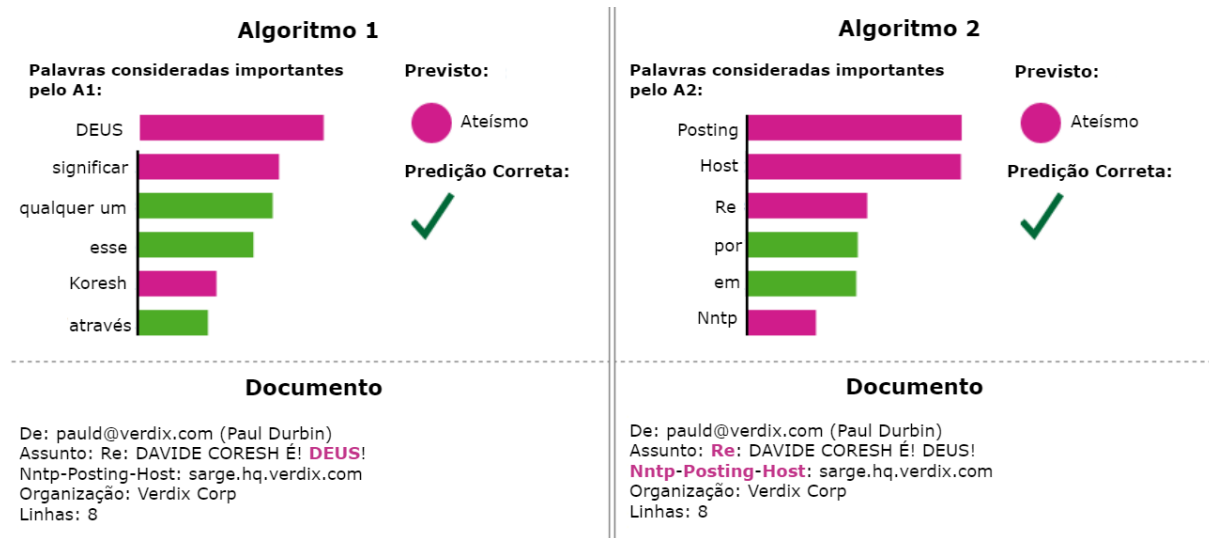
Na Figura 15 está ilustrado o processo de explicação de previsão individual. A explicação fornecida pelo algoritmo LIME, é uma lista de palavras com pesos relativos, que contribuem para a previsão representados na cor verde ou evidências contra ela representadas pela cor vermelha. Nesse caso, em específico, as palavras destacadas são sintomas, e por meio dessa explicação do LIME, os seres humanos, que comumente tem conhecimento sobre o domínio, podem avaliar se o modelo é confiável ou não, se conseguirem entender o raciocínio por trás da predição.

Com as perspectivas fornecidas por essas explicações, é possível identificar o que pode ser feito para transformar um modelo não confiável em um modelo confiável, em vista de que existem diferentes razões que podem fazer um modelo ou sua avaliação falhar, como por exemplo:

- **Vazamento de dados:** o vazamento não intencional de dados de treinamento ou validação, que não se esperaria que estivessem disponíveis no momento da previsão. O que pode fazer com que aumente potencialmente a pontuação da precisão, superestimando a utilidade que o modelo teria ao ser utilizado em um ambiente de produção (KAUFMAN et al., 2012). Este erro é particularmente difícil de ser detectado analisando apenas as previsões e os dados brutos.
- **Mudanças no conjunto de dados:** outro problema que é difícil de se detectar é a mudança do conjunto de dados; é um problema comum em modelagem preditiva que ocorre quando a distribuição conjunta de entradas e saídas difere entre as etapas de treinamento e teste. A mudança de co-variável, um caso particular de mudança do conjunto de dados, ocorre quando apenas a distribuição de entrada muda. A mudança do conjunto de dados está

presente na maioria das aplicações práticas, por razões que vão desde o viés introduzido pelo desenho experimental até a irreprodutibilidade das condições de teste no momento do treinamento.

Figura 16 – Análise individual realizada pelo LIME - Fonte: Adaptado de Ribeiro, Singh e Guestrin (2016)



Com o LIME é possível identificar esses dois problemas citados anteriormente e resolvê-los, como por exemplo removendo os dados vazados e alterando os dados de treinamento evitando a mudança do conjunto de dados solucionando possivelmente os dois problemas respectivamente. Na Figura 16, é demonstrado como as explicações de previsão individual, podem ser empregadas para escolher entre modelos, em conjunto com as outras métricas que são comumente capturadas dos modelos de linguagem como a acurácia.

3.3.2 Funcionamento do LIME

O objetivo principal do LIME é identificar um modelo interpretável baseado em uma representação que seja localmente fiel ao classificador. A seguir são explicados fatores importantes para o funcionamento do LIME.

O primeiro fator é a representação de dados interpretáveis, que diz respeito à ação de transformar os dados originais em um formato que seja mais intuitivo para um ser humano entender. Por exemplo, ao classificar textos, uma representação interpretável pode ser a contagem de ocorrências de palavras-chave relevantes em um documento, mesmo que o modelo utilize recursos mais complexos como *embeddings* de palavras.

O segundo fator é o problema de encontrar um modelo explicativo que seja simples o suficiente para ser interpretado por humanos, mas também fiel ao comportamento do modelo original. O LIME busca fornecer explicações que sejam tanto fiéis ao modelo original quanto interpretáveis pelos usuários.

A explicação do LIME é produzida pela Fórmula 3.1 (RIBEIRO; SINGH; GUESTRIN, 2016):

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (3.1)$$

Pela Fórmula 3.1 é possível ver que uma explicação é definida como um modelo $g \in G$, em que G é uma classe de modelos potencialmente interpretáveis como por exemplo modelos lineares ou árvores de decisão, e g representa o modelo explicativo que está sendo aprendido pelo método LIME. Um modelo $g \in G$ pode ser facilmente apresentado para o usuário por meio de recurso visuais e textuais.

No entanto, nem todos os modelos $g \in G$ podem ser suficientemente simples para que possam ser interpretáveis, portanto, é introduzido $\Omega(g)$, uma medida de complexidade que penaliza modelos explicativos muito complexos. É uma função que mapeia um modelo explicativo g para um valor real não negativo, onde valores maiores indicam modelos mais complexos.

A função $f(x)$ representa o modelo original que está sendo explicado pelo método LIME, em que x representa o vetor de entrada do modelo original. E π_x representa uma função de ponderação que mede a proximidade entre a instância sendo explicada (x) e as outras instâncias na vizinhança.

Assim $L(f, g, \pi_x)$ é a função de perda local que mede o quão infiel é o modelo explicativo g em aproximar o modelo original f na vizinhança do ponto sendo explicado, ponderada pela função de proximidade π_x .

Por fim para que seja possível garantir tanto a interpretabilidade quanto a fidelidade local, é minimizada a função $L(f, g, \pi_x)$ enquanto $\Omega(g)$ é mantido baixo o suficiente para que seja interpretável pelos humanos (RIBEIRO; SINGH; GUESTRIN, 2016).

3.3.3 Mapa de Saliência

No estudo de caso, outro método aplicado é o Integrated Gradient (IG) de Sundararajan, Taly e Yan (2017). O IG dá sequência à proposta de um trabalho anterior, em que é apresentado um método para a obtenção de mapa de saliência (SIMONYAN; VEDALDI; ZISSERMAN, 2014). Esse mapa foi originalmente aplicado à classificação de imagens, mas pode ser facilmente aplicado a textos como feito por Wallace et al. (2019). Esse método é baseado em gradiente e mostra quais partes de uma entrada são as mais relevantes para a previsão de uma determinada classe, permitindo ao usuário de um classificador entender quais recursos são mais importantes em nível de instância, ou seja, em relação a uma determinada entrada individual.

Neste método, uma entrada é dada à rede neural e o processo de propagação direta é realizado sobre ela. Isso permite que a importância de cada parte da entrada para uma classe seja obtida retropropagando o *logit* da classe específica até a entrada bruta. Para uma imagem, isso corresponde ao gradiente em cada pixel e, para textos, ao gradiente em cada token. O objetivo é que as partes que mais afetam o resultado da classificação, que possuem maiores gradientes, sejam as mais relevantes para a classificação.

Neste trabalho (SIMONYAN; VEDALDI; ZISSERMAN, 2014) os autores descrevem como uma rede neural convolucional (CNN) de classificação, pode ser usada para analisar o suporte espacial de uma classe específica da imagem. Os autores desejam classificar os pixels de uma imagem I_0 com base na sua influência na pontuação $S_c(I_0)$, considerando uma classe c e a função de pontuação de classe $S_c(I)$.

Para ilustrar o conceito, é utilizado um modelo de pontuação linear para a classe c , onde a pontuação $S_c(I)$ é dada por:

$$S_c(I) = w_c^T I + b_c \quad (3.2)$$

Nesse modelo, a imagem I é representada de forma vetorizada (unidimensional), sendo w_c e b_c o vetor de pesos e o viés do modelo, respectivamente. A magnitude dos elementos de w_c indica a importância dos pixels correspondentes de I para a classe c .

No entanto, nas CNNs profundas, a pontuação de classe $S_c(I)$ é uma função altamente não linear de I , o que impossibilita a aplicação direta do raciocínio anterior. Mas, é possível realizar uma aproximação linear da pontuação $S_c(I)$ na vizinhança da imagem I_0 usando a expansão de Taylor de primeira ordem:

$$S_c(I) \approx w^T I + b \quad (3.3)$$

Em que w é a derivada de S_c em relação à imagem I no ponto I_0 :

$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0} \quad (4) \quad (3.4)$$

Dessa maneira, é possível determinar quais são os elementos que, ao serem modificados, têm maior impacto na variação da pontuação para a classe desejada.

3.3.4 Integrated Gradients

Além do método de gradiente padrão apresentado na sessão anterior, conhecido como Vanilla Gradient, existem outras formas de obter mapas de saliência. Uma delas é o Integrated Gradients, que é o método aplicado neste trabalho.

O método de Integrated Gradients é uma técnica amplamente utilizada para interpretar e atribuir importância às características de um modelo de aprendizado de máquina. Esse método foi proposto por Sundararajan, Taly e Yan (2017) e tem sido aplicado em diversas áreas, como visão computacional, processamento de linguagem natural e bioinformática.

O objetivo principal desse método é atribuir contribuições individuais às características de entrada do modelo, fornecendo uma medida quantitativa da importância de cada característica em relação à predição final. Essa técnica baseia-se na ideia de que a importância de uma característica pode ser avaliada observando-se as mudanças no resultado do modelo quando essa característica é removida gradualmente.

O procedimento envolve o cálculo dos gradientes parciais em relação às características de entrada em diferentes pontos da linha de base (*baseline*) até o ponto de interesse. A linha de base pode ser definida como um ponto de referência onde todas as características têm um valor específico. Em seguida, é calculada a integral dos gradientes parciais ao longo de uma linha reta que liga a linha de base ao ponto de interesse.

O método Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017) é construído sobre o mapa de saliência *vanilla* (SIMONYAN; VEDALDI; ZISSERMAN, 2014). Em vez de obter um único gradiente para cada parte da entrada, múltiplas variações da entrada são alimentadas ao classificador e a saliência final é a média de todos os gradientes obtidos. Uma

variação é a entrada original e outra é uma entrada de linha de base, como uma imagem totalmente preta ou totalmente branca, o texto “desconhecido” ou uma incorporação contendo apenas zeros. Os gradientes são calculados ao longo de i dimensões para uma entrada (x) e um *baseline* (x'), de acordo com a Equação 3.5, onde F é a função da rede neural.

$$IntegratedGrads_i(x) ::= (x_i - x'_i) * \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} * d\alpha \quad (3.5)$$

O cálculo computacional do método de Integrated Gradients é realizado de acordo com a Equação 3.6. Este método acumula os gradientes de m diferentes formas da entrada, ponderando a importância de cada *feature*. O parâmetro m é escolhido pelo usuário e determina a quantidade de representações intermediárias utilizadas entre o *baseline* e a entrada original.

$$IntegratedGrads_i^{approx}(x) = (x_i - x'_i) * \sum_{k=1}^m \frac{\partial F(x' + \frac{k}{m}(x - x'))}{\partial x_i} * \frac{1}{m} \quad (3.6)$$

Na Figura 17 está um exemplo do método Integrated Gradients retirado do trabalho de Sundararajan, Taly e Yan (2017), aplicado para realizar a identificação dos componentes mais importantes para classificação de texto, no exemplo em questão o modelo foi treinado para classificar qual é o tipo de resposta esperada de acordo com a pergunta.

Figura 17 – Exemplo da detecção de componentes importantes para a classificação por meio do método de Integrated Gradients - Fonte: Sundararajan, Taly e Yan (2017)

how many townships have a population above 50 ? [prediction: NUMERIC]
 what is the difference in population between fora and masilo [prediction: NUMERIC]
 how many athletes are not ranked ? [prediction: NUMERIC]
 what is the total number of points scored ? [prediction: NUMERIC]
 which film was before the audacity of democracy ? [prediction: STRING]
 which year did she work on the most films ? [prediction: DATETIME]
 what year was the last school established ? [prediction: DATETIME]
 when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
 did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Uma das principais vantagens Integrated Gradients é a sua capacidade de fornecer interpretações individuais para cada característica de entrada, o que facilita a compreensão e a explicabilidade dos modelos de aprendizado de máquina. No entanto o Integrated Gradients também possui limitações, por exemplo, a escolha adequada da linha de base pode afetar os resultados,

e a computação dos gradientes em relação a todas as características pode ser computacionalmente custosa em modelos de alta dimensionalidade.

Mas apesar dessas limitações, o método de Integrated Gradients continua sendo uma técnica valiosa para a interpretação de modelos de aprendizado de máquina, auxiliando na compreensão do comportamento dos modelos e na identificação de características relevantes para a tomada de decisão. Sua aplicação em diversas áreas de pesquisa tem contribuído para avanços significativos e para a transparência dos modelos de aprendizado de máquina.

3.4 Trabalhos Correlatos

3.4.1 Trabalho de Kiefer (2022)

A tarefa de gerar explicações locais para modelos de classificação de texto frequentemente necessita de outra tarefa: a aproximação local. Para criar uma vizinhança de um documento a ser explicado, técnicas de amostragem são usadas, tratando características como semanticamente independentes, o que resulta na perda de informação contextual e semântica. Adicionalmente, conforme afirmado por Kiefer (2022), essas técnicas podem extrapolar em áreas com poucos dados, levando a interpretações errôneas, e ignoram o princípio de que “o todo é maior que a soma das partes” ao tratar palavras de forma isolada. Em vista disso, Kiefer (2022) desenvolveu uma arquitetura chamada Contextual and Semantic Explanations (CaSE), que permite a combinação de conhecimento semântico com qualquer classificador de texto.

A arquitetura CaSE, combina arranjos de recursos semânticos ou interrogações semânticas com uma versão modificada do LIME. Dentro do contexto textual, a abordagem semântica é expressa como um método de modelagem de texto, mais precisamente, uma abordagem de modelagem de tópicos, como o Latent Dirichlet Allocation (LDA). Essa técnica é capaz de extrair informações semânticas e contextuais do domínio de entrada de texto. É importante ressaltar que o CaSE não se limita a texto, podendo ser utilizado também para trabalhar com imagens.

O CaSE possui duas variantes, a primeira emprega a estrutura de tópicos do LDA para contextualmente organizar todos os elementos do documento que será classificado, e a segunda variante constitui uma versão revisada do LIME (topicLIME) que gera explicações na forma de atribuições de tópicos, incluindo as palavras coerentes correspondentes.

O LDA na modelagem de texto tem como propósito identificar representações concisas dos documentos contidos em um conjunto de dados textuais, mantendo simultaneamente as relações estatísticas fundamentais, seja dentro do mesmo documento ou entre documentos distintos.

Ao combinar um sistema de explicação com uma abordagem semântica como o LDA, é possível responder a perguntas como: se o classificador se comporta de maneira esperada pelos humanos, o quanto ele se assemelha à intuição humana e quais recursos ocultos ele aprendeu. Outra versão do CaSE chamada Semantic Interrogations, ajuda a responder a pergunta de quanto um classificador trabalha em conformidade semântica com o pensamento humano em relação aos aspectos aprendidos na cadeia indutiva do classificador. E ao integrar uma abordagem semântica

como o LDA e uma técnica de explicação agnóstica ao modelo como o LIME, o CaSE é capaz de fornecer essa funcionalidade além do escopo das redes neurais de maneira agnóstica ao modelo. Isso permite que os humanos identifiquem a presença ou ausência de recursos semânticos de alto nível.

3.4.2 Trabalho de Torcianti e Matzka (2021)

A técnica LIME tem sido amplamente adotada, porém não é possível encontrar muitos trabalhos utilizando essa técnica de explicação para conjuntos de dados em português do Brasil. Torcianti e Matzka (2021), apesar de não trabalharem com o mesmo idioma objetivo deste trabalho, apresentam em comum o objetivo de realizarem um trabalho que avalia a qualidade explicativa do LIME comparado a outros dois modelos explicativos, modelo de árvore de decisão explicável e desvio de recurso normalizado, para um conjunto de dados de manutenção preditiva sintético.

Os autores treinaram um modelo de classificação Bagged Trees Ensemble inicialmente para prever uma propriedade denominada “falha de máquina”. E dada a avaliação do modelo treinado considerada como satisfatória, são aplicados o LIME, um modelo de árvore de decisão explicável e desvio de recurso normalizado, e as explicações resultantes são avaliadas. Assim, os autores têm como objetivo fornecer uma explicação plausível para a classificação, permitindo ao usuário entender a previsão, pois cada modo de falha pode ser explicado por um conjunto específico de recursos. Sendo essas características:

- Falha por Desgaste da Ferramenta (TWF): causada por desgaste da ferramenta.
- Falha por Dissipação de Calor (HDF): causada por temperatura do ar, temperatura do processo e velocidade de rotação.
- Falha por Falha de Energia (PWF): causada por torque e velocidade de rotação.
- Falha por Sobrecarga (OSF): causada por torque, tipo e desgaste da ferramenta.

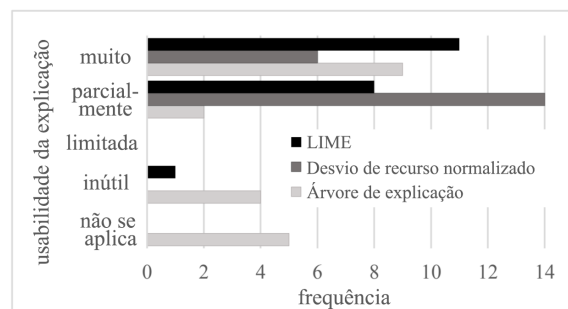
As explicações fornecidas pelo LIME são apresentadas como características com maior importância relativa e o correspondente valor de ativação da árvore simples para um determinado ponto de dados. Os valores de importância relativa das características são ajustados para 100% para todas as características. Isso é feito para restringir o número de características usadas na

explicação, tornando assim mais fácil para as pessoas compreenderem. Um limite de importância relativa das características de pelo menos 25% é aplicado com o objetivo de alcançar essa limitação. As análises foram pontuadas de acordo com a presença das palavras destacadas, e a usabilidade da explicação foi separada nas seguintes categorias:

- muito útil: pelo menos duas das três características essenciais devem ser incluídas (exceção para TWF: apenas o desgaste da ferramenta é essencial).
- parcialmente útil: um essencial ou, alternativamente, um essencial e um recurso não essencial (mas não enganoso).
- limitada: inclusão de uma característica essencial e uma característica enganadora.
- inútil: nenhuma explicação fornecida.

Na Figura 18 está a comparação quantitativa das pontuações de explicação entre as três técnicas: LIME, desvio de recurso normalizado e modelo de árvore de decisão explicável.

Figura 18 – Comparação quantitativa das pontuações de explicação entre árvores de decisão explicáveis, desvios normalizados de características e LIME - Fonte: Traduzido de Torcianti e Matzka (2021)



A Figura 18 indica que o LIME produziu principalmente explicações muito úteis, as árvores de decisão explicáveis produziram também algumas explicações muito úteis, no entanto falharam em alguns casos. Os desvios de recursos normalizados produzem principalmente explicações parcialmente úteis e poucas explicações muito úteis.

A qualidade explicativa geral do LIME é avaliada como superior à das árvores de decisão explicáveis e à da variação normalizada dos recursos. As características do LIME são especialmente úteis ao explicar um modo de falha que é definido por apenas uma característica essencial. Neste trabalho em questão, os autores afirmam que os resultados do LIME são bons e consistentes; no entanto, apontam que ele apresenta uma instabilidade em suas explicações.

3.4.3 Trabalho de Moradi e Samwald (2021)

O LIME se provou capaz de realizar boas avaliações, ficando pouco atrás de soluções para problemas específicos, como no trabalho de Moradi e Samwald (2021), que propõem um método de explicação de modelos de caixa opaca denominado Biomedical Confident Itemsets Explanation (BioCIE) especializado em textos biomédicos. No trabalho, os autores realizaram a comparação do BioCIE com outros dois métodos, o LIME e o Model Understanding through Subspace Explanations (MUSE) (LAKKARAJU et al., 2019).

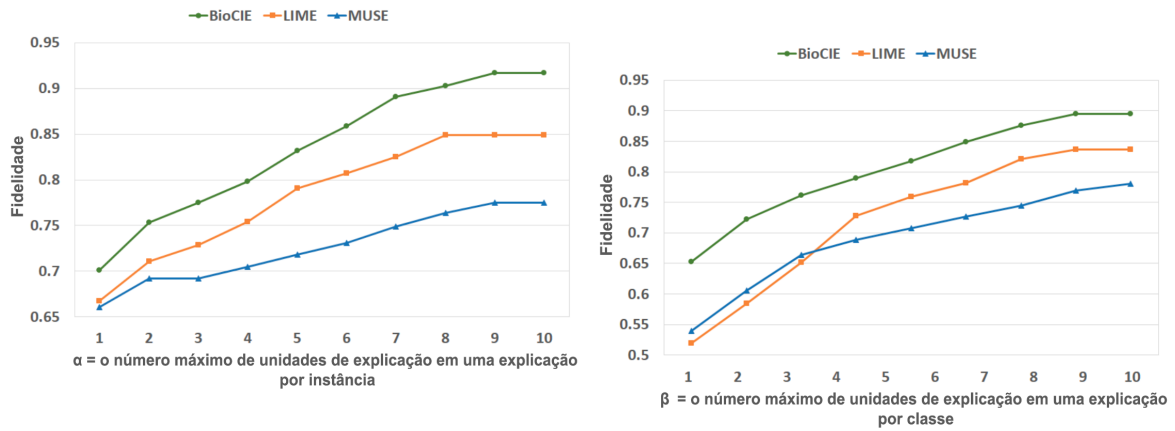
Os autores avaliaram dois aspectos das técnicas de explicabilidade: a fidelidade e a interpretabilidade. A fidelidade refere a capacidade de um método de explicação imitar o comportamento de um modelo caixa opaca, em relação a atribuição de rótulos às instâncias de dados. A interpretabilidade refere-se à facilidade com que uma explicação pode ser compreendida. Podendo ser avaliada de diferentes formas, os autores conduziram três experimentos para investigar a interpretabilidade analisando o comprometimento entre precisão descritiva e medidas de interpretabilidade.

Os autores conduziram três experimentos com o intuito de avaliar a capacidade dos explicadores em produzir explicações curtas e precisas. O Experimento 1 avalia uma explicação por instância, variando um parâmetro α que define o número máximo de unidades de explicação; o Experimento 2 avalia explicações de classe, variando o número máximo de unidades de explicação β ; e o Experimento 3 avalia explicações globais, variando o número máximo de unidades de explicação γ , otimizadas para cobrir o maior número de amostras de texto e minimizar sobreposições.

A precisão descritiva (ou fidelidade) refere-se à capacidade de um explicador imitar o comportamento de um modelo caixa opaca em termos de atribuição de rótulos de classe a instâncias de dados (MURDOCH et al., 2019).

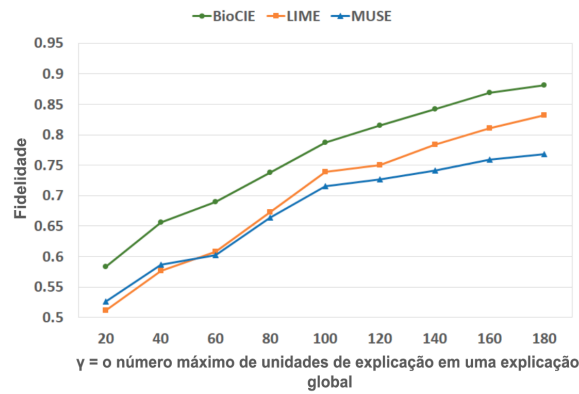
Na Figura 19 é possível ver a pontuação de fidelidade atingida pelos métodos de explicação nos três experimentos. Por meio dos gráficos, os autores concluíram que o método de explicação BioCIE superou os outros métodos de explicação em termos de fidelidade ao utilizar conjuntos menores de unidades de explicação, indicando que o BioCIE produz explicações mais interpretáveis. Em comparação, o LIME e o MUSE geralmente exigiram conjuntos maiores de unidades de explicação para alcançar a mesma precisão descritiva. Além disso, os resultados mostraram que, embora o aumento no tamanho das explicações geralmente resultasse em uma

Figura 19 – Número máximo de unidades de explicação em uma explicação por instância em comparação com os resultados de precisão descritiva - Fonte: Traduzido de Moradi e Samwald (2021)



(a) Experimento 1

(b) Experimento 2



(c) Experimento 3

melhora na fidelidade, essa relação não era linear, e após um certo limiar, não havia melhoria significativa na fidelidade, e a interpretabilidade diminuía.

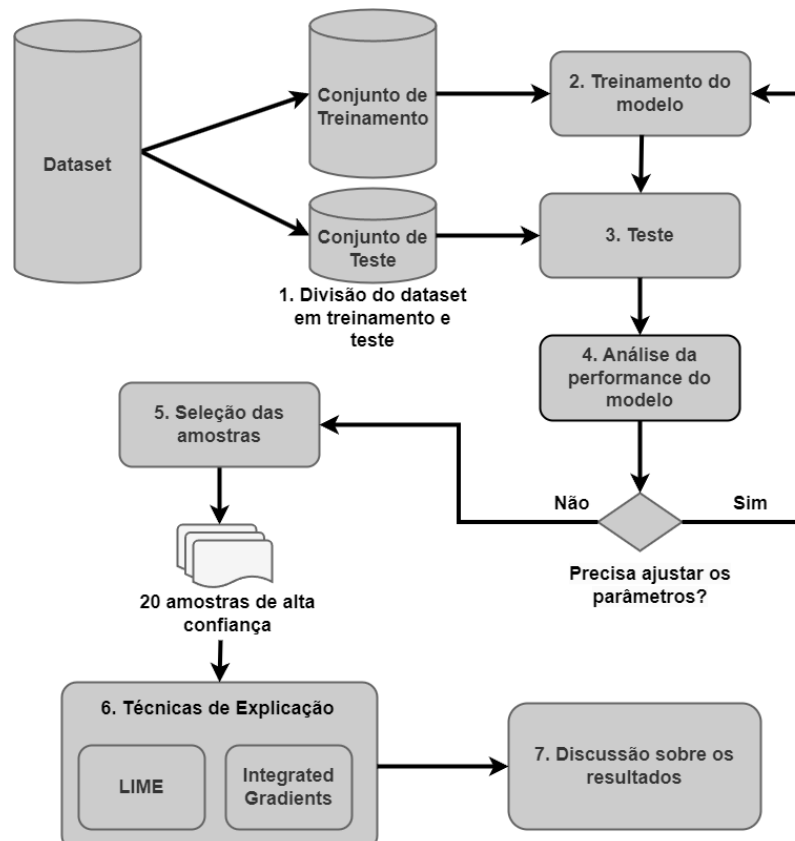
Como exemplo, o BioCIE alcançou uma precisão descritiva de 80% em explicações de instância, de classe e globais, utilizando, em média, 4, 22 e 110 unidades de explicação, respectivamente, como exibido na Figura 19. Em contraste, o LIME necessitou de uma média de 5,5, 35 e 150 unidades de explicação para atingir a mesma precisão em explicações de instância, de classe e globais. Por outro lado, o MUSE não conseguiu alcançar uma precisão descritiva de 80% dos experimentos.

4 Materiais e Métodos

Neste capítulo, foram realizados experimentos com dois métodos de XAI: LIME e IG, ambos amplamente utilizados na literatura para a interpretação de modelos de linguagem. Esses métodos foram aplicados individualmente sobre um conjunto de amostras para analisar modelos de linguagem treinados na classificação de texto de notícias falsas. Os métodos foram aplicados individualmente sobre um conjunto de amostras selecionadas. O modelo BERTimbau, introduzido anteriormente no Capítulo 2.3, Seção 2.3.5, foi utilizado para este fim, com a escolha da variante BERT-Base entre as disponíveis, devido à sua ampla utilização na literatura para o processamento de português brasileiro. Além disso, foram realizadas análises dos *datasets* utilizando o spaCy (HONNIBAL; MONTANI, 2017), uma biblioteca de processamento de linguagem natural.

A Figura 20 mostra as etapas seguidas para conduzir os experimentos, sendo que cada uma delas será explicada nas seções seguintes. O estudo de caso foi realizado com *datasets* no domínio de classificação de notícias falsas, os quais foram utilizados para treinar o modelo selecionado.

Figura 20 – Etapas seguidas para realização dos experimentos



4.1 Dataset

Antes de começar os experimentos, foi feita a seleção de dois *datasets* em português brasileiro para o estudo de caso. A disseminação de notícias falsas é um problema significativo em todo o mundo, principalmente nas redes sociais. Para combater essa prática, pesquisadores da área de informática têm se dedicado a desenvolver técnicas e ferramentas capazes de identificar e combater a disseminação dessas notícias.

Em vista deste contexto, foram selecionados o FakeRecogna (GARCIA; AFONSO; PAPA, 2022) e o Fake.Br Corpus (SANTOS; MONTEIRO; PARDO, 2018), dois *datasets* de notícias falsas para o português brasileiro que se destacam como importantes contribuição para o campo.

4.1.1 FakeRecogna

O FakeRecogna (GARCIA; AFONSO; PAPA, 2022) consiste em um *dataset* de notícias em português brasileiro em formato de texto simples, composto originalmente por 11.902 notícias, sendo 5.951 falsas e 5.951 verdadeiras; no entanto, durante os experimentos, foi encontrado um erro no *dataset*, uma das amostras do tipo notícia real, o conteúdo é apenas um número de telefone, então ela foi removida, e os experimentos seguiram com 5.951 amostras de notícias falsas e 5.950 amostras de notícias reais.

As notícias foram coletadas por *crawlers* desenvolvidos para minerar páginas de notícias, e pertencem ao período entre 2019 e 2021. As notícias falsas foram coletadas de páginas que verificam a veracidade das informações, sendo elas: Boatos.org, Fato ou Fake, E-farsas, UOL Confere, AFP Checamos e Projeto Comprova. Quanto às notícias reais, elas foram coletadas dos principais portais do Brasil, como G1, UOL, Extra, além da página inicial do Ministério da Saúde do Brasil.

O FakeRecogna, até o presente momento deste trabalho, é o maior *dataset* de notícias reais e falsas em português do Brasil, e o que apresenta as notícias mais recentes.

4.1.2 Fake.Br Corpus

Assim como o FakeRecogna, o *dataset* Fake.Br Corpus (SANTOS; MONTEIRO; PARDO, 2018) é composto de notícias em português brasileiro em formato de texto simples, contendo 7.200 notícias, sendo 3.600 falsas e 3.600 verdadeiras. As notícias coletadas neste trabalho pertencem ao período de janeiro de 2016 a janeiro de 2018.

As notícias falsas foram coletadas em quatro sites: Diário Brasil, A Folha do Brasil, The Jornal Brasil e Top Five TV, enquanto as notícias reais foram coletadas nas principais agências de notícias do Brasil: G1, Folha de São Paulo e Estadão. As notícias podem ser divididas em seis categorias de assuntos: TV e celebridades, política, sociedade e notícias diárias, ciência e tecnologia, economia e religião.

