UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA

PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

METODOLOGIA PARA DETECÇÃO DE NOTÍCIAS FALSAS USANDO RÓTULO DE VIÉS POLÍTICO

LUCAS A. LISBOA

Dissertação de Mestrado

RECIFE
2024

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA

LUCAS A. LISBOA

METODOLOGIA PARA DETECÇÃO DE NOTÍCIAS FALSAS USANDO RÓTULO DE VIÉS POLÍTICO

Trabalho apresentado ao Programa de PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO do CENTRO DE INFORMÁTICA da UNIVERSIDADE FEDERAL DE PERNAMBUCO como requisito parcial para obtenção do grau de Mestre em CIÊNCIA DA COMPUTAÇÃO.

Orientador: *GEORGE DARMITON DA CUNHA CAVALCANTI*
Co-orientadora: *FRANCIMARIA RAYANNE DOS SANTOS NASCIMENTO*

RECIFE
2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Lisboa, Lucas Albuquerque.
Metodologia para detecção de notícias falsas usando rótulo de viés político / Lucas Albuquerque Lisboa. - Recife, 2024.
111f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2024.
Orientação: George Darmiton da Cunha Cavalcanti.
Coorientação: Francimaria Rayanne dos Santos Nascimento.
Inclui referências e apêndices.

1. Notícias falsas; 2. Avaliação de classificadores; 3. Processamento de linguagem natural; 4. Viés político. I. Cavalcanti, George Darmiton da Cunha. II. Nascimento, Francimaria Rayanne dos Santos. III. Título.

UFPE-Biblioteca Central

Lucas Albuquerque Lisboa

“Metodologia para detecção de notícias falsas usando rótulo de viés político”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 29/10/2024.

BANCA EXAMINADORA

_____

Prof. Dr. George Darmiton da Cunha Cavalcanti
Centro de Informática / UFPE
(**orientador**)

_____

Profa. Dra. Nádia Félix Felipe da Silva
Instituto de Informática / UFG

_____

Prof. Dr. Fábio Manoel França Lobato
Instituto de Engenharia e Geociências / UFOPA

Agradecimentos

Agradeço ao professor George Darmiton pela orientação ao longo desta pesquisa. Ele foi bastante receptivo comigo sem me conhecer previamente e me deu liberdade de propor este tema, além de que foi essencial para o meu amadurecimento acadêmico. Agradeço à Francimaria Rayanne por ter aceitado co-orientar este trabalho e dedicado seu tempo a melhorá-lo. Também gostaria de agradecer aos professores da banca, Fábio Lobato e Nádia Félix, pelas contribuições ao trabalho.

Agradeço à minha noiva Edvania Fernandes, que está ao meu lado há quase dez anos e dividiu comigo todos os momentos ao longo desta pesquisa, incentivando-me a sempre dar o meu melhor. Sua mãe, Fernanda Fernandes, minha sogra, também sempre incentivou que eu alçasse voos na pós-graduação e me aprimorasse cada vez mais. Agradeço aos meus pais, Samuel Lisboa e Maria da Piedade, à minha vó, Maria de Lourdes, e à Zenilda (Patrícia) pelo carinho, suporte e incentivo em toda minha vida.

Agradeço aos meus amigos José Rubens e João Victor pelo apoio mútuo entre nós desde a graduação. Agradeço também à professora Roberta Lopes e ao professor Evandro Costa pelo apreço mesmo após a minha formação.

Agradeço aos colegas Francisco Vital, João Wojtyla, Othon Vinicius e Djayr Bispo pela colaboração durante o percurso.

Por fim, agradeço a você, leitor, pelo interesse em ler este trabalho.

Numa época de mentiras universais, dizer a verdade é um ato revolucionário.

—AUTOR DESCONHECIDO (Falsamente atribuída a George Orwell)

Resumo

A proliferação de notícias falsas se tornou um dos grandes dilemas da atualidade. Com a propagação em massa de material desinformativo em contextos eleitorais, o debate acerca de como o viés político impacta na produção e disseminação de *fake news* tem crescido. Por conta da grande quantidade de postagens e textos veiculados nos meios digitais, soluções de classificação automatizadas têm ganhado destaque. Grande parte das abordagens estabelecidas na literatura realizam o processamento e análise apenas do texto das notícias, ou de outras peças de mídia como imagens ou vídeos, desconsiderando que, em diversos contextos, a desinformação é associada a questões políticas de modo a induzir indivíduos a determinada opinião. Tendo em vista que o viés influencia nos processos de captação, redação e edição da notícia, há, então, uma escolha de palavras direcionada pelo viés por parte dos redatores das notícias falsas. Nesse sentido, este estudo visa avaliar como a incorporação do viés político em modelos de classificação pode contribuir na detecção de notícias falsas. Para isso, foi adotada uma metodologia para incutir o rótulo de viés aos textos correspondentes, a partir da concatenação das bases de notícias com a base de rótulos de viés extraídos de portais. Desse modo, foram comparados três cenários: um cenário em que apenas o texto é avaliado, um cenário em que apenas o rótulo do viés é avaliado e um cenário em que o texto é concatenado com o rótulo do viés. Em cada um dos cenários, foram utilizados sete algoritmos de aprendizagem de máquina e três extratores de características em três bases de *fake news* distintas. Constatou-se uma melhora significativa às abordagens tradicionais, com aumento de até 29,28% na acurácia e de 50,72% no *F1-Score* dos modelos a partir da rotulação, com a proposta tendo apresentado os melhores índices na maioria dos experimentos avaliados, indicando que o viés político pode ser um fator importante no processo de classificação de notícias falsas. Os resultados também apontam para o classificador *Support Vector Machine* (SVM) e para o extrator de características LLAMA 2 como aqueles que obtiveram melhor desempenho, além da proposta se mostrar eficiente tanto para o texto, quanto para o título da notícia.

Palavras-chave: Notícias Falsas. Avaliação de Classificadores. Processamento de Linguagem Natural. Viés Político.

Abstract

The proliferation of fake news has become one of the great dilemmas of the present time. With the mass spread of disinformative material in electoral contexts, the debate surrounding how political bias impacts the production and dissemination of fake news has grown. Due to the large amount of posts and texts circulating in digital media, automated classification solutions have gained prominence. Most of the approaches established in the literature only process and analyze the text of the news or other media pieces, such as images or videos, disregarding that, in many contexts, disinformation is associated with political issues to induce individuals toward a particular opinion. Given that bias influences the processes of gathering, writing and editing news, there is a biased word choice by the authors of fake news. In this sense, this study aims to evaluate how the incorporation of political bias into classification models can contribute to the detection of fake news. For this, a methodology was adopted to incorporate bias labels into the corresponding texts by merging the news datasets with the bias label dataset extracted from portals. Thus, three scenarios were compared: a scenario in which only the text is evaluated, a scenario in which only the bias label is evaluated, and a scenario in which the text is concatenated with the bias label. In each of the scenarios, seven machine learning algorithms and three feature extractors were used across three distinct fake news datasets. A significant improvement was observed compared to traditional approaches, with an increase of up to 29.28% in accuracy and 50.72% in the F1-Score of the models using bias labeling. The proposed method showed the best performance in most of the evaluated experiments, indicating that political bias can be an important factor in the process of classifying fake news. The results also point to the Support Vector Machine (SVM) classifier and the LLAMA 2 feature extractor as the ones that achieved the best performance. Furthermore, the proposed method proved to be efficient for both the text and the title of the news.

Keywords: Fake News. Classifier Evaluation. Natural Language Processing. Political Bias.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Um dos grandes problemas modernos enfrentados pela humanidade é a disseminação de notícias falsas, conforme apontado pelo Fórum Econômico Mundial, sendo incluídas, em 2013, na lista das maiores ameaças à civilização moderna [H+13]. Tendo em vista que as chamadas *fake news*, isto é, conteúdos contendo informações falsas ou manipuladas [PB20], tendem a serem mais acessadas e compartilhadas que publicações científicas online [BCD+15] e terem capacidade de propagação 70% maior que notícias reais [VRA18], é perceptível que o conteúdo desinformativo vinculado por elas pode alcançar muitos indivíduos. Assim, as mídias sociais têm sido usadas como mecanismo para propagação em massa das *fake news*, a fim de manipular a opinião pública [LBB+18].

A partir das eleições estadunidenses de 2016, em que houve o uso bastante expressivo de notícias falsas nas campanhas, surge uma grande preocupação na comunidade internacional dos riscos da desinformação em massa e de seu uso político [GJF+19]. Na Figura 1.1, estão descritos os números de publicações sobre *fake news* ao longo dos últimos anos, retirada do trabalhos de Farhangian, Cruz e Cavalvanti [FCC24]. Considerando que as eleições em questão ocorreram no final do ano de 2016, percebe-se que, logo após o pleito, houve um aumento exponencial do número de publicações sobre o tema. Constatação similar pode ser feita analisando o trabalho de Domenico et al. [DDSIN21], que também realizou um levantamento das publicações na temática, o qual pode ser conferido no Anexo I.

No que tange as postagens em rede sociais, é preciso levar em conta o volume exacerbado de publicações. Estima-se que, por minuto, 66 mil fotos e vídeos são compartilhados no Instagram, 510 mil comentários são publicados no Facebook e 350 mil tweets são postados no Twitter [Mar23]. Com esse fluxo intenso de postagens em todo o mundo, é impossível realizar uma verificação manual da veracidade de cada uma delas. Nessa perspectiva, há um crescente interesse na implementação de técnicas automatizadas para detectar notícias falsas, a exemplo da Declaração Ministerial do Grupo de Trabalho em Economia Digital do G20 [G2024], a qual aponta para adoção de modelos de inteligência artificial para contribuir para a solução deste problema.

As abordagens já estabelecidas dentro da literatura utilizam estratégias de Processamento de Linguagem Natural, processo em que o texto é tratado a fim de extrair características dele, resultando numa representação vetorial. Nesse tipo de abordagem, o processo leva em conta apenas as informações presentes no próprio texto, desconsiderando os aspectos de produção e distribuição do texto. No entanto, conforme constatado na revisão sistemática de Domenico et al. [DDSIN21] sobre *fake news*, há uma série de estudos que apontam um uso sistemático da desinformação para gerar impacto social e político, visando influenciar a opinião dos usuários leitores.

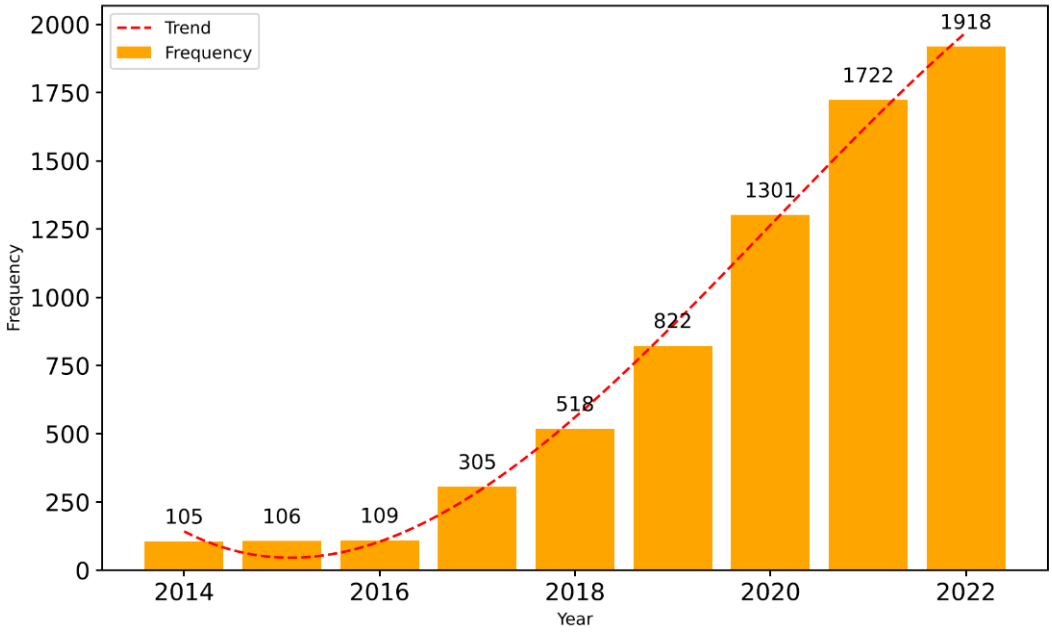


Figura (1.1) Número de Artigos Sobre Fake News ao Longo dos Anos (Farhangian, Cruz e Cavalvanti, 2024).

Nesse sentido, ao considerar apenas o texto, os modelos de aprendizagem de máquina não realizam as correlações entre os vieses dos produtores das notícias e a escolha das palavras usadas. Constatar esse tipo de correlação é importante, pois os produtores de notícias falsas formulam o texto de modo a manipular o leitor por um determinado viés. Em contextos políticos, o viés político das notícias é manifestado pela adesão ou reprovação a algum partido ou ideologia [STCL13]. Mesmo em outros tipos de abordagem para classificação de notícias falsas, como o processamento de imagens e vídeos, o foco reside majoritariamente nas peças de mídia analisadas, sem englobar informações relativa aos vieses dos produtores desse conteúdo. Desse modo, esta dissertação propõe incorporar os dados relativos ao viés político dos portais de notícias em classificadores de *fake news*.

1.1 Motivação

A relação entre notícias falsas e a política tem gerado bastante debate nas diversas áreas do conhecimento, tais como jornalismo, psicologia e sociologia. Tais áreas buscam analisar os efeitos da desinformação nos indivíduos e na sociedade, enquanto os estudos em computação buscam automatizar o processo de detecção de notícias falsas. No entanto, a literatura acerca da aplicação de aprendizagem de máquina ainda carece de abordagens que englobem teorias comunicacionais e/ou psicológicas que analisam a influência do viés político partidário na classificação de *fake news* [ZZ20]. Nesse sentido, a junção dos achados sobre o tema nas diversas áreas do conhecimento pode proporcionar o desenvolvimento de novas soluções para a propagação de notícias falsas.

Nessa perspectiva, há grande debate nas publicações de jornalismo e psicologia de como o viés político influencia na produção de notícias falsas e na veracidade de que os leitores inferem do material lido [Gaw21] [BR20]. Antecipando a discussão apresentada na Seção 2.2, o viés pode se manifestar na forma da redação do texto da notícia, com utilização de determinados termos ou expressões, que indicam preferência ou repulsa ao tema retratado. Há, então, uma relação entre a escolha das palavras utilizadas na redação da notícia com o viés do redator.

Tratando-se de notícias falsas, conforme será discutido na Seção 2.3, há o aspecto de manipulação, que busca influenciar o leitor a partir de determinado viés, recorrendo a estratégias de produção e reprodução de notícias que visam difamar adversários políticos. Além disso, a própria estrutura das redes sociais, por vezes, favorecem publicações associadas a determinados viés, o que foi utilizado nas últimas eleições ao redor do mundo para fins eleitorais. Isso é bastante perceptível no uso de *Clickbaits*, nos quais a formatação do texto é pensada para impulsionar cliques e replicações. Sendo assim, levando em consideração a discussão de como o viés político pode influenciar o comportamento dos indivíduos, bem como poder afetar o julgamento deles quanto à veracidade das notícias, ele, o viés político, pode ser considerado um dado importante a ser incluído nos modelos de classificação de notícias falsas.

Apesar desses apontamentos, as soluções automatizadas para classificação de *fake news* ainda não se debruçaram totalmente sobre a temática. As abordagens tradicionais de classificadores de notícias falsas levam em conta apenas a mídia ou o texto fornecido, desconsiderando como o viés do portal pode ter influenciado na redação final. Dessa forma, a motivação deste trabalho surge em buscar entender a relação entre o viés político e as notícias falsas. Levando em conta que toda notícia possui, em algum nível, determinado tipo de viés, levanta-se, aqui, a hipótese de que a rotulação do viés político do portal originário da notícia pode contribuir para a construção de modelos de classificação mais eficientes. Ao ser adicionado, então, a informação do viés dominante daquele veículo ao classificador, busca-se possibilitar uma estimativa que leve em conta a escolha das palavras utilizadas para cada espectro ideológico.

Assim, este estudo visa apresentar esta nova abordagem, a qual é adicionado o dado do viés político junto ao texto para o classificador. Ao final, tem-se uma proposta replicável e que apresenta melhores resultados frente às abordagens tradicionais. Nesse sentido, outros segmentos que lidam com o tema, como na atividade jornalistica, conseguem adotar a nova abordagem.