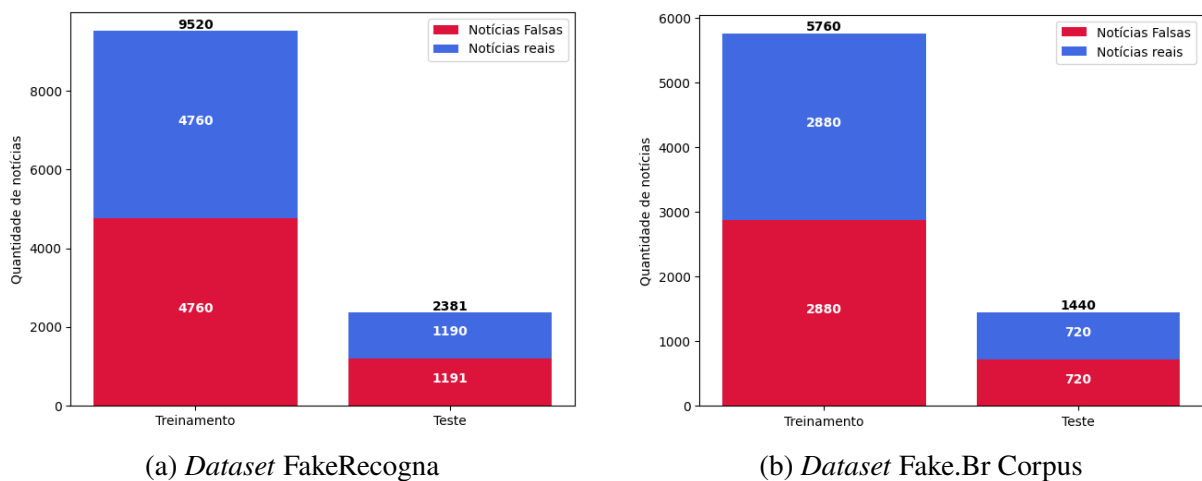
Em resumo, o Fake.Br Corpus é uma importante contribuição para o campo de detecção de notícias falsas em português brasileiro. Com este *corpus*, é possível desenvolver e avaliar sistemas de detecção de notícias falsas mais precisos e eficazes, que podem contribuir para a redução da disseminação de notícias falsas nas redes sociais e na mídia em geral.

4.2 Divisão do conjunto de dados em treinamento e teste

O modelo BERTimbau foi treinado no Fake.Br Corpus e no FakeRecogna de duas formas diferentes: com as notícias completas, incluindo as *stop words*, e com as notícias tratadas, sem *stop words* e caracteres numéricos. O objetivo é analisar se essas palavras influenciam no processo de decisão do modelo.

Na Figura 20, é possível observar a divisão de amostras que foi realizada em ambos os *datasets*. No caso, 20% das amostras foram separadas para o conjunto de teste e 80% para o conjunto de treinamento.

Figura 21 – Distribuição das amostras após a divisão nos conjuntos de treinamento e teste



4.3 Treinamento

O próximo passo neste experimento seguindo a Figura 20 foi o treinamento do modelo. Para isso, foi usada a biblioteca TensorFlow (ABADI et al., 2015) em uma GPU premium do Google Colaboratory (AI, 2018), especificamente a NVIDIA A100-SXM4 com 40 GB de RAM virtual e 80 multiprocessadores CUDA.

Para carregar o modelo BERTimbau-Base, foi utilizada a biblioteca Hugging Face Transformers (FACE, 2023). O tokenizador também foi carregado do modelo pré-treinado e usado para converter o texto de entrada em uma sequência de tokens que o modelo pode entender.

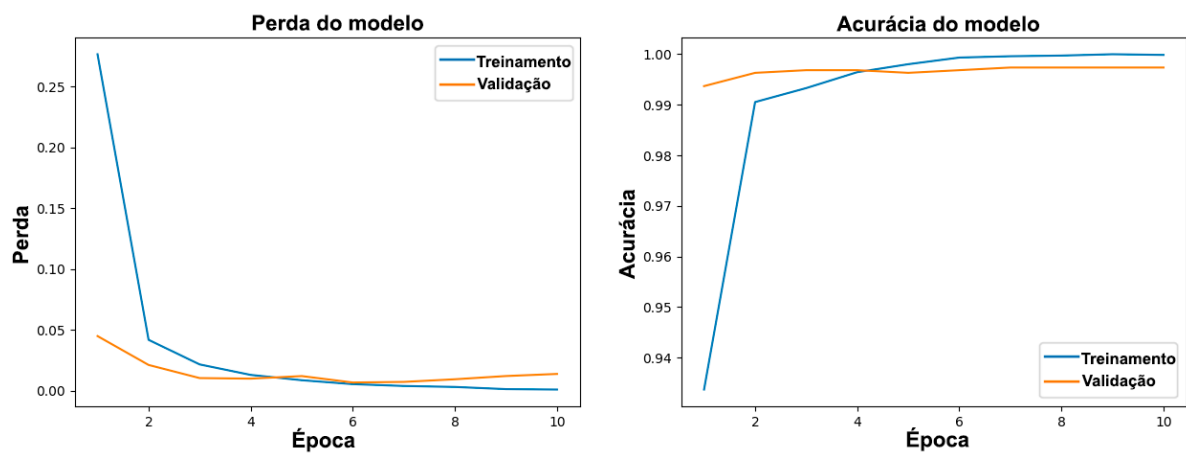
Foram treinados quatro modelos diferentes, dois com o *dataset* FakeRecogna, sendo um com as notícias em texto completo e outro com as *stop words* removidas. Da mesma forma, foram treinados mais dois modelos com o Fake.Br Corpus, seguindo o mesmo procedimento de manutenção e remoção das *stop words*. Na Tabela 6, é possível verificar informações sobre os modelos, como o uso de *stop words* e tempo de treinamento. Todos os modelos foram treinados ao longo de 10 épocas, utilizando a função de otimização Adam, com uma taxa de aprendizado de 1,00E-06 e *batch size* de 12.

Tabela 6 – Modelos treinados

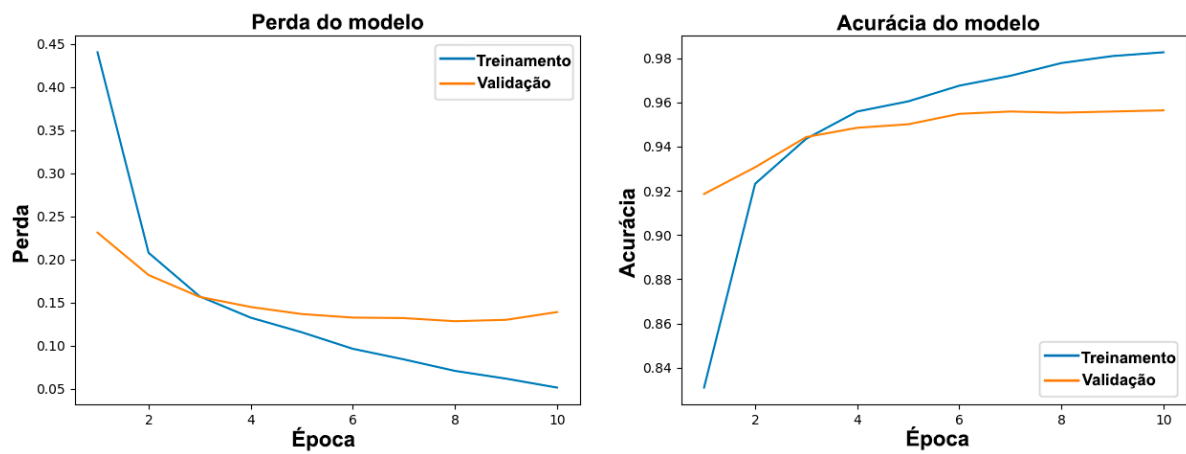
Modelo	Dataset	Tempo treinamento (min.)	Com <i>Stop words</i>
Modelo 1	FakeRecogna	59,22	Verdadeiro
Modelo 2	FakeRecogna	60,86	Falso
Modelo 3	Fake.Br Corpus	36,15	Verdadeiro
Modelo 4	Fake.Br Corpus	38,82	Falso

A acurácia e a taxa de perda foram registradas durante as 10 épocas de treinamento das amostras de treinamento e validação. As Figuras 22 e 23 mostram os gráficos com as curvas apresentadas pelos quatro modelos.

Figura 22 – Acurácia e Perda do Modelo 1 e 2 ao longo das 10 épocas de treinamento

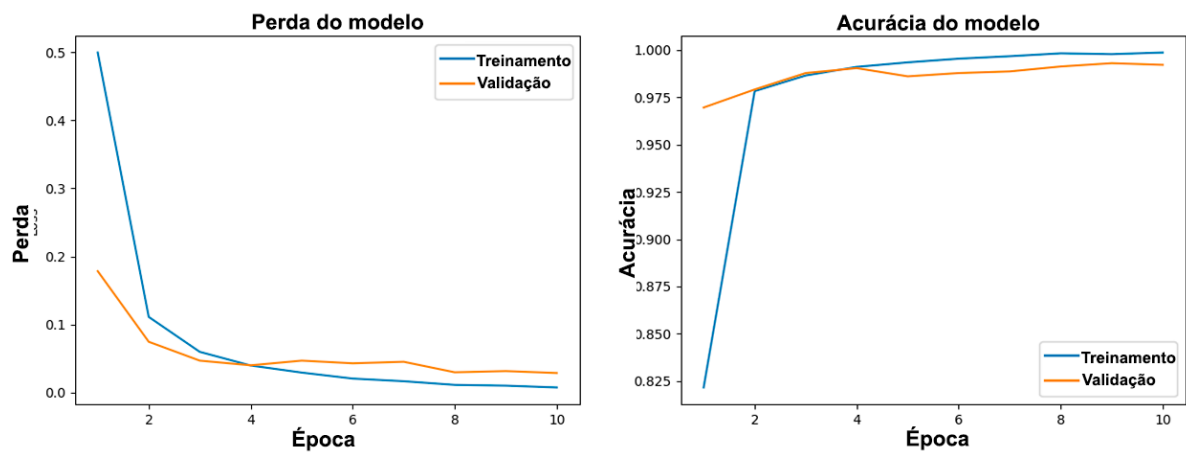


(a) Modelo 1

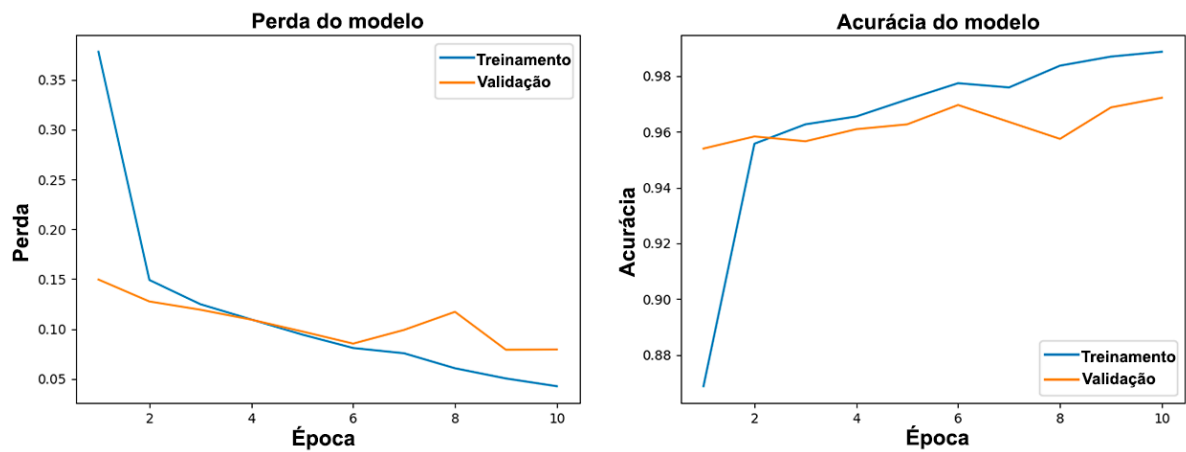


(b) Modelo 2

Figura 23 – Acurácia e Perda do Modelo 3 e 4 ao longo das 10 épocas de treinamento



(a) Modelo 3



(b) Modelo 4

5 Resultados

Neste capítulo, são apresentados os resultados dos experimentos conduzidos para investigar os conjuntos de dados e os modelos que foram treinados e apresentados no capítulo anterior.

5.1 Teste e Análise do modelo

Após a conclusão do treinamento, os pesos são testados no conjunto de teste, que foi dividido conforme explicado na Seção 4.2. Ao final de cada teste, foram armazenadas algumas métricas dos modelos, como a acurácia e a perda no conjunto de teste, que podem ser visualizadas na Tabela 7.

Tabela 7 – Taxa de perda e acurácia de teste dos modelos

Métrica	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Perda Teste	0,0190	0,1417	0,0379	0,1142
Acurácia Teste	0,9945	0,9551	0,9910	0,9604
Melhor Época	6	8	10	9
Perda Validação (Melhor Época)	0,0066	0,1286	0,0285	0,0792
Acurácia Validação (Melhor Época)	0,9968	0,9554	0,9922	0,9688

E também foram registradas métricas em relação à classe predita pelo modelo de linguagem, incluindo precisão, *recall* e *F1-Score*. Essas métricas podem ser visualizadas na Tabela 8.

Tabela 8 – Métricas dos modelos

Modelo	Classe	Precisão	Recall	F1-Score
Modelo 1	Falsa	1,00	0,99	0,99
	Real	0,99	1,00	0,99
Modelo 2	Falsa	0,98	0,94	0,96
	Real	0,93	0,98	0,95
Modelo 3	Falsa	0,99	0,99	0,99
	Real	0,99	0,99	0,99
Modelo 4	Falsa	0,98	0,94	0,96
	Real	0,94	0,98	0,96

De acordo com as Tabelas 7 e 8, é possível notar que entre os modelos treinados com o FakeRecogna, o Modelo 1, que inclui *stop words*, obteve um desempenho maior em relação ao Modelo 2, que não as contém. Como mostrado na Tabela 7, o Modelo 1 alcançou uma acurácia de 0,99 e uma taxa de perda de 0,01, enquanto o Modelo 2 obteve uma acurácia de 0,95 e uma taxa de perda de 0,14.

E o mesmo resultado foi notado entre os modelos treinados no conjunto de dados Fake.Br Corpus, com o Modelo 3, que possui *stop words*, tendo um desempenho maior que o Modelo 4, que não possui. Na Tabela 7, é possível ver que o Modelo 3 obteve 0,99 de acurácia e 0,03 de taxa de perda, enquanto o Modelo 4 obteve 0,96 de acurácia e 0,11 de taxa de perda.

Depois que o teste foi concluído, o desempenho do modelo foi analisado, sendo esta a etapa 4 da Figura 20. Com essa análise, é decidido se o modelo precisa ser ajustado. Caso seja decidido que sim, o processo volta para a etapa de treinamento, mas desta vez com os novos hiperparâmetros.

E nas Figuras 24 e 25 estão as matrizes de confusão dos modelos, que permitem observar como os modelos desempenharam com as amostras de teste. Nas matrizes, o “0” representa a classe de notícias falsas e o “1” representa as notícias reais.

Pela diagonal azul escura nas quatro matrizes, é possível notar que os modelos acertaram a classificação da maioria das amostras. Os Modelos 1, 2 e 4 tiveram mais erros ao classificar notícias reais, enquanto o Modelo 3 teve mais erros ao classificar as notícias falsas.

Figura 24 – Matrizes de Confusão do Modelo 1 e Modelo 2

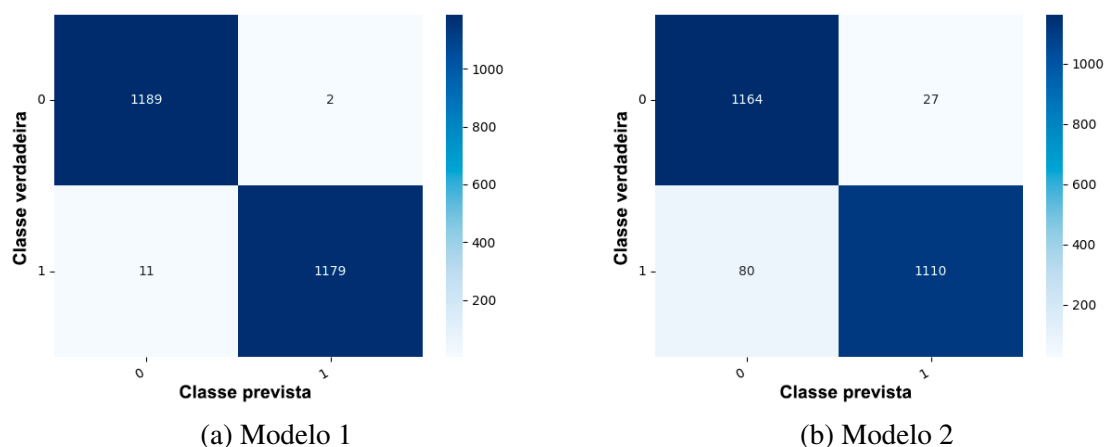
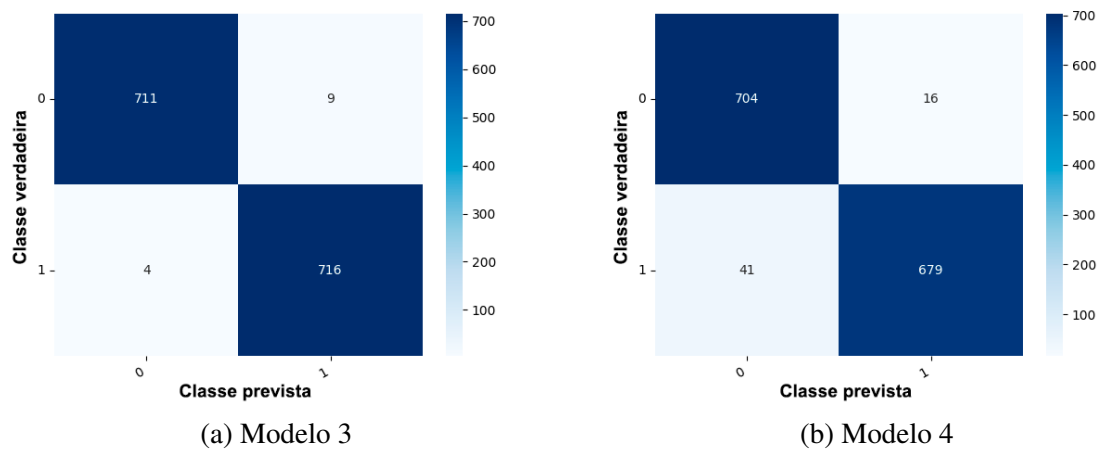


Figura 25 – Matrizes de Confusão do Modelo 3 e Modelo 4



5.2 Amostragem

Entre os modelos treinados, foram selecionados os modelos de melhor desempenho, treinados com o conjunto de dados completo e o conjunto de dados sem *stop words*, para serem analisados com os métodos de XAI selecionados.

Para analisar esses modelos, como exibido pela etapa 5 da Figura 20, foi planejada a seleção de um total de 20 amostras do conjunto de dados FakeRecogna e do Fake.Br Corpus, para terem suas previsões examinadas usando os algoritmos LIME e IG.

Dessas 20 amostras, tinha-se como objetivo selecionar de forma aleatória cinco verdadeiros positivos, cinco verdadeiros negativos, cinco falsos positivos e cinco falsos negativos com alta confiança. No entanto, como é possível observar na Figura 24, o Modelo 1 não obteve cinco falsos positivos, e na Figura 25, o Modelo 3 não obteve cinco falsos negativos. Desse modo, no Modelo 1, foram analisadas 17 amostras pelo LIME e IG, o Modelo 3 teve 19 amostras, enquanto os modelos 2 e 4 tiveram as 20 amostras planejadas explicadas pelos algoritmos.

Essa forma de amostragem foi escolhida para compreender as palavras que levam a acertos de alta confiança e erros com alta confiança.

5.3 Explicações geradas pelo LIME

Nesta seção, foram analisadas as explicações geradas pelo LIME. Quatro explicações das amostras classificadas por cada modelo foram destacadas e comentadas no texto, representando cada um dos critérios de seleção: uma amostra de notícia falsa classificada como falsa; uma

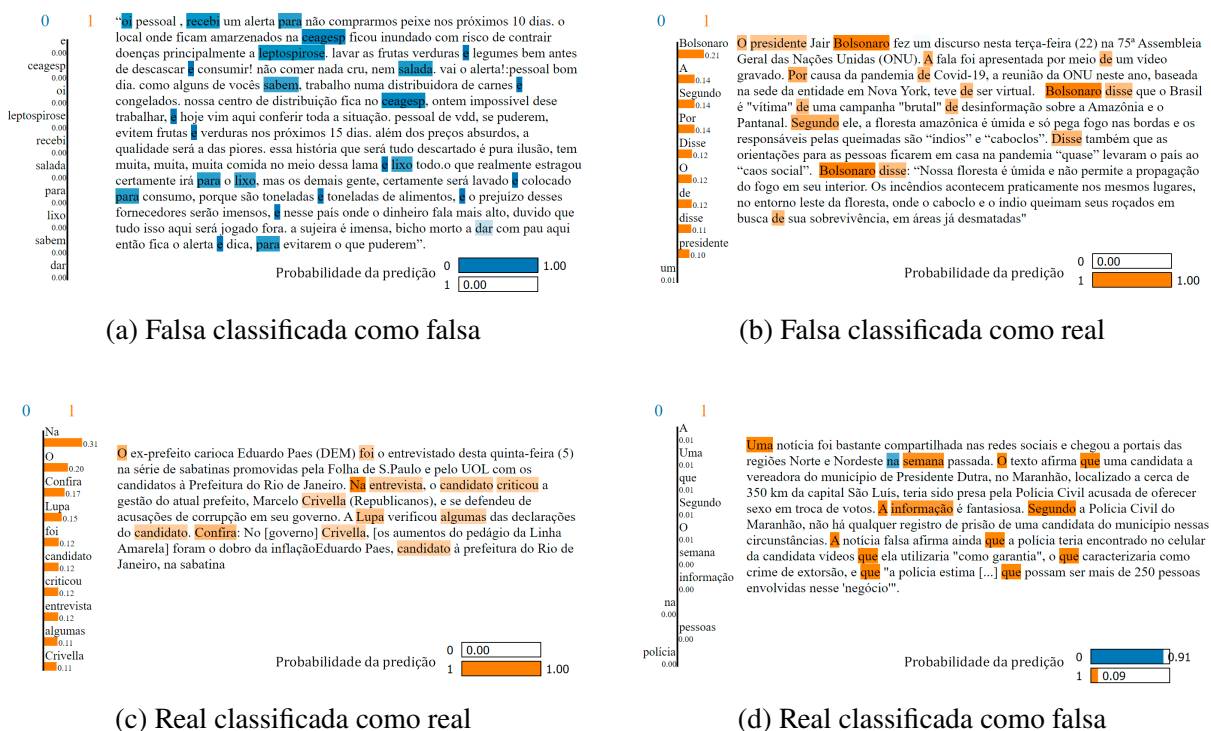
falsificação erroneamente classificada como real; uma notícia real classificada como real e, finalmente, uma notícia real classificada como falsa. As explicações geradas pelo IG foram analisadas seguindo a mesma estratégia e se encontram na Seção 5.4. Todas as outras explicações que foram geradas tanto pelo LIME quanto pelo IG podem ser encontradas na seção de apêndices.

Foi possível observar que as explicações do LIME apresentaram resultados mais interessantes e mais fáceis de serem interpretados, destacando as palavras mais importantes para a classificação de forma simples. Enquanto isso, nas explicações geradas pelo IG, as palavras apresentavam cores mais semelhantes. Uma elaboração mais detalhada dessas descobertas será fornecida na Seção 5.6.

Nas explicações geradas pelo LIME, as palavras são destacadas com duas cores, conforme mostrado por exemplo na Figura 26, azul para palavras com peso negativo (ou seja, contribuem para a previsão, como notícia falsa, também representadas por 0) e laranja para palavras com peso positivo (ou seja, contribuem para a previsão como notícia real, também representadas por 1).

5.3.1 Modelo 1: FakeRecognia incluindo *stop words*

Figura 26 – Explicações geradas pelo LIME do Modelo 1



5.3.1.1 Notícia falsa classificada como falsa

Na Figura 26a, há uma explicação do LIME para uma notícia falsa com *stop words*, classificada corretamente pelo Modelo 1. Com base nela, é possível observar os seguintes itens:

- Referência a saúde: as palavras “leptospirose”, “salada” e “lixo”, que foram destacadas com pontuações negativas, são palavras que abordam os tópicos de saúde e condições sanitárias. O uso dessas palavras em um contexto negativo pode aumentar a suspeita de que a notícia é falsa.
- Opiniões pessoais: as palavras “recebi” e “oi” foram destacadas com uma pontuação negativa, o uso delas podem indicar uma narrativa mais pessoal, e essas opiniões podem influenciar na interpretação da veracidade da notícia.
- “e”: é uma conjunção muito comum na língua portuguesa, considerada um *stop word*, geralmente não carrega significado semântico específico, mas apresentou uma pontuação negativa.
- “Para”: também é uma *stop word* e é uma preposição multifuncional em português e também recebeu uma pontuação negativa.
- “ceagesp ”: A pontuação negativa atribuída a esta palavra pode indicar que a menção da organização Ceagesp pode estar associada a eventos negativos.

5.3.1.2 Notícia falsa classificada como real

Analisando a explicação do LIME na Figura 26b de uma notícia falsa classificada erroneamente como real, é possível observar:

- “A”, “O”, “de”: neste exemplo, é possível notar que palavras comuns que compõem a língua portuguesa foram destacadas com um peso positivos. O peso positivo dessas palavras por si só não fornece muitas informações sobre a classificação, mas isso pode nos levar a considerar que notícias reais são melhor escritas e que, em muitas notícias falsas, essas palavras básicas consideradas *stop word* são escassas na estrutura do texto. A presença dessas palavras com pesos positivos também pode ser explicada pelo fato de serem consideradas palavras “neutras” em termos de conteúdo informativo.

- “Bolsonaro” e “presidente”: estas palavras tiveram pesos positivos, oque pode indicar uma inclinação política.
- “um”: recebe uma pontuação negativa, mas muito pequena em comparação com as outras palavras.

5.3.1.3 Notícia real classificada como real

Com base na explicação do LIME na Figura 26c, referente a uma amostra de notícia real classificado como real:

- “Na”: recebe uma pontuação positiva alta. Neste contexto, pode indicar o início de uma frase ou introduzir uma localização temporal, contribuindo para a estruturação e coesão da notícia.
- “O” e “foi”: novamente palavras comuns da língua portuguesa receberam pontuações positivas
- “Lupa”: recebe uma pontuação positiva. Isso pode indicar a verificação de fatos realizada por uma fonte confiável, adicionando credibilidade à notícia.
- “Candidato”: recebe uma pontuação positiva. Sua presença indica a identidade do entrevistado, fornecendo informações essenciais para o entendimento da notícia.
- “Crivella”: A pontuação positiva, pode ser um indicativo de inclinação política.

5.3.1.4 Notícia real classificada como falsa

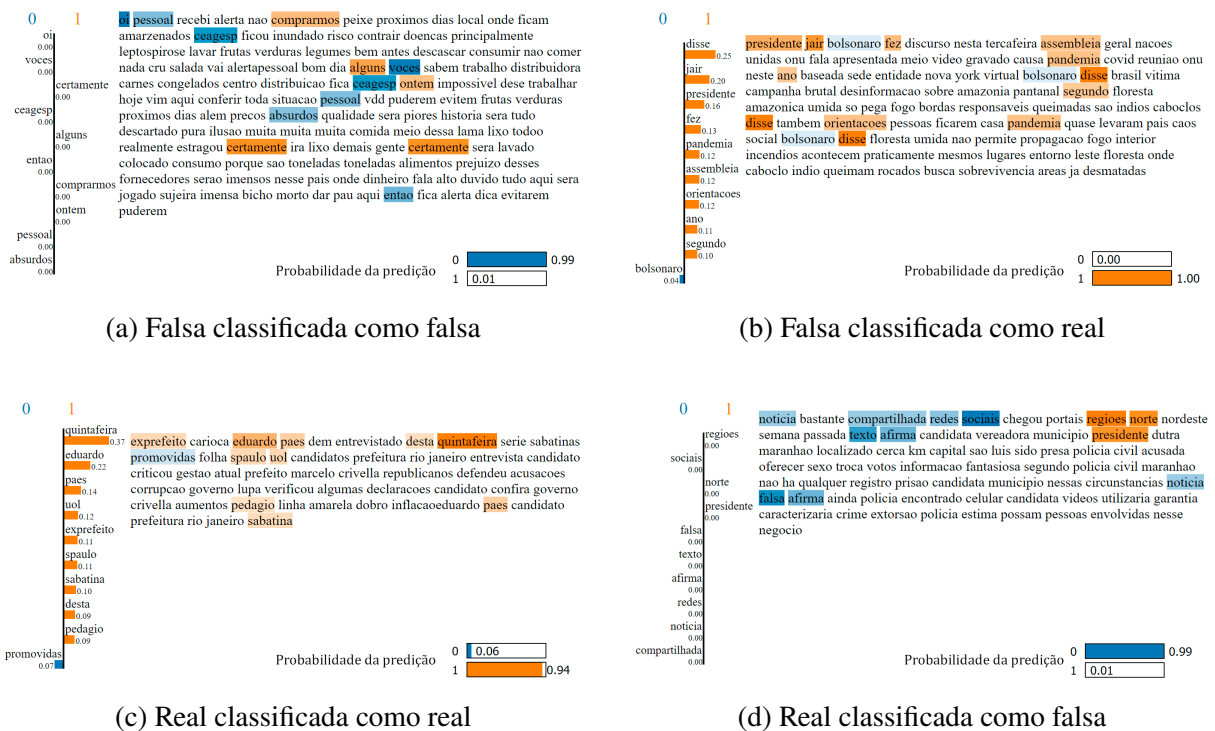
Finalmente, de acordo com a Figura 26d da explicação do LIME de uma notícia real classificada como falsa:

- “A”, “O”: dois artigos receberam uma pontuação positiva
- “na”: esta palavra é destacada mais uma vez no entanto agora recebeu uma pontuação negativa.
- “Segundo”: recebeu uma pontuação positiva, essa palavra geralmente introduz informações provenientes de fontes confiáveis, como no caso de declarações de autoridades policiais, aumentando a credibilidade da notícia.

- “Semana”: Essa palavra recebeu uma pontuação positiva, isso pode ser pelo fato de indicar um período de tempo em que a notícia foi compartilhada, fornecendo informações temporais a notícia a tornando mais verossímil.

5.3.2 Modelo 2: FakeRecogna sem *stop words*

Figura 27 – Explicações geradas pelo LIME do Modelo 2



5.3.2.1 Notícia falsa classificada como falsa

A explicação gerada pelo LIME na Figura 27a sobre a mesma amostra de notícia falsa classificada como falsa analisada anteriormente, no entanto sem o uso de *stop words* classificada pelo Modelo 2 gerou as seguintes observações:

- “ceagesp”: novamente o nome dessa organização recebeu uma pontuação negativa, o que pode indicar que essa organização seja alvos constantes de notícias falsas.
- “oi”: uma saudação comum da língua portuguesa recebeu uma grande pontuação negativa, neste contexto, isso pode se dar ao fato de que em uma notícia realizada por um canal responsável, não faz uso de linguagem informal.
- “pessoal” e “voces”: essa palavra também é utilizada em comunicação informal e também recebeu uma pontuação negativa.

- “certamente”: recebeu uma pontuação positiva, “certamente” é uma palavra de cunho positivo no entanto não é possível obter muitas informações do uso dessa palavra.

5.3.2.2 Notícias falsa classificada como real

Na figura 27b esta a análise realizada do LIME do uma notícia falsa que foi classificada erroneamente como real, é possível observar o seguinte:

- “presidente”: referência ao presidente da república teve um peso positivo.
- “jair” e “bolsonaro”: é curioso notar que o primeiro nome do ex-presidente “jair”, teve um peso positivo pra notícia, ou seja favorecendo a previsão como uma notícia real, enquanto o sobrenome “bolsonaro” teve um peso negativo. Isso talvez possa indicar que quando o ex presidente é alvo de notícias falsas, os veículos que as criam, utilizem mais o seu sobrenome apenas.
- “pandemia”: referência a pandemia do Covid teve um peso positivo para a notícia.
- “ano”: novamente uma referência temporal recebeu um peso positivo para a previsão.

5.3.2.3 Notícia real classificada como real

Na Figura 27c esta a análise realizada do LIME de uma notícia real que foi classificada como real, é possível observar o seguinte:

- “exprefeito”: referência a um político teve peso positivo para a notícia.
- “eduardo” e “paes”: referência ao nome do ex-prefeito do Rio de Janeiro obteve um peso positivo oque pode indicar certa inclinação política.
- “quinta-feira”: uma referência temporal obteve peso positivo para a previsão.
- “uol”: um veículo de notícias recebeu pontuação positiva.
- “promovidas”: recebeu um peso negativo bem pequeno mas não é possível tirar muitas conclusões dessa palavra.

5.3.2.4 Notícia real classificada como falsa

Analisando a figura 27d, que contem a análise realizada do LIME de uma notícia real foi classificada como falsa, é possível observar o seguinte:

- “compartilhada”, “redes”, “sociais”: referência a compartilhamento de notícias em redes sociais teve um peso negativo, isso pode se dar ao fato de que as notícias falsas são muito compartilhadas por meio dessas redes.
- “notícia”: a palavra notícia sozinha recebe um peso negativo, talvez isso possa indicar que as notícias falsas usem essa palavra explicitamente como forma de apelo para conseguirem atenção, no entanto neste caso a notícia é real e se refere a uma notícia falsa.
- “falsa”: essa palavra teve um alto peso negativo para a classificação, por se referir literalmente a notícia falsa, no entanto, a notícia é real e o assunto que ela abordava era uma notícia que foi confirmada como falsa.
- “presidente”: uma referência a figura política teve peso positivo novamente.
- “regiões” e “norte”: citar as regiões norte teve um peso positivo, isso pode indicar que informação de localidade pode trazer mais veracidade a notícia. Mas em contraponto a palavra nordeste não apresentou nenhum destaque.

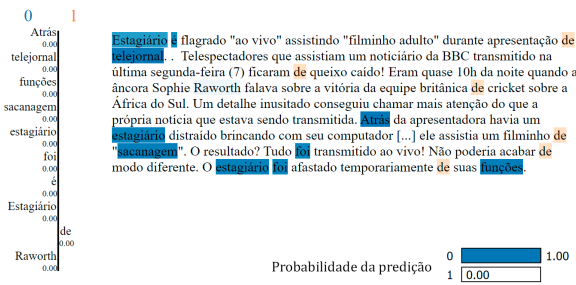
5.3.3 Modelo 3: Fake.Br incluindo *stop words*

5.3.3.1 Notícia falsa classificada como falsa

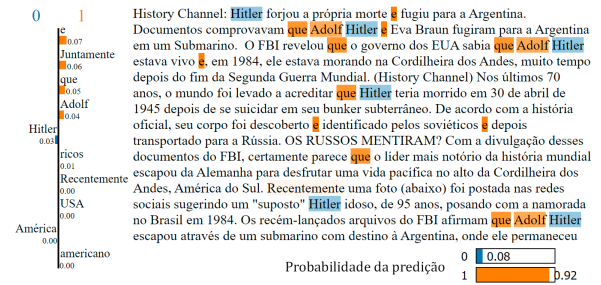
Analisando agora as explicações das amostras classificadas pelo Modelo 3, na Figura 28a esta a explicação gera pelo LIME de uma amostra de notícia falsa classificada como falsa do *dataset* Fake.Br Corpus com *stop words* gerou as seguintes observações:

- “estagiario”: a referência a essa entidade recebeu uma pontuação negativa no entanto não foi possível obter muitas percepções do porque disso.
- “telejornal”: recebe uma pontuação negativa, isso pode indicar que os criadores da notícias falsas podem ter incluído a palavra “telejornal” como parte de uma estratégia para enganar os leitores, induzindo-os a pensar que a informação é proveniente de uma fonte confiável.

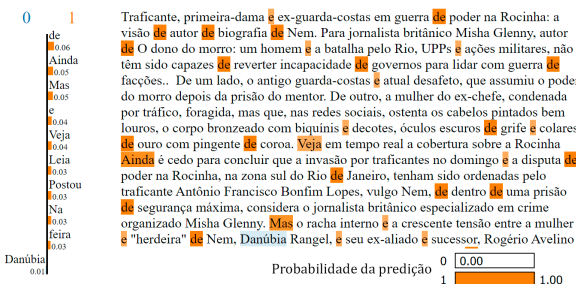
Figura 28 – Explicações geradas pelo LIME do Modelo 3



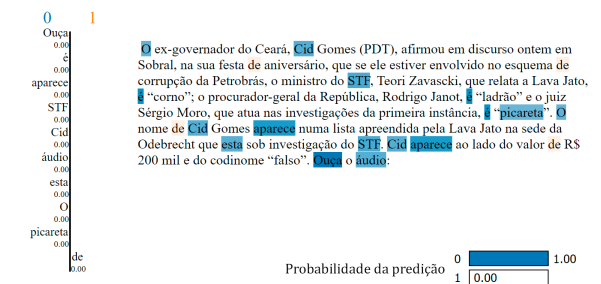
(a) Falsa classificada como falsa



(b) Falsa classificada como real



(c) Real classificada como real



(d) Real classificada como falsa

- “é”: uma palavra muito comum na língua portuguesa sendo considerada uma *stop word* que pode ser usada em uma variedade de contextos recebeu uma pontuação negativa. Em alguns casos, pode ser usada em afirmações que são factualmente verdadeiras, enquanto em outros casos pode ser usada para transmitir uma declaração falsa. A presença da palavra “é” em uma notícia falsa pode estar associada a afirmações enganosas que tentam parecer verdadeiras.
- “de”: Novamente uma *stop word* recebeu uma pontuação positiva.

5.3.3.2 Notícia falsa classificada como real

A explicação gerada pelo LIME na Figura 28b sobre uma amostra de notícia falsa classificada como real do *dataset* Fake.Br Corpus com *stop words* gerou as seguintes observações:

- “Adolf” e “Hitler”: a palavra “Hitler” recebeu uma pontuação negativa, essa palavra está fortemente associada a eventos históricos negativos. Portanto, sua presença em uma notícia pode ser interpretada como indicativa de conteúdo sensacionalista, polêmico ou potencialmente enganoso. Por outro lado, a palavra “Adolf”, por si só, pode não ter tantas conotações negativas e pode ser utilizada em contextos mais neutros ou até mesmo positivos, como referências a pessoas com esse nome que não estão relacionadas ao líder nazista.

- “que” e “e”: ambas as palavras são conectores comuns na língua portuguesa que são frequentemente usados para ligar frases e ideias em um texto. Sua presença pode indicar uma estrutura gramatical mais coesa e bem construída, o que é frequentemente associado a notícias reais e informativas.

5.3.3.3 Notícia real classificada como real

Na Figura 28c, está a explicação do LIME sobre a amostra de notícia real classificada como real:

- “de”, “e” e “na”: novamente palavras simples da língua portuguesa receberam uma pontuação positiva, o que pode reforçar a ideia de que as notícias reais são mais bem escritas gramaticalmente e fazem mais uso desses conectivos.
- “Veja”: a palavra “Veja” recebeu uma pontuação positiva, neste contexto é uma conjugação do verbo “ver”, mas a pontuação positiva devido ao fato de que “Veja” pode ser interpretado como uma referência à revista brasileira de mesmo nome, conhecida por seu jornalismo investigativo e confiável.
- “Danúbia”: um nome próprio recebeu uma pontuação negativa.

5.3.3.4 Notícia real classificada como falsa

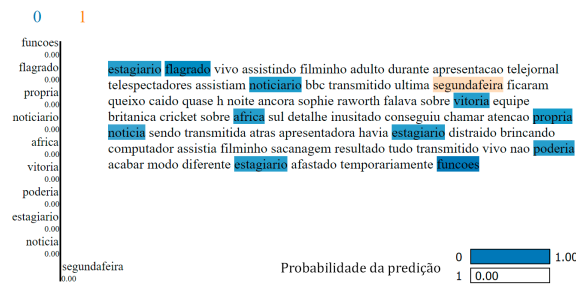
Analisando a amostra de notícia real classificada como falsa classificada pelo LIME na Figura 28d foi possível observar:

- “o”, “e”: neste caso a palavra “e” recebeu uma pontuação negativa, diferente da notícia anterior. A palavra “o” também recebeu uma pontuação negativa, sendo um artigo da língua portuguesa.
- “de”: como em outras notícias também recebeu uma pontuação positiva.
- “Cid”: referência a uma entidade política recebeu uma pontuação negativa.
- “STF”: referência ao Supremo Tribunal Federal, o que pode indicar que ele é constantemente cenário alvo de notícias falsas.

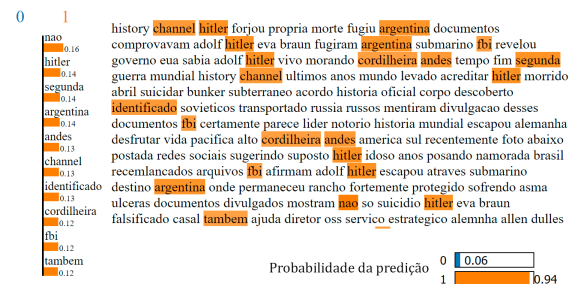
- “picareta”: recebeu peso negativo, isso pode se dar ao fato de que sua presença na notícia pode indicar um viés negativo ou hostilidade em relação ao assunto abordado, o que pode afetar a percepção da credibilidade da informação.

5.3.4 Modelo 4: Fake.Br sem stop words

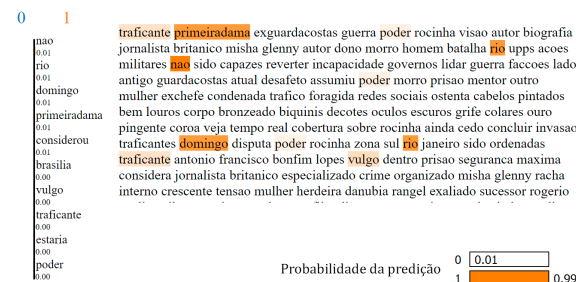
Figura 29 – Explicações geradas pelo LIME do Modelo 4



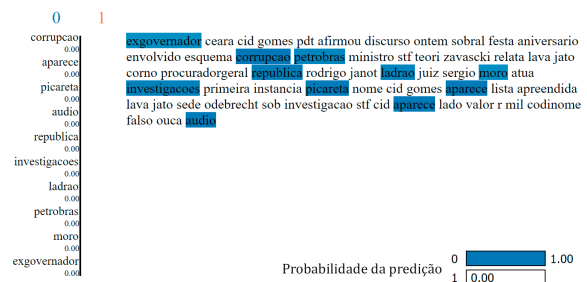
(a) Falsa classificada como falsa



(b) Falsa classificada como real



(c) Real classificada como real



(d) Real classificada como falsa

5.3.4.1 Notícia falsa classificada como falsa

Analisando as amostras classificadas pelo Modelo 4, na Figura 29a esta a explicação do LIME sobre a amostra de notícia falsa classificada corretamente, foi possível observar:

- “estagiário” e “telejornal”: novamente essas palavras receberam pontuações negativas como no modelo anterior.
- “noticiário”: é curioso a palavra “noticiário” assim como “telejournal” receber uma pontuação negativa, já que nesses veículos que as notícias reais costumam ser divulgadas.
- “africa”: a referência a um país recebeu uma pontuação negativa, isso pode indicar que esse país é constantemente alvo de notícias falsas.

- “segundafeira”: a palavra composta “segunda-feira”, com a remoção das *stop words* foi considerada como uma palavra simples “segundafeira”, essa palavra recebeu uma pontuação positiva, isso pode se dar ao fato de que notícias reais costumam ser mais precisas e informar quando um fato aconteceu.

5.3.4.2 Notícia falsa classificada como real

Com base na Figura 29b da explicação do LIME sobre a amostra de notícia falsa classificada como real, é possível observar:

- “hitler”: neste modelo a palavra Hitler recebeu uma pontuação positiva;
- “argentina”, “cordilheira” e “andes”: a referência a um país e a cordilheira dos andes recebeu pontuação positiva.
- “fbi”: referência ao serviço de inteligência e segurança doméstica dos Estados Unidos recebeu uma pontuação positiva.

5.3.4.3 Notícia real classificada como real

Na Figura 29c esta a explicação do LIME sobre a amostra de notícia real classificada como real:

- “primeiradama”: referência a primeira dama obteve uma pontuação positiva.
- “traficante”: é comum encontrar notícias divulgadas por fontes confiáveis sobre tráfico, por isso talvez essa palavra tenha um peso positivo, por se tratar de uma assunto que costuma estar presente em notícias reais.
- “não”: uma palavra de cunho negativo, recebeu uma pontuação positiva.

5.3.4.4 Notícia real classificada como falsa

A explicação do LIME sobre a amostra de notícia real classificada como real está na Figura 29c, com ela nota-se que:

- “exgovernador” e “moro”: referência a figuras políticas tiveram pesos negativos.

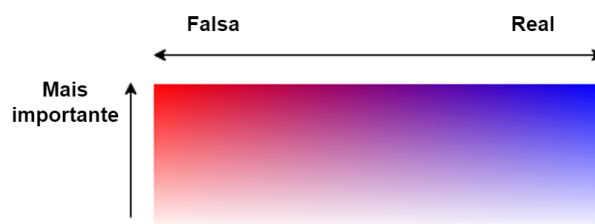
- “corrupção”: corrupção obteve um peso negativo, oque pode indicar que esse tema seja constantemente utilizado para espalhar notícias falsas sobre figuras políticas.
- “petrobras”: referência a empresa Petrobras obteve peso negativo oque pode indicar que ela é constante alvo de notícias falsas

5.4 Explicações geradas pelo IG

Nesta sessão encontram-se as explicações geradas pelo IG. Elas utilizam uma escala de cores para destacar o peso das palavras, como pode ser visto na Figura 30. Quanto mais vermelha a palavra, mais ela contribuiu para a classificação como notícia falsa, e quanto mais azul, mais ela contribuiu para a notícia ser classificada como real, sendo que quanto mais saturada a cor, mais importante era o peso da palavra para a classificação. A aplicação dessa escala de cores às amostras pode ser vista, por exemplo, na Figura 31.

É importante ressaltar que o IG retorna uma pontuação de relevância para cada subtoken na entrada. No entanto, para melhor visualização e comparação com o LIME, neste trabalho foi considerada uma pontuação por *token* (não *subtoken*). Para isso a máxima relevância entre todos os *subtokens* de um *token* foi considerada para representar o *token* inteiro.

Figura 30 – Escala de cores dos mapas de saliência



5.4.1 Modelo 1: FakeRecognia incluindo *stop words*

5.4.1.1 Notícia falsa classificada como falsa

Analisando as explicações das amostras classificadas pelo Modelo 1, na Figura 31a está uma notícia falsa classificada como falsa, é possível perceber por meio das suas cores e intensidades aplicadas pelo IG:

- “asbudos”: Essa palavra obteve um peso mais negativo, palavras que denotam exagero podem ser comuns em notícias falsas, pois muitas vezes elas tentam chamar a atenção do

Figura 31 – Explicações geradas pelo IG do Modelo 1

"oi pessoal, recebi um alerta para não comprarmos peixe nos próximos 10 dias, o local onde ficam armazenados na ceagesp ficou inundado com risco de contrair doenças principalmente a leptospirose. lavar as frutas verduras e legumes bem antes de descascar e consumir! não comer nada cru, nem salada. vai o alerta!: pessoal bom dia, como alguns de vocês sabem, trabalho numa distribuidora de carnes e congelados. nossa centro de distribuição fica no ceagesp, ontem impossível dese trabalhar, e hoje vim aqui conferir toda a situação. pessoal de vdd, se puderem, evitem frutas e verduras nos próximos 15 dias. além dos preços absurdos, a qualidade será a das piores. essa história que será tudo descartado é pura ilusão. tem muita, muita, muita comida no meio dessa lama e lixo todo. o que realmente estragou certamente irá para o lixo, mas os demais gente, certamente será lavado e colocado para consumo, porque são toneladas e toneladas de alimentos, e o prejuízo desses fornecedores serão imensos, e nesse país onde o dinheiro fala mais alto, duvido que tudo isso aqui será jogado fora. a sujeira é imensa, bicho morto a dar com pau aqui então fica o alerta e dica, para evitarem o que puderem".

(a) Falsa classificada como falsa

O presidente Jair Bolsonaro fez um discurso nesta terça - feira (22) na [UNK] Assembleia Geral das Nações Unidas (ONU). A fala foi apresentada por meio de um vídeo gravado. Por causa da pandemia de Covid - 19, a reunião da ONU neste ano, baseada na sede da entidade em Nova York, teve de ser virtual. Bolsonaro disse que o Brasil é " vítima " de uma campanha " brutal " de desinformação sobre a Amazônia e o Pantanal. Segundo ele, a floresta amazônica é úmida e só pega fogo nas bordas e os responsáveis pelas queimadas são " índios " e " caboclos ". Disse também que as orientações para as pessoas ficarem em casa na pandemia " quase " levaram o país ao " caos social ". Bolsonaro disse: " Nossa floresta é úmida e não permite a propagação do fogo em seu interior. Os incêndios acontecem praticamente nos mesmos lugares, no entorno leste da floresta, onde o caboclo e o índio queimam seus roçados em busca de sua sobrevivência, em áreas já desmatadas ".

(b) Falsa classificada como real

O ex - prefeito carioca Eduardo Paes (DEM) foi o entrevistado desta quinta - feira (5) na série de sabatinas promovidas pela Folha de S. Paulo e pelo UOL com os candidatos à Prefeitura do Rio de Janeiro. Na entrevista, o candidato criticou a gestão do atual prefeito, Marcelo Crivella (Republicanos), e se defendeu de acusações de corrupção em seu governo. A Lupa verificou algumas das declarações do candidato. Confira: No [governo] Crivella, [os aumentos do pedágio da Linha Amarela] foram o dobro da inflação Eduardo Paes, candidato à prefeitura do Rio de Janeiro, na sabatina

(c) Real classificada como real

Uma notícia foi bastante compartilhada nas redes sociais e chegou a portais das regiões Norte e Nordeste na semana passada. O texto afirma que uma candidata a vereadora do município de Presidente Dutra, no Maranhão, localizado a cerca de 350 km da capital São Luís, teria sido presa pela Polícia Civil acusada de oferecer sexo em troca de votos. A informação é fantasiosa. Segundo a Polícia Civil do Maranhão, não há qualquer registro de prisão de uma candidata do município nessas circunstâncias. A notícia falsa afirma ainda que a polícia teria encontrado no celular da candidata vídeos que ela utilizaria " como garantia ", o que caracterizaria como crime de extorsão, e que " a polícia estima [...] que possam ser mais de 250 pessoas envolvidas nesse ' negócio ' ".

(d) Real classificada como falsa

leitor com afirmações surpreendentes ou incríveis.

- “armazenados” e “descascar”: essas palavras tiveram grande destaque, no entanto, não é possível tirar muitas conclusões do uso dessas palavras
- “lavar”: esta palavra obteve um peso positivo, pode ser porque esta palavra é muito utilizada em notícias reais que falem sobre corrupção e lavagem de dinheiro, apesar de que nesta notícia essa palavra é aplicada a outro contexto
- “oi”: obteve um peso positivo importante, pode estar correlacionada com outras palavras ou estruturas linguísticas que são mais frequentes em notícias verdadeiras do que em falsas, no entanto não é possível tirar uma conclusão de qual correlação é essa.
- “prejuízo”: esta palavra obteve um peso positivo, e pode estar associada a notícias financeiras ou econômicas legítimas.
- “leptospirose”: assim como nas explicações do LIME essa palavra teve um pontuação negativa.

5.4.1.2 Notícia falsa classificada como real

Na Figura 31b esta uma notícia falsa que foi classificada como real, com a explicação gerada pelo IG é possível observar que a maioria das palavras tiveram pesos bem semelhantes, estando ao centro da escala de cores, apresentando uma cor roxa, mas foi possível destacar o seguinte:

- “O”: Essa palavra é uma *stop word* teve um peso positivo alto. Isso pode reforçar a teoria de que as notícias reais são melhor escritas e por isso possuem mais palavras como esta.
- “Segundo”: essa palavra também recebeu uma pontuação positiva maior. O termo “segundo” pode ser usado para introduzir citações ou informações provenientes de fontes confiáveis assim como foi utilizado nesta notícia. Desse modo pode ser que esta palavra seja utilizada em maior número por notícias reais.
- “Bolsonaro”: uma figura política teve um grau de importância um pouco maior em relação as outras palavras como é possível observar pela maior saturação de cor, tendo um peso mais positivo. Esta característica pode indicar uma inclinação política.

5.4.1.3 Notícia real classificada como real

Na Figura 31c esta uma notícia real classificada real, pode se observar a partir dela o seguinte:

- “inflaçãoEduardo”: duas palavras foram consideradas como um único *token* “inflação” e “Eduardo”, tendo um peso mais positivo e maior importância para a predição da amostra
- “UOL”: essa palavra também recebeu uma pontuação positiva e apresentou grande importância para a classificação assim como na explicação gerada pelo LIME anteriormente. Isso pode se dar ao fato que UOL é uma das maiores empresas de mídia online no Brasil e é conhecida por seu jornalismo confiável, logo citar essa empresa pode trazer credibilidade a notícia.
- “Confira”: a palavra “Confira” é comumente usado em chamadas de notícias, sua pontuação positiva pode indicar que usar este tipo de chamada seja mais utilizado em notícias legítimas.

5.4.1.4 Notícia real classificada como falsa

A notícia real classificada erroneamente na Figura 31d não apresentou muitos destaques na explicação gerada pelo IG, as palavras apresentaram tons muito semelhantes.

- “Uma”: esta palavra foi a que apresentou maior importância, e teve um peso positivo.
- A maioria das palavras apresentou uma pontuação mais negativa, com uma importância mediana como pode ser notado pelo tom mais claro.

5.4.2 Modelo 2: FakeRecogna sem *stop words*

Figura 32 – Explicações geradas pelo IG do Modelo 2

oi pessoal recebi alerta nao compramos peixe proximos dias local onde ficam
armazenados ceagesp ficou inundado risco contrair doencas principalmente
leptospirose lavar frutas verduras legumes bem antes descascar consumir nao comer
nada cru salada vai alertapessoal bom dia alguns voces sabem trabalho distribuidora
carnes congelados centro distribuicao fica ceagesp ontem impossivel dese trabalhar
hoje vim aqui conferir toda situacao pessoal vdd puderem evitem frutas verduras
proximos dias alem precos absurdos qualidade sera piores historia sera tudo
descartado pura ilusao muita muita comida meio dessa lama lixo todo
realmente estragou certamente ira lixo demais gente certamente sera lavado
colocado consumo porque sao toneladas toneladas alimentos prejuizo desses
fornecedores serao imensos nesse pais onde dinheiro fala alto duvido tudo aqui sera
jogado sujeira imensa bicho morto dar pau aqui entao fica alerta dica evitarem
puderem

(a) Falsa classificada como falsa

presidente jair bolsonaro fez discurso nesta tercafeira assembleia geral nacoes
unidas onu fala apresentada meio video gravado causa pandemia covid reuniao onu
neste ano baseada sede entidade nova york virtual bolsonaro disse brasil vitima
campanha brutal desinformacao sobre amazonia pantanal segundo floresta
amazonica umida so pega fogo bordas responsaveis queimadas sao indios caboclos
disse tambem orientacoes pessoas fiquem casa pandemia quase levaram pais caos
social bolsonaro disse floresta umida nao permite propagacao fogo interior
incendios acontecem praticamente mesmos lugares entorno leste floresta onde
caboclo indio queimam roçados busca sobrevivencia areas ja desmatadas

(b) Falsa classificada como real

exprefeito carioca eduardo paes dem entrevistado desta quintafeira serie sabatinas
promovidas folha spaulo uol candidatos prefeitura rio janeiro entrevista candidato
criticou gestao atual prefeito marcelo crivella republicanos defendeu acusacoes
corrupcao governo lupa verificou algumas declaracoes candidato confira governo
crivella aumentos pedagio linha amarela dobro inflacaoeduardo paes candidato
prefeitura rio janeiro sabatina

(c) Real classificada como real

noticia bastante compartilhada redes sociais chegou portais regioes norte nordeste
semana passada texto afirma candidata vereadora municipio presidente dutra
maranhao localizado cerca km capital sao luis sido presa policia civil acusada
oferecer sexo troca votos informacao fantasiosa segundo policia civil maranhao nao
ha qualquer registro prisao candidata municipio nessas circunstancias noticia falsa
afirma ainda policia encontrado celular candidata videos utilizaria garantia
caracterizaria crime extorsao policia estima possam pessoas envolvidas nesse
negocio

(d) Real classificada como falsa

5.4.2.1 Notícia falsa classificada como falsa

Analisando a amostra de notícia falsa classificada como falsa pelo Modelo 2 na Figura 32a foi possível observar:

- “alertapessoal”: esta palavra apresentou maior importância é a palavra com maior saturação, no entanto ela aparentou um tom roxo, ou seja, esta entre o peso falso e real.

- A maioria das palavras não apresentaram muita importância pra classificação, estando num nível de saturação muito próximo.

5.4.2.2 Notícia falsa classificada como real

Na Figura 32b esta a amostra de notícia falsa classificada como real, é possível notar:

- “video”: esta palavra recebeu uma pontuação mais positiva e grande importância. A presença da palavra “vídeo” pode indicar que a notícia inclui evidências visuais na forma de um vídeo e por isso recebe uma pontuação positiva.
- “york”: referência a uma cidade teve uma pontuação mais positiva, a importância atribuída a essa palavra, pode se dar ao fato de que notícias reais costumam ser mais completas, trazendo informações como local em que ocorreram.
- “brutal”: esta palavra teve um peso bem positivo, no entanto não apresentou muita importância.

5.4.2.3 Notícia real classificada como real

Na Figura 32c esta a notícia real classificada de forma correta, de acordo com ela é possível notar que:

- “exprefeito”: essa palavra foi a que apresentou maior importância, neste caso uma referência a um cargo político apresentou uma pontuação positiva para a classificação.
- “promovidas”: apresentou uma pontuação mais negativa, isso pode estar associado a narrativas que envolvem promoções ou avanços falsos ou enganosos.
- “quinta-feira” e “janeiro”: essas palavras não apresentaram grande importância, mas tiveram peso positivo, isso pode reforçar a ideia de que notícias reais trazem mais informações, logo uma informação temporal aparenta uma pontuação positiva,

5.4.2.4 Notícia real classificada como falsa

Na Figura 32d esta a notícia real classificada como falsa, é possível notar que:

- “compartilhada”: essa palavra foi umas das que apresentou maior importância, tendo uma pontuação mais negativa. Em muitos casos, notícias falsas ou desinformação são compartilhadas amplamente em redes sociais ou outras plataformas online. O termo “compartilhada” pode estar associado a esse tipo de disseminação de informações enganosas, levando o modelo a atribuir uma pontuação negativa à palavra.
- “envolvidas”: esta teve uma pontuação mais equilibrada entre real e falsa, no entanto apresentou alta importância.
- “votos”: teve um peso negativo, isso pode se dar ao fato de que em muitos casos, notícias falsas estão relacionadas a processos eleitorais, votações ou políticas.

5.4.3 Modelo 3: Fake.Br incluindo *stop words*

Figura 33 – Explicações geradas pelo IG do Modelo 3

Estagiário é flagrado "ao vivo" assistindo "filminho adulto" durante apresentação de telejornal. . . Telespectadores que assistiam um noticiário da BBC transmitido na última segunda - feira (7) ficaram de queixo caído ! Eram quase 10h da noite quando a âncora Sophie Raworth falava sobre a vitória da equipe britânica de cricket sobre a África do Sul . Um detalhe inusitado conseguiu chamar mais atenção do que a própria notícia que estava sendo transmitida . Atrás da apresentadora havia um estagiário distraído brincando com seu computador [. . .] ele assistia um filminho de " sacanagem " . O resultado ? Tudo foi transmitido ao vivo ! Não poderia acabar de modo diferente . O estagiário foi afastado temporariamente de suas funções .

(a) Falsa classificada como falsa

History Channel : Hitler forjou a própria morte e fugiu para a Argentina . Documentos comprovavam que Adolf Hitler e Eva Braun fugiram para a Argentina em um Submarino . O FBI revelou que o governo dos EUA sabia que Adolf Hitler estava vivo e , em 1984 , ele estava morando na Cordilheira dos Andes , muito tempo depois do fim da Segunda Guerra Mundial . (History Channel) Nos últimos 70 anos , o mundo foi levado a acreditar que Hitler teria morrido em 30 de abril de 1945 depois de se suicidar em seu bunker subterrâneo . De acordo com a história oficial , seu corpo foi descoberto e identificado pelos soviéticos e depois transportado para a Rússia . OS RUSSOS MENTIRAM ? Com a divulgação desses documentos do FBI , certamente parece que o líder mais notório da história mundial escapou da Alemanha para desfrutar uma vida pacífica no alto da Cordilheira dos Andes , América do Sul . Recentemente uma foto (abaixo) foi postada nas redes sociais sugerindo um " suposto " Hitler idoso , de 95 anos , posando com a namorada no Brasil em 1984 . Os recém - lançados arquivos do FBI afirmam que Adolf Hitler escapou através de um submarino com destino à Argentina , onde ele

(b) Falsa classificada como real

Traficante , primeira - dama e ex - guarda - costas em guerra de poder na Rocinha : a visão de autor de biografia de Nem . Para jornalista britânico Misha Glenny , autor de O dono do morro : um homem e a batalha pelo Rio , UPPs e ações militares , não têm sido capazes de reverter incapacidade de governos para lidar com guerra de facções . De um lado , o antigo guarda - costas e atual desafeto , que assumiu o poder do morro depois da prisão do mentor . De outro , a mulher do ex - chefe , condenada por tráfico , foragida , mas que , nas redes sociais , ostenta os cabelos pintados bem louros , o corpo bronzeado com biquínis e decotes , olhos escuros de grife e colares de ouro com pingente de coroa . Veja em tempo real a cobertura sobre a Rocinha Ainda é cedo para concluir que a invasão por traficantes no domingo e a disputa de poder na Rocinha , na zona sul do Rio de Janeiro , tenham sido ordenadas pelo traficante Antônio Francisco Bonfim Lopes , vulgo Nem , de dentro de uma prisão de segurança máxima , considera o jornalista britânico especializado em crime organizado Misha Glenny . Mas o racha interno e a crescente tensão entre a mulher e " herdeira " de Nem , Danúbia Rangel , e seu ex -

(c) Real classificada como real

O ex - governador do Ceará , Cid Gomes (PDT) , afirmou em discurso ontem em Sobral , na sua festa de aniversário , que se ele estiver envolvido no esquema de corrupção da Petrobrás , o ministro do STF , Teori Zavascki , que relata a Lava Jato , é " corno " ; o procurador - geral da República , Rodrigo Janot , é " ladrão " e o juiz Sérgio Moro , que atua nas investigações da primeira instância , é " picareta " . O nome de Cid Gomes aparece numa lista apreendida pela Lava Jato na sede da Odebrecht que esta sob investigação do STF . Cid aparece ao lado do valor de R \$ 200 mil e do codinome " falso " . Ouça o áudio :

(d) Real classificada como falsa

5.4.3.1 Notícia falsa classificada como falsa

Analisando agora as amostras classificadas pelo Modelo 3, começando pela notícia falsa classificada como falsa na Figura 33a foi possível notar:

- As palavras tiveram um nível de importância muito próximos.

- “telejornal”: essa palavra recebeu uma pontuação negativa, isso pode indicar que o modelo associou “telejornal” a fontes de informação que foram desacreditadas no passado por disseminar desinformação.

5.4.3.2 Notícia falsa classificada como real

Analisando a notícia falsa classificada como real na Figura 33b foi possível notar:

- “History”: é estranho notar que esta palavra teve importância no entanto de duas formas, em uma teve um peso mais positivo e em outra um peso mais negativo.
- “FBI”: novamente esta agência teve uma pontuação positiva, a citação dessa organização entrega mais confiabilidade a notícias.
- “forjou” e “Mentiram”: duas palavras que estão associadas a atividades enganosas tiveram importância na predição, no entanto tiveram pontuações positivas, isso pode ser ao fato de que em algumas situações esses termos podem indicar transparência e responsabilidade jornalística, indicando que a história está expondo atividades fraudulentas ou enganosas.

5.4.3.3 Notícia real classificada como real

Analisando na Figura 33c a notícia real classificada como real foi possível notar:

- Nesta notícia as palavras apresentaram pesos bem semelhantes,
- “dama”: a pontuação negativa atribuída a esta palavra pode ser explicada pela associação com o termo “primeira-dama” (que é o termo utilizado na amostra), que frequentemente figura em notícias sensacionalistas. Essa associação sugere que a figura política em questão pode ser alvo de rumores ou falsas narrativas.
- “Misha Glenny”, “Danúbia Rangel”, “Antonio Francisco Lopes”: esses nomes foram associados a pontuações negativas, pode ser que eles sejam constantemente alvos de notícias falsas, ou que normalmente notícias falsas tem como alvos pessoas de forma geral, e por isso o uso de nomes próprios recebe esta pontuação.