1.2 Objetivos

O objetivo principal deste trabalho é avaliar o impacto da inclusão de rótulo de viés político na classificação de *fake news*. Como objetivos específicos, tem-se:

- Descrever e propor um processo de rotulação de bases com viés político;
- Analisar os resultados e métricas dos modelos com e sem rotulação, tanto a nível de texto completo, quanto a nível de título;
- Avaliar a metodologia propostas em diferentes bases e algoritmos a fim de verificar a robustez do procedimento aplicado;

- Verificar o potencial de novas tecnologias, tais como o LLAMA 2 [FCC24], neste tipo de abordagem.

1.3 Contribuições

Dentre as contribuições deste estudo, estão:

- Melhora na acurácia dos modelos avaliados de até 30%;
- Melhora no *F1-Score* nos modelos avaliados de até 51%;
- Proposta metodológica reproduzível;
- Incorporação de conceitos interdisciplinares em soluções de aprendizagem de máquina.

1.4 Estrutura da Dissertação

Esta dissertação está estruturada em cinco capítulos, incluindo esta Introdução. O Capítulo 2 discute a fundamentação teórica, aprofundando a explanação dos conceitos que embasaram esta pesquisa, bem como dos trabalhos correlatos, com enfoque nos estudos que analisam relação entre modelos de detecção de *fake news* e viés político. O Capítulo 3 aborda a metodologia adotada, detalhando aspectos relativos ao protocolo experimental. No Capítulo 4, há a apresentação dos resultados encontrados e uma análise comparativa dos cenários propostos. Por fim, o Capítulo 5 conclui este trabalho e propõe diretivas futuras para a pesquisa.

CAPÍTULO 2

Fundamentação Teórica

Este capítulo apresenta os conceitos que fornecem as bases teóricas da pesquisa realizada. A Seção 2.1 apresenta as definições da literatura de *fake news*. A Seção 2.2 discorre sobre as manifestações do viés em notícias e seus efeitos no público leitor. A Seção 2.3 demonstra como as notícias falsas podem ser utilizadas com objetivos políticos. A Seção 2.4 descreve conceitos bases do processamento de linguagem natural e da aprendizagem de máquina envolvendo classificadores. Por fim, a Seção 2.5 apresenta trabalhos correlatos envolvendo classificadores de notícias falsas e viés político.

2.1 Definição de Fake News

Visando lidar com a problemática das notícias falsas, a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) lançou, em 2018, um manual para auxiliar na formação de jornalistas no que tange o combate à desinformação [IP18]. Neste material, Ireton e Posetti definem a desinformação como:

> [...] tentativas deliberadas (frequentemente orquestradas) para confundir ou manipular pessoas por meio de transmissão de informações desonestas. Isso geralmente é combinado com estratégias de comunicação paralelas e cruzadas e um conjunto de outras táticas, como hackear ou comprometer pessoas. [...] Os provedores da desinformação atacam a vulnerabilidade ou o potencial partidário dos destinatários esperando que eles se alistem como amplificadores e multiplicadores. Desta forma, eles procuram encorajar-nos para nos tornarmos condutores de suas mensagens, explorando nossas propensões para compartilhar informações por múltiplas razões.

Percebe-se como elemento central nesta caracterização a intencionalidade de enganar o público. Tal elemento, isto é, a intencionalidade de gerar engano, também é apontado por Bernecker, Flowerree e Grundmann [BFG21]. Segundo os autores, há três visões possíveis na literatura para definir o fenômeno das *fake news*:

- **Híbrida:** Entendimento predominante na literatura. Nele, os autores citam dois conceitos importantes para entender o funcionamento desse tipo de conteúdo: a **verdade** e a **veracidade**. No primeiro, para que a notícia seja considerada verdadeira, não pode haver a veiculação de dados falsos; já o segundo versa sobre a intenção de vincular dados fabricados; Ou seja, a verdade analisa se, objetivamente, as informações contidas no material da notícia condizem com a realidade, enquanto a veracidade engloba a intenção de

propagar determinada desinformação. Assim, tem-se o conjunto de conteúdo falso com intenção fraudulenta. Esta visão coincide com a apresentada por Ireton e Posetti [IP18];

- **Privativa:** Neste entendimento, as *fake news* não apresentam o formalismo de notícias da mídia tradicional. Nesse sentido, as *fake news* não seguem a metodologia e as práticas jornalisticas de boa apuração e veiculação da notícia, como, por exemplo, a qualidade das fontes envolvidas;

- **Centrada no Consumidor:** Aqui, entende-se que uma *fake news* é produzida visando seu consumo, bem como os efeitos desse consumo, por determinado público.

As três visões não se anulam, podendo ser complementares entre si, pois é possível analisar uma mesma notícia por cada uma dessas perspectivas, podendo ela apresentar dados falsos, ter intenção de enganar, não apresentar boa metodologia jornalística e também ter sido produzida visando determinado público. No entanto, por se atentar tanto ao conteúdo, quanto à intencionalidade na produção da notícia, a visão híbrida demonstra ser mais completa que as demais, pois as demais visões se limitam à metodologia para confecção daquele material, o que explica sua adesão ser maior dentro da literatura.

Outra importante definição é a encontrada no trabalho de Tandoc, Lim e Ling [TJLL18], no qual eles analisaram 34 artigos publicados entre 2003 e 2017, condensando o significado atribuído ao termo *"fake news"* e suas características. São apresentados seis tipos de notícias falsas:

- **Sátira:** Utiliza do sarcasmo para gerar humor, a partir do exagero de fatos correntes. Usada com bastante frequência para críticas sociais e políticas;

- **Paródia:** Próxima da sátira, pois também tem como objetivo a geração do humor. Sua principal diferença é o uso de eventos ficcionais para tecer comentários sobre a realidade;

- **Fabricadas:** Notícias inteiramente fabricadas, emulando estilo de verdadeiras, a fim de enganar o público;

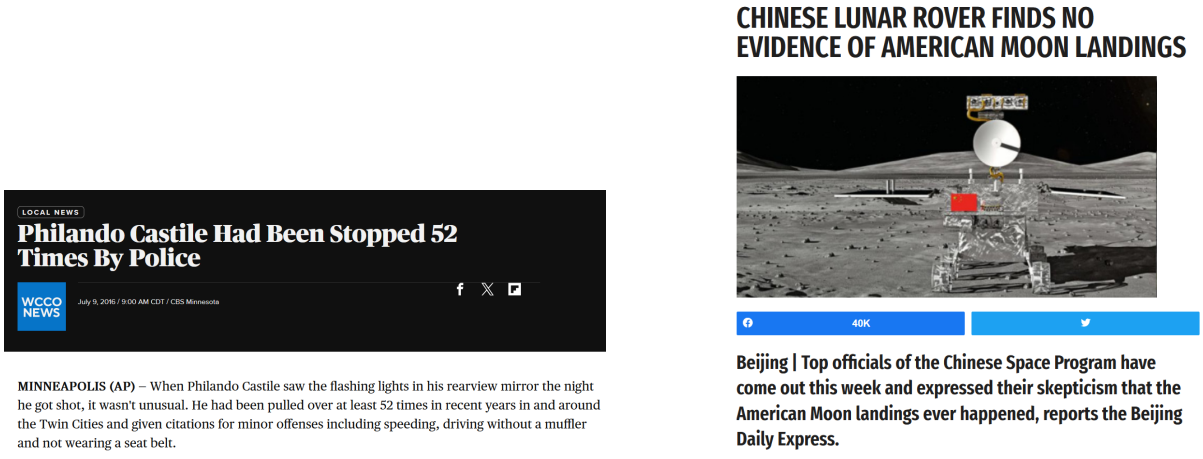
- **Manipulação de Fotos:** Utilização de técnicas de manipulação de imagem ou vídeo para criar uma falsa narrativa;

- **Publicidade:** Materiais publicitários que se passam por notícias;

- **Propaganda:** Apesar do nome, está mais voltada ao uso político, no qual notícias são usadas para favorecimento de uma entidade ou figura partidária.

Os autores destacam que os dois primeiros tipos não possuem intenção de enganar. Apesar de se referir a fatos ficcionais ou exagerados, há o uso de uma linguagem irônica o suficiente para que o público constate que se trata de uma peça humorística, enquanto nos demais grupos há a intenção em emular uma linguagem objetiva, a fim de manipular a opinião pública.

Na Figura 2.1, há dois exemplos extraídos da base FakeNewsNet - Politifact [SMW+20], sendo um de notícia verdadeira, proveniente do portal CBS News, e um de notícia falsa, proveniente do portal World News Daily Reports. Neste exemplo de *fake news*, é perceptível a emulação do formato de portais com credibilidade a fim de enganar o leitor.



(a) Exemplo de Notícia Verdadeira (b) Exemplo de Notícia Falsa

Figura (2.1) Exemplos de Notícia Verdadeira e Notícia Falsa extraídas da base FakeNewsNet - Politifact.

2.2 Notícias e Viés

Há diversos apontamentos na literatura que indicam a inevitabilidade de toda notícia possuir algum tipo de viés. Park et al. [PKCS12] afirmam que *"o viés na mídia é uma falha inerente ao processo de produção de notícias."* [1]. Isso ocorre pois, seja na captação da informações, na redação da notícia ou na edição do texto, há a avaliação subjetiva dos sujeitos envolvidos na atividade realizada ou a influência de fatores externos à produção da notícia [Arr15].

Nessa perspectiva, Saez-Trumper, Castillo e Lalmas [STCL13] consideram que existem três tipos de vieses em notícias, sendo elas:

- **Viés de Seleção:** A preferência em selecionar determinadas histórias e assuntos;
- **Viés de Cobertura:** O tamanho da atenção a determinadas histórias e assuntos;
- **Viés de Afirmação:** A emissão de opiniões favoráveis em determinadas histórias e assuntos.

Em notícias políticas, tais vieses se manifestam no alinhamento ou rejeição a determinado partido ou espectro político. Tal definição, foi também adotada em outros trabalho envolvendo classificadores de notícias, a exemplo de Arruda [Arr15]. O trabalho de Goldman, Gupta e Israelsen [GGI24] ilustra como esse processo se dá na prática. Os autores avaliaram 25 anos de publicações de notícias financeiras de dois importantes jornais dos Estados Unidos, o *New York Times*, alinhado ao Partido Democrata, e o *Wall Street Jornal*, alinhado ao Partido Republicano. Em notícias envolvendo empresas alinhadas ao espectro político do portal em questão, houve um aumento de cerca de 15% de palavras positivas em relação ao concorrente. Além disso, os autores também apontam consequências práticas da divergência de cobertura no mídia. Em

[1] Em tradução livre.

dias em que eram publicadas notícias com potencial polarizador, havia um aumento de 30% no volume de negociações nas empresas consideradas politicamente extremas.

Nesse sentido, a influência do viés político na forma em que o leitor avalia se o conteúdo é verdeiro ou não ainda é debatida. Há duas correntes de pensamento dentro os estudos de psicologia: *Motivated System 2 Reasoning (MS2R)* e a Teoria Clássica do Raciocínio [BRP20]. A primeira defende a tese de que as pessoas deliberam a partir de suas crenças pré-estabelecidas, geralmente políticas, tendendo a aceitar ideias que favoreçam seus posicionamentos, influenciando, assim, na avaliação da veracidade das notícias consumidas. Já a segunda corrente de pensamento afirma que a ausência de pensamento analítico por si só influencia no julgamento da veracidade da notícia, ou seja, uma análise mais criteriosa da notícia levaria o público a identificar se ela é verdadeiro ou não, independente do viés político.

Há estudos experimentais que reforçam cada uma das correntes apresentadas. O trabalho de Calvillo et al. [CGBM21] realizou um teste com uma amostra de 346 participantes, os quais se identificavam pró-Democratas, pró-Republicanos e sem identificação partidária. A partir do uso de regressão estatística, foi identificado que participantes mais liberais perceberam manchetes pró-liberais como mais precisas do que manchetes pró-conservadoras, enquanto participantes mais conservadores fizeram o contrário, ou seja, os participantes atribuíram maior veracidade às manchetes que corroboravam com seus vieses. Nessa mesma linha, o trabalho de Vegetti e Mancosu [VM20] realizou uma pesquisa online com 3.005 participantes. Os autores constataram que há sim uma forte tendência dos indivíduos analisados a acreditarem em notícias enviesadas ao seu viés político.

Já na segunda linha de pensamento, o estudo apresentado por Ross, Rand e Pennycook [RRP19] realizou um experimento com 1.977 participantes, categorizando-os pela proximidade entre Republicanos e Democratas. Os apontamentos indicam que o pensamento analítico é mais influente na constatação da veracidade das notícias do que o viés político, de forma que, se o indivíduo buscar avaliar criticamente a notícia, seu viés político será minimizado em sua classificação. Outro trabalho que traz achados similares é o de Bago, Rando e Pennycook [BRP20], o qual avaliou 1.635 pessoas, em dois cenários: no primeiro pediam para os candidatos classificarem uma notícia como verdadeira ou falsa de maneira instintiva, sem muito tempo para pensarem, enquanto no segundo cenário pediam para o participante analisar cautelosamente. Os achados apontam uma melhora significativa no segundo cenário, mesmo destacando que as respostas instintivas tinham influência do viés político dos participantes.

O estudo de Bradley, Pantzalis e Yuan [BPY16] avaliou como o viés político poderia influenciar em operações envolvendo fundos estatais. Os autores constataram que, por vezes, investidores optavam por empresas politicamente alinhadas a eles. Os autores apontam o envolvimento entre empresas e políticos congressistas locais num processo de *lobby*, gerando, inclusive, liquidez negativa para os fundos envolvidos.

No que tange ao comportamento nas redes sociais, destaca-se o trabalho de Osmundsen et al. [OBV$^+$21], o qual avaliou o comportamento de 2.337 contas estadunidense do Twitter. A partir da análise da interação dessas contas com o conjunto de cerca de 500 mil manchetes de notícias, foi constatado que há forte motivação política para compartilhamento de notícias falsas, de modo que os indivíduos que relatam odiar seus oponentes políticos são os mais propensos a compartilhar notícias políticas falsas e compartilhar seletivamente conteúdo que é útil

para denegrir esses oponentes. Dessa forma, para efeito desta pesquisa aqui realizada, é considerado que há indícios da influência do viés político na análise dos usuários das notícias lidas, em especial nas manchetes, não desconsiderando a importância do pensamento analítico.

Diversos estudos avaliam como o comportamento social pode ser afetado pelas redes sociais [GBL23]. Bond et al. [BFJ+12] realizaram um experimento envolvendo cerca de 61 milhões de usuários do Facebook em que três grupos foram separados: um grupo receberia uma mensagem incentivando a irem votar; um segundo grupo receberia a mesma mensagem com a adição da informação de quantos amigos do usuários já tinham votado, junto de seus nomes e fotos; e um grupo controle que não recebeu nenhuma mensagem de incentivo. Os resultados mostram que o grupo que recebeu a mensagem de apoio com a mensagem do indicativo de seus amigos teve um aumento na autoexpressão política, busca de informações e a efetivamente irem votar, em especial quando haviam amigos com forte interação dentre os indicados na mensagem. Anspach [Ans21] realizou um experimento, no qual três grupos foram avaliados, os quais foram expostos a *tweets* do então presidente Donald Trump. O primeiro grupo, tido como controle, foi exposto a um *tweet* sobre economia; o segundo grupo foi exposto a um *tweet* implicitamente racista; e o terceiro grupo foi exposto a um *tweet* explicitamente racista. O autor constatou que os participantes do segundo e do terceiro grupo que demonstravam sinais de ressentimento racial passaram a adotar cada vez mais termos explícitos de preconceito.

A análise de como o viés político pode influenciar no ranqueamento de resultados de buscas online também é um importante objeto de estudo. Nessa perspectiva, Epstein e Robertson [ER15] analisaram os resultados de pesquisa relativos aos cenários eleitorais dos Estados Unidos e da Índia. A partir de cinco experimentos randomizados e duplo-cegos com 4.556 participantes, constataram que resultados de buscas enviesados podem influenciar 20% do eleitorado indeciso. Partindo desse achado, Kulshrestha et al. [KEM+19] desenvolveram um *framework* para avaliar o viés em busca envolvendo tópicos políticos no Google e Twitter. No caso do Google, os autores apontam que os resultados tendem a trazer um viés relativo ao objeto da busca, isto é, quando é feita uma busca sobre um candidato, os resultados retornados privilegiam links de fontes do próprio candidato. Já no Twitter, foram analisados dois tipos de filtros, *"top"* e *"news"*. Os resultados retornados no filtro *"top"* possuem viés republicano, enquanto os do filtro *"news"* possuem viés democrata. Os autores destacam que a falta de transparência nos viés dos itens retornados pode ser prejudicial para os usuários. Huszár et al. [HKO+22] realizaram um experimento em que o algoritmo de seleção do Feed de Notícias do Twitter era desativado, mantendo apenas a ordenação cronológica das notícias. O experimento foi realizado em sete países, Estados Unidos, Japão, Reino Unido, França, Espanha, Canadá e Alemanha, e em todos eles foi constatado que o algoritmo de curadoria amplifica a visibilidade de postagens de agentes políticos de direita.