5.4.3.4 Notícia real classificada como falsa

Analisando na Figura 33d a notícia real classificada como real foi possível notar:

- “Sobral”: esta palavra teve uma pontuação negativa, ela se refere a uma cidade com este nome na notícia, no entanto Sobral é um sobrenome comum no Brasil, e pode ser que o modelo associou um nome próprio a uma notícia falsa.
- “investigação”: a menção desta palavra em uma notícia pode sugerir que há uma busca ativa pela verdade e uma abordagem responsável para descobrir informações precisas. Isso pode aumentar a confiança dos leitores na veracidade da notícia, pois sugere que há um esforço diligente para verificar os fatos, por isso a palavra recebe uma pontuação positiva.
- “ladrão”: uma palavra de conotação negativa, teve uma pontuação negativa, isso pode indicar que essa palavra seja muito utilizada para acusar injustamente pessoas em notícias falsas. No entanto, crimes são comumente noticiados, e a palavra ladrão não deveria apresentar uma pontuação tao negativa.

5.4.4 Modelo 4: Fake.Br sem *stop words*

Figura 34 – Explicações geradas pelo IG do Modelo 4



5.4.4.1 Notícia falsa classificada como falsa

Analisando agora as amostras classificadas pelo Modelo 4, começando pela notícia falsa classificada como falsa na Figura 34a foi possível notar:

- “atenção”: essa palavra teve uma pontuação negativa, se a palavra “atenção” é usada em um contexto sensacionalista ou alarmante, isso pode indicar que a notícia é falsa. Notícias que buscam chamar a atenção do leitor de forma exagerada ou enganosa podem ser vistas como menos confiáveis.
- “não”: essa palavra recebeu uma pontuação negativa, é uma palavra de negação no entanto o seu destaque com um peso negativo pode não significar muita coisa.

5.4.4.2 Notícia falsa classificada como real

Analisando uma notícia falsa classificada como falsa na Figura 34b foi possível observar:

- “forjou”, “mentiram”: neste modelo essas palavras também receberam uma pontuação negativa.
- “governo”: essa palavra recebeu uma pontuação positiva, isso pode se dar pelo fato de que o termo “governo” muitas vezes se refere às instituições governamentais responsáveis por fornecer informações oficiais e confiáveis sobre política.

5.4.4.3 Notícia real classificada como real

Analisando uma notícia falsa classificada como falsa na Figura 34c foi possível observar:

- “primeiradama”: neste modelo sem *stop words* é interessante notar que a palavra “primeiradama” (primeira dama) teve um peso positivo.
- “exguardacostas”: essa palavra recebeu uma pontuação positiva, o uso de descrições detalhadas e específicas pode contribuir para uma análise mais precisa e informativa.
- “domingo”: novamente uma palavra de referência temporal recebeu uma pontuação positiva.
- “glenny”: dessa vez um sobrenome apenas recebeu uma pontuação negativa.

5.4.4.4 Notícia real classificada como falsa

Na Figura 34d esta a notícia real classificada de forma errada, com ela foi possível observar:

- “exgovernador”: referência a uma figura política teve uma pontuação negativa.

- “corrupção”: essa palavra teve grande importância e apresentou peso negativo, isso pode se dar ao fato de que muitas notícias falsas acusam políticos de corrupção devido inclinações políticas.
- “odebrecht”: essa palavra recebeu uma pontuação positiva e um grande nível de importância, isso pode ser porque a Odebrecht é uma empresa que esteve envolvida em diversos escândalos de corrupção em vários países, o que pode gerar desconfiança em relação a notícias que mencionam o nome da empresa.

5.5 Análises dos datasets

LIME e IG têm algumas limitações quando usados com textos, pois focam apenas nos níveis morfológico e lexical do texto. Outros níveis também são importantes, como o nível morfossintático, que nos permite identificar se algumas etiquetas ocorrem com mais frequência. Além disso, o nível semântico nos permite entender que algumas entidades são distribuídas de maneira diferente. Além disso, o nível retórico pode considerar que textos mais longos são menos propensos a serem notícias falsas.

Em vista dessa limitação nesta seção, serão comentadas as análises realizadas sobre cada um dos *datasets*, utilizando a biblioteca spaCy (HONNIBAL; MONTANI, 2017) para verificar se algumas das hipóteses levantadas são verdadeiras, sendo elas: notícias reais possuem citam mais organizações, e utilizam mais *stop words*, enquanto notícias falsas citam mais indivíduos principalmente políticos.

Por meio do spaCy, foi possível obter dos textos das notícias as quantidades de entidades, de classes gramaticais e as palavras mais frequentes. Com estes dados, gráficos foram gerados para os conjuntos de dados FakeRecogna e Fake.Br Corpus, considerando as variações com e sem *stop words*, e podem ser vistos nas Figuras 35 à 38

As figuras incluem gráficos com essas informações normalizadas pela média de palavras presentes nas notícias de cada classe do conjunto de dados em questão. É importante enfatizar que as notícias reais são em média, mais longas do que as falsas como exibido na Tabela 9. A média de palavras em notícias falsas que incluem *stop words* é de 96,24, e em notícias reais é de 166,89, enquanto a média de palavras em notícias falsas sem *stop words* é 56,06 e nas notícias reais é de 97,20 (média dos dois conjuntos de dados combinados).

Tabela 9 – Tamanho médio das notícias dos *datasets*

Dataset	Com Stop words	Classe	Média de caracteres	Média de palavras
FakeRecogna	Sim	Falsa	569,70	95,53
		Real	1010,45	163,42
Fake.Br Corpus	Sim	Falsa	1924,84	317,84
		Real	7233,30	1200,15
FakeRecogna + Fake. Br Corpus	Sim	Falsa	574,01	96,24
		Real	1031,29	166,89
FakeRecogna	Não	Falsa	420,53	55,66
		Real	756,90	95,29
Fake.Br Corpus	Não	Falsa	1387,16	180,53
		Real	4997,71	636,52
FakeRecogna + Fake. Br Corpus	Não	Falsa	423,60	56,06
		Real	771,81	97,20

Na Tabela 10 se encontram as siglas dos tipos de entidades identificadas pelo spaCy, juntamente com seus respectivos significados, e na Tabela 11 estão as siglas das classes gramaticais.

Tabela 10 – Sigla das entidades do spaCy

Sigla	Significado
LOC	Locais que nao sejam entidades geopolíticas, cadeias de montanhas, corpos de água
ORG	
PER	
MISC	Entidades diversas, por ex. eventos, nacionalidades, produtos ou obras de arte

Tabela 11 – Siglas das classes gramaticais do spaCy

Sigla	Significado	Sigla	Significado
PROUN	Pronome	NUM	Número
VERB	Verbo	PROPN	Substantivo Próprio
ADV	Adverbo	ADJ	Adjetivo
DET	Determinantes	CCONJ	Conjunção Coordenada
NOUN	Substantivo	X	Outros
AUX	Verbo auxiliar	SYM	Símbolo
SCONJ	Conjunção Subordinada	INTJ	Interjeição
PUNCT	Pontuação	PART	Partícula
ADP	Adposição		

5.5.1 FakeRecogna incluindo *stop words*

No *dataset* FakeRecogna com *stop words*, a suposição de que as notícias reais mencionam mais organizações se mostra verdadeira, como evidenciado no gráfico 35b. É possível observar que essa distribuição não se limita apenas às organizações, mas também abrange todos os outros tipos de entidades, tendo uma maior diferença de distribuição nas entidades do tipo LOC e ORG.

No entanto, de acordo com o gráfico de classes gramaticais 35d, é mostrado que a quantidade de substantivos (NOUN) utilizados pelas notícias falsas é um pouco maior, enquanto a quantidade de substantivos próprios (PROPN) é maior nas notícias reais. No mesmo gráfico, é possível ver que as notícias falsas também fazem um uso maior de adjetivos e verbos.

Quanto às palavras mais frequentes no *dataset* sem tratamento, são as *stop words*. Havia a suposição de que as notícias reais possuíam mais *stop words*, no entanto, pode-se perceber que o uso delas nas notícias reais e falsas não apresentou grande diferença.

5.5.2 FakeRecogna sem *stop words*

No *dataset* em que foi realizada a remoção das *stop words*, é possível notar no gráfico 36b que foi registrado um numero maior de organizações (ORG) e de entidades diversas (MISC) nas notícias falsas, apresentando uma diferença mínima em relação as notícias reais.

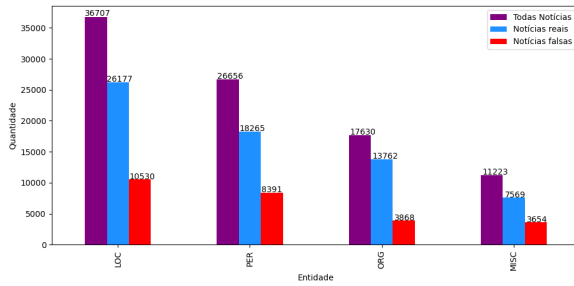
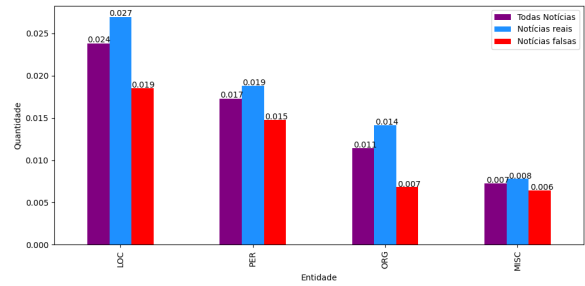
O gráfico 36d, que apresenta a quantidade de classes gramaticais, também mostrou uma diferença mínima entre as notícias reais e falsas.

No gráfico 36f de palavras mais frequentes, é possível notar que a palavras “não” esta muito presente em notícias falsas, enquanto as palavras “covid” e “saúde” esta mais presente nas notícias reais. Isso pode sinalizar que no *dataset*, as notícias reais trazem mais temas relacionados a saúde, enquanto as notícias falsas focam mais em abordar o cenário político.

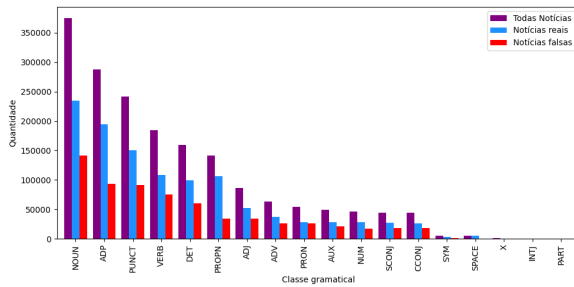
5.5.3 Fake.Br incluindo *stop words*

No *dataset* Fake.Br Corpus com *stop words*, no gráfico 37b, é possível notar que as notícias falsas fazem maior uso de entidades do tipo pessoa ou família nomeada (PER), o que pode indicar que as notícias falsas normalmente se tratam de pessoas. Também houve um maior uso de entidades diversas.

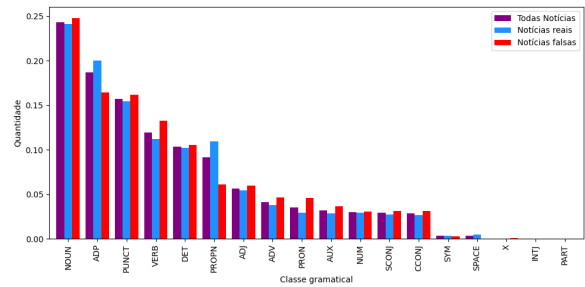
E no gráfico 37d de quantidade de classes gramaticais, é possível ver um maior uso de

Figura 35 – Métricas *Dataset FakeRecogna* com *stop words*(a) Quantidade de entidades em todo o *dataset*

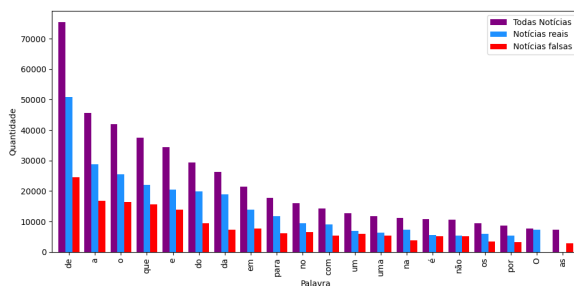
(b) Quantidade de entidades normalizada



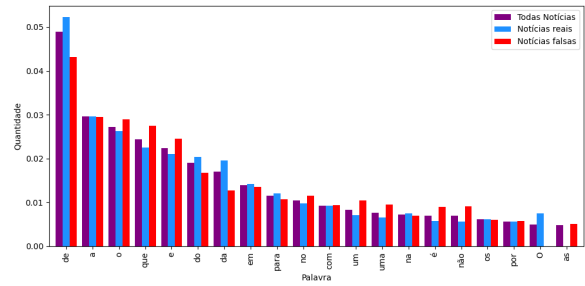
(c) Quantidade Classes Gramaticais



(d) Quantidade Classes Gramaticais Normalizada



(e) Quantidade das palavras mais frequentes



(f) Quantidade das palavras mais frequentes normalizada

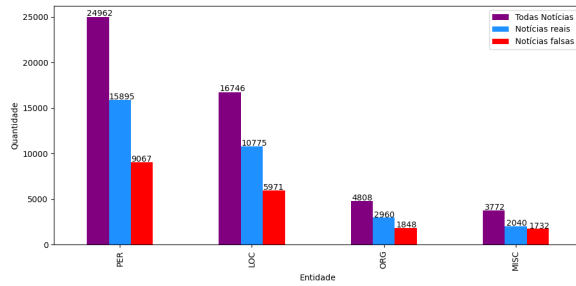
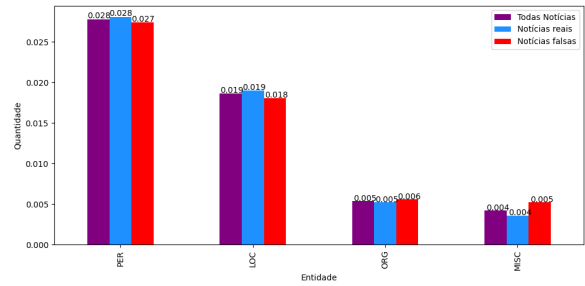
substantivos próprios (PROPN), o que reforça a afirmação anterior.

Quanto ao uso das *stop words* entre notícias reais e falsas, aparentou estar bem equilibrado, como pode ser visto no Gráfico 37f.

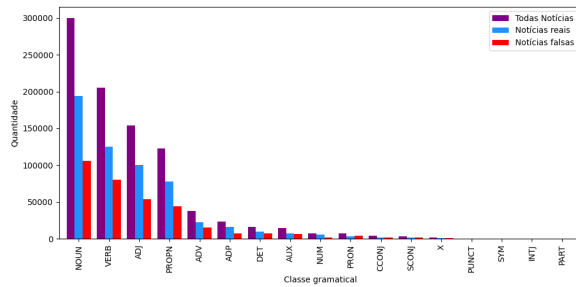
5.5.4 Fake.Br sem *stop words*

Quanto ao *dataset* Fake.Br Corpus sem as *stop words*, o uso das entidades do tipo locais (LOC) nas notícias falsas teve um aumento, como exibido no gráfico 38b.

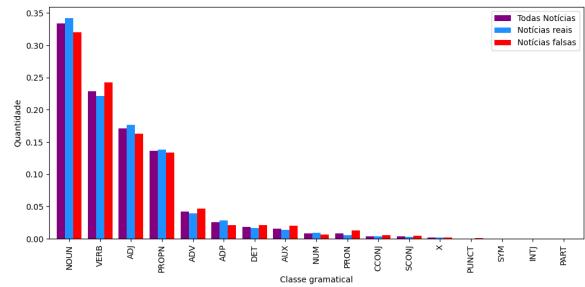
O uso das classes gramaticais manteve-se bem distribuído, como pode ser visualizado no gráfico 38d. Quanto às palavras mais frequentes, diferente do *dataset* FakeRecogna, ambas as categorias de notícias apresentaram uma quantidade similar da palavra “não”, que foi a palavra

Figura 36 – Métricas *Dataset FakeRecogna* sem *stop words*(a) Quantidade de entidades em todo o *dataset*

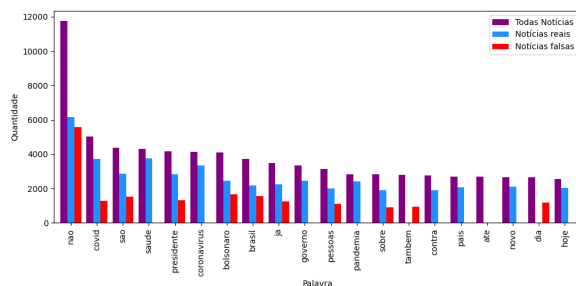
(b) Quantidade de entidades normalizada



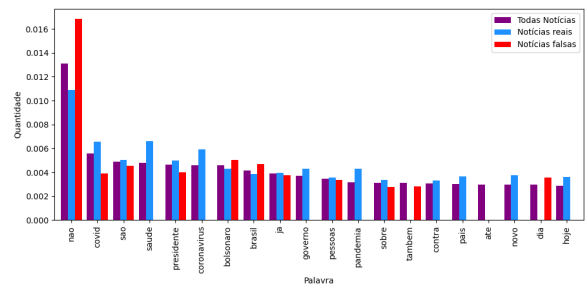
(c) Quantidade Classes Gramaticais



(d) Quantidade Classes Gramaticais Normalizada



(e) Quantidade das palavras mais frequentes

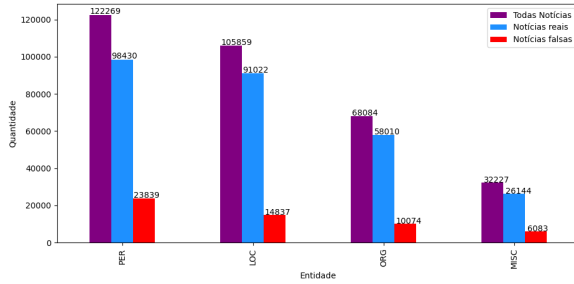
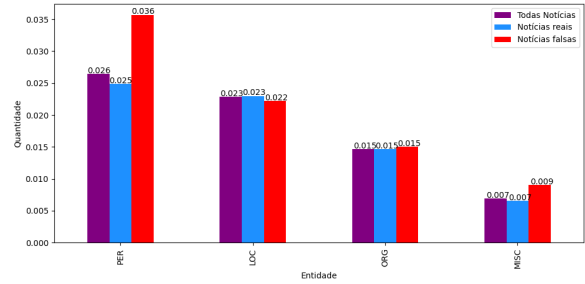


(f) Quantidade das palavras mais frequentes normalizada

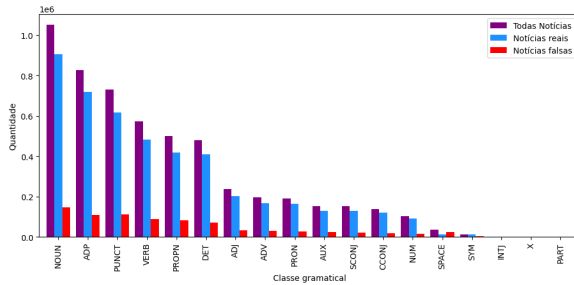
mais utilizada. Neste *dataset*, as palavras “lula” e “temer” tiveram um uso muito maior nas notícias falsas, o que pode indicar uma inclinação política. A palavra “presidente” também demonstrou mais uso nas notícias falsas. Outras palavras de cunho político também apresentaram uma frequência maior em notícias falsas, como “Brasil” e “país”. No entanto, as palavras “governo” e “federal” demonstraram um uso bem parecido entre notícias reais e falsas.

5.6 Considerações Finais

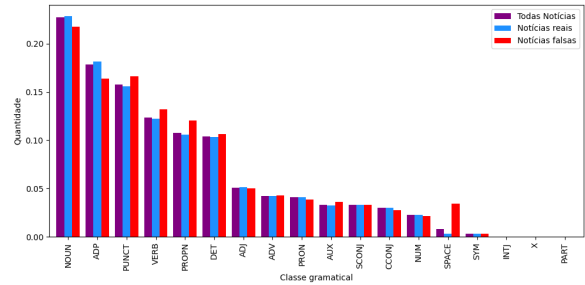
No estudo realizado, observou-se que tanto o IG quanto o LIME exibem comportamentos semelhantes ao destacar palavras significativas nas amostras. No entanto, o IG mostrou-se mais desafiador na interpretação textual, devido às sutis diferenças de importância entre as palavras

Figura 37 – Métricas *Dataset Fake.Br* Corpus com *stop words*(a) Quantidade de entidades em todo o *dataset*

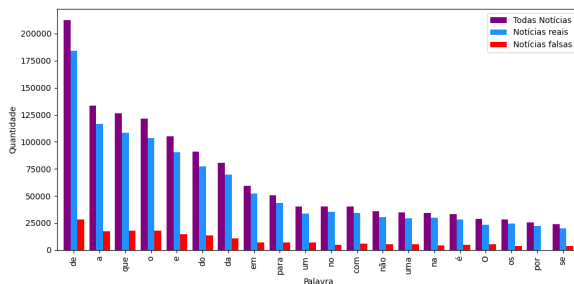
(b) Quantidade de entidades normalizada



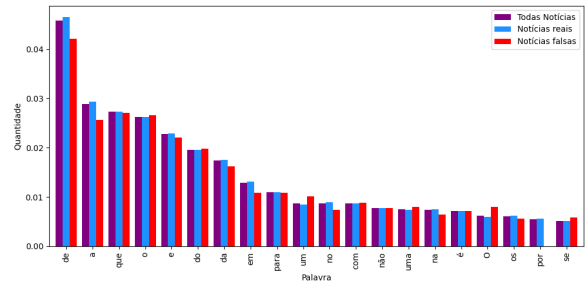
(c) Quantidade Classes Gramaticais



(d) Quantidade Classes Gramaticais Normalizada



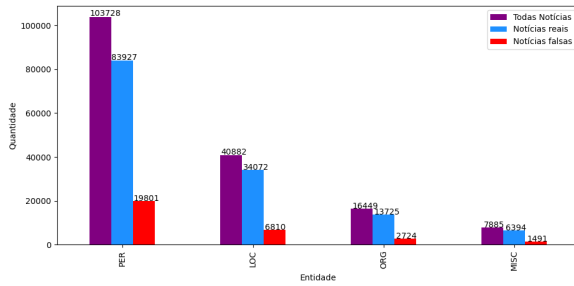
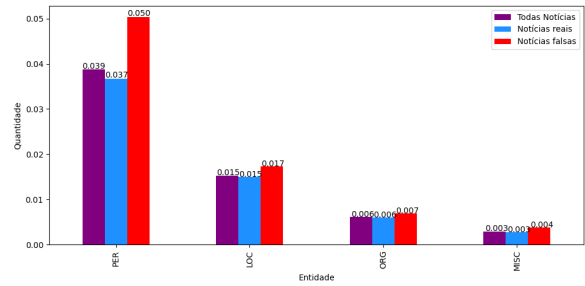
(e) Quantidade das palavras mais frequentes



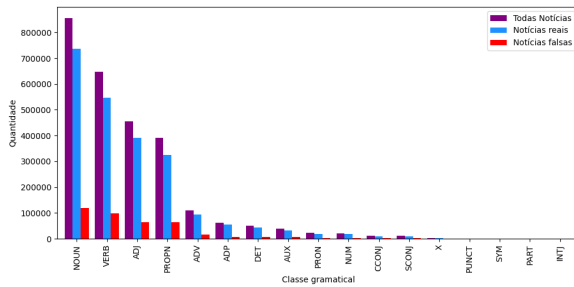
(f) Quantidade das palavras mais frequentes normalizada

destacadas pelos tons de cores. Por outro lado, o LIME destacou-se por sua interpretação direta, identificando de forma sucinta as palavras-chave para as previsões, tornando mais fácil visualizar os elementos mais relevantes para as previsões e simplificando a análise.

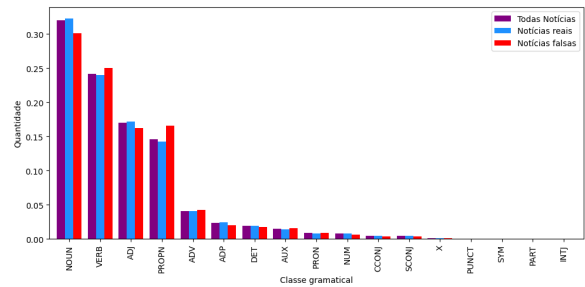
As análises do LIME destacaram a importância das *stop words* na classificação precisa das amostras pelos modelos de linguagem. Inicialmente, supôs-se que as notícias reais, devido à sua qualidade superior de redação, conteriam mais *stop words*. As atribuições de pesos positivos (peso que influência para que a predição seja notícia real) das *stop words* exibidas pelo LIME pareciam confirmar essa suposição. No entanto, ao analisar os gráficos normalizados na Seção 5.5, verificou-se que, proporcionalmente ao tamanho das notícias, a distribuição das *stop words* entre notícias reais e falsas é muito semelhante, contrariando a teoria inicial. Portanto, o LIME expôs

Figura 38 – Métricas *Dataset Fake.Br Corpus sem stop words*(a) Quantidade de entidades em todo o *dataset*

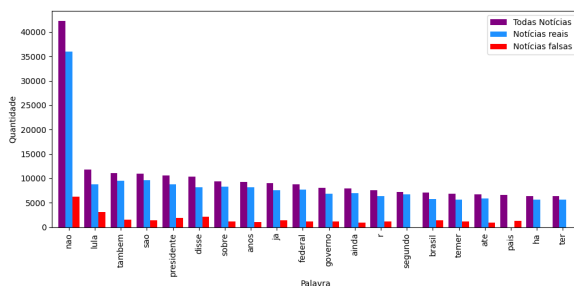
(b) Quantidade de entidades normalizada



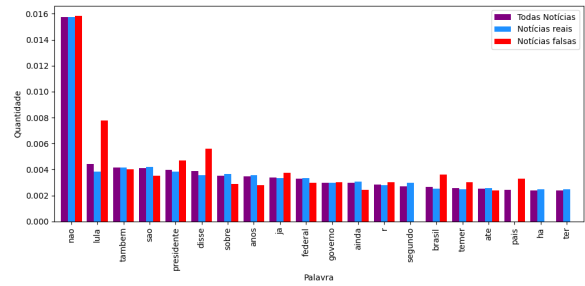
(c) Quantidade Classes Gramaticais



(d) Quantidade Classes Gramaticais Normalizada



(e) Quantidade das palavras mais frequentes



(f) Quantidade das palavras mais frequentes normalizada

que as *stop words* são relevantes para a predição correta de notícias como falsas ou reais, devido à forma como elas estão distribuídas no texto e não pela frequência. Devido a isso, os Modelos 1 e 3, que mantiveram as *stop words* em seu treinamento, obtiveram as melhores métricas.

Com as explicações geradas também foi possível notar que, nos modelos treinados com o FakeRecogna, as palavras relacionadas a políticos tiveram em sua maioria pontuações positivas (presidente, Bolsonaro, candidato, Crivella, jair, eduardo), pontuação dada à palavras que impactam na predição da notícia como real, isso pode indicar que talvez as fontes de notícias das quais o *dataset* foi construído contém em sua maioria notícias relacionadas a política.

Quanto aos modelos treinados utilizando o Fake.Br Corpus as palavras relacionadas à política que foram destacadas pelo LIME contribuíram para a classificação das notícias como

sendo falsas (Cid, STF, ex-governador, corrupção, Moro) ou seja obtiveram pontuações negativas.

No entanto se observado os gráficos da Sessão 5.5, a distribuição do uso das palavras relacionadas a política que foram muito utilizadas apresentaram uma distribuição equilibrada entre notícias reais e notícias falsas no *dataset* FakeRecogna, é possível notar que uma das palavras (“bolsonaro”) teve um uso levemente maior em notícias falsas, enquanto no Fake.Br Corpus a distribuição da palavra “lula”, “temer”, “presidente” apresentaram maior uso nas notícias falsas. A diferença na atribuição de pesos às palavras relacionadas à política entre os *datasets* pode estar relacionada com a gestão ativa do Brasil no momento em que as notícias foram coletadas. As notícias que compõem o *dataset* FakeRecogna pertencem ao período de 2019 a 2021, quando Jair Bolsonaro era o presidente da República do Brasil. Enquanto, as notícias do Fake.Br Corpus pertencem ao período de 2016 a 2018, que teve Michel Temer como presidente.

Os gráficos gerados revelaram que, no *dataset* FakeRecogna, as notícias reais frequentemente mencionam palavras relacionadas à saúde. Esse comportamento pode estar relacionado às fontes das quais as notícias foram coletadas e também pode ter sido influenciado pelo período em que foram coletadas, marcado pela pandemia de COVID-19.

Em resumo, ao analisar os resultados, foi possível perceber alguns comportamentos em relação aos modelos. O Modelo 1, treinado FakeRecogna incluindo *stop words*, foi o que obteve melhor desempenho dentre os outros modelos, como pode ser visto na Tabela 7, atingindo uma precisão de 0,99 e uma perda de 0,01 e em segundo lugar foi o Modelo 3 treinado com o Fake.Br Corpus incluindo *stop words*, com uma acurácia de 0,99 e taxa de perda de 0,03 sendo assim pode-se concluir que para este caso de estudo as *stop words* apresentaram grande importância para a classificação correta dos modelos, devido a forma em que são distribuídas ao longo dos textos e não a frequência.

Neste trabalho percebeu-se que os métodos de explicabilidade, originalmente projetos no contexto de visão computacional e aprendizado de máquina clássico, apresentam algumas limitações quando aplicados à textos. Essas limitações se dão devido a língua ser estrutura em níveis e os métodos serem mais eficientes em evidenciar os níveis morfológico e léxico (palavras e partes de palavras). Porém, outros níveis também tem um papel importante no processo de classificação sendo necessário recorrer ao spaCy. Níveis como por exemplo:

- **Morfossintático:** as categorias gramaticais podem variar entre as classes. Neste trabalho, verificou-se que no Fake.Br Corpus substantivos próprios são mais utilizados por notícias

falsas enquanto no FakeRecogna são mais utilizados por notícias reais;

- **Semântico:** evidenciou-se que alguns nomes próprios são mais comuns que outros em notícias reais dependendo do *dataset* analisado. No FakeRecogna, entidades como pessoas e organizações aparecem com maior frequência em notícias reais do que em notícias falsas. Por outro lado, no Fake.Br, as notícias falsas fazem maior uso de entidades do tipo pessoa e família nomeada;
- **Retórico:** foi observado que notícias reais são mais longas do que notícias falsas. Possivelmente essa diferença se dê porque notícias reais tem o propósito de informar e tendem a apresentar diversos nuances para a melhor compreensão de um fato. Já as notícias falsas tendem a ser menores, trazendo menos detalhes, possivelmente pela falta de informações reais de apoio, ao mesmo tempo que se tornam mais fáceis de serem compreendidas e se propagarem mais rapidamente.

5.7 Limitações

A principal limitação deste trabalho foi a escassez de estudos que aplicam e comparam métodos de explicabilidade, como LIME e IG, em modelos de linguagem voltados para o português do Brasil. A maioria das investigações sobre XAI concentra-se em modelos de visão computacional, sendo menos frequente a aplicação em modelos de linguagem, especialmente em textos em português. Além disso, a escolha do algoritmo de XAI adequado no campo do processamento de linguagem natural em português ainda é um desafio, dado o número limitado de referências. Essas restrições reforçam a importância de ampliar as pesquisas sobre explicabilidade em modelos de linguagem no contexto da língua portuguesa.

Outro ponto é que o estudo foi focado em um domínio específico, o de identificação de notícias falsas, o que pode não ser diretamente aplicável a outros domínios textuais. Também é importante considerar que o trabalho foi desenvolvido para a língua portuguesa, e características de notícias falsas em outras línguas podem diferir significativamente, o que restringe a aplicabilidade dos resultados para outros idiomas.

Além disso, o trabalho comparou apenas dois métodos de explicabilidade o LIME e o IG, o que limita a generalização dos achados para outros métodos disponíveis. Por fim, nossa análise foi predominantemente qualitativa e subjetiva, o que destaca o desafio de desenvolver métricas

objetivas para comparar os métodos de explicabilidade de forma mais rigorosa.