Um novo tipo de tecnologia que ainda demanda mais estudo de seu impacto social são os *Large Language Models*, sendo o mais conhecido o ChatGPT. Buscando avaliar o viés político nele, Motoki, Neto e Rodrigues [MPNR24] realizaram um experimento em que foram escolhidos três cenários: solicitar que o *bot* respondesse como um republicano; solicitar que o *bot* respondesse como um democrata; e solicitar uma resposta neutra. Para cada cenário, foram realizadas 62 perguntas, replicadas 100 vezes cada, sendo aplicado o teste do *Political Compass* [Com24] para avaliar o espectro político de cada resposta. Os resultados indicam que as respos-

tas neutras se aproximam ao espectro democrata. Quando replicado o experimento em outros países, o ChatGPT demonstrou viés mais alinhado ao presidente Lula do que ao ex-presidente Bolsonaro, no caso brasileiro, e viés mais alinhado ao Partido Trabalhista do que ao Partido Conservador, no caso inglês.

Ainda sobre o comportamento nos meios digitais, constata-se uma relação que pauta a formatação de notícia que é o fenômeno do *Clickbait*, o qual consiste na adoção de estratégias para capturar o clique do usuário, como a adoção de manchetes sensacionalistas. Nesse sentido, Delmazo e Valente [DV18] apontam uma relação entre esse processo e as notícias falsas, de modo de que há uma estratégia na construção do texto a fim de, além de enganar, capturar a atenção do leitor, em especial em temas políticos. Uma estratégia adotada em manchetes é justamente o uso de termos e expressões associadas a grupos políticos, com objetivo de atingir determinada identidade social. Hopkins, Lelkes e Wolken [HLW22] constataram que esse tipo de estratégia pode aumentar o número de cliques em cerca de 0,79 em mil impressões, valor semelhante a palavras apelativas, como *"sex"*.

A fim de ilustrar esse processo na prática, foi selecionado aqui uma notícia do *American Action News*, portal classificado como de direita e de baixa credibilidade pelo *Media Bias / Fact Check* [VZ22], publicada 07/04/2024. A manchete dela é: *"There Is No Place In Our Justice System For Communist-Style Wealth Confiscation"* [2]. Nela, é possível verificar o uso de termo de identidade para o grupo adversário, *"Communist"*, associado a ações negativas, *"Confiscation"*, além de expressões enfáticas, *"There Is No Place"*.

Assim, é nítido que o viés influencia os processos de produção, de distribuição e de leitura de uma notícia. Em notícias falsas, a construção do texto da notícia é feito visando influenciar o usuário por determinado viés.

2.3 Fake News e a Política

A influência das redes sociais no debate político tem sido amplamente discutida nos últimos anos, em especial no aumento da polarização e de visões extremistas. Nesse sentido, as últimas eleições ao redor do mundo trazem preocupação sobre riscos ao regime democrático, com uso de Inteligência Artificial para criar propaganda personalizada em larga escala ou a propagação massiva de conteúdo enganoso [FMA+23]. Santini et al. [SAB+18] realizaram uma revisão da literatura para avaliar o uso de perfis em redes sociais para manipular o debate público em questões políticas. Dentre os apontamentos realizados, está o uso de *bots* para gerar engajamento em candidatos a eleições, difamar opositores e fazer propaganda de suas pautas, sendo comum a veiculação de desinformação no processo.

Nesse contexto, há um conceito bastante importante para compreender essa relação entre as *fake news* e a política: as chamadas "Operações de Influência". Santini [San21] traz a seguinte definição:

> [...] se refere à coleta de informações táticas sobre um adversário, bem como a disseminação de propaganda e informações falsas em busca de vantagem competitiva sobre um oponente. Considerada como fator fundamental para ganhar uma

[2] "Não há lugar em nosso sistema de justiça para o confisco de riquezas ao estilo comunista", em tradução livre

> guerra, as operações de influência usam métodos multidimensionais de persuasão. O objetivo é afetar as vulnerabilidades de comunidades e indivíduos em sua estrutura social para gerar mudanças de comportamento, percepção e atitude-desejadas do(s) público(s)-alvo.

Esse processo pode ser percebido nas constatações de Nithyanand, Schaffner e Gill [NSG17]. Os autores analisaram publicações no Reddit de 2005 até 2017, percebendo que o número de compartilhamento de teorias da conspiração, notícias sensacionalistas e desinformação em fóruns republicanos foi 16 vezes maior durante a eleição de 2016 do que antes dela. Além disso, foi apontado um aumento da hostilidade entre adversários políticos nos últimos anos.

Outra importante revisão de literatura foi feita por Lorenz-Spreen et al. [LSOLH23], o qual avaliou cerca de 496 artigos a fim de verificar como os meios digitais podem influenciar a democracia. Os autores constatam que há uma forte correlação dos meios digitais com a participação popular, sendo esse fator benéfico para as democracias. No entanto, há também forte correlação para o aumento do discurso de ódio, da polarização, populismo e desinformação, sendo estes fatores maléficos à democracia.

Essas constatações são reforçadas por outro trabalho de Santini et al. [SST21], o qual analisou a dinâmica de menções no Twitter a figuras políticas no contexto da eleição da prefeitura do Rio de Janeiro em 2016. Neste trabalho, identificaram que o maior número de menções não era direcionado a um candidato da disputa, mas sim ao então deputado federal Jair Bolsonaro, com uso massivos de *bots* e veiculação de material desinformativo. Os autores apontam uma tendência de influenciar a opinião pública para a eleição presidencial de 2018, a qual Bolsonaro seria efetivamente candidato.

Lisboa et al. [LFBL20] constataram que notícias falsas proferidas por líderes estatais durante a pandemia da COVID-19 geravam picos de buscas no Google. Os autores apontam que falas polarizadoras geram tal efeito, o qual é potencializado pela figura de autoridade. Para sustentar seus apontamentos, os autores evocam a teoria de Koopmans [Koo04]. Nela, são descritos três importantes conceitos: **legitimidade**, **ressonância** e **visibilidade**. Legitimidade é o quanto determinada mensagem é considerada legítima perante a sociedade, isto é, total legitimidade significa apoio total, já zero legitimidade significa total rejeição. Assim, é importante observar as mensagens que ficam no meio do espectro da legitimidade, pois são as mensagens consideradas polarizadoras. Desse modo, ressonância significa a capacidade de uma mensagem gerar debates em torno de si mesma. Assim, mensagens polarizadoras têm grande ressonância e, consequentemente, visibilidade, a qual é considerada extensão da cobertura da mensagem pelos meios de comunicação em massa.

Vosoughi, Roy e Aral [VRA18], ao analisar notícias vinculadas no Twitter entre 2006 e 2017, perceberam que notícias envolvendo assuntos políticos possuem um amplo efeito cascata, isto é, um intenso processo de compartilhamento e difusão dessas postagens. Este fato ressalta certa característica apontada nas definições listadas de *fake news*: há um componente político muito forte nelas. Nesse sentido, Flynn, Nyhan e Reifler [FNR17] analisaram os mecanismos psicológicos que levam as pessoas a acreditarem em informações políticas desinformativas. Neste trabalho, foi constatado a capacidade das percepções errôneas (*misperceptions*) afetar a opinião pública, em especial pela perseverança de crença (*Belief Perseverance*), isto é, a resistência em mudar sua opinião mesmo após a constatação de uma nova informação que

contradiga o que já se acredita previamente. Outro ponto destacado é que há um senso de identidade criado a partir das crenças políticas, havendo uma certa pressão social para reafirmação de posicionamentos. Por fim, foi citado que há uma tendência entre pesquisadores em defender a tese de que determinadas ideologias afetam a percepção de mundo, sendo o exemplo levantado de que os conservadores pontuam menos em avaliações sobre abertura a experiências, além de ser observado com maior frequência neste grupo a ocorrência de instâncias do *Backfire Effects*, efeito em que uma pessoa reforça determinada crença ao entrar em contato com uma evidência que a contraria.

Chen et al. [CPYM21] realizaram um experimentos com 15 *bots* no Twitter, a fim de avaliar os vieses da rede social em questão. Os autores apontaram que as conexões iniciais influenciavam fortemente o conteúdo e sugestões que viriam posteriormente para o perfil, de forma que, se a primeira conexão formada fosse de esquerda, por exemplo, os conteúdos potencializados pela plataforma seriam de esquerda. Os autores também constataram que perfis que começavam seguindo páginas de direita recebiam mais material desinformativo, com quase 15% dos links recebidos, enquanto perfis que começavam seguindo páginas de esquerda tinham cerca de 1% dos links recebidos classificados como desinformativo.

Há diversos estudos que apontam o uso das notícias falsas com fins eleitorais e partidários, em especial envolvendo a extrema-direita. O estudo de Allcott e Matthew [AG17] constatou que a maioria dos estadunidenses teve contato dentre um a três materiais desinformativos no mês anterior às eleições presidenciais de 2016, predominantemente a favor do então candidato Donald Trump. Van Der Linden, Panagopoulos e Roozenbeek [VdLPR20] realizaram uma pesquisa com mil indivíduos dos Estados Unidos acerca da percepção deles sobre jornalismo e desinformação, obtendo como resultado que conservadores estão mais propensos a desacreditar na grande mídia e acreditar em teorias conspiratórias, de forma que *"ter votado em Trump aumenta as chances de associar mentalmente a mídia mainstream às notícias falsas em 187%"*[3]. Resultado similar foi encontrado por Baptista et al. [BCGPN21] que analisaram de maneira experimental o contexto de Portugal. O estudo analisou 712 participantes a partir de um questionário, sendo constatado que pessoas de direita possuem maior tendência a acreditarem em *fake news*, pois enquanto a média para participantes de esquerda acreditarem em notícias falsas de direita foi 1,50 pontos, os de centro foi de 1,79 e os de direita 2,01, representando um aumento de até 34%. Até mesmo para notícias falsas de esquerda, participantes de direita pontuaram mais, obtendo 2,22, contra 2,09 dos de centro e 1,99 dos de esquerda, sendo um aumento de cerca de 11,5%. A *BBC News* realizou um estudo, liderado por Chakrabarti, Stengel e Solanki [CSS18], acerca da disseminação de *fake news* na Índia. Nele, foram selecionados cerca de 16 mil perfis do Twitter e 3.200 páginas no Facebook, sendo constatado um ecossistema organizado de produção e replicação de notícias falsas alinhados à extrema-direita. Inclusive, cerca de 34,28% das contas seguidas pelo perfil oficial do Primeiro Ministro, Narendra Modi, integram esse *cluster* pró-BJP, partido conservador do país. Os autores destacam que o nível alto de atividade dessas contas em relação ao tempo de existência dos perfis é um forte indício de que não são de pessoas reais. Aqui, no Brasil, Ramos, Machado e Cerqueira-Santos [RMCS22], a partir de uma pesquisa com 1.328 pessoas, constataram que indivíduos com maior autoritarismo e conservadorismo eram mais propensos a acreditar em informações advindas do *WhatsApp* do

[3]Em tradução livre

que em veículos da mídia tradicional, com correlação positiva e moderada de Spearman para tais índices (0,229 para autoritarismo e 0,219 para conservadorismo).

Nessa perspectiva, destaca-se também o trabalho de Ribeiro et al. [RSB+19], o qual analisou os anúncios vinculados durante a campanha eleitoral presidencial dos Estados Unidos de 2016. Nele, os autores constataram que os anúncios tinham conteúdo personalizado para cada tipo de usuário a qual a propaganda seria direcionada, aproveitando-se de critérios raciais e viés político.

2.4 Processamento de Linguagem Natural e Classificadores

As abordagens tradicionais na literatura para detecção automatizada de notícias falsas utilizam métodos de processamento de linguagem natural conjuntamente com algoritmos de aprendizagem de máquina em seus classificadores. Nessa perspectiva, a Figura 2.2, extraída do trabalho de Coutinho [COU23], ilustra as etapas desse tipo de abordagem. Nela, os dados textuais passam pela etapa de processamento de linguagem natural resultando em uma representação numérica. Esta nova representação é fornecida como dados de entrada juntamente das respectivas classes para o classificador. Com isso, há o treinamento do modelo, o qual será usado para classificar novas instâncias.

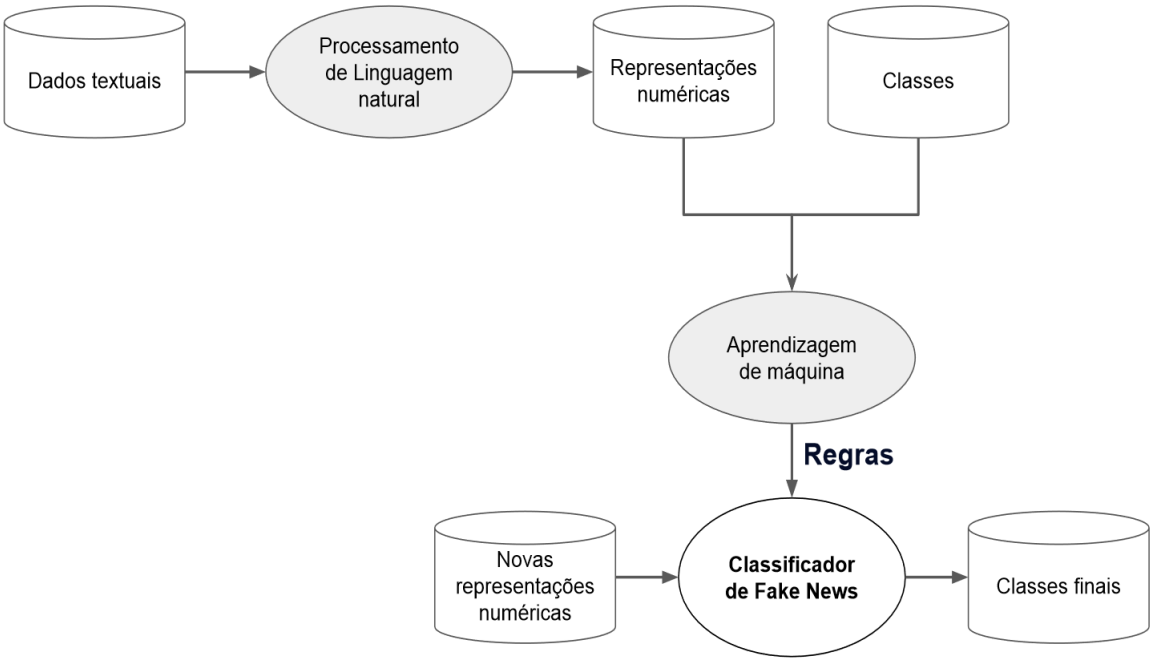


Figura (2.2) Fluxograma de Classificação de Notícias Falsas (Coutinho, 2023).

No que tange a etapa de processamento de linguagem natural, antes de converter o texto original em dados numéricos, é necessário realizar o pré-processamento dele, também chamado de limpeza. Este procedimento consiste em normalizar o texto, buscando eliminar símbolos gráficos e palavras que não trazem valor semântico para a análise dos dados. Os detalhes do pré-processamento de texto adotado neste estudo estão descritos na Seção 3.4.

Após o pré-processamento, há a extração de características do texto, no qual há a conversão de dados textuais para numéricos. Há diferentes métodos para esta finalidade, com representações baseadas em **contagem** e representações baseadas em **predição** [FCC24]. O primeiro tipo de representação é baseado no número de vezes que as palavras ocorrem no texto, enquanto o segundo tipo é baseado em prever a próxima palavra baseado nas anteriores, englobando os aspectos semânticas e sintáticas do texto. Os detalhes dos métodos de extração de características adotados neste estudo estão descritos na Seção 3.5.

No que tange a aprendizagem de máquina, é adotada a abordagem de aprendizagem supervisionada, de forma que um conjunto de dados de entradas são fornecidos juntamente com a classe correspondente a cada instância a fim de obter regras para classificar novas entradas. Os detalhes dos algoritmos de aprendizagem de máquina adotados neste estudo estão descritos na Seção 3.6.