6 Conclusões

A criação de um modelo de linguagem é uma tarefa predecessora à realização de diversas outras, como classificação de texto, tradução automática e reconhecimento de entidades nomeadas. Esta tarefa vem se tornando cada vez mais relevante e desafiadora com o objetivo de aprimorar os sistemas de processamento de linguagem natural para atingirem um novo nível na compreensão da linguagem.

Diante do crescente uso dos modelos de linguagem e algoritmos de aprendizado de máquina em diferentes aplicações, a questão da opacidade dos modelos de linguagem, popularmente conhecidos como “caixa opaca”, tornou-se uma preocupação significativa. A confiabilidade e compreensão dos modelos são fundamentais, especialmente quando eles são empregados em processos de tomada de decisão na vida real. Neste trabalho, foi explorada a importância da compreensão do comportamento dos modelos de linguagem, bem como a necessidade de entender suas previsões individuais.

Os objetivos deste trabalho foram alcançados com sucesso. Primeiramente, foi realizado um estudo abrangente e apresentação dos tipos de métodos de XAI existentes, com ênfase no LIME e IG. Este estudo proporcionou uma base sólida para entender as capacidades e limitações desses métodos, preparando o terreno para a aplicação prática nas etapas subsequentes do trabalho.

Em seguida, um estudo de caso foi conduzido no domínio de identificação de notícias falsas em português brasileiro. Este estudo de caso foi essencial para demonstrar a aplicação dos métodos de explicabilidade, proporcionando percepções valiosas sobre como esses métodos ajudam a interpretar os padrões identificados pelo modelo BERTimbau. A eficácia do modelo BERTimbau foi confirmada pela sua capacidade de classificar as notícias de forma precisa.

Com o estudo de caso, como expresso anteriormente, as técnicas LIME e IG apresentam comportamentos muito semelhantes, porém, nos experimentos realizados neste estudo de caso, as explicações geradas pelo LIME se mostraram mais fáceis de interpretar, destacando de forma simples as palavras significativas para a previsão da amostra. Enquanto, as explicações geradas pelo IG apresentaram uma interpretação mais difícil na tarefa empregada, pois em muitas das explicações geradas a diferença nos pesos representados pelas cores foi mínima, dificultando a visualização do que é importante para a previsão.

A hipótese levantada inicialmente de que *stop words* são mais frequentes em notícias reais foi rechaçada, pois como analisado nos gráficos normalizados gerados pelo spaCy na Seção 5.5, descobriu-se que, em proporção ao comprimento das notícias, a distribuição de *stop words* entre notícias reais e falsas não apresenta uma diferença significativa. Mas verificou-se que as *stop words* são importantes mesmo assim para a classificação, devido à forma com que se distribuem ao longo do texto, as explicações geradas pelo IG e principalmente pelo LIME, comprovam pois elas foram altamente destacadas nas explicações das amostras classificadas pelos Modelos 1 e 3, que alcançaram o melhor desempenho.

O intuito com estas ferramentas era identificar quais palavras ou partes do texto são mais importantes para a classificação, com elas juntamente do spaCy foi possível compreender melhor o comportamento dos modelos que foram treinados. Por meio de técnicas como estas é possível desenvolver estratégias que aumentam a confiabilidade das previsões do modelo, proporcionando uma visão mais transparente de seu funcionamento interno.

Neste estudo além das percepções fornecidas anteriormente, constatou-se que os métodos de explicabilidade apresentam limitação ao serem utilizados para trabalhar com textos devido à estrutura multi-nível da linguagem, isso pode se dar pelo fato deles serem desenvolvidos originalmente para aplicações de visão computacional e para modelos de aprendizado de máquina mais simples. Eles concentram-se nos níveis léxico e morfológico, perdendo oportunidades de captar aspectos significativos que poderiam ser observados nos demais níveis linguísticos.

No nível morfossintático, observou-se por meio do spaCy que nomes próprios são mais frequentes em notícias reais. No nível semântico, certas categorias, como organizações, são mais comuns em notícias reais, enquanto a frequência de nomes de pessoas permanece constante. No nível retórico, foi observado que em ambos os *datasets* analisados, as notícias reais tendem a ser mais extensas e minuciosas, buscando informar diversas camadas informativas para seus leitores. Em contrapartida, as notícias falsas são mais concisas, oferecendo menos detalhes e sendo assim mais fáceis de compreender possibilitando uma rápida disseminação.

Para futuras investigações, é desejável explorar métodos adicionais de explicabilidade além do LIME e do IG, como por exemplo o SHapley Additive exPlanations (SHAP) (LUNDBERG; LEE, 2017). Além disso, seria vantajoso conduzir estudos de caso em domínios diversos que não estejam restritos ao contexto de notícias falsas. Diversificar os domínios de aplicação poderia fornecer uma compreensão mais abrangente sobre como diferentes métodos de explicabi-

lidade se comportam em contextos variados, ampliando assim a aplicabilidade e a generalização das técnicas investigadas.

Outra direção futura interessante seria investigar como os modelos se comportam sob diferentes condições de entrada. Isso poderia incluir, por exemplo, usar um modelo treinado sem *stop words* para realizar a previsão de amostras contendo *stop words*, ou aplicar um modelo treinado com o FakeRecogna para prever amostras do corpus Fake.Br Corpus, e vice-versa, e poderiam ser aplicadas abordagens robustas de ataque adversário nos modelos.

Também seria relevante realizar testes removendo das amostras as palavras que os métodos de XAI consideraram importantes, para observar como a remoção dessas palavras afeta as previsões dos modelos. Essas abordagens permitiriam avaliar a robustez dos modelos e sua capacidade de manter a precisão sob diferentes condições. O desenvolvimento de estratégias de explicabilidade que considerem essas variáveis seria uma contribuição valiosa para o campo, aumentando a confiabilidade das previsões e a transparência dos modelos em contextos críticos.

Por fim, seria interessante desenvolver um novo *framework* de explicabilidade para textos que vá além da simples distribuição de palavras, superando as limitações apresentadas pelos métodos examinados neste trabalho (LIME e IG).

Referências

- ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <https://www.tensorflow.org/>.
- AHMED, I.; JEON, G.; PICCIALLI, F. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, IEEE, v. 18, n. 8, p. 5031–5042, 2022.
- AI, G. *Google Colaboratory*. 2018. Disponível em: <https://colab.research.google.com/>.
- ALAMMAR, J. *The illustrated transformer*. 2018. Disponível em: <https://jalamar.github.io/illustrated-transformer/>.
- ALLEN, J. *Natural language understanding*. [S.l.]: Benjamin-Cummings Publishing Co., Inc., 1988.
- BOJAR, O. et al. Findings of the 2014 workshop on statistical machine translation. In: *Proceedings of the ninth workshop on statistical machine translation*. [S.l.: s.n.], 2014. p. 12–58.
- CHAMOLA, V. et al. A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access*, IEEE, 2023.
- CHOMSKY, N. Syntactic structures. In: *Syntactic Structures*. [S.l.]: De Gruyter Mouton, 2009.
- CONNEAU, A. et al. Xnli: Evaluating cross-lingual sentence representations. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. [S.l.]: Association for Computational Linguistics, 2018.
- DALE, R. Classical approaches to natural language processing. In: *Handbook of natural language processing, second edition*. [S.l.]: CRC Press, Taylor & Francis Group, 2010. p. 3–7.
- DAS, A.; RAD, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.
- DENG, L.; YU, D. et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, Now Publishers, Inc., v. 7, n. 3–4, p. 197–387, 2014.
- DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- DSA, D. S. A. *Data Science Academy. Deep Learning Book*. 2022. Disponível em: <https://www.deeplearningbook.com.br/>.
- FACE, H. *Hugging Face*. 2023. Disponível em: <https://huggingface.co/>.
- FAN, X. et al. Comparative study of decision tree with different evidential uncertainty measures. In: IEEE. *2018 Chinese Automation Congress (CAC)*. [S.l.], 2018. p. 3104–3108.
- FILHO, J. A. W. et al. The brwac corpus: a new open resource for brazilian portuguese. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. [S.l.: s.n.], 2018.

- GARCIA, G.; AFONSO, L.; PAPA, J. Fakerecogna: A new brazilian corpus for fake news detection. In: _____. [S.l.: s.n.], 2022. p. 57–67. ISBN 978-3-030-98304-8.
- GOBINDA, G. C. Natural language processing. *Annual Review of Information Science and Technology*, v. 37, p. 51–89, 2003.
- GOHEL, P.; SINGH, P.; MOHANTY, M. Explainable ai: current status and future directions. *arXiv preprint arXiv:2107.07045*, 2021.
- HAI, S. *AI Index Report 2023 Artificial Intelligence Index*. 2021. Disponível em: <https://aiindex.stanford.edu/report/>.
- HAN, X.; EISENSTEIN, J. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 4238–4248. Disponível em: <https://aclanthology.org/D19-1433>.
- HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2001.
- HONNIBAL, M.; MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 2017.
- INDURKHYA, N.; DAMERAU, F. J. *Handbook of Natural Language Processing*. 2nd. ed. [S.l.]: Chapman amp; Hall/CRC, 2010. ISBN 1420085921.
- JARRETT, K. et al. What is the best multi-stage architecture for object recognition? In: *IEEE. 2009 IEEE 12th international conference on computer vision*. [S.l.], 2009. p. 2146–2153.
- JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. [S.l.]: Pearson/Prentice Hall, 2009.
- KAUFMAN, S. et al. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, ACM New York, NY, USA, v. 6, n. 4, p. 1–21, 2012.
- KIEFER, S. Case: Explaining text classifications by fusion of local surrogate explanation models with contextual and semantic knowledge. *Information Fusion*, Elsevier, v. 77, p. 184–195, 2022.
- LAKKARAJU, H. et al. Faithful and customizable explanations of black box models. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. [S.l.: s.n.], 2019. p. 131–138.
- LIDDY, E. D. *Natural language processing*. 2001.
- LOGESWARAN, L. et al. Zero-shot entity linking by reading entity descriptions. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 3449–3460. Disponível em: <https://aclanthology.org/P19-1335>.
- LUGER, G. F. *Inteligência Artificial-: Estruturas e estratégias para a solução de problemas complexos*. [S.l.]: Bookman, 2004.

- LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, v. 30, 2017.
- MEHRI, S. et al. Pretraining methods for dialog context representation learning. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 3836–3845. Disponível em: <https://aclanthology.org/P19-1373>.
- MLADENOVA, T.; VALOVA, I. Comparative analysis between the traditional k-nearest neighbor and modifications with weight-calculation. In: IEEE. *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. [S.l.], 2022. p. 961–965.
- MORADI, M.; SAMWALD, M. Explaining black-box models for biomedical text classification. *IEEE journal of biomedical and health informatics*, IEEE, v. 25, n. 8, p. 3112–3120, 2021.
- MURDOCH, W. J. et al. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.
- MURPHY, K. P. *Probabilistic machine learning: an introduction*. [S.l.]: MIT press, 2022.
- PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2002. p. 311–318.
- PARIKH, A. P. et al. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.
- PHANG, J.; FÉVRY, T.; BOWMAN, S. R. *Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks*. [S.l.]: arXiv, 2018. Disponível em: <https://arxiv.org/abs/1811.01088>.
- PRÖLLOCHS, N.; FEUERRIEGEL, S.; NEUMANN, D. Learning interpretable negation rules via weak supervision at document level: A reinforcement learning approach. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.], 2019. v. 1, p. 407–413.
- RADFORD, A. et al. Improving language understanding by generative pre-training. OpenAI, 2018.
- REUTERS. *União Europeia chega a acordo Pioneiro sobre Regulamentação de Inteligência Artificial*. 2023. Disponível em: <https://g1.globo.com/tecnologia/noticia/2023/12/09/uniao-europeia-chega-a-acordo-pioneiro-sobre-regulamentacao-de-inteligencia-artificial.gh.html>.
- RIBEIRO, M. T.; SINGH, S.; GUESTIN, C. “why should i trust you?” explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144.
- RODRIGUES, R. C. et al. Portuguese language models and word embeddings: evaluating on semantic similarity tasks. In: SPRINGER. *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*. [S.l.], 2020. p. 239–248.

RUDER, S. *Recent Advances in Language Model Fine-tuning*. 2021. Disponível em: <http://ruder.io/recent-advances-lm-fine-tuning>.

SANTOS, D. et al. Harem: An advanced ner evaluation contest for portuguese. In: *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*. [S.l.: s.n.], 2006.

SANTOS, R. L.; MONTEIRO, R. A.; PARDO, T. A. The fake. br corpus-a corpus of fake news for brazilian portuguese. In: *Latin American and Iberian Languages Open Corpora Forum (OpenCor)*. [S.l.: s.n.], 2018. p. 1–2.

SELLAM, T.; DAS, D.; PARIKH, A. BLEURT: Learning robust metrics for text generation. In: JURAFSKY, D. et al. (Ed.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 7881–7892. Disponível em: <https://aclanthology.org/2020.acl-main.704>.

SEO, M. et al. Bidirectional attention flow for machine comprehension iclr. 2017.

SHEVSKAYA, N. V. Explainable artificial intelligence approaches: challenges and perspectives. In: IEEE. *2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*. [S.l.], 2021. p. 540–543.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian conference on intelligent systems*. [S.l.], 2020. p. 403–417.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. [S.l.]: JMLR.org, 2017. (ICML'17), p. 3319–3328.

TORCIANTI, A.; MATZKA, S. Explainable artificial intelligence for predictive maintenance applications using a local surrogate model. In: IEEE. *2021 4th international conference on artificial intelligence for industries (ai4i)*. [S.l.], 2021. p. 86–88.

UFRGS. *Livro de Probabilidade e Estatística EAD*. [S.l.]: UFRGS, 2022. Disponível em: https://www.ufrgs.br/probabilidade-estatistica/livro/livro_completo/ch7-reg-simples.html.

VASWANI, A. et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017. Disponível em: <http://arxiv.org/abs/1706.03762>.

WALLACE, E. et al. AllenNLP interpret: A framework for explaining predictions of NLP models. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 7–12. Disponível em: <https://www.aclweb.org/anthology/D19-3002>.

WANG, A. et al. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

WILLIAMS, A.; NANGIA, N.; BOWMAN, S. A broad-coverage challenge corpus for sentence understanding through inference. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. [S.l.]: Association for Computational Linguistics, 2018. p. 1112–1122. Disponível em: <http://aclweb.org/anthology/N18-1101>.

WU, Y. et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

ZHANG, A. et al. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.

Apêndices

APÊNDICE A – Explicações geradas pelo LIME - Modelo 1: FakeRecogna incluindo *stop words*

Neste capítulo estão todas as explicações geradas pelo LIME, sobre as amostras classificadas pelo Modelo 1 treinado com o FakeRecogna incluindo *stop words*.

A.1 Modelo 1 - Explicações do LIME de amostras Falsas classificadas como Falsas

Figura 39 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.

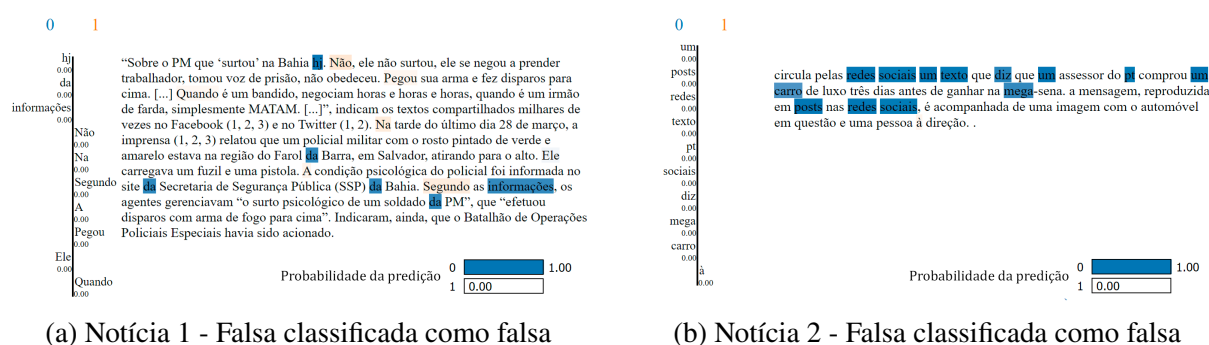


Figura 40 – Modelo 1 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.

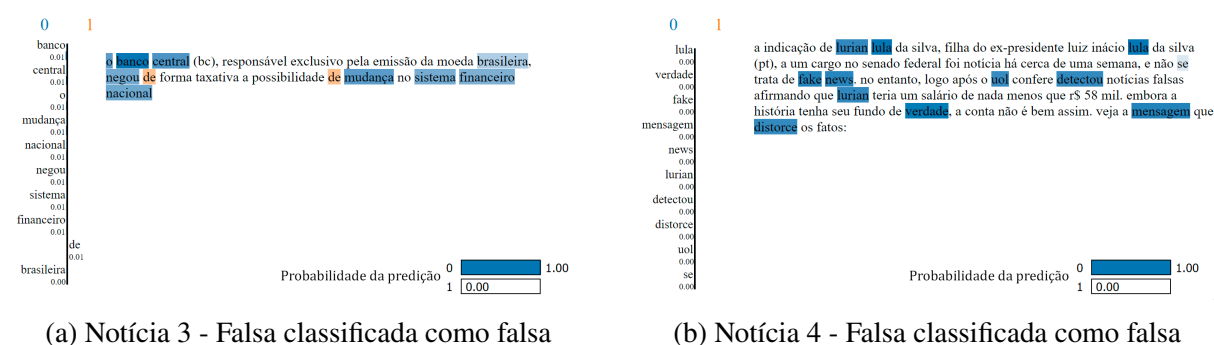
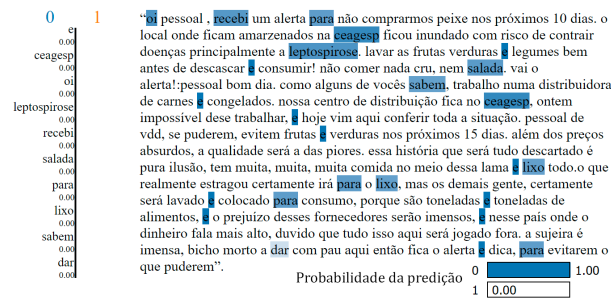
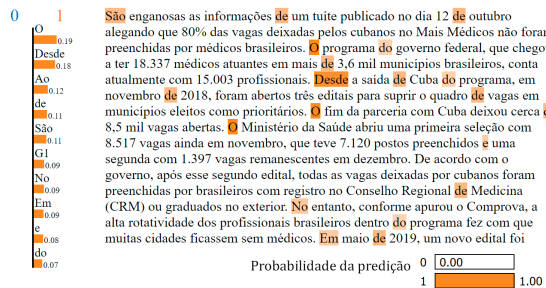


Figura 41 – Modelo 1 - Explicação da notícias 5 falsa classificada como falsa gerada pelo LIME.

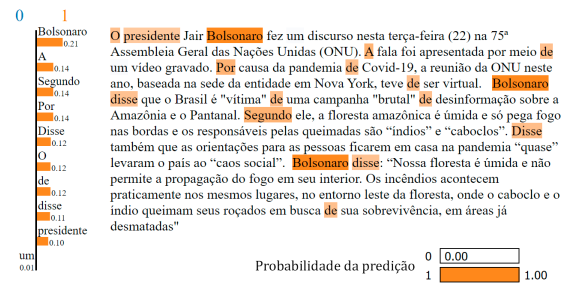


A.2 Modelo 1 - Explicações do LIME de amostras Falsas classificadas como Reais

Figura 42 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.



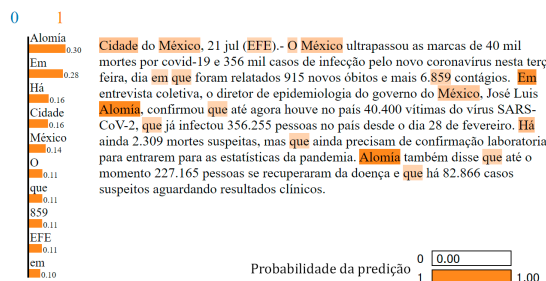
(a) Notícia 1 - Falsa classificada como real



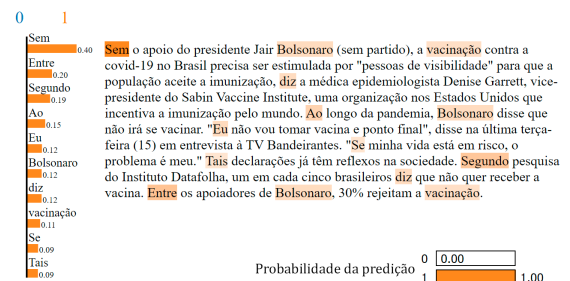
(b) Notícia 2 - Falsa classificada como real

A.3 Modelo 1 - Explicações do LIME de amostras Reais classificadas como Reais

Figura 43 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.



(a) Notícia 1 - Real classificada como real



(b) Notícia 2 - Real classificada como real

Figura 44 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.

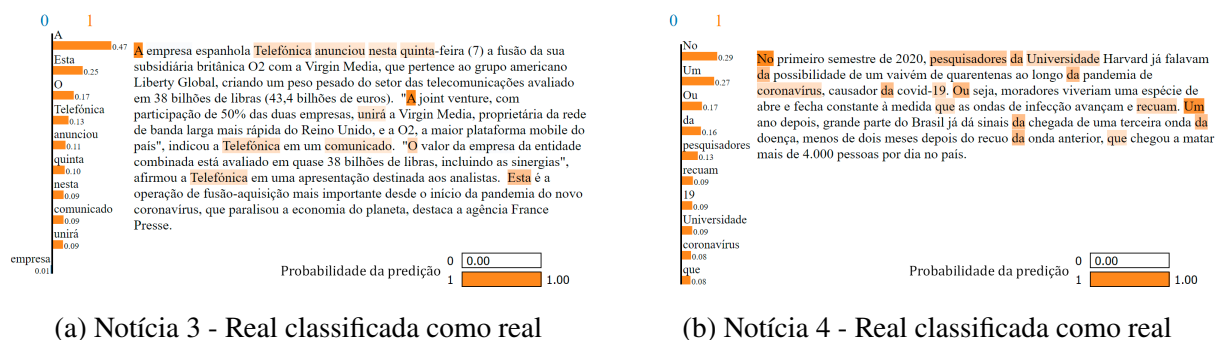
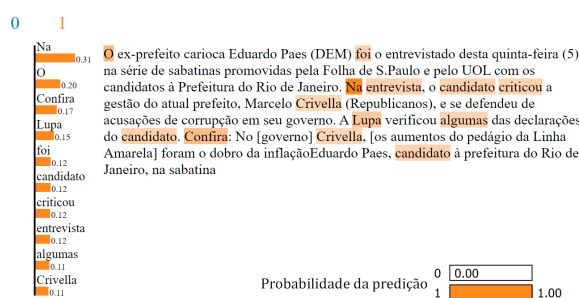


Figura 45 – Modelo 1 - Explicação da notícias 5 real classificada como real gerada pelo LIME.



A.4 Modelo 1 - Explicações do LIME de amostras Reais classificadas como Falsas

Figura 46 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.

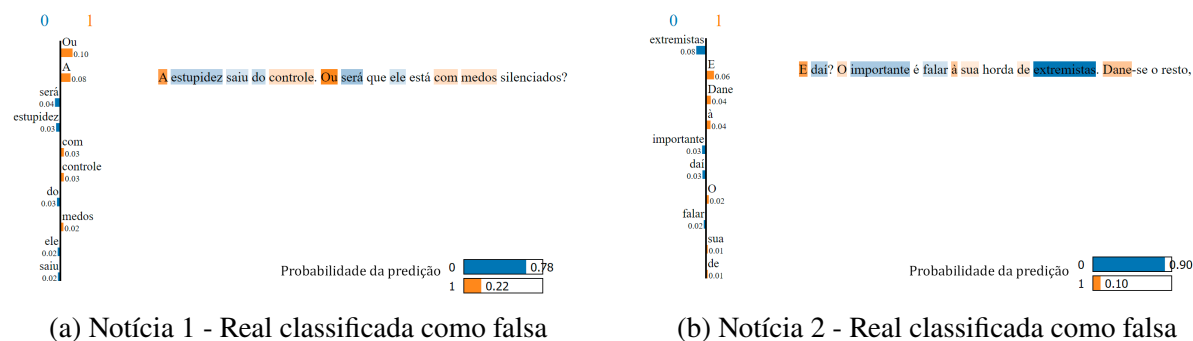


Figura 47 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.

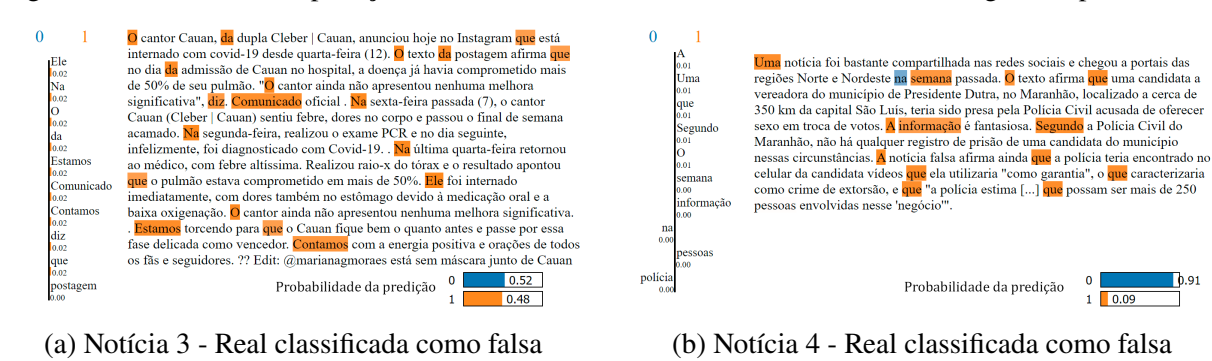
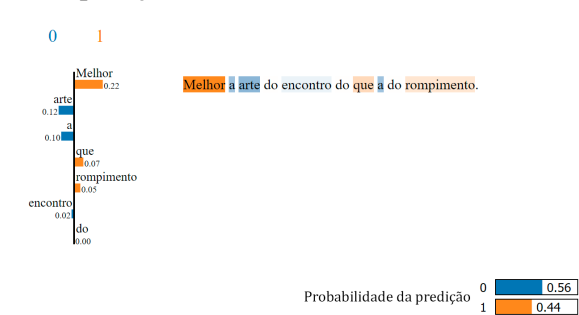


Figura 48 – Modelo 1 - Explicação da notícias 5 real classificada como falsa gerada pelo LIME.



APÊNDICE B – Explicações geradas pelo LIME - Modelo 2: FakeRecogna sem *stop words*

Neste capítulo estão todas as explicações geradas pelo LIME, sobre as amostras classificadas pelo Modelo 2 treinado com o FakeRecogna sem *stop words*.

B.1 Modelo 2 - Explicações do LIME de amostras Falsas classificadas como Falsas

Figura 49 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.

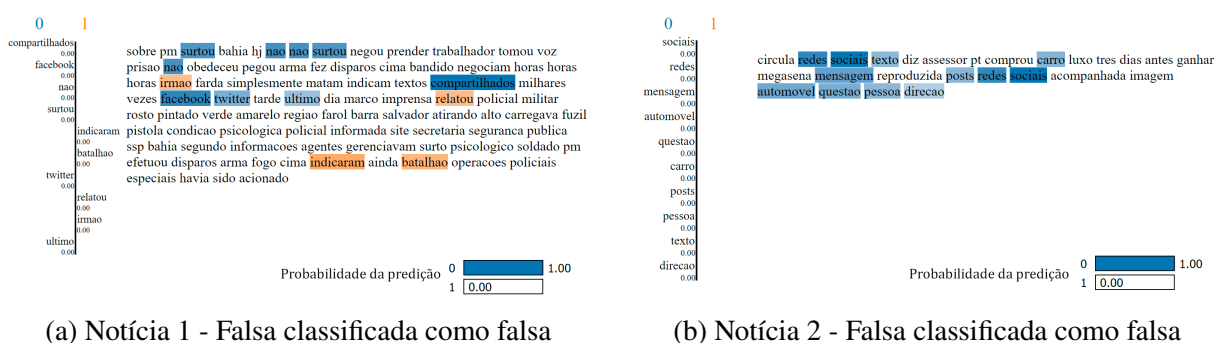


Figura 50 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.

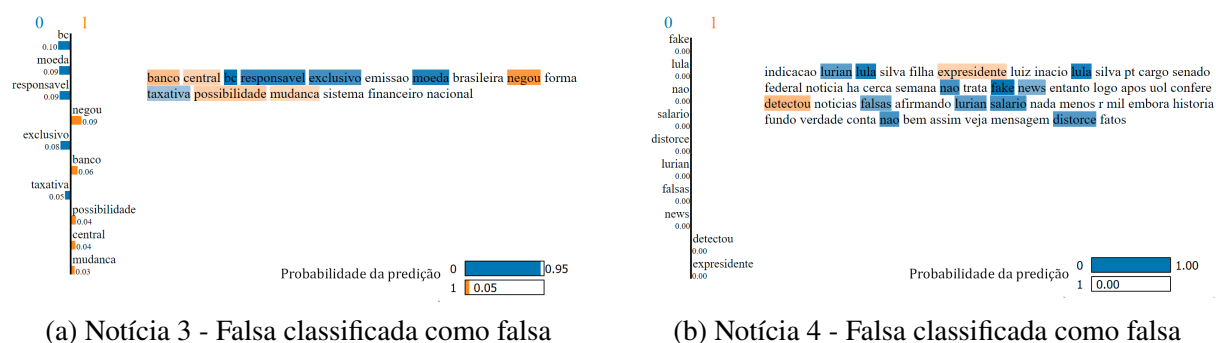
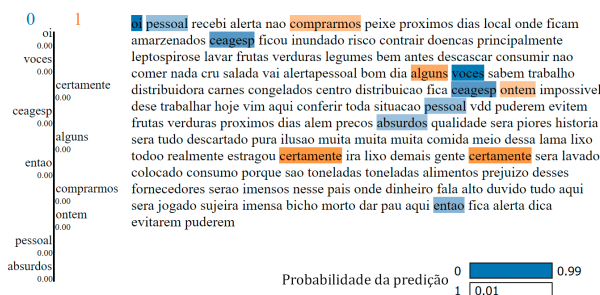
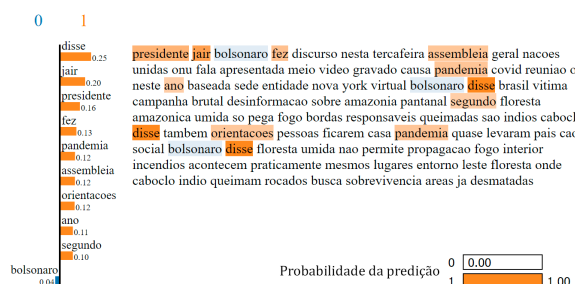


Figura 51 – Modelo 2 - Explicação da notícia 5 falsa classificada como falsa gerada pelo LIME.

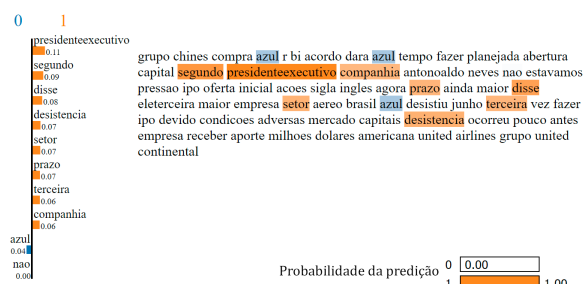


B.2 Modelo 2 - Explicações do LIME de amostras Falsas classificadas como Reais

Figura 52 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.

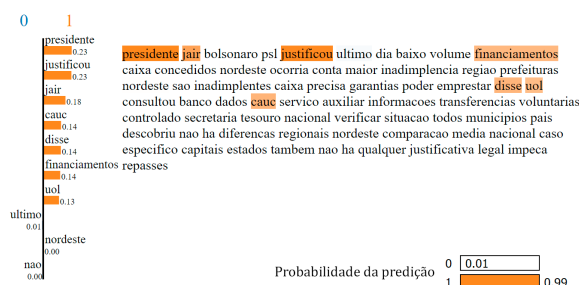


(a) Notícia 1 - Falsa classificada como real

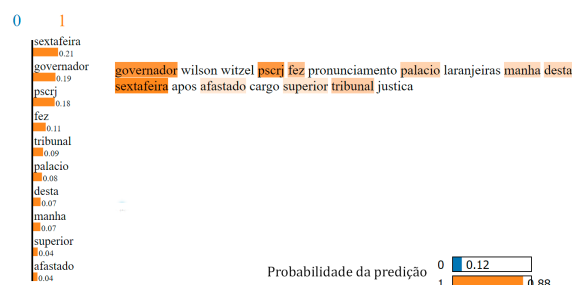


(b) Notícia 2 - Falsa classificada como real

Figura 53 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo LIME.