2.5 Trabalhos Relacionados

Ainda há poucos trabalhos na literatura que abordam questões relacionadas ao viés político na classificação de notícias falsas. Nesse sentido, ALDayel e Magdy [AM21] realizaram um levantamento acerca de *stance detection* [4] em redes sociais. Nele, há apenas uma breve subseção que discorre sobre como a análise dos posicionamentos pode contribuir para problemas de verificação da veracidade de textos. É possível caracterizar os trabalhos listados pelos autores em três tipos de abordagens.

O primeiro tipo de trabalho citado é o que analisa o perfil nas redes sociais dos usuários que compartilham notícias falsas ou teorias da conspiração, com destaque para o de Allcott e Gentzkow [AG17], que avaliou a dinâmica das notícias falsas nas redes sociais durante as eleições estadunidenses de 2016, constatando que a maioria das *fake news* disseminadas no período contribuíam para a campanha do então candidato Donald Trump, bem como os usuários tinham tendência a compartilhar notícias que favoreciam seu candidato preferido. Já o segundo tipo de publicação consistiu na proposição de *datasets* para desenvolvimento de métodos de verificação de notícias falsas, tendo como importante recurso a detecção de posicionamento. Dentre eles, estão os estudos de: Derczynski et al. [DBL$^+$14] que desenvolveu o projeto PHEME; Derczynski et al. [DBL$^+$17] com a proposta da SemEval-2017 Task 8 (RumourEval); Gorrell et al. [GKL$^+$19] com a SemEval-2019 Task 7, um avanço na proposta RumourEval de 2017; Ferreira e Vlachos [FV16] com a base Emergent. Em tais abordagens, os autores começaram a observar como os posicionamentos dos usuários poderiam ser incorporados em modelos de classificação. Nelas, o foco está nos rumores gerados nos comentários e nos posicionamentos dos usuários, isto é, se eles estão favoráveis ou não a determinada matéria.

O terceiro tipo de trabalho foca em modelos de classificação de *fake news* a partir de detecção de posicionamento, com foco na iniciativa *Fake News Challenge* (FNC-1). Nesse tipo de abordagem, os dados referentes à opinião dos usuários sobre determinado tema eram utilizados para alimentar os modelos. A proposta de Baird, Sibley, e Pan [BSP17] fez a combinação de Árvore de Decisão com Redes Neurais Convolucionais, alimentando o modelo com dados

[4] "Detecção de posicionamento", em tradução livre.

referentes a se os usuários concordavam ou não com os títulos das notícias. Mohtarami et al. [MBG+18] obtiveram resultados parecidos usando *Long Short Term Memory* (LSTM) e Redes Neurais Convolucionais. Também nessa linha, Ghanem, Rosso e Rangel [GRR18] realizaram a combinação de representação de conhecimento léxico, *word embedding* e *n-gram*; Já Borges, Martins e Calado [BMC19] propuseram um novo método com Redes Neurais Recorrentes Bidirecionais com informações acerca do título e as duas primeiras frases das notícias.

As principais referências adotadas aqui são os trabalhos de Kaushal et al. [KSG21], Shu et al. [SZW+19], ambos citados por ALDayel e Magdy [AM21], e Shu, Wang e Liu [SWL17]. No primeiro, os autores constataram uma correlação entre o viés dos usuários, a escolha léxica e os sentimentos deles. Dessa forma, ao realizar o comparativo do modelo proposto com o uso do BERT, com e sem a rotulação de vieses, constatou-se uma melhora no *F1-Score*. O foco não foi em notícias falsas, mas em como o cruzamento de variáveis de posicionamento fora do domínio de classificação pode contribuir para uma melhora do modelo.

No segundo trabalho, foram elaboradas as seguintes perguntas de pesquisa:

1. Quais usuários são mais propícios a compartilhar notícias falsas e notícias reais?

2. Quais características influenciam mais nos usuários para compartilhar notícias falsas ou notícias reais?

3. É possível usar as características dos usuários para identificar notícias falsas? Como?

Para responder essas perguntas, eles utilizam o algoritmo do *Random Forest* e verificam o Índice de Gini de algumas *features*. Dentre as *features* mais importantes pelo Índice de Gini, o viés político foi uma delas.

Por fim, no terceiro trabalho, os autores desenvolveram um modelo tripartite para detecção de *fake news*, apelidado de TriFN, levando em conta notícia, editor e usuário. Cerca de cinco componentes formaram o modelo: Conteúdo das Notícias, Perfil do Usuário, Interação entre Usuário-Notícia, Classificação de Notícias, e Relação entre Editor e Notícia. Neste último ponto, os autores consideraram as rotulações de *left*, *left-center*, *least-biased*, *right-center* e *right* presente na base Media Bias/Fact Check [VZ22]. Em outro trabalho [SWL19], os mesmo autores realizaram a comparação do modelo proposto com outros modelos de detecção de notícias falsas, como *Rhetorical Structure Theory* (RST), *Linguistic Inquiry and Word Count* (LIWC), Castillo, RST + Castillo e LIWC + Castillo, tendo o TriFN apresentado boas performances em relação aos demais. Os autores destacaram a importância do viés político para a construção do modelo.

Dentre outros trabalhos de destaque, Borenstein et al. [BSR+23] analisaram como vieses de raça e gênero podem ser detectados em documentos históricos a partir de modelos de *Word Embedding*. Os autores constataram que aspectos léxicos e semânticos apontam indícios dos viés relativos à produção desses documentos. Apontamentos similares foram feitos por Reddy, Duggenpudi e Mamidi [GDM19], em que desenvolveram um modelo de detecção de viés político em manchete de notícias a partir de mecanismos linguísticos para despertar a atenção do leitor. Nesse caso, o estudo constatou que o uso de determinadas palavras é um forte indicativo do viés político do redator.

Um trabalho que aborda a dinâmica de difusão de notícias falsas é o de Srinivas, Das e Pulabaigari [SDP22]. Nele, os autores analisam como as três dimensões do comportamento humano influenciam no compartilhamento de *fake news* nas redes sociais. Utilizando a base FakeNewsNet, a ser descrita no Capítulo 3, o modelo desenvolvido constatou dinâmicas diferentes entre os usuários falsos que compartilham conteúdo desinformativo de fofoca e os mesmos tipos de usuários que compartilham desinformação de cunho político, identificando que estes propagam notícias falsas que dão suporte aos seus pontos de vista. Apesar do modelo não ser focado na classificação da notícia em si, mas em predizer o alcance dela, o trabalho corrobora com a hipótese levantada da influência do viés político na disseminação de desinformação.

Nesse sentido, os trabalhos apontam para a importância do conhecimento acerca do viés político do emissor da mensagem para modelos de predição, mas ainda carecem de análises mais aprofundadas com relação a ele. O primeiro ponto a ser avaliado é um estudo comparativo visando apenas o texto base e o rótulo do viés. Apesar de apresentarem resultados satisfatórios e destacarem a importância desse dado nos modelos gerados, os autores alimentaram seus modelos com outros dados e informações no treinamento, não sendo possível estipular o impacto dessa característica isolada.

O segundo ponto é verificar a generalidade dessa constatação. Os autores fizeram os testes em apenas uma ou duas bases em seus estudos, sendo usada majoritariamente a FakeNewsNet [SMW+20], base criada pelos próprios autores. Então, ainda há a necessidade de verificar se em outras bases de notícias a importância do viés político também é verificada.

Destaca-se, também, a falta de trabalhos voltados para explicabilidade acerca das classificações de veracidade de informações, sendo um dos únicos o de Atanasova et al. [ASLA20]. Nele, os autores elencam três características que uma explicação de classificação de notícias deve ter:

1. **Cobertura:** A explicação contém informações relevantes e não deixa nenhum ponto importante de fora;

2. **Não-Redundância:** Não há repetição de informações já fornecidas na explicação;

3. **Não-Contradição:** As informações fornecidas na explicação não entram em contradição.

Percebe-se, então, que ainda não há nesse tipo de trabalho a análise de formatação do texto com o viés do emissor da mensagem.

Conforme será abordado no Capítulo 3, além do texto completo da notícia, serão avaliados os títulos. Nesse sentido, é importante frisar que os trabalhos relativos à classificação de texto concentram-se nos textos completos em detrimentos dos trechos mais curtos, resultando num volume menor de publicações sobre as manchetes das notícias, conforme constatado por Rana, Khalid e Akbar [RKA14]. Nessa perspectiva, Shu et al. [SSW+17] aponta que classificadores de manchetes para a questão do *clickbait* podem ser utilizados para classificação de notícias falsas, reforçando os apontamentos de Delmazo e Valente [DV18].

Em suma, os diferenciais deste trabalho em relação aos demais trabalhos presentes na literatura envolvendo viés político e modelos de classificação estão descritos na Tabela 2.1. Tais diferenciais consistem em:

Tabela (2.1) Diferenciais deste Trabalho para a Literatura.

Trabalhos	**Variáveis**	**Campo da Notícia**	**Bases**
Nosso Trabalho	Notícia e Viés	Título, Texto e Título + Texto	Politifact, KaggleFN e LIAR
Shu [SZW+19] [SWL17]	Notícia, Usuário, Interação e Portal	Texto	Politifact e BuzzFeed
Kaushal [KSG21]	Análise de Sentimento		
Citados por AlDayel [AM21]	Desenvolvimento de Dataset e Análise de Usuários		

- Propor abordagem que difere da análise do perfil de usuários e da criação de base de dados, buscando enfoque e avanço na detecção de notícias falsas;
- Apresentar uma proposta para modelos de classificação que foque no viés político dos portais de notícias e não na opinião ou sentimentos dos usuários de redes sociais;
- Realizar uma análise tanto a nível do texto, quanto a nível do título da notícia;
- Desenvolver uma metodologia que englobe apenas os dados referentes aos textos / títulos e o viés do portal da notícia, desconsiderando outras variáveis, a fim de avaliar o impacto direto do viés na classificação;
- Realizar testes em mais de uma base, compostas de notícias coletadas de portais.

CAPÍTULO 3

Metodologia

A metodologia adotada nos experimentos está ilustrada no fluxograma da Figura 3.1. Nela, há dois momentos, que consistem no treinamento e no teste do experimento, demarcados pelas áreas em branco e cinza, respectivamente. O conjunto de dados Γ corresponde aos dados de notícias para treinamento, o γ aos dados de notícias para teste e o Δ aos dados de vieses. A primeira etapa foi a seleção das bases de *fake news*, a qual está descrita na Subseção 3.1. A segunda etapa foi a seleção da base que contenha o viés político de cada portal de notícias, sendo abordada na Subseção 3.2. O terceiro passo consistiu na junção das duas bases, removendo os registros que não estão referenciados na base de vieses, processo definido na Figura 3.1 como Verificação do Viés, o qual foi detalhado na Subseção 3.3. Com isso, os textos das notícias da base reduzida, expressa na figura como Γ', passam pelo pré-processamento, chamado de Limpeza do Texto, especificado na Subseção 3.4, resultando em um novo vetor de textos de notícias, definido aqui como Γ''. Paralelamente, os dados de vieses, após a redução, são definidos como Δ'. Esse conjunto de etapas é definido como Geração, pois corresponde à geração dos vetores que serão utilizados nos experimentos.

A partir desse ponto, para avaliar o efeito da rotulação do viés político nos *datasets* de notícias falsas, este trabalho adotou uma metodologia comparativa entre os três cenários selecionados: bases rotuladas com viés (texto + viés), bases sem o rótulo de viés (apenas texto) e bases com apenas o rótulo (apenas viés). Para ilustrar cada cenário, há o detalhamento da etapa de Concatenação na Figura 3.2. Essa etapa corresponde à seleção de qual base será usada para o treinamento do classificador. No Cenário 1, é utilizado apenas a base de rótulo de vieses, de forma que cada rótulo é tratado como uma variável categórica, especificado como Categorização na Figura 3.2. No Cenário 2, é avaliado apenas a base de texto da notícias, o qual passa pela etapa de Extração de Características, detalhada na Subseção 3.5. No Cenário 3, são utilizadas ambas as bases, sendo aplicada a Rotulação de viés aos textos das notícias, os quais seguem para a Extração de Características. Por fim, os algoritmos de aprendizagem (especificados na Subseção 3.6) são escolhidos para os três cenários descritos. Após o treinamento, o modelo é definido como λ na Figura. Ao final, há o teste do modelo, no qual as bases passam pelas mesmas etapas de Geração e Concatenação descritas e as predições são avaliadas, sendo esta etapa discutida na Subseção 3.7.

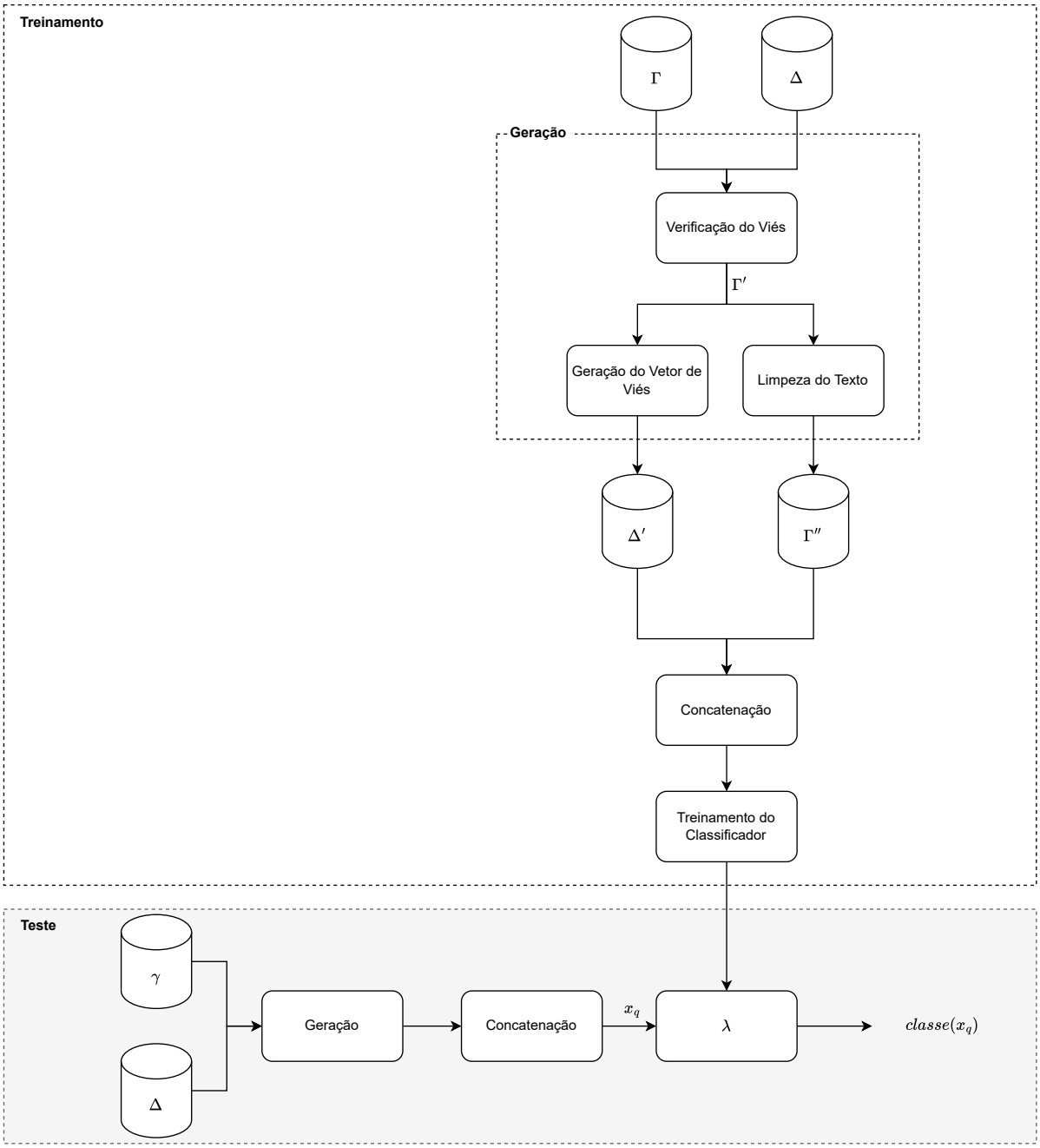


Figura (3.1) Fluxograma das Etapas do Experimento.

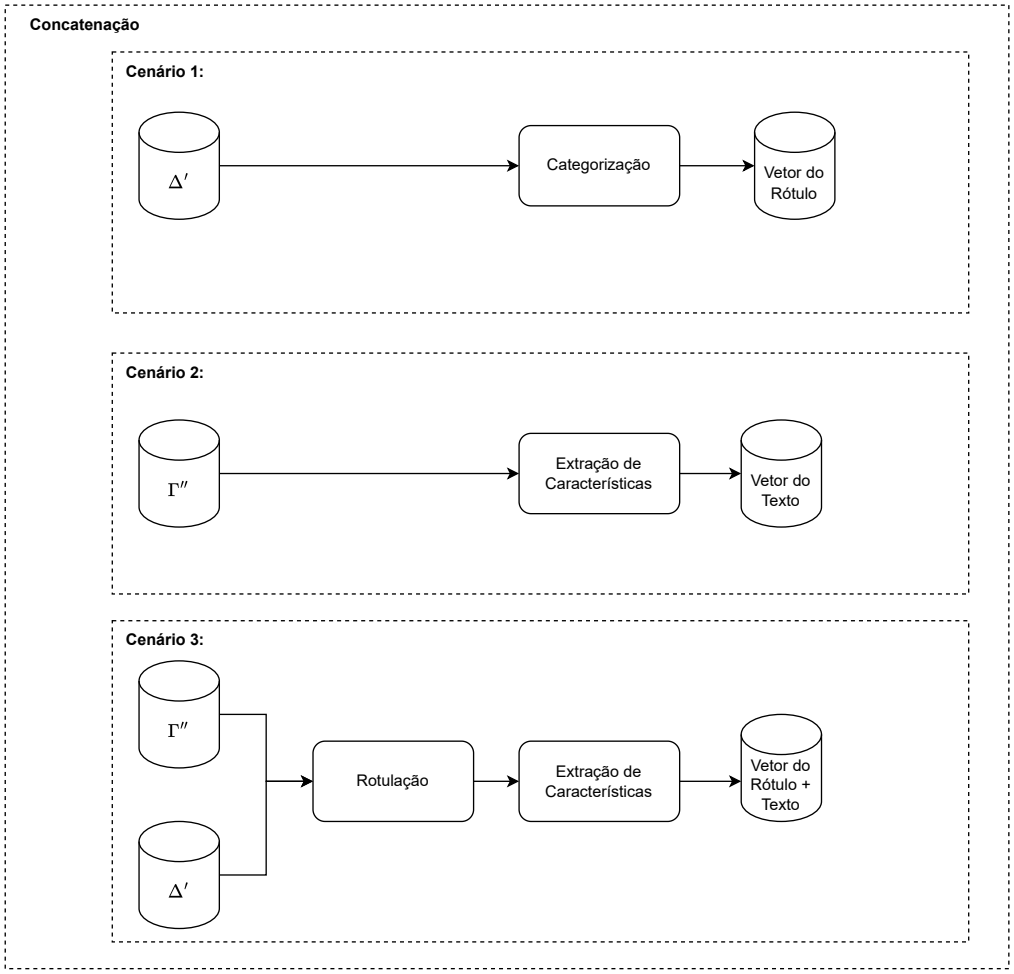


Figura (3.2) Detalhamento da Etapa de Concatenação.

3.1 Dataset de Notícias

Considerando o trabalho de D'Ulizia et al. [DCFG21], o qual realizou um levantamento das bases de *fake news* disponíveis, é possível identificar dois tipos: bases de notícias publicadas em algum portal e bases de declarações. No primeiro tipo, cada instância corresponde a uma matéria veiculada como notícia, possuindo título e estrutura dissertativa em sua escrita; Enquanto no segundo tipo, cada instância da base consiste em alguma entrevista ou postagem de determinado indivíduo, não possuindo estrutura jornalística. Essa diferença é importante pois a estrutura dos textos avaliados são diferentes, além de que, para este estudo, é preciso ajustar a rotulação para cada um desses tipos, como será abordado na Seção 3.2.

Nesse sentido, foi necessário definir critérios para formalizar a escolha das bases de notícias, a fim de selecionar os *datasets* que permitissem uma melhor avaliação da proposta deste trabalho. Nesse sentido, foram elencados os seguintes critérios:

- **Critério 1:** Ser composto de notícias em inglês;
- **Critério 2:** Possuir um campo para corpo do texto da notícia e um campo para o título / manchete, no caso das bases de notícias publicadas;
- **Critério 3:** Ser composto de notícias que abordem questões explicitamente políticas;
- **Critério 4:** Ter classe alvo binária (Falsa ou Verdadeira).

O Critério 1 foi adotado para possibilitar o processo de comparação com os vieses dos portais de notícias, o qual, conforme descrito na Seção 3.2, analisam portais de língua inglesa. O Critério 2 foi escolhido pois, como descrito nos Capítulos 1 e 2.5, há apontamentos na literatura para o efeito *Clickbait* nos títulos de notícias, no entanto, há escassez de trabalhos que analisem as diferenças de resultados entre modelos aplicados ao corpo do texto, modelos aplicados ao título da notícia e modelos aplicados à junção de título com o texto, sendo este um diferencial do trabalho aqui desenvolvido. Este critério só pôde ser adotado para bases de notícias publicadas. O Critério 3 foi escolhido para delimitar o escopo dos tipos de notícias a serem avaliados, escolhendo o tipo de notícia que mais pode sofrer influência do viés político do portal em questão. O Critério 4 foi escolhido para que o problema de classificação fosse mais homogêneo, facilitando a análise entre cenários. Neste quesito, um dos fatores escolhidos para eliminação de bases com multi-classes foi o trabalho de Farhangian, Cruz e Cavalvanti [FCC24], o qual aponta desafios para classificação desse tipo do problema, por conta da sobreposição de classes.

Tabela (3.1) Bases Escolhidas para os Experimentos deste Trabalho.

Dataset	**Ano**	**Tamanho Original**	**Verdadeiros**	**Falsos**
FakeNewsNet - PolictFact	2020	1056	624	432
KaggleFN	2020	2095	801	432
LIAR	2017	12791	4507	8284

Dessa forma, foram selecionadas três bases para realizar os testes, descritas na Tabela 3.1, as quais atendiam os critérios estabelecidos. Uma delas faz parte do trabalho de Shu et al. [SMW+20]. Nele, os autores coletaram diversas notícias vinculadas no Twitter, sendo rotuladas a partir de agências de *Fact Checking*, as quais são instituições jornalísticas voltadas para checagem de dados e informações já veiculadas, seja por agentes ou pela própria imprensa, a fim de validá-las ou não [dSdAV19]. Nesse sentido, essas agências atestam se as peças comunicativas analisadas apresentam ou não verdade e veracidade, sendo rotuladas, ao final do processos, como verdadeiras ou falsas. Do trabalho em questão, foi escolhida a base de nome Politifact[1], na qual os autores coletaram 1.056 notícias. Destaca-se que a base se constitui de *links* para as postagens originais, sendo executado um *script* de coleta das notícias. No entanto, muitos *links* não estão mais ativos, resultando em número menor de registros: 709. Além dela, havia ainda a base GossipCop, também presente no trabalho em questão, a qual não foi escolhida pois não atendeu ao critério de notícias relacionadas a temas políticos.

[1] https://github.com/KaiDMML/FakeNewsNet

A segunda base de notícias utilizada foi a presente no trabalho de Bharadwaj et al. [BAB+20], intitulada de KaggleFN, disposibilizada na plataforma de mesmo nome [2]. A base foi construída a partir da coleta de notícias em diversos portais, com 2.095 registros. Por fim, para contemplar as bases do tipo de declarações, foi selecionada a base LIAR [3] [Wan17], a qual consiste em diversas declarações envolvendo figuras políticas estadunidenses. No entanto, para contemplar o Critério 4, a base foi ajustada. Originalmente, ela possui seis classes, sendo nesse estudo convertidas para apenas duas classes, da seguinte forma: as classes *pants-fire*, *false*, *barely-true* e *half-true* representam a classe de notícias falsas, enquanto que *mostly-true* e *true*, as verdadeiras.

3.2 Dataset de Vieses

Para rotular os vieses dos portais de notícias, foi utilizada a base *Media Bias/Fact Check* [VZ22], a qual foi utilizada no trabalho de Shu, Wang e Liu [SWL19]. A base foi desenvolvida por um portal homônimo independente que realiza o processo de *Fact Checking* e classificação de viés de diversas fontes de mídia. Tendo em vista que a base é periodicamente atualizada, este trabalho adotou a versão de 2022. Nesse sentido, para definir qual espectro político cada portal estava mais alinhado, a equipe do *Media Bias/Fact Check* adotou os seguintes critérios:

- **Omissão:** Análise de quais conteúdos foram omitidos ou ignorados nas publicações;
- **Rotulação:** Uso de termos pejorativos ou elogiosos para definir determinado agente ou grupo;
- **Posicionamento:** Adoção de posicionamento em determinado tema;
- **Fontes:** Escolha de determinadas fontes em detrimento de outras;
- **Comentários:** Emissão de comentários por parte do jornalista;
- **Seleção de História:** Escolha de quais histórias serão contadas ou não;
- **Viés de Confirmação:** Tendência a privilegiar informações que validem posições já definidas;
- **Conotação:** Interpretação aplicada em torno de um termo usado;
- **Denotação:** Significado de Dicionário das palavras usadas;
- **Linguagem Carregada:** Uso de linguagem carregada emocionalmente ou com esteriótipos com intuito de influenciar;
- ***Purr Words*:** Termos que indicam posicionamento favorável;

[2] https://www.kaggle.com/datasets/ruchi798/source-based-news-classification
[3] https://paperswithcode.com/dataset/liar

- ***Snarl Words*:** Termos que indicam posicionamento contrário.

Tais critérios coadunam com as definições de vieses estipuladas por Saez-Trumper, Castillo e Lalmas [STCL13], conforme discutido na Seção 2.2. Dessa forma, cada portal de notícias recebe um dos seguintes rótulos:

- **Pouco Tendencioso / Centro (*center*):** Portais com pouco uso de linguagem carregada, em que não há posicionamento muito aparente sobre os temas expostos;
- **Centro-Esquerda (*left-center*):** Portais inclinados levemente a moderadamente ao viés liberal;
- **Esquerda (*left*):** Portais inclinados moderadamente a fortemente ao viés liberal;
- **Centro-Direita (*right-center*):** Portais inclinados levemente a moderadamente ao viés conservador;
- **Direita (*right*):** Portais inclinados moderadamente a fortemente ao viés conservador;
- **Conspiração-Pseudociência (*conspiracy-pseudoscience*):** Portais com baixa credibilidade e carência de fontes consistente. Aqui, estão presentes portais acerca de temas como alienígenas, propaganda anti-vacina e demais teorias das conspirações;
- ***Fake News*:** Portais identificados como extremamente enviesados, que deliberadamente produziram desinformação. O que difere do rótulo anterior é que as notícias dos portais daqui foram verificadas como falsas, enquanto as de Conspiração-Pseudociência utilizam de elementos que não podem ser verificados;
- **Pró-Ciência (*pro-science*):** Portais que seguem o método científico e realizam divulgação científica;
- **Sátira (*satire*):** Portais que utilizam do humor e da ironia para comentar notícias.

É perceptível que os parâmetros para vieses políticos levam em conta os espectros do liberalismo e do conservadorismo. Isso ocorre pois, como ressaltado pelos autores, os principais partidos políticos dos Estados Unidos são tomados como referência, isto é, o Partido Democrata é considerado Centro-Esquerda por seu viés liberal, enquanto o Partido Republicano é considerado Direita por seu viés conservador.

Para ilustrar como ocorre tal classificação na prática, foram selecionados, como exemplos, dois portais de espectros políticos distintos: *CNN* e *The Sun*. No caso da *CNN*, a equipe do *Media Bias / Fact Checking* avaliou que o portal assumia uma postura mais progressistas em seus debates, tecendo críticas ao Partido Republicano. Analisando a cobertura do portal em relação aos 100 primeiros dias do governo de Donald Trump, constatou-se que 93% de suas notícias eram negativas ao governo [Pat17]. Já o *The Sun*, tanto a versão norte-americana, quanto a do Reino Unido, foi constatado um alinhamento do portal com pautas conservadoras desde 2010, com frequentes comentários negativos à gestão de Joe Biden.

No caso da base LIAR, o próprio *dataset* possui um campo com a afiliação partidária do indivíduo que proferiu aquela declaração. Então, por exemplo, uma declaração proferida por Donald Trump, tem o rótulo de *"republican"*, enquanto uma proferida por Barack Obama, tem o rótulo *"democrat"*.

3.3 Fusão das Bases

O próximo passo foi a fusão das duas bases, a de notícias e a de vieses. No entanto, há portais que não constam nas duas bases, sendo necessário, então, buscar a interseção das bases. Na Figura 3.3, é possível visualizar em forma de Diagrama de Venn o processo de fusão das bases. Nesse sentido, esta pesquisa seleciona as instâncias da interseção, descartando as demais instâncias.

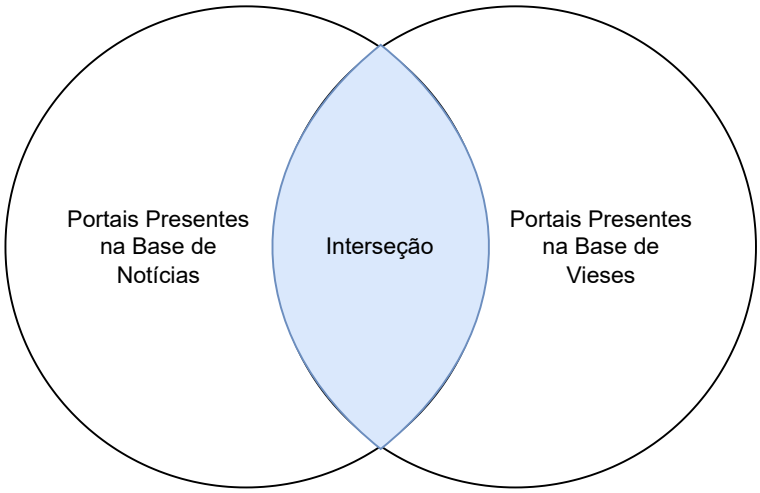


Figura (3.3) Diagrama de Venn das Bases.

O processo comparativo entre bases avaliou se o portal do item do *dataset* de notícias existia na base de vieses, de forma que foi comparada a coluna da origem da notícia, a qual consistia no link do site em questão, com a coluna de URL da base de vieses. Para isso, houve uma limpeza nas *strings* do *dataset* de notícias para deixar no padrão do *dataset* de vieses, removendo assim o *"https:"*, *"http:"*, *"/"* e *"www."* das URLs.

Após esse processo, foi percorrido o *dataset* de notícias, buscando o viés do portal de cada item. Caso o portal também estivesse listado no *dataset* de vieses, seu viés era adicionado ao *dataset* de notícias em uma nova coluna intitulada *"bias"*. O valor adicionado corresponde a um dos nove vieses descritos na Subseção 3.2. Dessa forma, por exemplo, notícias do portal *cnn.com* recebem o rótulo *left*, enquanto as notícias do *thehill.com* recebem o rótulo *center*. Caso não possuísse, o item era removido do *dataset* de notícias. Ao final, o KaggleFN foi o que teve menor taxa de redução. No caso da base LIAR, foram removidos afiliações partidárias que não fossem democratas ou republicanos, pois eram afiliações que não apontavam viés partidário, como *"independent"*. Os valores resultantes da redução das bases são mostrados na Tabela 3.2.

Tabela (3.2) Redução das Bases.

Dataset	Classe	Antes	Depois	Redução (%)
FakeNewsNet - Politifact	Verdadeiro	394	193	51,01
	Falso	315	156	50,47
KaggleFN	Verdadeiro	801	714	10,86
	Falso	1294	1073	17,07
LIAR	Verdadeiro	4507	3513	22,05
	Falso	8284	6289	24,08

Considerando o tamanho dos *datasets* na literatura, D'Ulizia et al. [DCFG21] realizaram uma revisão sistemática das bases de notícias falsas disponíveis, avaliando 27 bases. Foram apontados um amplo espectro no tamanho das bases, indo de base com 75 instâncias até base com 60 milhões de instâncias. Os autores destacam que a maioria das bases podem ser consideradas pequenas, com menos de 15 mil notícias. Dessa forma, o resultado da redução ainda mantêm-se dentro da média presente na literatura.