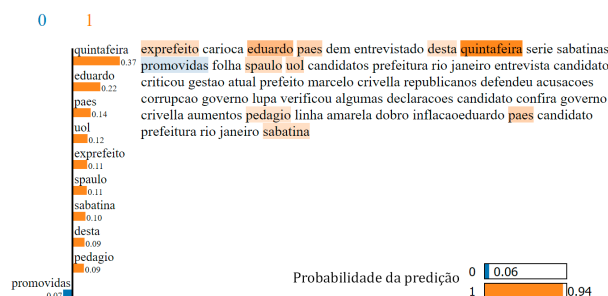


(a) Notícia 3 - Falsa classificada como real



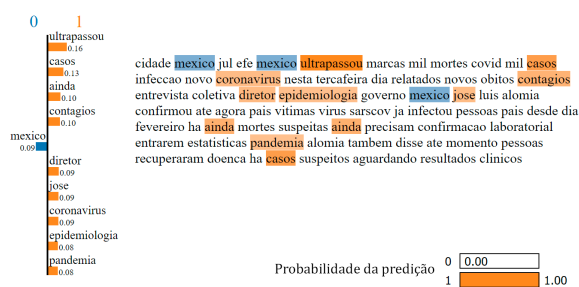
(b) Notícia 4 - Falsa classificada como real

Figura 54 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo LIME.

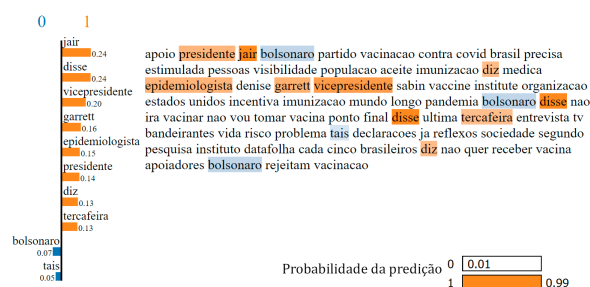


B.3 Modelo 2 - Explicações do LIME de amostras Reais classificadas como Reais

Figura 55 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.

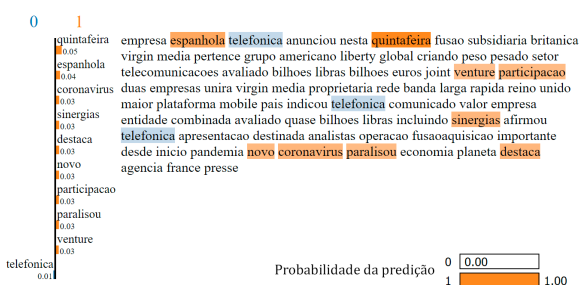


(a) Notícia 1 - Real classificada como real

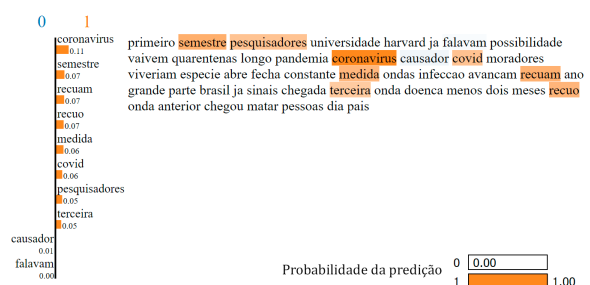


(b) Notícia 2 - Real classificada como real

Figura 56 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.

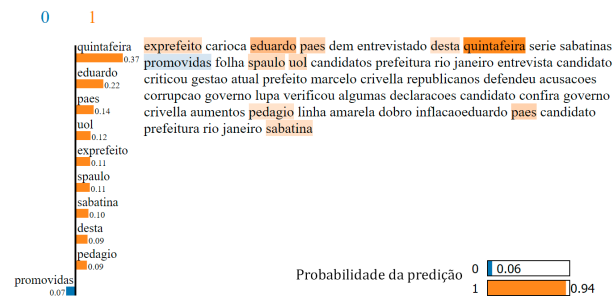


(a) Notícia 3 - Real classificada como real



(b) Notícia 4 - Real classificada como real

Figura 57 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo LIME.



B.4 Modelo 2 - Explicações do LIME de amostras Reais classificadas como Falsas

Figura 58 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.

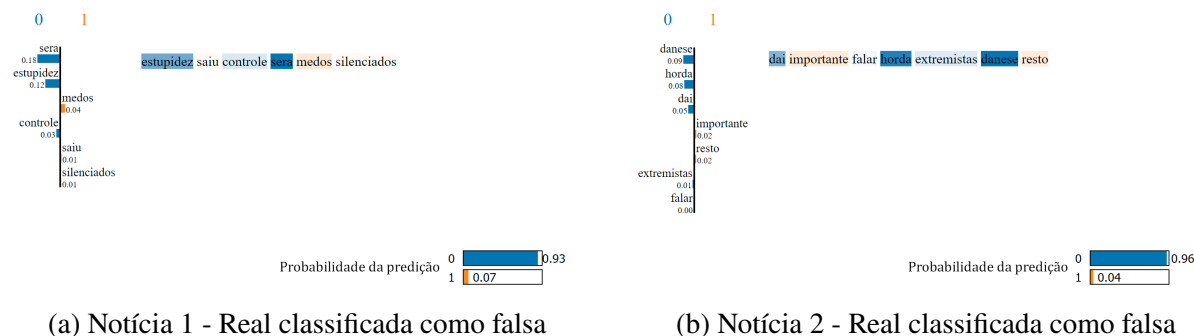


Figura 59 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.

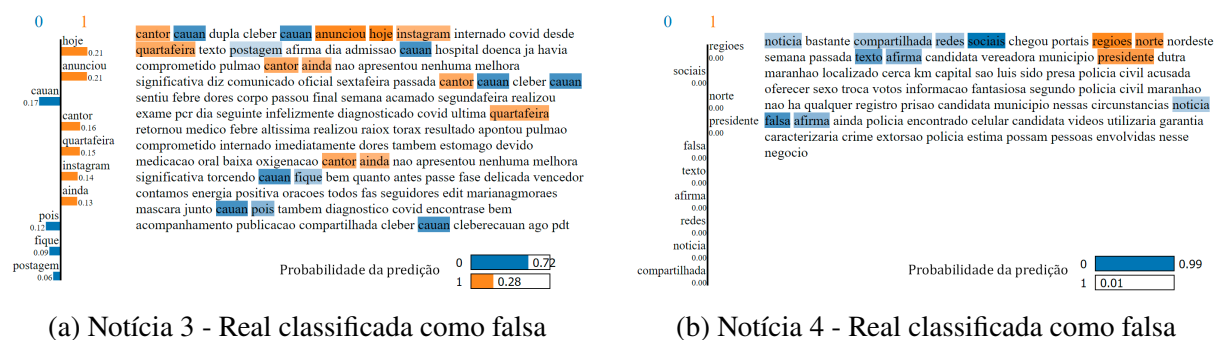
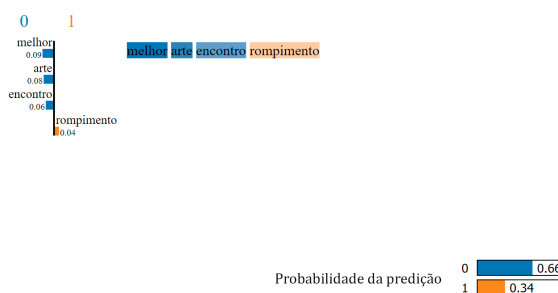


Figura 60 – Modelo 2 - Explicação da notícia 5 real classificada como falsa gerada pelo LIME.



APÊNDICE C – Explicações geradas pelo LIME - Modelo 3: Fake.Br Corpus incluindo *stop words*

Neste capítulo estão todas as explicações geradas pelo LIME, sobre as amostras classificadas pelo Modelo 3 treinado com o Fake.Br Corpus incluindo *stop words*.

C.1 Modelo 3 - Explicações do LIME de amostras Falsas classificadas como Falsas

Figura 61 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.

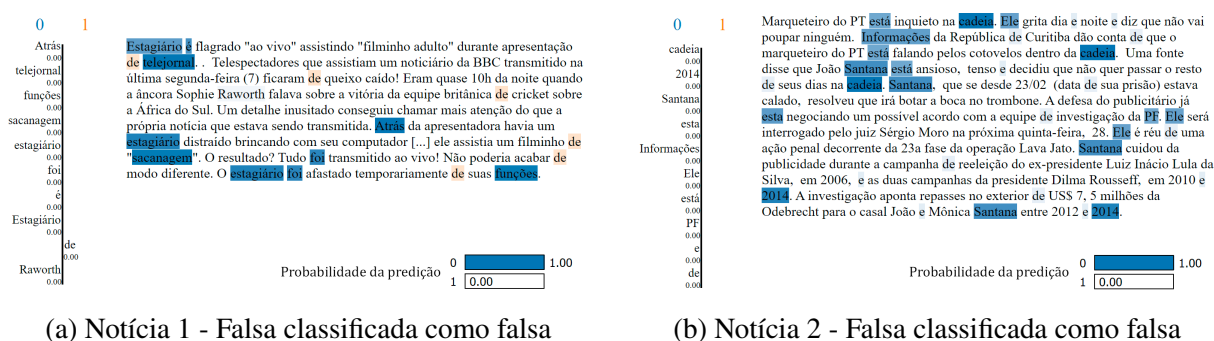


Figura 62 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.

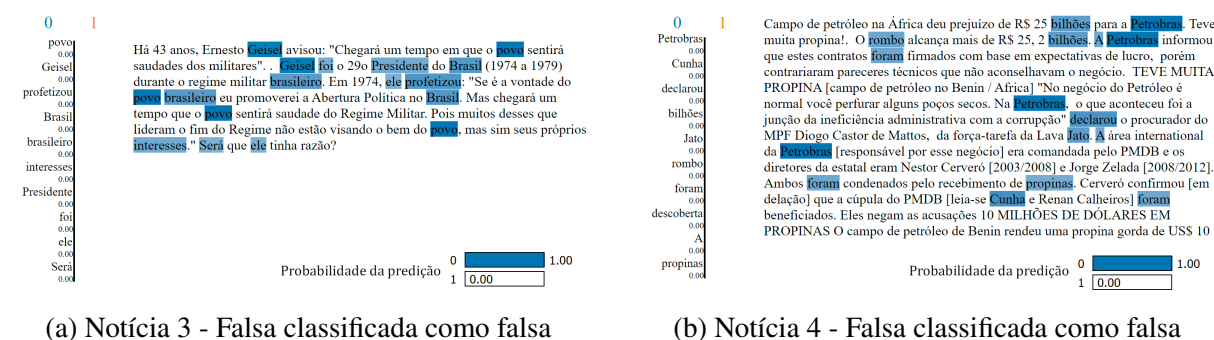
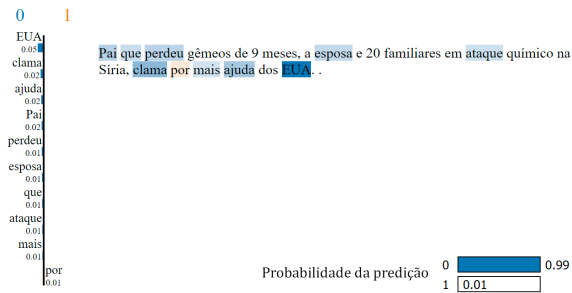
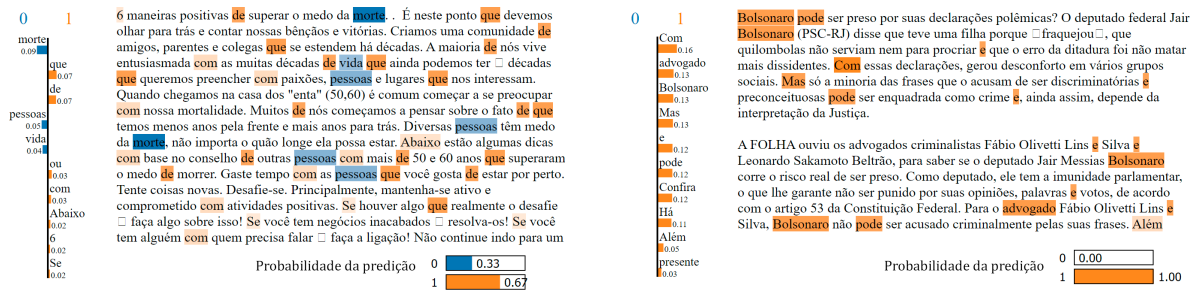


Figura 63 – Modelo 3 - Explicação da notícia 5 falsa classificada como falsa gerada pelo LIME.



C.2 Modelo 3 - Explicações do LIME de amostras Falsas classificadas como Reais

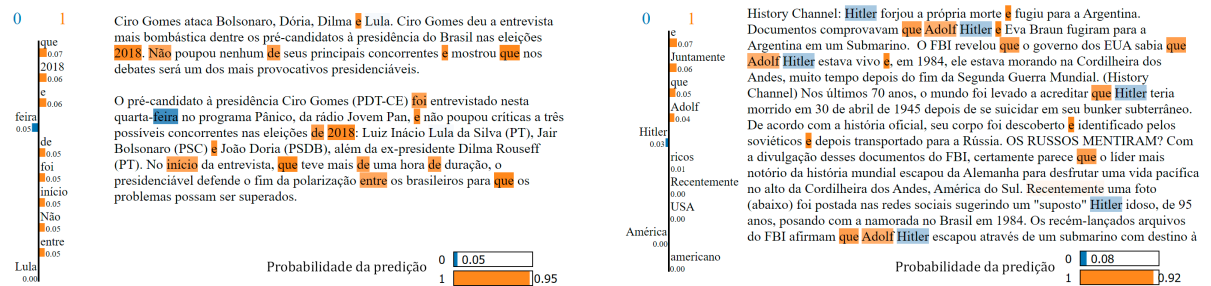
Figura 64 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.



(a) Notícia 1 - Falsa classificada como real

(b) Notícia 2 - Falsa classificada como real

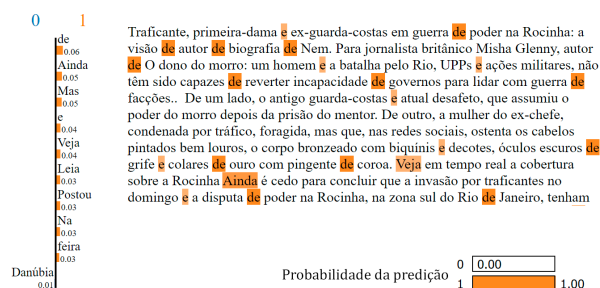
Figura 65 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo LIME.



(a) Notícia 3 - Falsa classificada como real

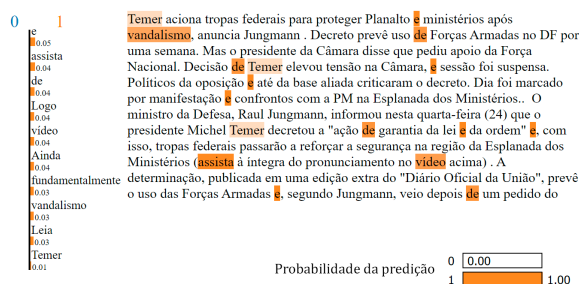
(b) Notícia 4 - Falsa classificada como real

Figura 66 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo LIME.

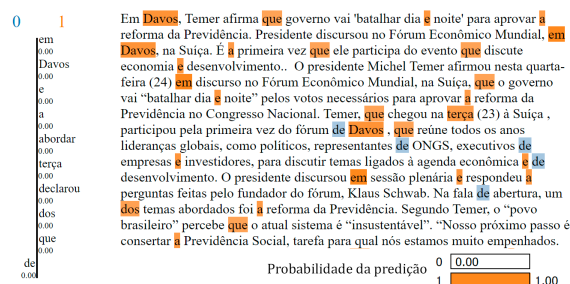


C.3 Modelo 3 - Explicações do LIME de amostras Reais classificadas como Reais

Figura 67 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.

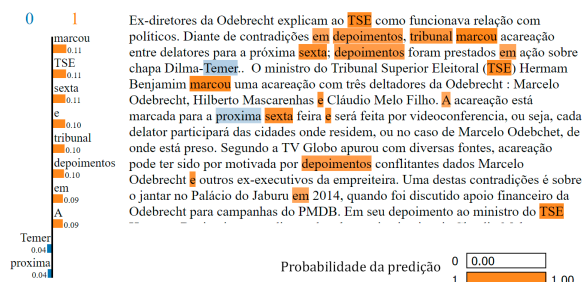


(a) Notícia 1 - Real classificada como real

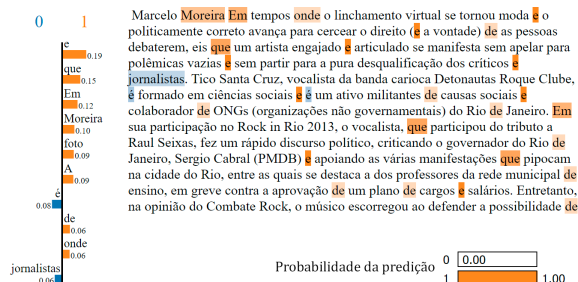


(b) Notícia 2 - Real classificada como real

Figura 68 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.

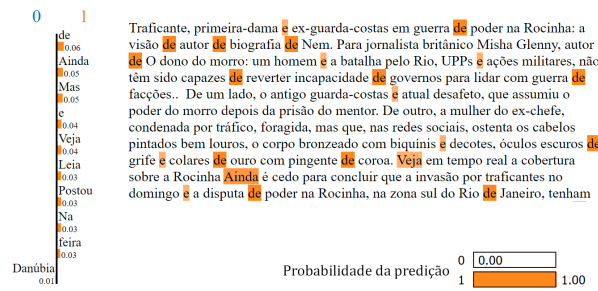


(a) Notícia 3 - Real classificada como real



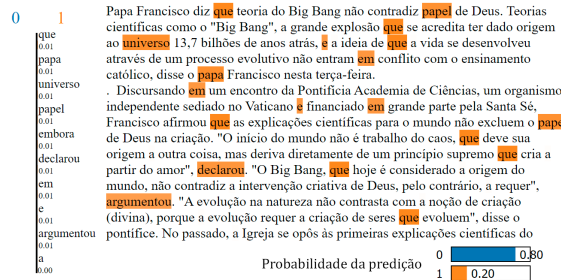
(b) Notícia 4 - Real classificada como real

Figura 69 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo LIME.

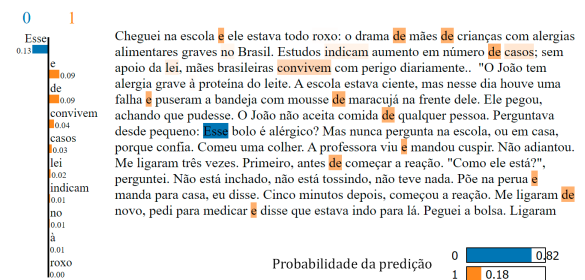


C.4 Modelo 3 - Explicações do LIME de amostras Reais classificadas como Falsas

Figura 70 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.

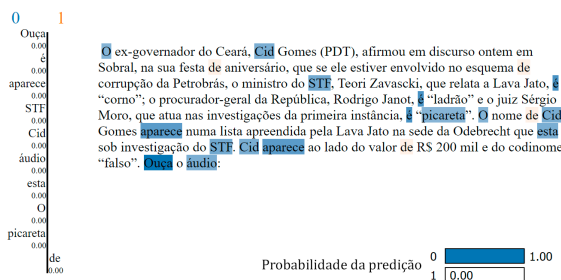


(a) Notícia 1 - Real classificada como falsa

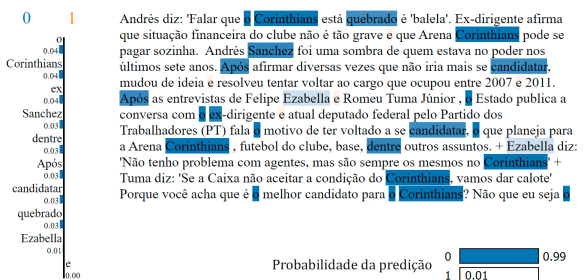


(b) Notícia 2 - Real classificada como falsa

Figura 71 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.



(a) Notícia 3 - Real classificada como falsa



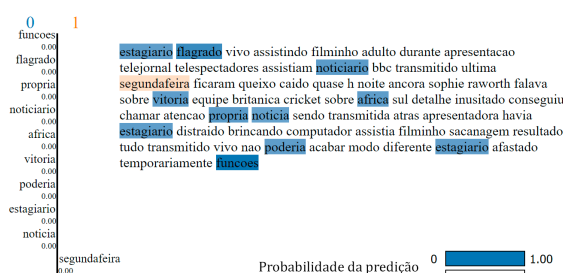
(b) Notícia 4 - Real classificada como falsa

APÊNDICE D – Explicações geradas pelo LIME - Modelo 4: Fake.Br Corpus sem *stop words*

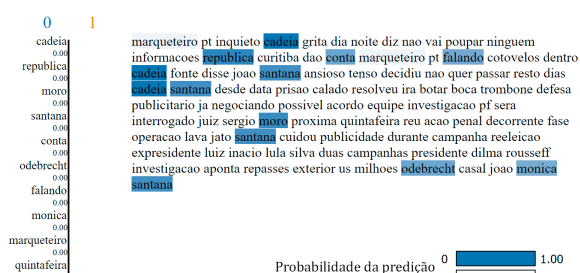
Neste capítulo estão todas as explicações geradas pelo LIME, sobre as amostras classificadas pelo Modelo 4 treinado com o Fake.Br Corpus sem *stop words*.

D.1 Modelo 4 - Explicações do LIME de amostras Falsas classificadas como Falsas

Figura 72 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo LIME.

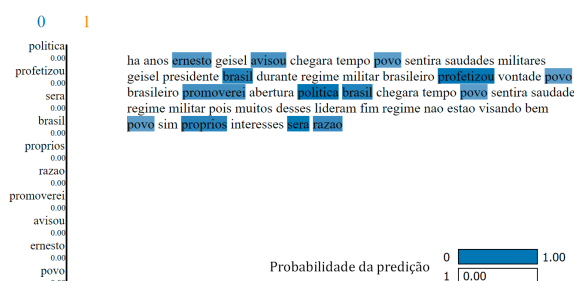


(a) Notícia 1 - Falsa classificada como falsa

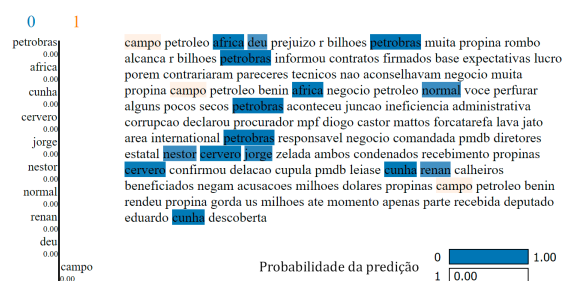


(b) Notícia 2 - Falsa classificada como falsa

Figura 73 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo LIME.

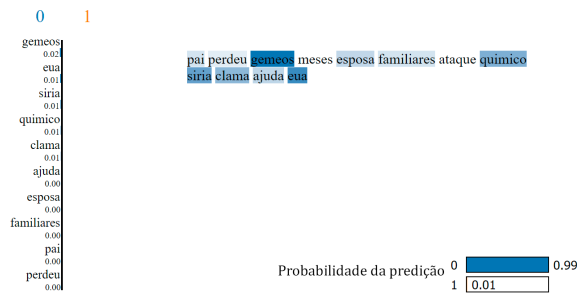


(a) Notícia 3 - Falsa classificada como falsa



(b) Notícia 4 - Falsa classificada como falsa

Figura 74 – Modelo 4 - Explicação da notícia 5 falsa classificada como falsa gerada pelo LIME.



D.2 Modelo 4 - Explicações do LIME de amostras Falsas classificadas como Reais

Figura 75 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo LIME.

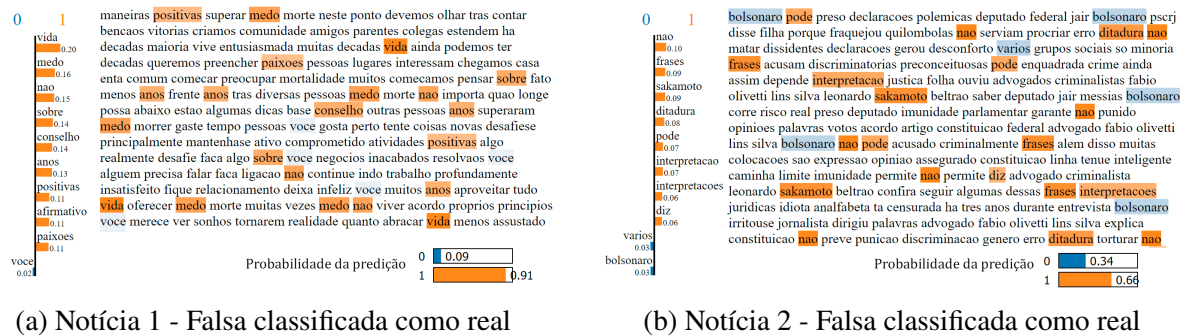


Figura 76 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo LIME.

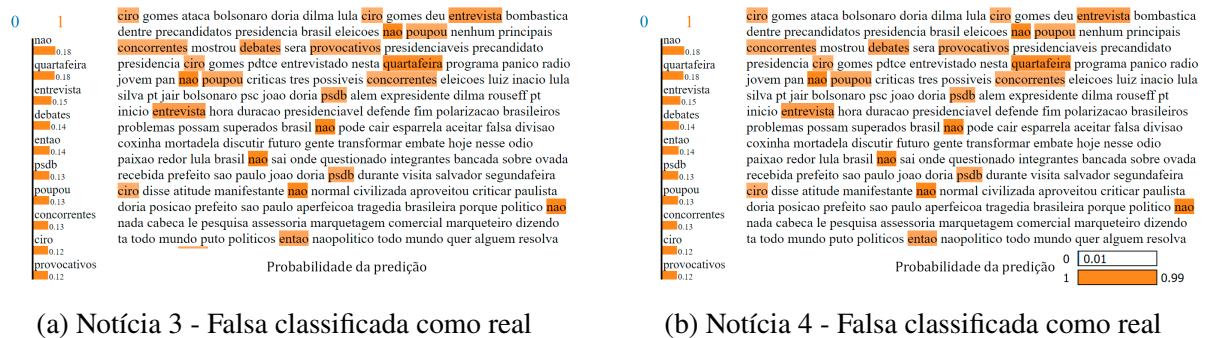
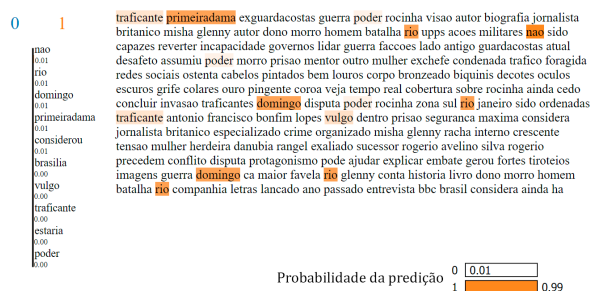
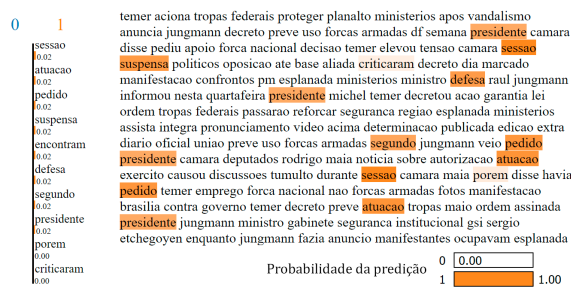


Figura 77 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo LIME.

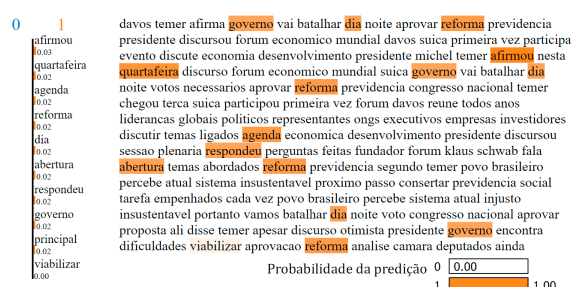


D.3 Modelo 4 - Explicações do LIME de amostras Reais classificadas como Reais

Figura 78 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo LIME.

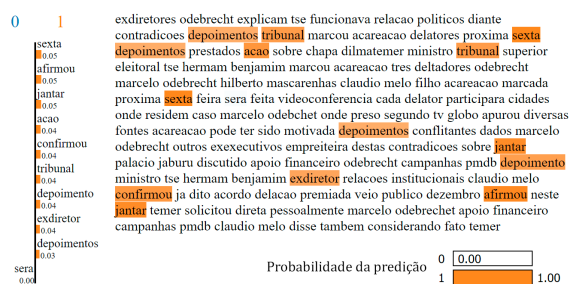


(a) Notícia 1 - Real classificada como real

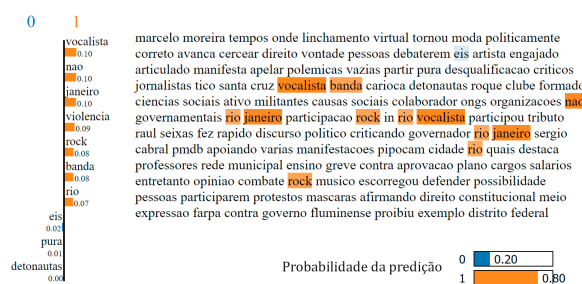


(b) Notícia 2 - Real classificada como real

Figura 79 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo LIME.

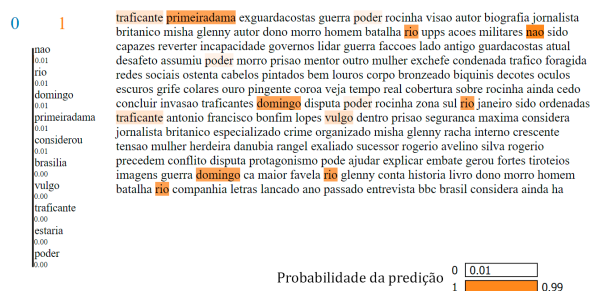


(a) Notícia 3 - Real classificada como real



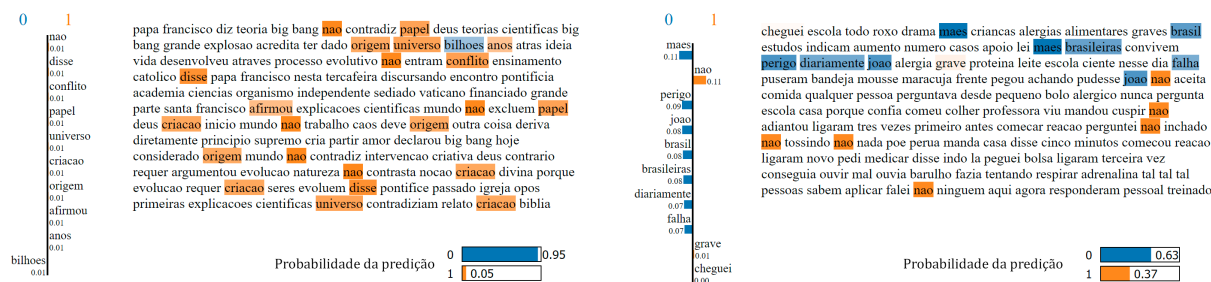
(b) Notícia 4 - Real classificada como real

Figura 80 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo LIME.



D.4 Modelo 4 - Explicações do LIME de amostras Reais classificadas como Falsas

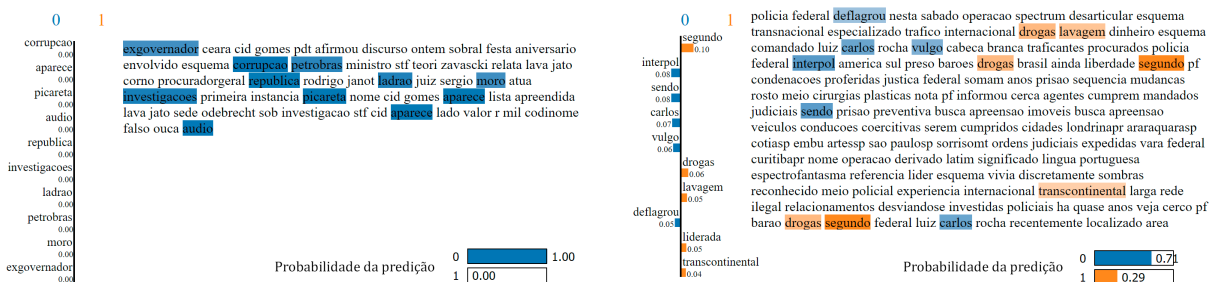
Figura 81 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo LIME.