Na Figura 3.4, está representado o número de instâncias por viés na base do Politifact, sendo a coluna azul para notícias verdeiras e a coluna vermelha para notícias falas. Percebe-se uma predominância do rótulo de *left-center* nas notícias verdadeiras, enquanto nas notícias falsas predomina o rótulo de *fake-news*. Já a Figura 3.5 ilustra a distribuição de rótulos de vieses na base do KaggleFN para notícias verdadeiras e para notícias falsas, seguindo a mesma distribuição de cores. Nas notícias verdadeiras, os casos de *fake-news* são os que têm maior número de instâncias, enquanto nas notícias falsas, os rótulos de *conspiracy-pseudoscience* possuem mais ocorrências. Por fim, a Figura 3.6 ilustra a distribuição de filiações partidárias em relação a cada tipo de notícia, em que percebe-se que o rótulo *republican* possui uma predominância de notícias falsas.

É importante destacar que a rotulação dos vieses ocorre a nível de portal e não a nível de notícia. Dessa forma, um portal classificado como *"fake-news"* foi considerado como o propagador de notícias falsas, no entanto, ele pode, eventualmente, publicar notícias verdadeiras, como pode ser visto na base KaggleFN. Além disso, apesar da base de vieses prever cerca de nove tipos de rótulos, pode ocorrer da base de notícias avaliada não apresentar exemplares de todos os tipos de vieses, novamente como no caso da KaggleFN. Nesse exemplo, não há notícias para os rótulo de *center*, *left-center*, *right-center* e *pro-science*.

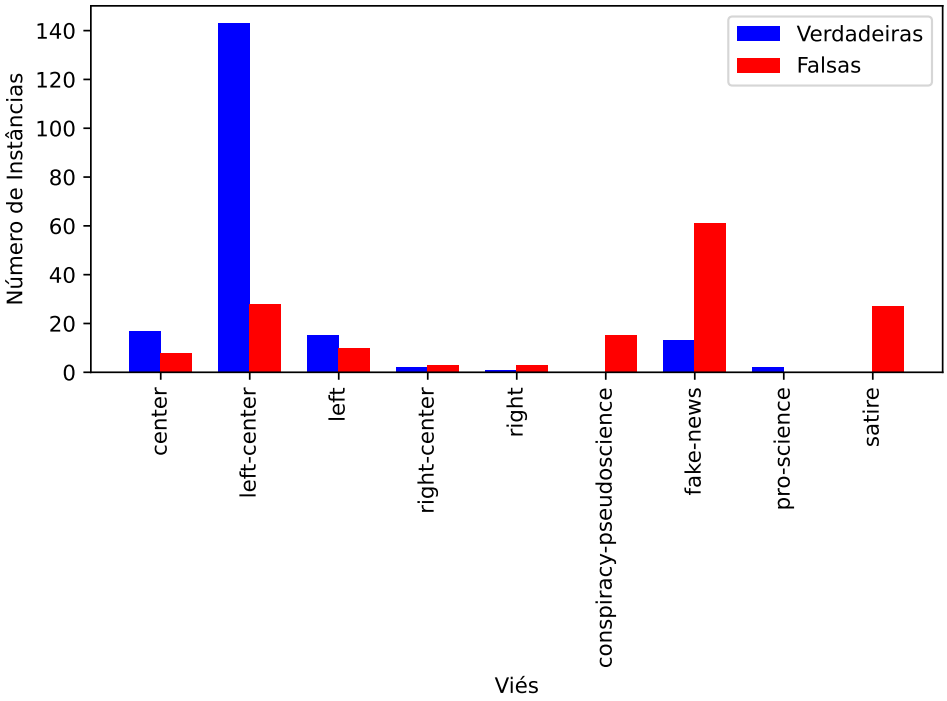


Figura (3.4) Distribuição dos Vieses - Politifact

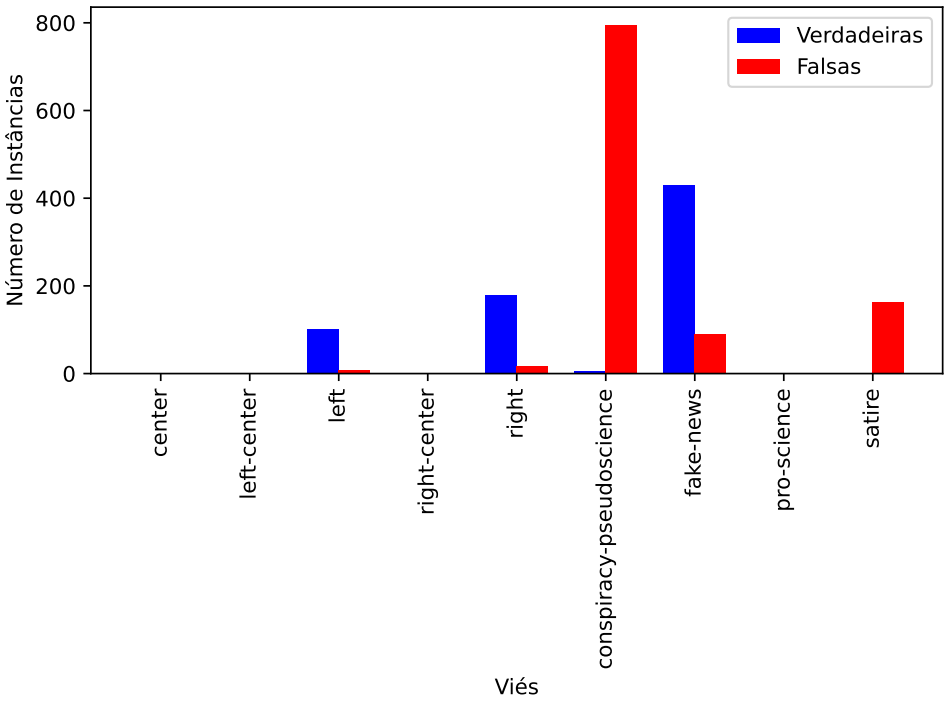


Figura (3.5) Distribuição dos Vieses - KaggleFN

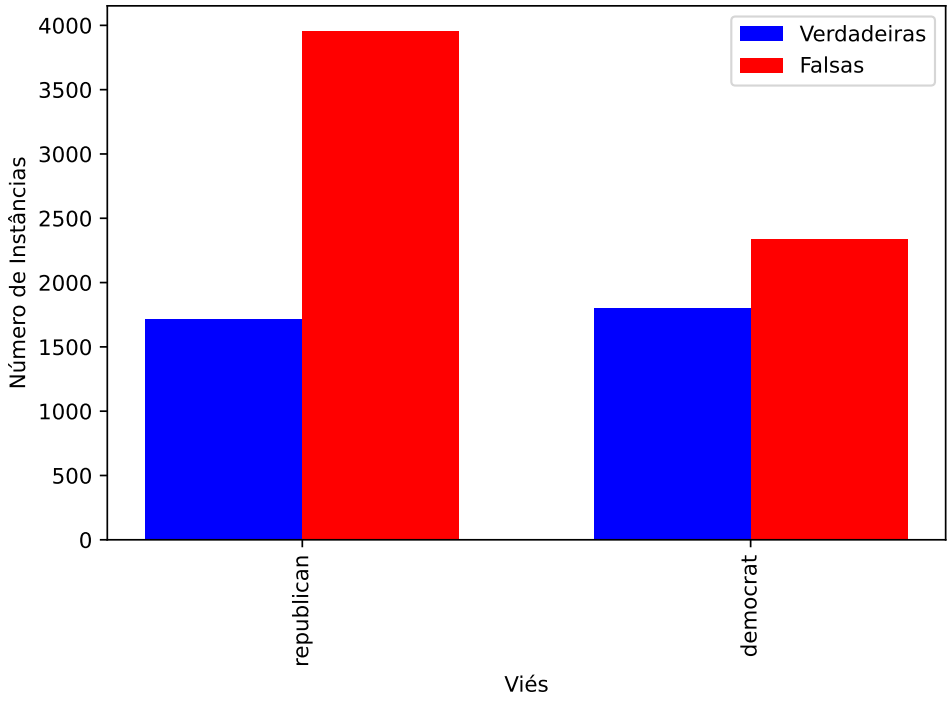


Figura (3.6) Distribuição dos Vieses - Liar

Além da *Media Bias/Fact Check*, também foi avaliada a base da *AllSides* [All23], no entanto, ela não foi inclusa no estudo em virtude da alta redução das bases de notícias, chegando a mais de 90%.

3.4 Pré-processamento

O pré-processamento dos dados é a fase em que há o tratamento do texto a fim de possibilitar seu processamento. Nesse sentido, este estudo se baseou no trabalho de Farhangian, Cruz e Cavalcanti [FCC24], o qual aponta cinco etapas importantes para o Processamento de Linguagem Natural, sendo elas:

1. **Normalização:** Processo de converter texto para sua forma padrão. Aqui, foi adotada a conversão das maiúsculas para minúsculas;

2. **Tokenização:** Processo de separar o texto bruto em unidades de análise. Aqui, cada palavra se tornou um *token*;

3. **Remoção de *Stop Words*:** Processo de remover palavras consideradas de pouca utilidade semântica, isto é, que não possuem sentido próprio, como artigos;

4. **Remoção de Pontuação e Símbolos:** Processo de remover sinais gráficos, ".", ",", "@", entre outros;

5. ***Stemming:*** Processo de retornar uma palavra à sua forma radical, removendo prefixos e sufixos. Aqui, foi adotada uma variação, a **Lematização**, na qual uma palavra volta a sua forma original, no caso de verbos o infinitivo, por exemplo. Foi escolhida essa variação pois no *stemming* pode acontecer de ser gerada uma palavra que não existe no dicionário, enquanto na lematização isso não ocorre.

3.5 Extração de Características

Nesta etapa, é feita a aplicação de algoritmos para extração de características dos textos. Nesse sentido, são criados três cenários para o teste: apenas rótulo de viés, apenas texto e texto + rótulo. No primeiro caso, os modelos são treinados apenas com os vieses de cada instância da base, não passando pelo estágio de extração de *features*, pois são tratadas como variáveis categóricas. No segundo caso, apenas o texto original tem suas *features* extraídas. Já no terceiro caso, o valor do rótulo é adicionado ao início do texto que será avaliado. Sendo assim, há a concatenação do *Viés + Caractere Espaço (' ') + Texto*. Esse procedimento está ilustrado na Figura 3.7, o qual demonstra o resultado da rotulação adotando como exemplo o texto da notícia do portal *haaretz.com*, rotulado como *left*.

Figura (3.7) Exemplo de Rotulação

Após isso, é aplicado a extração de *features* nos cenários de texto e texto + rótulo. Nos testes realizados, foi utilizado o TF-IDF, o qual consiste em um algoritmo que gera uma matriz com os valores da razão da frequência de determinado termo naquele texto dividido pela frequência desse termo nos textos da base. Essa escolha consiste no fato de que esse método é bastante comum na extração de *features* textuais na literatura [ZG20]. Além disso, os cenários também foram avaliados com o uso de dois métodos dependentes do contexto, DistilBERT e o LLAMA 2. O primeiro é uma variação do BERT [SDCW19], o qual consiste numa rede neural com modelo de incorporação de palavras pré-treinado baseado na arquitetura de codificação de *transformers*, também sendo bastante utilizado na literatura [KGN21]. O segundo consiste num modelo, desenvolvido em parceria entre Meta e Microsoft, sendo adotada nesta pesquisa a sua versão pré-treinada e ajustada com 7 bilhões de parâmetros [TMS$^+$23]. Os parâmetros utilizados para configurar o DistilBERT e o LLAMA 2 podem ser vistos no Apêndice A, sendo utilizado o *[CLS] token* no DistilBERT e o *Weighted-Mean-Pooling* no LLAMA 2. O *[CLS] token* consiste numa camada que codifica a média dos demais *tokens* de forma que cada sentença tenha seu valor único; no caso do LLAMA 2, essa camada não existe na arquitetura do modelo, sendo necessário realizar a aplicação de pesos nos *tokens*, de forma que os últimos possuem maior peso que os do início da sentença, pois estão mais contextualizados que os demais, isto é, uma palavra posicionada no final de uma sentença possui todo o contexto semântico do restante

da frase, enquanto a palavra posicionada no início não possui esse referencial.

Tabela (3.3) Tamanho dos Vetores por Método de Extração. Entre parenteses, está ilustrado quanto o vetor aumentou com a rotulação, para os casos em que houve mudança no tamanho.

Dataset	**Campo**	**TF-IDF**	**DistilBERT**	**LLAMA 2**
FakeNewsNet - Politifact	Texto	21.260 (+1)	768	4.096
	Título	1.491 (+2)		
	Título + Texto	21.302 (+1)		
KaggleFN	Texto	39.744 (+1)		
	Título	4.973 (+1)		
	Título + Texto	40.172 (+1)		
LIAR	Texto	10.558		

Na Tabela 3.3, estão distribuídos os tamanhos de cada vetor por método de extração de *feature*. A arquitetura do DistilBERT e do LLAMA 2 utilizam um número fixo de *tokens*, independente do texto de entrada do modelo. No caso do DistilBERT é gerada uma matriz de 768 colunas, enquanto no LLAMA 2 é uma matriz de 4.096 colunas. Já para o método TF-IDF, há uma alteração no tamanho dos vetores. Para a base do Politifact, foi gerada uma matriz esparsa para o texto com cerca de 21.260 colunas antes da rotulação, enquanto a rotulação gerou 1 coluna extra, totalizando 21.261. Já para os títulos desta base, antes da rotulação, a matriz esparsa tinha 1.491 e, depois da rotulação, passou a ter 1.493, ou seja, adição de duas colunas. No caso da junção de título + texto, a matriz passou de 21.302 para 21.303 a partir da adição de uma coluna. Apesar de à primeira vista, ser uma adição mínima, isso representa um acréscimo de informação relevante em cada instância da análise. Pegando como exemplo o título da instância 144, antes da rotulação apenas duas colunas possuíam valor significativo, isto é, diferente de zero, sendo elas a do termo *czars* (valor de 0,707) e *romanovs* (valor de 0,707), e ao ser rotulada passou a ter três colunas com valor significativo, *czars* (valor de 0,696), *left* (valor de 0,178) e *romanovs* (valor de 0,696). Analisando a proporção entre zeros e valores significativos da matriz, para o vetor de texto, tem-se 98,14% de valores zeros e 1,86% de valores significativos; para o vetor de título, 99,54% de valores zeros e 0,45% de valores significativos; e para o vetor de título + texto, 98,13% de valores zeros e 1,87% de valores significativos. A rotulação gerou uma aumento de 0,01% nos valores significativos no vetor de texto, 0,13% no vetor de título e não apresentou alteração percentual no vetor de título + texto.

O mesmo acontece na base do KaggleFN, a qual sua matriz esparsa para o texto da notícia tinha cerca de 39.744 colunas, passando a ter 39.745 depois da rotulação, enquanto a matriz dos títulos passou de 4.973 para 4.974 e a matriz de texto + título passou de 40.172 para 40.173. Analisando a instância 76, o título possuía inicialmente duas colunas significativas *estate* (valor de 0,750) e *master* (valor de 0,662), e após a inclusão do rótulo passou a ter quatro colunas significativas *conspiracy* (valor de 0,168), *estate* (valor de 0,728), *master* (valor de 0,643) e *pseudoscience* (valor de 0,168). Analisando a proporção entre zeros e valores significativos da matriz, para o vetor de texto, tem-se 99,53% de valores zeros e 0,47% de valores significativos; para o vetor de título, 99,86% de valores zeros e 0,14% de valores significativos; e para o vetor de título + texto, 99,52% de valores zeros e 0,48% de valores significativos. A rotulação gerou

um aumento de 0,03% de valores significativos no vetor de título e não apresentou alteração percentual nos vetores de texto e título + texto.

A base LIAR não alterou o número de colunas após a inclusão do rótulo, permanecendo em ambos os cenários com 10.558. Analisando a proporção entre zeros e valores significativos da matriz, tem-se 99,91% de valores zeros e 0,09% de valores significativos, com aumento de 0,01% na rotulação.

Percebe-se uma redução nos valores das colunas que já haviam valores significativos, além da adição de um novo valor relativo ao viés político. Pelo cálculo do TF-IDF, quanto maior o número de instâncias em que determinado termo está inserido, menor a pontuação final daquela coluna, o que faz com que boa parte dos rótulos caia a pontuação. No entanto, a presença de determinada coluna de viés com valor significativo contribui positivamente para a classificação da veracidade da notícia, conforme será discutido no Capítulo 4.

3.6 Algoritmos de Aprendizagem de Máquina

O trabalho de Farhangian, Cruz e Cavalvanti [FCC24] aponta três tipos de algoritmos de aprendizagem de máquina utilizados para classificação de *fake news*, sendo eles: Algoritmos Clássicos de Aprendizagem de Máquina, Métodos de *Ensemble* e Modelos de *Deep Learning*. Este estudo selecionou 7 algoritmos de aprendizagem de máquina para realizar a etapa de classificação, correspondendo a todos da categoria de algoritmos clássicos e a todos da categoria de métodos *ensemble* descritos no estudo em questão. Dos algoritmos escolhidos, três deles que foram utilizados tanto no trabalho de Shu et al. [SMW$^+$20], quanto no de Bharadwaj et al. [BAB$^+$20], sendo eles: Regressão Logística, *Support Vector Machine (SVM)* e Naive Bayes. Outros dois adotados neste trabalho também estão presentes no trabalho de Bharadwaj et al. [BAB$^+$20], *Random Forest* e *AdaBoost*. Ainda, há dois que foram adições deste estudo, como *K-Nearest Neighbors (KNN)* e *XGBoost*.

Destaca-se também que foi adotado o *Grid-Search* para avaliar quais melhores parâmetros para cada algoritmo avaliado. Os parâmetros utilizados podem ser observados na Tabela 3.4. Assim, com a definição dos parâmetros por algoritmo, são realizadas as etapas de treinamento e teste (descritas na Seção 3.7) e, ao final, são extraídos a média e o desvio padrão das medidas de desempenho de todas as execuções. Para os experimentos realizados, foram adotados a acurácia e o *F1-Score* como medidas de desempenho.

3.7 Teste e Validação

No estudo de Shu et al. [SMW$^+$20], foi utilizado a divisão de 80% da base para treinamento e 20% para teste. Já no trabalho de Bharadwaj et al. [BAB$^+$20], são adotados três cenários: 80% da base para treinamento e 20% para teste; 70% da base para treinamento e 30% para teste; e 60% da base para treinamento e 40% para teste. Para os experimentos realizados neste estudo, foi utilizado a separação de 80% para treinamento e 20% para teste, com estratificação em função da proporção de instâncias por classe, sendo que cada algoritmo foi avaliado 30 vezes, gerando uma nova configuração de treinamento e teste a cada execução.

Tabela (3.4) Parâmetros do Grid-Search

Algoritmo	**Hiper-parâmetros**
Regressão Logística	penalty : [l1, l2] C : [0.1,1, 10, 100] solver : [liblinear]
SVM	C: [0.1,1, 10, 100], gamma: [1,0.1,0.01,0.001] kernel: [rbf, poly, sigmoid]
Naive Baayes	var_smoothing: np.logspace(0,-9, num=100)
KNN	n_neighbors=[3,5,7,9]
Random Forest	max_depth:[3,5,10], n_estimators:[10,100,200], max_features:[1,3,5,7], min_samples_leaf:[1,2,3], min_samples_split:[1,2,3]
AdaBoost	n_estimators: [10, 50, 100], learning_rate : [0.01, 0.1, 1.0]
XGBoost	n_estimators: [10, 50, 100], learning_rate: [0.01, 0.1, 1.0]

Para avaliar se há significância estatística entre os resultados encontrados, foi adotada a metodologia proposta por Demšar [Dem06], abordagem já utilizada em trabalhos de classificação de texto, tais como o de Sousa [SOU20]. Nesse sentido, primeiro é realizado o Teste de Friedman, no qual é verificado se o *p-value* encontrado é menor que 0,05 para assim rejeitar a hipótese nula, isto é, a hipótese de que não há diferenças entre as amostras analisadas. Sendo rejeitada a hipótese nula, é aplicado o Pós-teste de Nemenyi para ilustrar onde estão essas diferenças.

Todos os testes foram feitos no ambiente do *Google Colab* na linguagem Python 3, sendo alocado a opção de CPU para os testes com TF-IDF como acelerador de *hardware* e a opção de T4 GPU para o DistilBERT e LLAMA 2. A escolha deste ambiente se deu em conta da versatilidade e praticidade do mesmo, tornando fácil a reprodutibilidade dos experimentos aqui apresentados. O código utilizado nos experimentos será disponibilizado na plataforma Github[4].

Tabela (3.5) Resumo dos Experimentos Realizados.

Datasets	**Campos**	**Extratores**	**Classificadores**	**Cenários**	**Execuções**	**Total (9.450)**
2 (Politifact e KaggleFN)	3 (Título, Texto e Título + Texto)	3 (TF-IDF, DistilBERT e LLAMA 2)	7 (Regressão Logística, SVM, Naive Bayes, KNN, Random Forest, AdaBoost e XGBoost)	2 (Com e Sem Rotulação)	30	7.560
1 (LIAR)	1 (Texto)					1.260
3 (Politifact, KaggleFN e LIAR)	-	-		1 (Apenas Rótulo)		630

Na Tabela 3.5, há um resumo dos experimentos realizados, com o número total das configurações avaliadas. Dessa forma, considerando: 2 bases de notícias publicadas $\times$ 3 campos

[4]https://github.com/lucasalisboa/Rotulacao_Fake_News

textuais avaliados (título, corpo da notícia e título + corpo da notícia) $\times$ 3 extratores de *features* $\times$ 7 classificadores $\times$ 2 cenários (rotulado e não rotulado) $\times$ 30 execuções = 7.560 configurações avaliadas, enquanto a LIAR avaliou 3 extratores de *features* $\times$ 7 classificadores $\times$ 2 cenários (rotulado e não rotulado) x 30 execuções = 1.260 configurações. Considerando o contexto de apenas rótulo (sem texto), tem-se 3 bases $\times$ 7 classificadores $\times$ 30 execuções = 630 configurações avaliadas. Ao total, o estudo verificou 9.450 configurações experimentais, desconsiderando as avaliações internas do *Grid-Search*.

CAPÍTULO 4

Resultados

Este capítulo apresenta os resultados obtidos nos experimentos propostos, detalhando as métridas de acurácia e *F1-Score* encontradas, bem como a validação estatística de cada cenário avaliado. Na Seção 4.1, estão dispostos os resultados para a base FakeNewsNet - Politifact. Na Seção 4.2, estão dispostos os resultados para a base KaggleFN. Na Seção 4.3, estão dispostos os resultados para a base LIAR. Por fim, a Seção 4.4 sintetiza os principais achados obtidos neste estudo.

4.1 FakeNewsNet - Politifact

Tabela (4.1) Resultados da Acurácia para o Texto da Politifact. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	55,61 (0,51)	87,28 (3,26)	81,42 (4,04)	75,42 (4,53)	**89,19 (3,35)**	84,19 (3,17)	86,04 (3,58)
SVM	82,14 (4,45)	86,99 (3,83)	81,57 (4,64)	75,80 (4,66)	**88,95 (2,70)**	84,33 (3,22)	86,09 (4,40)
Naive Bayes	66,66 (4,27)	80,04 (5,23)	75,71 (4,18)	75,66 (3,48)	80,47 (3,38)	79,47 (4,49)	**81,95 (3,80)**
KNN	79,33 (5,40)	69,00 (4,44)	73,61 (4,73)	70,00 (4,51)	71,85 (3,72)	77,85 (3,53)	**80,99 (3,94)**
Random Forest	80,61 (4,40)	76,52 (5,21)	77,04 (5,01)	74,28 (4,13)	77,61 (5,93)	82,19 (4,55)	**83,95 (4,13)**
AdaBoost	81,04 (3,47)	76,00 (3,94)	76,47 (4,66)	71,52 (4,10)	**88,19 (3,23)**	79,66 (4,21)	82,99 (4,04)
XGBoost	82,00 (2,98)	78,42 (4,88)	76,28 (4,81)	72,90 (5,13)	**84,04 (4,51)**	78,61 (3,98)	81,04 (4,44)

Na Tabela 4.1 estão dispostos os resultados da acurácia nos experimentos com o texto da FakeNewsNet - Politifact. Conforme descrito no Capítulo 3, foram avaliados três cenários: treinar os classificadores apenas com o rótulo de viés, apenas com o texto e com a concatenação de texto e rótulo de viés. Cada célula da tabela representa a média das 30 execuções, estando os valores do desvio padrão entre parênteses. Os melhores resultados por algoritmo estão destacados em negrito, sendo todos eles com a inclusão do viés. Na Figura 4.1, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

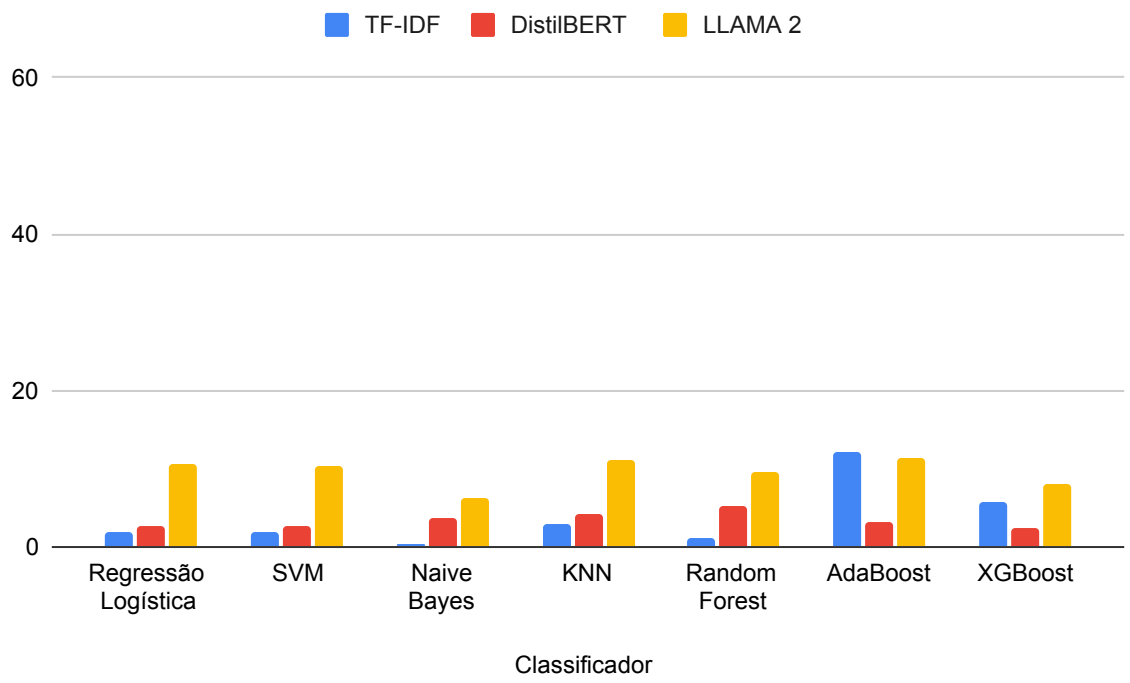
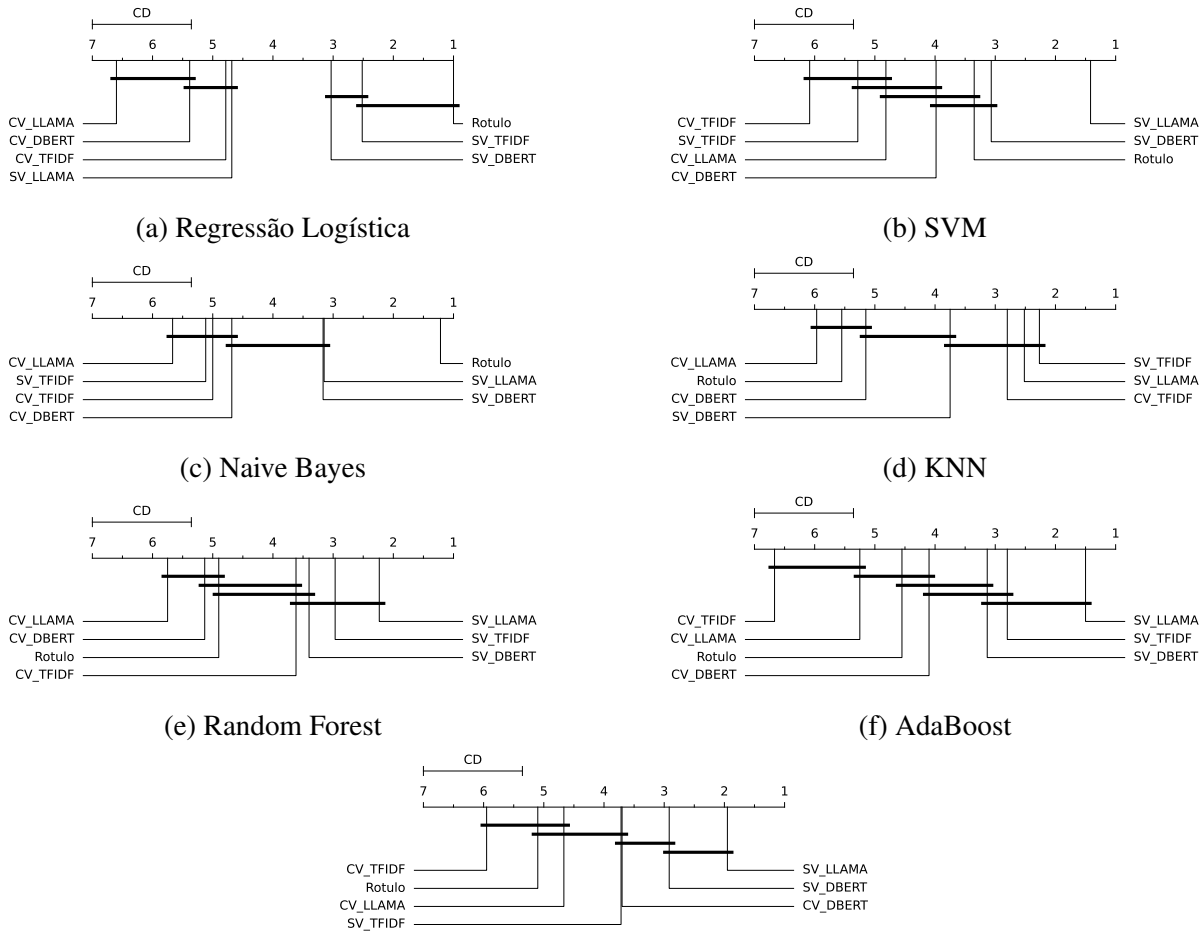


Figura (4.1) Percentuais de Ganhos com a Rotulação na Politifact - Texto (Acurácia)

Conforme é possível constatar, o extrator de característica LLAMA 2 apresentou um percentual de aumento acima dos demais, com todos os classificadores desse extrator obtendo resultados de ao menos 80,99% (no caso do KNN), enquanto com apenas texto o melhor classificador desse extrator obteve 75,80% (no caso do SVM), ou seja, o resultado menos expressivo do LLAMA 2 com a rotulação ainda é melhor que o resultado mais expressivo do LLAMA 2 sem a rotulação. Por sua vez, o DistilBERT e o TF-IDF também apresentaram resultados interessantes com a rotulação, com este último chegando no maior valor constatado, de 89,19% (no caso do Regressão Logística). É interessante pontuar que o cenário de apenas o rótulo, apesar de performar mal com alguns algoritmos, como Regressão Logística e Naive Bayes, apresentou na maioria do casos valores elevados, próximos ou superiores do cenário de apenas texto, indicando que a variável do viés político pode ter importante peso no processo de classificação.