(a) Notícia 1 - Real classificada como falsa

(b) Notícia 2 - Real classificada como falsa

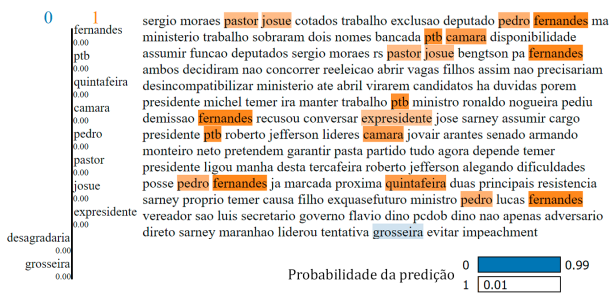
Figura 82 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo LIME.



(a) Notícia 3 - Real classificada como falsa

(b) Notícia 4 - Real classificada como falsa

Figura 83 – Modelo 4 - Explicação da notícia 5 real classificada como falsa gerada pelo LIME.



APÊNDICE E – Explicações geradas pelo IG -

Modelo 1: FakeRecogna incluindo *stop words*

Neste capítulo estão todas as explicações geradas pelo IG, sobre as amostras classificadas pelo Modelo 1 treinado com o FakeRecogna incluindo *stop words*.

E.1 Modelo 1 - Explicações do IG de amostras Falsas classificadas como Falsas

Figura 84 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.

“ Sobre o PM que ‘ surtou ’ na Bahia hj . Não , ele não surtou , ele se negou a prender trabalhador , tomou voz de prisão , não obedeceu . Pegou sua arma e fez disparos para cima . [...] Quando é um bandido , negociam horas e horas e horas , quando é um irmão de farda , simplesmente MATAM . [...] ” , indicam os textos compartilhados milhares de vezes no Facebook (1 , 2 , 3) e no Twitter (1 , 2) . Na tarde do último dia 28 de março , a imprensa (1 , 2 , 3) relatou que um policial militar com o rosto pintado de verde e amarelo estava na região do Farol da Barra , em Salvador , atirando para o alto . Ele carregava um fuzil e uma pistola . A condição psicológica do policial foi informada no site da Secretaria de Segurança Pública (SSP) da Bahia . Segundo as informações , os agentes gerenciavam “ o surto psicológico de um soldado da PM ” , que “ efetuou disparos com arma de fogo para cima ” . Indicaram , ainda , que o Batalhão de Operações Policiais Especiais havia sido acionado .

(a) Notícia 1 - Falsa classificada como falsa

circula pelas redes sociais um texto que diz que um assessor do pt comprou um carro de luxo três dias antes de ganhar na mega - sena . a mensagem , reproduzida em posts nas redes sociais , é acompanhada de uma imagem com o automóvel em questão e uma pessoa a direção .

(b) Notícia 2 - Falsa classificada como falsa

Figura 85 – Modelo 1 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.

o banco central (bc) , responsável exclusivo pela emissão da moeda brasileira , negou de forma taxativa a possibilidade de mudança no sistema financeiro nacional

(a) Notícia 3 - Falsa classificada como falsa

a indicação de Lurian Lula da Silva , filha do ex - presidente Luiz Inácio Lula da Silva (pt) , a um cargo no Senado Federal foi notícia há cerca de uma semana , e não se trata de fake news . no entanto , logo após o uol confere detectou notícias falsas afirmando que Lurian teria um salário de nada menos que R\$ 58 mil , embora a história tenha seu fundo de verdade , a conta não é bem assim . veja a mensagem que distorce os fatos :

(b) Notícia 4 - Falsa classificada como falsa

Figura 86 – Modelo 1 - Explicação da notícias 5 falsa classificada como falsa gerada pelo IG.

"oi pessoal , recebi um alerta para não comprarmos peixe nos próximos 10 dias . o local onde ficam armazenados na ceagesp ficou inundado com risco de contrair doenças principalmente a leptospirose . lavar as frutas verduras e legumes bem antes de descascar e consumir ! não comer nada cru , nem salada . vai o alerta ! ; pessoal bom dia , como alguns de vocês sabem , trabalho numa distribuidora de carnes e congelados . nossa centro de distribuição fica no ceagesp . ontem impossível dese trabalhar , e hoje vim aqui conferir toda a situação . pessoal de vdd , se puderem , evitem frutas e verduras nos próximos 15 dias . além dos preços absurdos , a qualidade será a das piores . essa história que será tudo descartado é pura ilusão . tem muita , muita , muita comida no meio dessa lama e lixo todo . o que realmente estragou certamente irá para o lixo , mas os demais gente . certamente será lavado e colocado para consumo , porque são toneladas e toneladas de alimentos . e o prejuízo desses fornecedores serão imensos , e nesse país onde o dinheiro fala mais alto , duvido que tudo isso aqui será jogado fora . a sujeira é imensa , bicho morto a dar com pau aqui então fica o alerta e dica , para evitarem o

E.2 Modelo 1 - Explicações do IG de amostras Falsas classificadas como Reais

Figura 87 – Modelo 1 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.

São enganosas as informações de um tuite publicado no dia 12 de outubro alegando que 80 % das vagas deixadas pelos cubanos no Mais Médicos não foram preenchidas por médicos brasileiros . O programa do governo federal , que chegou a ter 18 . 337 médicos atuantes em mais de 3 . 6 mil municípios brasileiros , conta atualmente com 15 . 003 profissionais . Desde a saída de Cuba do programa , em novembro de 2018 , foram abertos três editais para suprir o quadro de vagas em municípios eleitos como prioritários . O fim da parceria com Cuba deixou cerca de 8 , 5 mil vagas abertas . O Ministério da Saúde abriu uma primeira seleção com 8 . 517 vagas ainda em novembro , que teve 7 . 120 postos preenchidos e uma segunda com 1 . 397 vagas remanescentes em dezembro . De acordo com o governo , após esse segundo edital , todas as vagas deixadas por cubanos foram preenchidas por brasileiros com registro no Conselho Regional de Medicina (CRM) ou graduados

(a) Notícia 1 - Falsa classificada como real

O presidente Jair Bolsonaro fez um discurso nesta terça - feira (22) na [UNK] Assembleia Geral das Nações Unidas (ONU) . A fala foi apresentada por meio de um vídeo gravado . Por causa da pandemia de Covid - 19 , a reunião da ONU neste ano , baseada na sede da entidade em Nova York , teve de ser virtual . Bolsonaro disse que o Brasil é " vítima " de uma campanha " brutal " de desinformação sobre a Amazônia e o Pantanal . Segundo ele , a floresta amazônica é úmida e só pega fogo nas bordas e os responsáveis pelas queimadas são " índios " e " caboclos " . Disse também que as orientações para as pessoas ficarem em casa na pandemia " quase " levaram o país ao " caos social " . Bolsonaro disse : " Nossa floresta é úmida e não permite a propagação do fogo em seu interior . Os incêndios acontecem praticamente nos mesmos lugares , no entorno leste da floresta , onde o caboclo e o índio queimam seus roçados em busca de sua sobrevivência , em áreas já desmatadas " .

(b) Notícia 2 - Falsa classificada como real

E.3 Modelo 1 - Explicações do IG de amostras Reais classificadas como Reais

Figura 88 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.

Cidade do México , 21 jul (EFE) . - O México ultrapassou as marcas de 40 mil mortes por covid - 19 e 356 mil casos de infecção pelo novo coronavírus nesta terça - feira , dia em que foram relatados 915 novos óbitos e mais 6 . 859 contágios . Em entrevista coletiva , o diretor de epidemiologia do governo do México , José Luis Alomia , confirmou que até agora houve no país 40 . 400 vítimas do vírus SARS - CoV - 2 , que já infectou 356 . 255 pessoas no país desde o dia 28 de fevereiro . Há ainda 2 . 309 mortes suspeitas , mas que ainda precisam de confirmação laboratorial para entrarem para as estatísticas da pandemia . Alomia também disse que até o momento 227 . 165 pessoas se recuperaram da doença e que há 82 . 866 casos suspeitos aguardando resultados clínicos .

(a) Notícia 1 - Real classificada como real

Sem o apoio do presidente Jair Bolsonaro (sem partido) , a vacinação contra a covid - 19 no Brasil precisa ser estimulada por " pessoas de visibilidade " para que a população aceite a imunização , diz a médica epidemiologista Denise Garrett , vice - presidente do Sabin Vaccine Institute , uma organização nos Estados Unidos que incentiva a imunização pelo mundo . Ao longo da pandemia , Bolsonaro disse que não irá se vacinar . " Eu não vou tomar vacina e ponto final " , disse na última terça - feira (15) em entrevista à TV Bandeirantes . " Se minha vida está em risco , o problema é meu . " Tais declarações já têm reflexos na sociedade . Segundo pesquisa do Instituto Datafolha , um em cada cinco brasileiros diz que não quer receber a vacina . Entre os apoiadores de Bolsonaro , 30 % rejeitam a vacinação .

(b) Notícia 2 - Real classificada como real

Figura 89 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.

A empresa espanhola Telefónica anunciou nesta quinta - feira (7) a fusão da sua subsidiária britânica O2 com a Virgin Media , que pertence ao grupo americano Liberty Global , criando um peso pesado do setor das telecomunicações avaliado em 38 bilhões de libras (43 , 4 bilhões de euros) . " A joint venture , com participação de 50 % das duas empresas , unirá a Virgin Media , proprietária da rede de banda larga mais rápida do Reino Unido , e a O2 , a maior plataforma mobile do país " , indicou a Telefónica em um comunicado . " O valor da empresa da entidade combinada está avaliado em quase 38 bilhões de libras , incluindo as sinergias " , afirmou a Telefónica em uma apresentação destinada aos analistas . Esta é a operação de fusão - aquisição mais importante desde o início da pandemia do novo coronavírus , que paralisou a economia do planeta , destaca a agência France Presse

No primeiro semestre de 2020 , pesquisadores da Universidade Harvard já falavam da possibilidade de um vaivém de quarentenas ao longo da pandemia de coronavírus , causador da covid - 19 . Ou seja , moradores viveriam uma espécie de abre e fecha constante à medida que as ondas de infecção avançam e recuam . Um ano depois , grande parte do Brasil já dá sinais da chegada de uma terceira onda da doença , menos de dois meses depois do recuo da onda anterior , que chegou a matar mais de 4 . 000 pessoas por dia no país .

(a) Notícia 3 - Real classificada como real

(b) Notícia 4 - Real classificada como real

Figura 90 – Modelo 1 - Explicação da notícias 5 real classificada como real gerada pelo IG.

O ex - prefeito carioca Eduardo Paes (DEM) foi o entrevistado desta quinta - feira (5) na série de sabatinas promovidas pela Folha de S . Paulo e pelo UOL com os candidatos à Prefeitura do Rio de Janeiro . Na entrevista , o candidato criticou a gestão do atual prefeito , Marcelo Crivella (Republicanos) , e se defendeu de acusações de corrupção em seu governo . A Lupa verificou algumas das declarações do candidato . Confira : No [governo] Crivella , [os aumentos do pedágio da Linha Amarela] foram o dobro da inflação Eduardo Paes , candidato à prefeitura do Rio de Janeiro , na sabatina

E.4 Modelo 1 - Explicações do IG de amostras Reais classificadas como Falsas

Figura 91 – Modelo 1 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.

A estupidez saiu do controle . Ou será que ele está com medos silenciados ?

E daí ? O importante é falar à sua horda de extremistas . Dane - se o resto .

(a) Notícia 1 - Real classificada como falsa

(b) Notícia 2 - Real classificada como falsa

Figura 92 – Modelo 1 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.

O cantor Cauan , da dupla Cleber & Cauan , anunciou hoje no Instagram que está internado com covid - 19 desde quarta - feira (12) . O texto da postagem afirma que no dia da admissão de Cauan no hospital , a doença já havia comprometido mais de 50 % de seu pulmão . " O cantor ainda não apresentou nenhuma melhora significativa " , diz . Comunicado oficial . Na sexta - feira passada (7) , o cantor Cauan (Cleber & Cauan) sentiu febre , dores no corpo e passou o final de semana acamado . Na segunda - feira , realizou o exame PCR e no dia seguinte , infelizmente , foi diagnosticado com Covid - 19 . Na última quarta - feira retornou ao médico , com febre altíssima . Realizou raio - x do tórax e o resultado apontou que o pulmão estava comprometido em mais de 50 % . Ele foi internado imediatamente , com dores também no estômago devido à medicação oral e a baixa oxigenação . O cantor ainda não apresentou nenhuma melhora significativa . . Estamos torcendo para que o Cauan fique bem o quanto antes e passe por essa fase delicada como vencedor . Contamos com a energia positiva e orações de todos os fãs e seguidores . ? ? Edit : @ marianagmoraes está sem máscara junto de Cauan

(a) Notícia 3 - Real classificada como falsa

Uma notícia foi bastante compartilhada nas redes sociais e chegou a portais das regiões Norte e Nordeste na semana passada . O texto afirma que uma candidata a vereadora do município de Presidente Dutra , no Maranhão , localizado a cerca de 350 km da capital São Luís , teria sido presa pela Polícia Civil acusada de oferecer sexo em troca de votos . A informação é fantasiosa . Segundo a Polícia Civil do Maranhão , não há qualquer registro de prisão de uma candidata do município nessas circunstâncias . A notícia falsa afirma ainda que a polícia teria encontrado no celular da candidata vídeos que ela utilizaria " como garantia " , o que caracterizaria como crime de extorsão , e que " a polícia estima [. . .] que possam ser mais de 250 pessoas envolvidas nesse ' negócio ' " .

(b) Notícia 4 - Real classificada como falsa

Figura 93 – Modelo 1 - Explicação da notícias 5 real classificada como falsa gerada pelo IG.

Melhor a arte do encontro do que a do rompimento .

APÊNDICE F – Explicações geradas pelo IG -

Modelo 2: FakeRecogna sem *stop words*

Neste capítulo estão todas as explicações geradas pelo IG, sobre as amostras classificadas pelo Modelo 2 treinado com o FakeRecogna sem *stop words*.

F.1 Modelo 2 - Explicações do IG de amostras Falsas classificadas como Falsas

Figura 94 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.

sobre pm surtou bahia hj nao nao surtou negou prender trabalhador tomou voz
prisao nao obedeceu pegou arma fez disparos cima bandido negociam horas horas
horas irmao farda simplesmente matam indicam textos compartilhados milhares
vezes facebook twitter tarde ultimo dia marco imprensa relatou policial militar rosto
pintado verde amarelo regioa farol barra salvador atirando alto carregava fuzil
pistola condicao psicologica policial informada site secretaria seguranca publica ssp
bahia segundo informacoes agentes gerenciavam surto psicologico soldado pm
efetuou disparos arma fogo cima indicaram ainda batalhao operacoes policiais
especiais havia sido acionado

(a) Notícia 1 - Falsa classificada como falsa

circula redes sociais texto diz assessor pt comprou carro luxo tres dias antes ganhar
megasena mensagem reproduzida posts redes sociais acompanhada imagem
automovel questao pessoa direcao

(b) Notícia 2 - Falsa classificada como falsa

Figura 95 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.

banco central bc responsavel exclusivo emissao moeda brasileira negou forma
taxativa possibilidade mudanca sistema financeiro nacional

(a) Notícia 3 - Falsa classificada como falsa

indicacao lurian lula silva filha expresidente luiz inacio lula silva pt cargo senado
federal noticia ha cerca semana nao trata fake news entanto logo apos uol confere
detectou noticias falsas afirmando lurian salario nada menos r mil embora historia
fundo verdade conta nao bem assim veja mensagem distorce fatos

(b) Notícia 4 - Falsa classificada como falsa

Figura 96 – Modelo 2 - Explicação da notícia 5 falsa classificada como falsa gerada pelo IG.

oi pessoal recebi alerta nao compramos peixe proximos dias local onde ficam
 armazenados ceagesp ficou inundado risco contrair doencas principalmente
 leptospirose lavar frutas verduras legumes bem antes descascar consumir nao comer
 nada cru salada vai alertapessoal bom dia alguns voces sabem trabalho distribuidora
 carnes congelados centro distribuicao fica ceagesp ontem impossivel dese trabalhar
 hoje vim aqui conferir toda situacao pessoal vdd puderem evitem frutas verduras
 proximos dias alem precos absurdos qualidade sera piores historia sera tudo
 descartado pura ilusao muita muita comida meio dessa lama lixo todoo
 realmente estragou certamente ira lixo demais gente certamente sera lavado
 colocado consumo porque sao toneladas toneladas alimentos prejuizo desses
 fornecedores serao imensos nesse pais onde dinheiro fala alto duvido tudo aqui sera
 jogado sujeira imensa bicho morto dar pau aqui entao fica alerta dica evitarem
 puderem

F.2 Modelo 2 - Explicações do IG de amostras Falsas classificadas como Reais

Figura 97 – Modelo 2 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.

presidente jair bolsonaro fez discurso nesta tercafeira assembleia geral nacoes
 unidas onu fala apresentada meio video gravado causa pandemia covid reuniao onu
 neste ano baseada sede entidade nova york virtual bolsonaro disse brasil vitima
 campanha brutal desinformacao sobre amazonia pantanal segundo floresta
 amazonica umida so pega fogo bordas responsaveis queimadas sao indios caboclos
 disse tambem orientacoes pessoas ficarem casa pandemia quase levaram pais caos
 social bolsonaro disse floresta umida nao permite propagacao fogo interior
 incendios acontecem praticamente mesmos lugares entorno leste floresta onde
 caboclo indio queimam roçados busca sobrevivencia areas ja desmatadas

grupo chines compra azul r bi acordo dara azul tempo fazer planejada abertura
 capital segundo presidenteexecutivo companhia antonoaldo neves nao estavamos
 pressao ipo oferta inicial acoes sigla ingles agora prazo ainda maior disse eleterceira
 maior empresa setor aereo brasil azul desistiu junho terceira vez fazer ipo devido
 condicoes adversas mercado capitais desistencia ocorreu pouco antes empresa
 receber aporte milhoes dolares americana united airlines grupo united continental

(a) Notícia 1 - Falsa classificada como real

(b) Notícia 2 - Falsa classificada como real

Figura 98 – Modelo 2 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo IG.

presidente jair bolsonaro psl justificou ultimo dia baixo volume financiamentos
 caixa concedidos nordeste ocorria conta maior inadimplencia regioao prefeituras
 nordeste sao inadimplentes caixa precisa garantias poder emprestar disse uol
 consultou banco dados cauc servico auxiliar informacoes transferencias voluntarias
 controlado secretaria tesouro nacional verificar situacao todos municipios pais
 descobriu nao ha diferencas regionais nordeste comparacao media nacional caso
 especifico capitais estados tambem nao ha qualquer justificativa legal impeca
 repasses

governador wilson witzel pscrj fez pronunciamento palacio laranjeiras manha desta
 sexta-feira apos afastado cargo superior tribunal justica

(a) Notícia 3 - Falsa classificada como real

(b) Notícia 4 - Falsa classificada como real

Figura 99 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo IG.

exprefeito carioca eduardo paes dem entrevistado desta quinta-feira serie sabatinas promovidas folha spaulo uol candidatos prefeitura rio janeiro entrevista candidato criticou gestao atual prefeito marcelo crivella republicanos defendeu acusacoes corrupcao governo lupa verificou algumas declaracoes candidato confira governo crivella aumentos pedagio linha amarela dobro inflacaoeduardo paes candidato prefeitura rio janeiro sabatina

F.3 Modelo 2 - Explicações do IG de amostras Reais classificadas como Reais

Figura 100 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.

cidade mexico jul efe mexico ultrapassou marcas mil mortes covid mil casos infeccao novo coronavirus nesta terca-feira dia relatados novos obitos contagios entrevista coletiva diretor epidemiologia governo mexico jose luis alomia confirmou ate agora pais vitimas virus sarscov ja infectou pessoas pais desde dia fevereiro ha ainda mortes suspeitas ainda precisam confirmacao laboratorial entrarem estatisticas pandemia alomia tambem disse ate momento pessoas recuperaram doenca ha casos suspeitos aguardando resultados clinicos

(a) Notícia 1 - Real classificada como real

apoio presidente jair bolsonaro partido vacinacao contra covid brasil precisa estimulada pessoas visibilidade populacao aceite imunizacao diz medica epidemiologista denise garrett vicepresidente sabin vaccine institute organizacao estados unidos incentiva imunizacao mundo longo pandemia bolsonaro disse nao ira vacinar nao vou tomar vacina ponto final disse ultima terca-feira entrevista tv bandeirantes vida risco problema tais declaracoes ja reflexos sociedade segundo pesquisa instituto datafolha cada cinco brasileiros diz nao quer receber vacina apoiadores bolsonaro rejeitam vacinacao

(b) Notícia 2 - Real classificada como real

Figura 101 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.

empresa espanhola telefonica anunciou nesta quinta-feira fusao subsidiaria britanica virgin media pertence grupo americano liberty global criando peso pesado setor telecomunicacoes avaliado bilhoes libras bilhoes euros joint venture participacao duas empresas unira virgin media proprietaria rede banda larga rapida reino unido maior plataforma mobile pais indicou telefonica comunicado valor empresa entidade combinada avaliado quase bilhoes libras incluindo sinergias afirmou telefonica apresentacao destinada analistas operacao fusaoaquisicao importante desde inicio pandemia novo coronavirus paralisou economia planeta destaca agencia france presse

(a) Notícia 3 - Real classificada como real

primeiro semestre pesquisadores universidade harvard ja falavam possibilidade vaivem quarentenas longo pandemia coronavirus causador covid moradores viveriam especie abre fecha constante medida ondas infeccao avancam recuam ano grande parte brasil ja sinais chegada terceira onda doenca menos dois meses recuo onda anterior chegou matar pessoas dia pais

(b) Notícia 4 - Real classificada como real

Figura 102 – Modelo 2 - Explicação da notícia 5 real classificada como real gerada pelo IG.

exprefeito carioca eduardo paes dem entrevistado desta quinta-feira serie sabatinas promovidas folha spaulo uol candidatos prefeitura rio janeiro entrevista candidato criticou gestao atual prefeito marcelo crivella republicanos defendeu acusacoes corrupcao governo lupa verificou algumas declaracoes candidato confira governo crivella aumentos pedagio linha amarela dobro inflacaoeduardo paes candidato prefeitura rio janeiro sabatina

F.4 Modelo 2 - Explicações do IG de amostras Reais classificadas como Falsas

Figura 103 – Modelo 2 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.

estupidez saiu controle sera medos silenciados

dai importante falar horda extremistas danese resto

(a) Notícia 1 - Real classificada como falsa

(b) Notícia 2 - Real classificada como falsa

Figura 104 – Modelo 2 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.

cantor cauan dupla cleber cauan anunciou hoje instagram internado covid desde quarta-feira texto postagem afirma dia admissao cauan hospital doenca ja havia comprometido pulmao cantor ainda nao apresentou nenhuma melhora significativa diz comunicado oficial sexta-feira passada cantor cauan cleber cauan sentiu febre dores corpo passou final semana acamado segunda-feira realizou exame per dia seguinte infelizmente diagnosticado covid ultima quarta-feira retornou medico febre altissima realizou raiox torax resultado apontou pulmao comprometido internado imediatamente dores tambem estomago devido medicacao oral baixa oxigenacao cantor ainda nao apresentou nenhuma melhora significativa torcendo cauan fique bem quanto antes passe fase delicada vencedor contamos energia positiva oracoes todos fas seguidores edit marianagmoraes mascara junto cauan pois tambem diagnostico covid encontrase bem acompanhamento publicacao compartilhada cleber cauan cleberecauan ago pdt

noticia bastante compartilhada redes sociais chegou portais regioes norte nordeste semana passada texto afirma candidata vereadora municipio presidente dutra maranhao localizado cerca km capital sao luis sido presa policia civil acusada oferecer sexo troca votos informacao fantasiosa segundo policia civil maranhao nao ha qualquer registro prisao candidata municipio nessas circunstancias noticia falsa afirma ainda policia encontrado celular candidata videos utilizaria garantia caracterizaria crime extorsao policia estima possam pessoas envolvidas nesse negocio

(a) Notícia 3 - Real classificada como falsa

(b) Notícia 4 - Real classificada como falsa

Figura 105 – Modelo 2 - Explicação da notícia 5 real classificada como falsa gerada pelo IG.

melhor arte encontro rompimento

APÊNDICE G – Explicações geradas pelo IG

- Modelo 3: Fake.Br Corpus incluindo *stop words*

Neste capítulo estão todas as explicações geradas pelo IG, sobre as amostras classificadas pelo Modelo 3 treinado com o Fake.Br Corpus incluindo *stop words*.

G.1 Modelo 3 - Explicações do IG de amostras Falsas classificadas como Falsas

Figura 106 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.

Estagiário é flagrado " ao vivo " assistindo " filminho adulto " durante apresentação de telejornal . . Telespectadores que assistiam um noticiário da BBC transmitido na última segunda - feira (7) ficaram de queixo caído ! Eram quase 10h da noite quando a âncora Sophie Raworth falava sobre a vitória da equipe britânica de cricket sobre a África do Sul . Um detalhe inusitado conseguiu chamar mais atenção do que a própria notícia que estava sendo transmitida . Atrás da apresentadora havia um estagiário distraído brincando com seu computador [. . .] ele assistia um filminho de " sacanagem " . O resultado ? Tudo foi transmitido ao vivo ! Não poderia acabar de modo diferente . O estagiário foi afastado temporariamente de suas funções .

(a) Notícia 1 - Falsa classificada como falsa

Marqueteiro do PT está inquieto na cadeia . Ele grita dia e noite e diz que não vai poupar ninguém . Informações da República de Curitiba dão conta de que o marqueteiro do PT está falando pelos cotovelos dentro da cadeia . Uma fonte disse que João Santana está ansioso , tenso e decidiu que não quer passar o resto de seus dias na cadeia . Santana , que se desde 23 / 02 (data de sua prisão) estava calado , resolveu que irá botar a boca no trombone . A defesa do publicitário já está negociando um possível acordo com a equipe de investigação da PF . Ele será interrogado pelo juiz Sérgio Moro na próxima quinta - feira , 28 . Ele é réu de uma ação penal decorrente da 23a fase da operação Lava Jato . Santana cuidou da publicidade durante a campanha de reeleição do ex - presidente Luiz Inácio Lula da Silva , em 2006 , e as duas campanhas da presidente Dilma Rousseff , em 2010 e 2014 . A investigação aponta repasses no exterior de US \$ 7 , 5 milhões da Odebrecht para o casal João e Mônica Santana entre 2012 e 2014 .

(b) Notícia 2 - Falsa classificada como falsa

Figura 107 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.

Há 43 anos , Ernesto Geisel avisou : " Chegará um tempo em que o povo sentirá saudades dos militares " . Geisel foi o 29o Presidente do Brasil (1974 a 1979) durante o regime militar brasileiro . Em 1974 , ele profetizou : " Se é a vontade do povo brasileiro eu promoverei a Abertura Política no Brasil . Mas chegará um tempo que o povo sentirá saudade do Regime Militar . Pois muitos desses que lideram o fim do Regime não estão visando o bem do povo , mas sim seus próprios interesses . " Será que ele tinha razão ?

(a) Notícia 3 - Falsa classificada como falsa

Campo de petróleo na África deu prejuízo de R \$ 25 bilhões para a Petrobras . Teve muita propina ! . O rombo alcança mais de R \$ 25 , 2 bilhões . A Petrobras informou que estes contratos foram firmados com base em expectativas de lucro , porém contrariaram pareceres técnicos que não aconselhavam o negócio . TEVE MUITA PROPINA [campo de petróleo no Benin / África] " No negócio do Petróleo é normal você perfurar alguns poços secos . Na Petrobras , o que aconteceu foi a junção da ineficiência administrativa com a corrupção " declarou o procurador do MPF Diogo Castor de Mattos , da força - tarefa da Lava Jato . A área internacional da Petrobras [responsável por esse negócio] era comandada pelo PMDB e os diretores da estatal eram Nestor Cerveró [2003 / 2008] e Jorge Zelada [2008 / 2012] . Ambos foram condenados pelo recebimento de propinas . Cerveró confirmou [em delação] que a cúpula do PMDB [leia - se Cunha e Renan

(b) Notícia 4 - Falsa classificada como falsa

Figura 108 – Modelo 3 - Explicação da notícia 5 falsa classificada como falsa gerada pelo IG.

Pai que perdeu gêmeos de 9 meses , a esposa e 20 familiares em ataque químico na Síria , clama por mais ajuda dos EUA .

G.2 Modelo 3 - Explicações do IG de amostras Falsas classificadas como Reais

Figura 109 – Modelo 3 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.

6 maneiras positivas de superar o medo da morte . É neste ponto que devemos olhar para trás e contar nossas bênçãos e vitórias . Criamos uma comunidade de amigos , parentes e colegas que se estendem há décadas . A maioria de nós vive entusiasmada com as muitas décadas de vida que ainda podemos ter décadas que queremos preencher com paixões , pessoas e lugares que nos interessam . Quando chegamos na casa dos " enta " (50 , 60) é comum começar a se preocupar com nossa mortalidade . Muitos de nós começamos a pensar sobre o fato de que temos menos anos pela frente e mais anos para trás . Diversas pessoas têm medo da morte , não importa o quão longe ela possa estar . Abaixo estão algumas dicas com base no conselho de outras pessoas com mais de 50 e 60 anos que superaram o medo de morrer . Gaste tempo com as pessoas que você gosta de estar por perto . Tente coisas novas . Desafie - se . Principalmente , mantenha - se ativo e comprometido com atividades positivas . Se houver algo que realmente o desafie faça algo sobre

(a) Notícia 1 - Falsa classificada como real

Bolsonaro pode ser preso por suas declarações polêmicas ? O deputado federal Jair Bolsonaro (PSC - RJ) disse que teve uma filha porque fraquejou , que quilombolas não serviam nem para procriar e que o erro da ditadura foi não matar mais dissidentes . Com essas declarações , gerou desconforto em vários grupos sociais . Mas só a minoria das frases que o acusam de ser discriminatórias e preconceituosas pode ser enquadrada como crime e , ainda assim , depende da interpretação da Justiça . A FOLHA ouviu os advogados criminalistas Fábio Olivetti Lins e Silva e Leonardo Sakamoto Beltrão , para saber se o deputado Jair Messias Bolsonaro corre o risco real de ser preso . Como deputado , ele tem a imunidade parlamentar , o que lhe garante não ser punido por suas opiniões , palavras e votos , de acordo com o artigo 53 da Constituição Federal . Para o advogado Fábio Olivetti Lins e Silva , Bolsonaro não pode ser acusado criminalmente pelas suas frases . Além disso , muitas de suas colocações são expressão de opinião , o que é assegurado pela Constituição . Mas a linha é tênue . Ele é inteligente e caminha no limite entre o que

(b) Notícia 2 - Falsa classificada como real

Figura 110 – Modelo 3 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo IG.

Ciro Gomes ataca Bolsonaro , Dória , Dilma e Lula . Giro Gomes deu a entrevista mais bombástica dentre os pré - candidatos à presidência do Brasil nas eleições 2018 . Não poupou nenhum de seus principais concorrentes e mostrou que nos debates será um dos mais provocativos presidenciais . O pré - candidato à presidência Giro Gomes (PDT - CE) foi entrevistado nesta quarta - feira no programa Pânico , da rádio Jovem Pan , e não poupou críticas a três possíveis concorrentes nas eleições de 2018 : Luiz Inácio Lula da Silva (PT) , Jair Bolsonaro (PSC) e João Doria (PSDB) , além da ex - presidente Dilma Rousseff (PT) . No início da entrevista , que teve mais de uma hora de duração , o presidencial defende o fim da polarização entre os brasileiros para que os problemas possam ser superados . O Brasil não pode cair na esparrela de aceitar essa falsa divisão entre coxinha e mortadela . Tem que discutir o futuro . Se a gente for transformar o embate que nós temos hoje nesse ódio ou paixão ao redor do Lula , o Brasil não sai de onde está . Ao ser questionado pelos integrantes da bancada sobre a avada

(a) Notícia 3 - Falsa classificada como real

History Channel : Hitler forjou a própria morte e fugiu para a Argentina . Documentos comprovavam que Adolf Hitler e Eva Braun fugiram para a Argentina em um Submarino . O FBI revelou que o governo dos EUA sabia que Adolf Hitler estava vivo e , em 1984 , ele estava morando na Cordilheira dos Andes , muito tempo depois do fim da Segunda Guerra Mundial . (History Channel) Nos últimos 70 anos , o mundo foi levado a acreditar que Hitler teria morrido em 30 de abril de 1945 depois de se suicidar em seu bunker subterrâneo . De acordo com a história oficial , seu corpo foi descoberto e identificado pelos soviéticos e depois transportado para a Rússia . OS RUSSOS MENTIRAM ? Com a divulgação desses documentos do FBI , certamente parece que o líder mais notório da história mundial escapou da Alemanha para desfrutar uma vida pacífica no alto da Cordilheira dos Andes , América do Sul . Recentemente uma foto (abaixo) foi postada nas redes sociais sugerindo um " suposto " Hitler idoso , de 95 anos , posando com a namorada no Brasil em 1984 . Os recém - lançados arquivos do FBI afirmam que

(b) Notícia 4 - Falsa classificada como real

Figura 111 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo IG.

Traficante , primeira - dama e ex - guarda - costas em guerra de poder na Rocinha , a visão de autor de biografia de Nem . Para jornalista britânico Misha Glenny , autor de O dono do morro : um homem e a batalha pelo Rio , UPPs e ações militares , não têm sido capazes de reverter incapacidade de governos para lidar com guerra de facções . De um lado , o antigo guarda - costas e atual desafeto , que assumiu o poder do morro depois da prisão do mentor . De outro , a mulher do ex - chefe , condenada por tráfico , foragida , mas que , nas redes sociais , ostenta os cabelos pintados bem louros , o corpo bronzeado com biquínis e decotes , óculos escuros de grife e colares de ouro com pingente de coroa . Veja em tempo real a cobertura sobre a Rocinha Ainda é cedo para concluir que a invasão por traficantes no domingo e a disputa de poder na Rocinha , na zona sul do Rio de Janeiro , tenham sido ordenadas pelo traficante Antônio Francisco Bonfim Lopes , vulgo Nem , de dentro de uma prisão de segurança máxima , considera o jornalista

G.3 Modelo 3 - Explicações do IG de amostras Reais classificadas como Reais

Figura 112 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.

Temer aciona tropas federais para proteger Planalto e ministérios após vandalismo , anuncia Jungmann . Decreto prevê uso de Forças Armadas no DF por uma semana . Mas o presidente da Câmara disse que pediu apoio da Força Nacional . Decisão de Temer elevou tensão na Câmara , e sessão foi suspensa . Políticos da oposição e até da base aliada criticaram o decreto . Dia foi marcado por manifestação e confrontos com a PM na Esplanada dos Ministérios . O ministro da Defesa , Raul Jungmann , informou nesta quarta - feira (24) que o presidente Michel Temer decretou a " ação de garantia da lei e da ordem " e , com isso , tropas federais passarão a reforçar a segurança na região da Esplanada dos Ministérios (assista à íntegra do pronunciamento no vídeo acima) . A determinação , publicada em uma edição extra do " Diário Oficial da União " , prevê o uso das Forças Armadas e , segundo Jungmann , veio depois de um pedido do presidente da Câmara dos Deputados , Rodrigo Maia . O notícia sobre a autorização para atuação do Exército causou

Em Davos , Temer afirma que governo vai ' batalhar dia e noite ' para aprovar a reforma da Previdência . Presidente discursou no Fórum Econômico Mundial , em Davos , na Suíça . É a primeira vez que ele participa do evento que discute economia e desenvolvimento . O presidente Michel Temer afirmou nesta quarta - feira (24) em discurso no Fórum Econômico Mundial , na Suíça , que o governo vai " batalhar dia e noite " pelos votos necessários para aprovar a reforma da Previdência no Congresso Nacional . Temer , que chegou na terça (23) à Suíça , participou pela primeira vez do fórum de Davos , que reúne todos os anos lideranças globais , como políticos , representantes de ONGS , executivos de empresas e investidores , para discutir temas ligados à agenda econômica e de desenvolvimento . O presidente discursou em sessão plenária e respondeu a perguntas feitas pelo fundador do fórum , Klaus Schwab . Na fala de abertura , um

(a) Notícia 1 - Real classificada como real

(b) Notícia 2 - Real classificada como real

Figura 113 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.

Ex - diretores da Odebrecht explicam ao TSE como funcionava relação com políticos . Diante de contradições em depoimentos , tribunal marcou acareação entre delatores para a próxima sexta ; depoimentos foram prestados em ação sobre chapa Dilma - Temer . O ministro do Tribunal Superior Eleitoral (TSE) Hermam Benjamin marcou uma acareação com três delatores da Odebrecht : Marcelo Odebrecht , Hilberto Mascarenhas e Cláudio Melo Filho . A acareação está marcada para a próxima sexta feira e será feita por videoconferência , ou seja , cada delator participará das cidades onde reside , ou no caso de Marcelo Odebrecht , de onde está preso . Segundo a TV Globo apurou com diversas fontes , acareação pode ter sido por motivada por depoimentos conflitantes dados Marcelo Odebrecht e outros ex - executivos da empreiteira . Uma destas contradições é sobre o jantar no Palácio do Jaburu em 2014 , quando foi discutido apoio financeiro da Odebrecht para campanhas do PMDB . Em seu depoimento ao ministro do TSE Hermam Benjamin , o ex - diretor de relações institucionais Claudio Melo confirmou o que já tinha dito

Marcelo Moreira Em tempos onde o linchamento virtual se tornou moda e o politicamente correto avança para cercar o direito (e a vontade) de as pessoas debaterem , eis que um artista engajado e articulado se manifesta sem apelar para polêmicas vazias e sem partir para a pura desqualificação dos críticos e jornalistas . Tico Santa Cruz , vocalista da banda carioca Detonautas Roque Clube , é formado em ciências sociais e é um ativo militante de causas sociais e colaborador de ONGs (organizações não governamentais) do Rio de Janeiro . Em sua participação no Rock in Rio 2013 , o vocalista , que participou do tributo a Raul Seixas , fez um rápido discurso político , criticando o governador do Rio de Janeiro , Sergio Cabral (PMDB) e apoiando as várias manifestações que pipocam na cidade do Rio , entre as quais se destaca a dos professores da rede municipal de ensino , em greve contra a aprovação de um plano de cargos e salários . Entretanto , na opinião do Combate

(a) Notícia 3 - Real classificada como real

(b) Notícia 4 - Real classificada como real

Figura 114 – Modelo 3 - Explicação da notícia 5 real classificada como real gerada pelo IG.

Traficante , primeira - dama e ex - guarda - costas em guerra de poder na Rocinha ; a visão de autor de biografia de Nem . Para jornalista britânico Misha Glenny , autor de O dono do morro : um homem e a batalha pelo Rio , UPPs e ações militares , não têm sido capazes de reverter incapacidade de governos para lidar com guerra de facções . De um lado , o antigo guarda - costas e atual desafeto , que assumiu o poder do morro depois da prisão do mentor . De outro , a mulher do ex - chefe , condenada por tráfico , foragida , mas que , nas redes sociais , ostenta os cabelos pintados bem louros , o corpo bronzeado com biquínis e decotes , óculos escuros de grife e colares de ouro com pingente de coroa . Veja em tempo real a cobertura sobre a Rocinha Ainda é cedo para concluir que a invasão por traficantes no domingo e a disputa de poder na Rocinha , na zona sul do Rio de Janeiro , tenham sido ordenadas pelo traficante Antônio Francisco Bonfim Lopes , vulgo Nem , de dentro de uma prisão de segurança máxima , considera o jornalista

G.4 Modelo 3 - Explicações do IG de amostras Reais classificadas como Falsas

Figura 115 – Modelo 3 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.

Papa Francisco diz que teoria do Big Bang não contradiz papel de Deus . Teorias científicas como o " Big Bang " , a grande explosão que se acredita ter dado origem ao universo 13 , 7 bilhões de anos atrás , e a ideia de que a vida se desenvolveu através de um processo evolutivo não entram em conflito com o ensinamento católico , disse o papa Francisco nesta terça - feira . Discursando em um encontro da Pontifícia Academia de Ciências , um organismo independente sediado no Vaticano e financiado em grande parte pela Santa Sé , Francisco afirmou que as explicações científicas para o mundo não excluem o papel de Deus na criação . " O início do mundo não é trabalho do caos , que deve sua origem a outra coisa , mas deriva diretamente de um princípio supremo que cria a partir do amor " , declarou . " O Big Bang , que hoje é considerado a origem do mundo , não contradiz a intervenção criativa de Deus , pelo contrário , a requer " , argumentou . " A evolução na natureza não contrasta com a noção de criação (divina) , porque a

(a) Notícia 1 - Real classificada como falsa

Cheguei na escola e ele estava todo roxo : o drama de mães de crianças com alergias alimentares graves no Brasil . Estudos indicam aumento em número de casos : sem apoio da lei , mães brasileiras convivem com perigo diariamente . " O João tem alergia grave à proteína do leite . A escola estava ciente , mas nesse dia houve uma falha e puseram a bandeja com mousse de maracujá na frente dele . Ele pegou , achando que pudesse . O João não aceita comida de qualquer pessoa . Perguntava desde pequeno : Esse bolo é alérgico ? Mas nunca pergunta na escola , ou em casa , porque confia . Comeu uma colher . A professora viu e mandou cuspir . Não adiantou . Me ligaram três vezes . Primeiro , antes de começar a reação . " Como ele está ? " , perguntei . Não está inchado , não está tossindo , não teve nada . Põe na perua e manda para casa , eu disse . Cinco minutos depois , começou a reação . Me ligaram de novo , pedi para medicar e disse que estava indo para lá . Peguei a bolsa . Ligaram pela terceira vez . Eu conseguia ouvir que ele estava muito mal , ouvia o barulho que ele fazia , tentando respirar . Dá a adrenalina . Tal , tal e tal pessoas sabem aplicar , falei . Não tem ninguém aqui agora , responderam . O

(b) Notícia 2 - Real classificada como falsa

Figura 116 – Modelo 3 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.

O ex - governador do Ceará , Cid Gomes (PDT) , afirmou em discurso ontem em Sobral , na sua festa de aniversário , que se ele estiver envolvido no esquema de corrupção da Petrobrás , o ministro do STF , Teori Zavascki , que relata a Lava Jato , é " corno " ; o procurador - geral da República , Rodrigo Janot , é " ladrão " e o juiz Sérgio Moro , que atua nas investigações da primeira instância , é " picareta " . O nome de Cid Gomes aparece numa lista apreendida pela Lava Jato na sede da Odebrecht que esta sob investigação do STF . Cid aparece ao lado do valor de R \$ 200 mil e do codinome " falso " . Ouça o áudio :

(a) Notícia 3 - Real classificada como falsa

Andrés diz : Falar que o Corinthians está quebrado é ' balela ' . Ex - dirigente afirma que situação financeira do clube não é tão grave e que Arena Corinthians pode se pagar sozinha . Andrés Sanchez foi uma sombra de quem estava no poder nos últimos sete anos . Após afirmar diversas vezes que não iria mais se candidatar , mudou de ideia e resolveu tentar voltar ao cargo que ocupou entre 2007 e 2011 . Após as entrevistas de Felipe Ezabella e Romeu Tuma Júnior , o Estado publica a conversa com o ex - dirigente e atual deputado federal pelo Partido dos Trabalhadores (PT) fala o motivo de ter voltado a se candidatar , o que planeja para a Arena Corinthians , futebol do clube , base , dentre outros assuntos . Ezabella diz : Não tenho problema com agentes , mas são sempre os mesmos no Corinthians . Tuma diz : Se a Caixa não aceitar a condição do Corinthians , vamos dar calote . Porque você acha que é o melhor candidato para o Corinthians ?

(b) Notícia 4 - Real classificada como falsa

APÊNDICE H – Explicações geradas pelo IG

- Modelo 4: Fake.Br Corpus sem *stop words*

Neste capítulo estão todas as explicações geradas pelo IG, sobre as amostras classificadas pelo Modelo 4 treinado com o Fake.Br Corpus sem *stop words*.

H.1 Modelo 4 - Explicações do IG de amostras Falsas classificadas como Falsas

Figura 117 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como falsas geradas pelo IG.

estagiario flagrado vivo assistindo filminho adulto durante apresentacao telejornal telespectadores assistiam noticiario bbc transmitido ultima segunda-feira ficaram queixo caído quase h noite ancora sophie raworth falava sobre vitoria equipe britanica cricket sobre africa sul detalhe inusitado conseguiu chamar atencao propria noticia sendo transmitida atras apresentadora havia estagiario distraido brincando computador assistia filminho sacanagem resultado tudo transmitido vivo nao poderia acabar modo diferente estagiario afastado temporariamente funcoes

(a) Notícia 1 - Falsa classificada como falsa

marqueteiro pt inquieto cadeia grita dia noite diz nao vai poupar ninguém informacoes republica curitiba dao conta marqueteiro pt falando cotovelos dentro cadeia fonte disse joao santana ansioso tenso decidiu nao quer passar resto dias cadeia santana desde data prisao calado resolveu ira botar boca trombone defesa publicitario ja negociando possivel acordo equipe investigacao pf sera interrogado juiz sergio moro proxima quinta-feira reu acao penal decorrente fase operacao lava jato santana cuidou publicidade durante campanha reeleicao expresidente luiz inacio lula silva duas campanhas presidente dilma rousseff investigacao aponta repasses exterior us milhoes odebrecht casal joao monica santana

(b) Notícia 2 - Falsa classificada como falsa

Figura 118 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como falsas geradas pelo IG.

ha anos ernesto geisel avisou chegara tempo povo sentira saudades militares geisel presidente brasil durante regime militar brasileiro profetizou vontade povo brasileiro promoverei abertura politica brasil chegara tempo povo sentira saudade regime militar pois muitos desses lideram fim regime nao estao visando bem povo sim proprios interesses sera razao

(a) Notícia 3 - Falsa classificada como falsa

campo petroleo africa deu prejuizo r bilhoes petrobras muita propina rombo alcanca r bilhoes petrobras informou contratos firmados base expectativas lucro porem contrariaram pareceres tecnicos nao aconselhavam negocio muita propina campo petroleo benin africa negocio petroleo normal voce perfurar alguns pocos secos petrobras aconteceu juncao ineficiencia administrativa corrupcao declarou procurador mpf diogo castor mattos forcatarefa lava jato area international petrobras responsavel negocio comandada pmdb diretores estatal nestor certero jorge zelada ambos condenados recebimento propinas cervero confirmou delacao cupula pmdb leiaze cunha renan calheiros beneficiados negam acusacoes milhoes dolares propinas campo petroleo benin rendeu propina gorda us milhoes ate momento apenas parte recebida deputado eduardo cunha descoberta

(b) Notícia 4 - Falsa classificada como falsa

Figura 119 – Modelo 4 - Explicação da notícia 5 falsa classificada como falsa gerada pelo IG.

pai perdeu gêmeos meses esposa familiares ataque químico siria clama ajuda eua

H.2 Modelo 4 - Explicações do IG de amostras Falsas classificadas como Reais

Figura 120 – Modelo 4 - Explicações das notícias 1 e 2 falsas classificadas como reais geradas pelo IG.

maneiras positivas superar medo morte neste ponto devemos olhar tras contar
bencãos vitórias criamos comunidade amigos parentes colegas estendem ha decadas
maioria vive entusiasmada muitas decadas vida ainda podemos ter decadas
queremos preencher paixoes pessoas lugares interessam chegamos casa enta
comum comecar preocupar mortalidade muitos comecamos pensar sobre fato
menos anos frente anos tras diversas pessoas medo morte nao importa quao longe
possa abaixo estao algumas dicas base conselho outras pessoas anos superaram
medo morrer gaste tempo pessoas voce gosta perto tente coisas novas desafiese
principalmente mantenha ativo comprometido atividades positivas algo realmente
desafie faca algo sobre voce negocios inacabados resolvaos voce alguem precisa
falar faca ligacao nao continue indo trabalho profundamente insatisfeito fique
relacionamento deixa infeliz voce muitos anos aproveitar tudo vida oferecer medo
morte muitas vezes medo nao viver acordo proprios principios voce merece ver
sonhos tomarem realidade quanto abraçar vida menos assustado voce ficara chegar
hora ajuda reconhecer parte ciclo evolutivo podemos encontrar conforto fato todos

(a) Notícia 1 - Falsa classificada como real

bolsonaro pode preso declaracoes polemicas deputado federal jair bolsonaro pscij
disse filha porque fraquejou quilombolas nao serviam procriar erro ditadura nao
matar dissidentes declaracoes gerou desconforto varios grupos sociais so minoria
frases acusam discriminatorias preconceituosas pode enquadrada crime ainda assim
depende interpretacao justica folha ouviu advogados criminalistas fabio olivetti lins
silva leonardo sakamoto beltrao saber deputado jair messias bolsonaro corre risco
real preso deputado imunidade parlamentar garante nao punido opinioes palavras
votos acordo artigo constituicao federal advogado fabio olivetti lins silva bolsonaro
nao pode acusado criminalmente frases alem disso muitas colocacoes sao expressao
opiniao assegurado constituicao linha tenue inteligente caminha limite imunidade
permite nao permite diz advogado criminalista leonardo sakamoto beltrao confira
seguir algumas dessas frases interpretacoes juridicas idiota analfabeta ta censurada
ha tres anos durante entrevista bolsonaro irritouse jornalista dirigiu palavras
advogado fabio olivetti lins silva explica constituicao nao preve punicao
discriminacao genero erro ditadura torturar nao matar passagem servico militar

(b) Notícia 2 - Falsa classificada como real

Figura 121 – Modelo 4 - Explicações das notícias 3 e 4 falsas classificadas como reais geradas pelo IG.

ciro gomes ataca bolsonaro doria dilma lula ciro gomes deu entrevista bombastica
dentre precandidatos presidencia brasil eleicoes nao poupou nenhum principais
concorrentes mostrou debates sera provocativos presidenciais precandidato
presidencia ciro gomes pdice entrevistado nesta quarta-feira programa panico radio
jovem pan nao poupou criticas tres possiveis concorrentes eleicoes luiz inacio lula
silva pt jair bolsonaro psc joao doria psdb alem expresidente dilma rouseff pt inicio
entrevista hora duracao presidenciael defende fim polarizacao brasileiros
problemas possam superados brasil nao pode cair esparrela aceitar falsa divisao
coxinha mortadela discutir futuro gente transformar embate hoje nesse odio paixao
redor lula brasil nao sai onde questionado integrantes bancada sobre ovada recebida
prefeito sao paulo joao doria psdb durante visita salvador segunda-feira ciro disse
atitude manifestante nao normal civilizada aproveitou criticar paulista doria posicao
prefeito sao paulo aperfeicoa tragedia brasileira porque politico nao nada cabeça le
pesquisa assessoria marquetagem comercial marqueteiro dizendo ta todo mundo

(a) Notícia 3 - Falsa classificada como real

history channel hitler forjou propria morte fugiu argentina documentos
comprovavam adolf hitler eva braun fugiram argentina submarino fbi revelou
governo eua sabia adolf hitler vivo morando cordilheira andes tempo fim segunda
guerra mundial history channel ultimos anos mundo levado acreditar hitler morrido
abril suicidar bunker subterraneo acordo historia oficial corpo descoberto
identificado sovieticos transportado russia russos mentiram divulgacao desses
documentos fbi certamente parece lider notorio historia mundial escapou alemanha
desfrutar vida pacifica alto cordilheira andes america sul recentemente foto abaixo
postada redes sociais sugerindo suposto hitler idoso anos posando namorada brasil
reclamancados arquivos fbi afirmam adolf hitler escapou atraves submarino destino
argentina onde permaneceu rancho fortemente protegido sofrendo asma ulceras
documentos divulgados mostram nao so suicidio hitler eva braun falsificado casal
tambem ajuda diretor oss servico estrategico alemnha allen dundes possibilitar fuga
documentos liberados revelase fbi ciente misterioso submarino alemao navegava
rumo costa argentina misterioso informante carta los angeles bureau fbi datada

(b) Notícia 4 - Falsa classificada como real

Figura 122 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo IG.

traficante primeiradama exguardacostas guerra poder rocinha visao autor biografia
 jornalista britanico misha glenny autor dono morro homem batalha rio upps acoes
 militares nao sido capazes reverter incapacidade governos lidar guerra faccoes lado
 antigo guardacostas atual desafeto assumiu poder morro prisao mentor outro mulher
 exchefe condenada trafico foragida redes sociais ostenta cabelos pintados bem
 louros corpo bronzeado biquinis decotes olhos escuros grife colares ouro pingente
 coroa veja tempo real cobertura sobre rocinha ainda cedo concluir invasao
 traficantes domingo disputa poder rocinha zona sul rio janeiro sido ordenadas
 traficante antonio francisco bonfim lopes vulgo dentro prisao seguranca maxima
 considera jornalista britanico especializado crime organizado misha glenny racha
 interno crescente tensao mulher herdeira danubia rangel exaliado sucessor rogerio
 avelino silva rogerio precedem conflito disputa protagonismo pode ajudar explicar
 embate gerou fortes tiroteios imagens guerra domingo ca maior favela rio glenny
 conta historia livro dono morro homem batalha rio companhia letras lancado ano

H.3 Modelo 4 - Explicações do IG de amostras Reais classificadas como Reais

Figura 123 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como reais geradas pelo IG.

temer aciona tropas federais proteger planalto ministerios apos vandalismo anuncia
 jungmann decreto preve uso forcas armadas df semana presidente camara disse
 pediu apoio forca nacional decisao temer elevou tensao camara sessao suspensa
 politicos oposicao ate base aliada criticaram decreto dia marcado manifestacao
 confrontos pm esplanada ministerios ministro defesa raul jungmann informou nesta
 quarta-feira presidente michel temer decretou acao garantia lei ordem tropas federais
 passaramo reforcar seguranca regio esplanada ministerios assista integra
 pronunciamento video acima determinacao publicada edicao extra diario oficial
 uniao preve uso forcas armadas segundo jungmann veio pedido presidente camara
 deputados rodrigo maia noticia sobre autorizacao atuacao exercito causou
 discussoes tumulto durante sessao camara maia porem disse havia pedido temer
 emprego forca nacional nao forcas armadas fotos manifestacao brasil contra
 governo temer decreto preve atuacao tropas maio ordem assinada presidente

(a) Notícia 1 - Real classificada como real

davos temer afirma governo vai batalhar dia noite aprovar reforma previdencia
 presidente discursou forum economico mundial davos suica primeira vez participa
 evento discute economia desenvolvimento presidente michel temer afirmou nesta
 quarta-feira discurso forum economico mundial suica governo vai batalhar dia noite
 votos necessarios aprovar reforma previdencia congresso nacional temer chegou
 terca suica participou primeira vez forum davos reúne todos anos liderancas globais
 politicos representantes ongs executivos empresas investidores discutir temas
 ligados agenda economica desenvolvimento presidente discursou sessao plenaria
 respondeu perguntas feitas fundador forum klaus schwab fala abertura temas
 abordados reforma previdencia segundo temer povo brasileiro percebe atual sistema
 insustentavel proximo passo consertar previdencia social tarefa empenhados cada
 vez povo brasileiro percebe sistema atual injusto insustentavel portanto vamos
 batalhar dia noite voto congresso nacional aprovar proposta ali disse temer apesar
 discurso otimista presidente governo encontra dificuldades viabilizar aprovacao

(b) Notícia 2 - Real classificada como real

Figura 124 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como reais geradas pelo IG.

exdiretores odebrecht explicam tse funcionava relacao politicos diante contradicoes
 depoimentos tribunal marcou acareaao delatores proxima sexta depoimentos
 prestados acao sobre chapa dilmatemer ministro tribunal superior eleitoral tse
 herman benjamin marcou acareaao tres delatores odebrecht marcelo odebrecht
 hilberto mascarenhas claudio melo filho acareaao marcada proxima sexta feira sera
 feita videoconferencia cada delator participara cidades onde residem caso marcelo
 odebrecht onde preso segundo tv globo apurou diversas fontes acareaao pode ter
 sido motivada depoimentos conflitantes dados marcelo odebrecht outros
 exexecutivos empreiteira destas contradicoes sobre jantar palacio jaburu discutido
 apoio financeiro odebrecht campanhas pmdb depoimento ministro tse herman
 benjamin exdiretor relacoes institucionais claudio melo confirmou ja dito acordo
 delacao premiada veio publico dezembro afirmou neste jantar temer solicitou direta
 pessoalmente marcelo odebrecht apoio financeiro campanhas pmdb claudio melo
 disse tambem considerando fato temer vicepresidente epoca presidente partido

(a) Notícia 3 - Real classificada como real

marcelo moreira tempos onde linchamento virtual tornou moda politicamente
 correto avanca cercar direito vontade pessoas debaterem eis artista engajado
 articulado manifesta apelar polemicas vazias partir pura desqualificacao criticos
 jornalistas tico santa cruz vocalista banda carioca detonautas roque clube formado
 ciencias sociais ativo militantes causas sociais colaborador ongs organizacoes nao
 governamentais rio janeiro participacao rock in rio vocalista participou tributo raul
 seixas fez rapido discurso politico criticando governador rio janeiro sergio cabral
 pmdb apoiando varias manifestacoes pipocam cidade rio quais destaca professores
 rede municipal ensino greve contra aprovacao plano cargos salarios entretanto
 opiniao combate rock musico escorregou defender possibilidade pessoas
 participarem protestos mascaras afirmando direito constitucional meio expressao
 farpa contra governo fluminense proibiu exemplo distrito federal pessoas
 mascaradas participem protestos tentativa coibir vandalismo promovido grupelhos
 politicos infiltram manifestacoes pacificas critiquei neste espaco disse fazia
 apologia violencia defender mascarados protestos forma educada incisiva tico santa

(b) Notícia 4 - Real classificada como real

Figura 125 – Modelo 4 - Explicação da notícia 5 real classificada como real gerada pelo IG.

traficante primeiradama exguardacostas guerra poder rocinha visao autor biografia
 jornalista britanico misha glenny autor dono morro homem batalha rio upps acoes
 militares nao sido capazes reverter incapacidade governos lidar guerra faccoes lado
 antigo guardacostas atual desafeto assumiu poder morro prisao mentor outro mulher
 chefe condenada trafico foragida redes sociais ostenta cabelos pintados bem
 louros corpo bronzeado biquinis decotes olhos escuros grife colares ouro pingente
 coroa veja tempo real cobertura sobre rocinha ainda cedo concluir invasao
 traficantes domingo disputa poder rocinha zona sul rio janeiro sido ordenadas
 traficante antonio francisco bonfim lopes vulgo dentro prisao seguranca maxima
 considera jornalista britanico especializado crime organizado misha glenny racha
 interno crescente tensao mulher herdeira danubia rangel exaliado sucessor rogerio
 avelino silva rogerio precedem conflito disputa protagonismo pode ajudar explicar
 embate gerou fortes tiroteios imagens guerra domingo ca maior favela rio glenny
 conta historia livro dono morro homem batalha rio companhia letras lancado ano

H.4 Modelo 4 - Explicações do IG de amostras Reais classificadas como Falsas

Figura 126 – Modelo 4 - Explicações das notícias 1 e 2 reais classificadas como falsas geradas pelo IG.

papa francisco diz teoria big bang nao contradiz papel deus teorias cientificas big
 bang grande explosao acredita ter dado origem universo bilhoes anos atras ideia
 vida desenvolveu atraves processo evolutivo nao entram conflito ensinamento
 catolico disse papa francisco nesta tercafeira discursando encontro pontificia
 academia ciencias organismo independente sediado vaticano financiado grande
 parte santa francisco afirmou explicacoes cientificas mundo nao excluem papel
 deus criacao inicio mundo nao trabalho caos deve origem outra coisa deriva
 diretamente principio supremo cria partir amor declarou big bang hoje considerado
 origem mundo nao contradiz intervencao criativa deus contrario requer argumentou
 evolucao natureza nao contrasta nocao criacao divina porque evolucao requer
 criacao seres evoluem disse pontifice passado igreja opoe primeiras explicacoes
 cientificas universo contradiziam relato criacao biblia condenando astronomo

(a) Notícia 1 - Real classificada como falsa

cheguei escola todo roxo drama maes criancas alergias alimentares graves brasil
 estudos indicam aumento numero casos apoio lei maes brasileiras convivem perigo
 diariamente joao alergia grave proteina leite escola ciente nesse dia falha puseram
 bandeja mousse maracuja frente pegou achando pudesse joao nao aceita comida
 qualquer pessoa perguntava desde pequeno bolo alergico nunca pergunta escola
 casa porque confia comeu colher professora viu mandou cuspir nao adiantou
 ligaram tres vezes primeiro antes comecar reacao perguntei nao inchado nao
 tossindo nao nada poe perua manda casa disse cinco minutos comecou reacao
 ligaram novo pedi medicar disse indo la peguei bolsa ligaram terceira vez conseguia
 ouvir mal ouvia barulho fazia tentando respirar adrenalina tal tal tal pessoas sabem
 aplicar falei nao ninguem aqui agora responderam pessoal treinado mim ido embora
 medicamento instrucao ingles figuras coordenadora comecou perguntar fazia fechei
 olho tentar concentrar expliquei vez voltou fazer perguntas nao responder nao
 tempo le aplica desliguei telefone sai dirigindo cheguei escola nao sabia ia

(b) Notícia 2 - Real classificada como falsa

Figura 127 – Modelo 4 - Explicações das notícias 3 e 4 reais classificadas como falsas geradas pelo IG.

exgovernador ceara cid gomes pdt afirmou discurso ontem sobral festa aniversario
 envolvido esquema corrupcao petrobras ministro stf teori zavascki relata lava jato
 como procuradorgeral republica rodrigo janot ladrao juiz sergio moro atua
 investigacoes primeira instancia picareta nome cid gomes aparece lista apreendida
 lava jato sede odebrecht sob investigacao stf cid aparece lado valor r mil codinome
 falso ouca audio

(a) Notícia 3 - Real classificada como falsa

policia federal deflagrou nesta sabado operacao spectrum desarticular esquema
 transnacional especializado trafico internacional drogas lavagem dinheiro esquema
 comandado luiz carlos rocha vulgo cabeça branca traficantes procurados policia
 federal interpol america sul preso baroes drogas brasil ainda liberdade segundo pf
 condenacoes proferidas justica federal somam anos prisao sequencia mudancas
 rosto meio cirurgias plasticas nota pf informou cerca agentes cumprem mandados
 judiciais sendo prisao preventiva busca apreensao imoveis busca apreensao veiculos
 conducoes coercitivas serem cumpridos cidades londrinapr araraquarasp cotiasp
 embu artessp sao paulosp sorrisomt ordens judiciais expedidas vara federal
 curitibapr nome operacao derivado latim significado lingua portuguesa
 espectrofantasma referencia lider esquema vivia discretamente sombras
 reconhecido meio policial experiencia internacional transcontinental larga rede
 ilegal relacionamentos desviandose investidas policiais ha quase anos veja cerco pf
 barao drogas segundo federal luiz carlos rocha recentemente localizado area

(b) Notícia 4 - Real classificada como falsa

Figura 128 – Modelo 4 - Explicação da notícia 5 real classificada como falsa gerada pelo IG.

sergio moraes pastor josue cotados trabalho exclusao deputado pedro fernandes ma
 ministerio trabalho sobraram dois nomes bancada ptb camara disponibilidade
 assumir funcao deputados sergio moraes rs pastor josue bengtson pa fernandes
 ambos decidiram nao concorrer reeleicao abrir vagas filhos assim nao precisariam
 desincompatibilizar ministerio ate abril virarem candidatos ha duvidas porem
 presidente michel temer ira manter trabalho ptb ministro ronaldo nogueira pediu
 demissao fernandes recusou conversar expresidente jose sarney assumir cargo
 presidente ptb roberto jefferson lideres camara jovair arantes senado armando
 monteiro neto pretendem garantir pasta partido tudo agora depende temer
 presidente ligou manha desta tercafeira roberto jefferson alegando dificuldades
 posse pedro fernandes ja marcada proxima quintafeira duas principais resistencia
 sarney proprio temer causa filho exquasefuturo ministro pedro lucas fernandes
 vereador sao luis secretario governo flavio dino pedob dino nao apenas adversario
 direto sarney maranhao liderou tentativa grosseira evitar impeachment expresidente