O processo de validação estatística foi realizando a partir do teste de Friedman, no qual, em todos os casos, a hipótese nula foi rejeitada. Os resultados do teste para cada cenário podem ser verificados no Apêndice B. Diante disso, foi aplicado o pós-teste de Nemenyi com significância de 0,05, ilustrado na Figura 4.2. Nele, estão ilustrados as diferenças significativas entre os classificadores, de forma que cada linha perpendicular vertical representa um cenário avaliado daquele classificador, enquanto as linhas horizontais representam os grupos estatisticamente semelhantes, de forma que se dois cenários estão na mesma linha horizontal, significa que são estatisticamente semelhantes. Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente. Dessa forma, confirma-se, aqui, os ganhos encontrados no LLAMA 2, de forma que em todos os classificadores esse extrator de características apresentou diferença significativa entre o cenário

com rotulação e o cenário sem rotulação. No caso do DistilBERT houve diferença significativa na Regressão Logística, e no caso do TF-IDF na Regressão Logística, AdaBoost e XGBoost.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.2) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Texto da base Politifact (Acurácia). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Desempenho similar também foi observado com os resultados obtidos com métrica do F1-Score, os quais podem ser observados na Tabela 4.2. Novamente, os melhores resultados por algoritmo foram com a inclusão do rótulo do viés. Na Figura 4.3, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

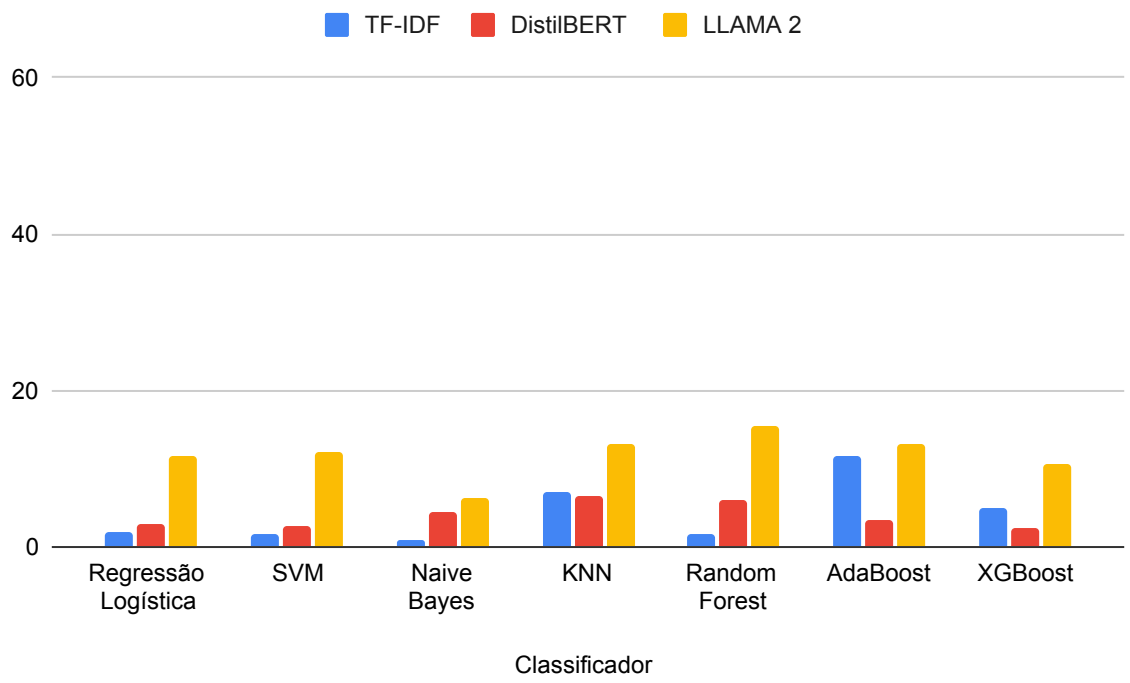


Figura (4.3) Percentuais de Ganhos com a Rotulação na Politifact - Texto (F1-Score)

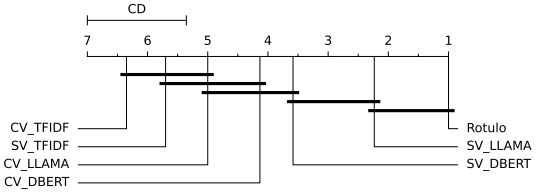
Tabela (4.2) Resultados da F1-Score para o Texto da Politifact. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	0,65 (3,50)	86,47 (3,16)	78,78 (4,93)	72,27 (4,77)	**88,27 (3,51)**	81,69 (4,03)	83,95 (4,23)
SVM	77,90 (5,72)	86,39 (3,75)	79,33 (5,06)	71,67 (5,32)	**87,92 (2,97)**	82,07 (3,79)	83,85 (5,37)
Naive Bayes	49,38 (8,09)	74,71 (7,27)	73,02 (5,19)	72,36 (4,35)	75,48 (5,19)	77,56 (5,14)	**78,69 (5,19)**
KNN	74,75 (7,69)	50,08 (9,30)	66,95 (6,97)	61,74 (6,46)	57,17 (7,10)	73,35 (4,57)	**74,91 (6,28)**
Random Forest	74,97 (6,31)	75,52 (5,18)	72,63 (6,31)	64,74 (6,86)	77,14 (6,37)	78,52 (5,98)	**80,09 (6,06)**
AdaBoost	75,84 (4,61)	74,55 (3,85)	73,02 (5,54)	66,93 (4,42)	**86,11 (3,94)**	76,44 (5,31)	80,00 (5,15)
XGBoost	77,36 (4,43)	76,60 (5,66)	72,25 (6,37)	67,22 (6,80)	**81,68 (5,47)**	74,77 (4,64)	77,86 (5,10)

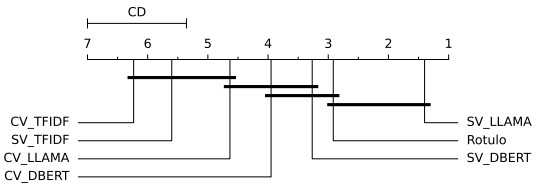
LLAMA 2 continuou apresentando os melhores percentuais de melhora após a rotulação, com o valor menos expressivo sendo de 74,91% (no caso do KNN), enquanto o valor máximo atingindo antes da rotulação foi de 72,36% (no caso do Naive Bayes). Destaca-se também que em todos os casos o LLAMA 2 com texto e rótulo foi melhor que apenas o rótulo. DistilBERT e TF-IDF apresentaram bons resultados com a rotulação, com este último obtendo o maior valor, com 88,27% (no caso da Regressão Logística). Novamente, os algoritmos de Regressão Logística e Naive Bayes apresentaram resultados baixos para o cenário de apenas rótulo, enquanto os demais obtiveram valores melhores.

Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada. Os valores do teste estão descritos no Apêndice B. Na Figura 4.4, estão ilustrados os Diagramas de

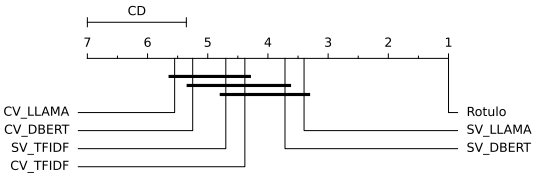
Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. Constata-se que em todos os casos, o cenário do LLAMA 2 com rotulação de viés foi significativamente diferente do cenário sem rotulação, enquanto para o TF-IDF a rotulação apresentou diferença estatística para os casos de Regressão Logísitca, AdaBoost e XGBoost e para DistilBERT houve diferença nos casos de Regresssão Logísitca e Random Forest.



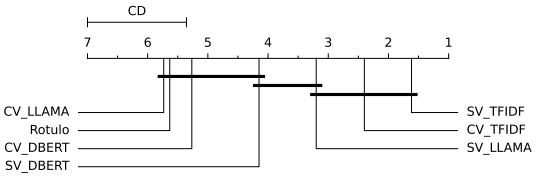
(a) Regressão Logística



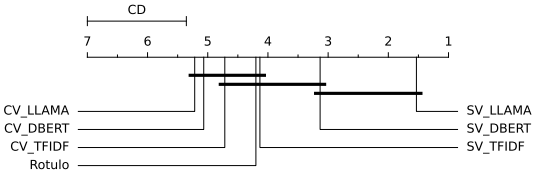
(b) SVM



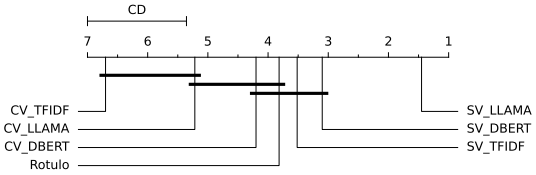
(c) Naive Bayes



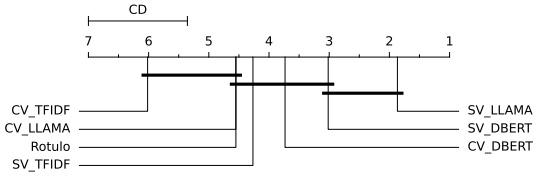
(d) KNN



(e) Random Forest



(f) AdaBoost



(g) XGBoost

Figura (4.4) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Texto da base Politifact (F1-Score). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Tabela (4.3) Resultados da Acurácia para o Título da Politifact. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	55,61 (0,51)	75,95 (4,15)	78,52 (3,64)	84,99 (4,21)	84,47 (3,79)	86,38 (3,10)	**90,90 (3,52)**
SVM	82,14 (4,45)	72,80 (4,80)	81,00 (5,25)	84,14 (4,07)	85,23 (4,01)	86,14 (3,72)	**90,95 (3,44)**
Naive Bayes	66,66 (4,27)	70,71 (4,52)	75,90 (5,99)	81,95 (4,03)	72,71 (4,12)	85,23 (4,21)	**86,04 (3,14)**
KNN	79,33 (5,40)	70,66 (4,60)	76,38 (4,42)	80,71 (4,55)	74,52 (4,36)	85,52 (4,00)	**86,09 (3,51)**
Random Forest	80,61 (4,40)	56,76 (2,08)	78,33 (4,50)	78,85 (3,69)	67,33 (7,18)	**84,95 (4,51)**	84,76 (4,33)
AdaBoost	81,04 (3,47)	67,04 (3,70)	75,47 (4,66)	76,28 (4,45)	81,76 (3,80)	82,42 (3,63)	**85,66 (4,08)**
XGBoost	82,00 (2,98)	65,42 (4,61)	77,95 (3,82)	77,71 (3,12)	80,71 (4,11)	84,76 (3,96)	**87,52 (3,53)**

Passando para a análise do título da FakeNewsNet - Politifact, os resultados da acurácia estão dispostos na Tabela 4.3. Novamente, os melhores resultados por algoritmo estão destacados em negrito, sendo todos eles com a inclusão do viés. Na Figura 4.5, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

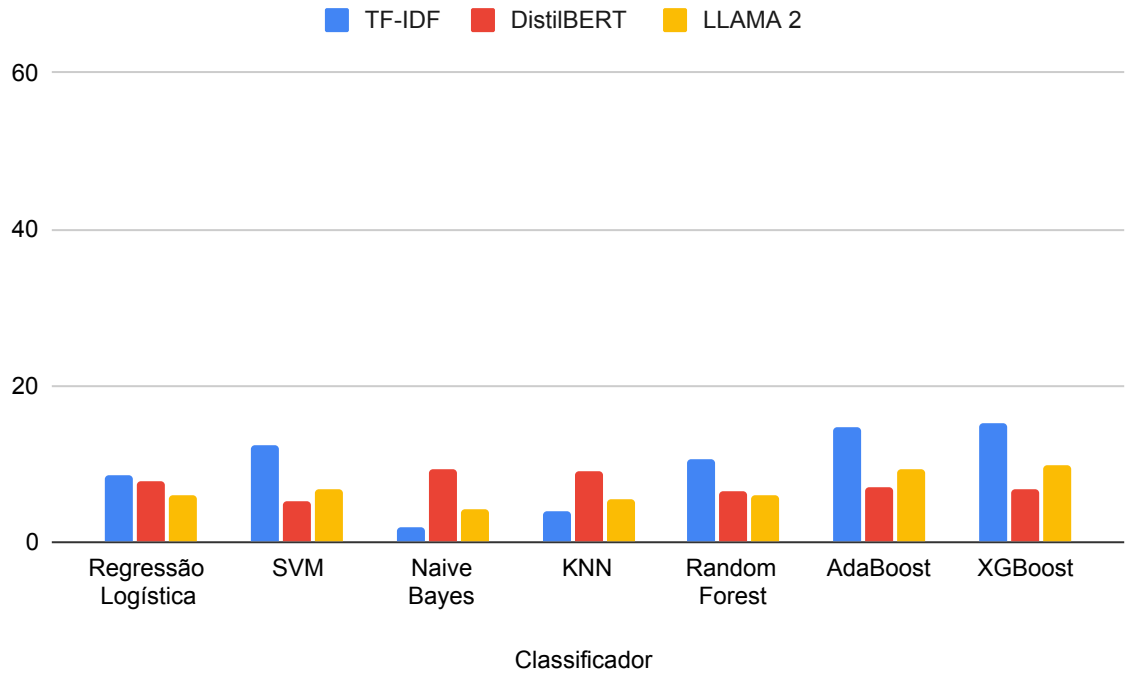
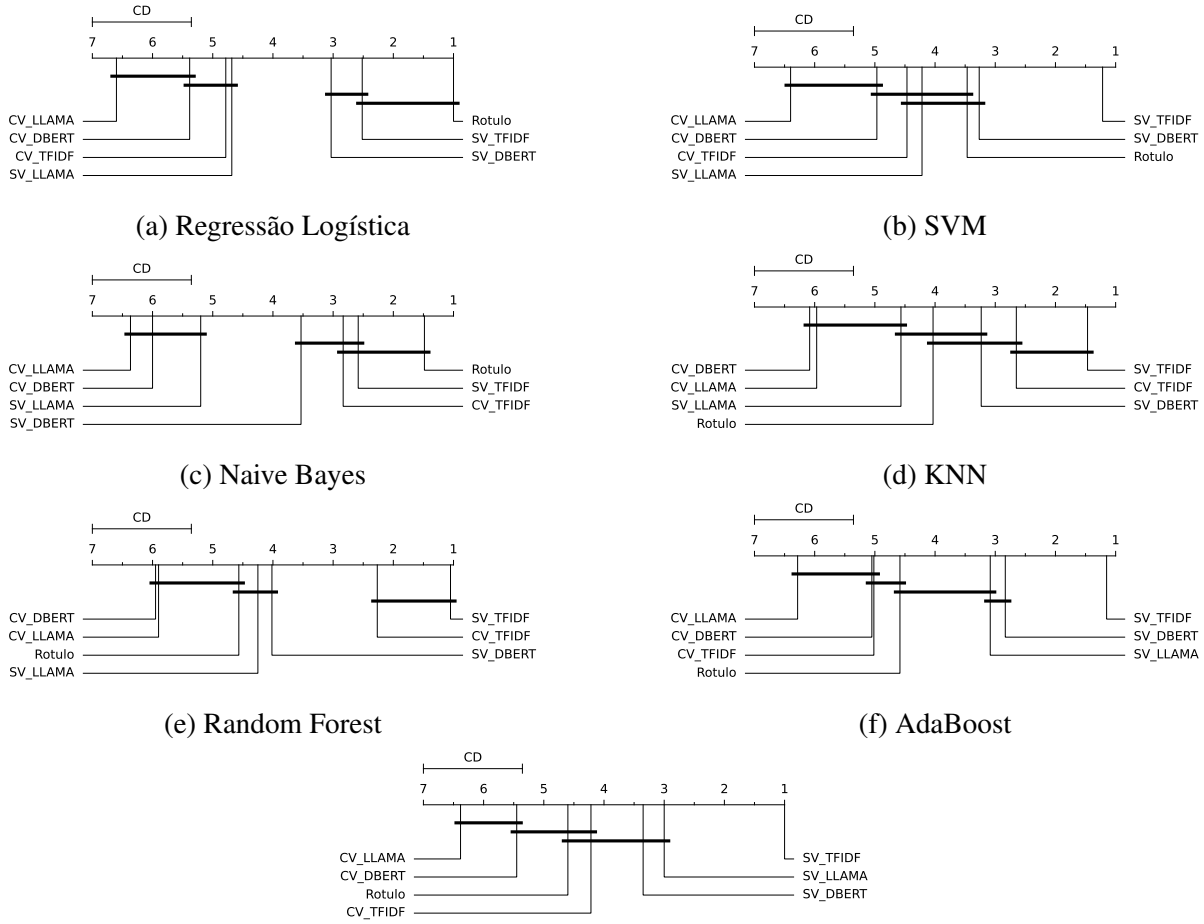


Figura (4.5) Percentuais de Ganhos com a Rotulação na Politifact - Título (Acurácia)

Aqui, o TF-IDF foi o extrator com o percentual de melhora, chegando mais de 15% de ganho com a rotulação. Enquanto isso, o LLAMA 2 obteve os melhores resultado, superando os 90% de acurácia, nos casos de Regressão Logística e SVM. Os testes de validação estatística confirmam os ganhos potenciais. Destaque também que o LLAMA 2 e o DistilBERT obtiveram sempre resultados melhores com a junção de rótulo e texto do que com apenas o rótulo.

Em todos os casos, a hipótese nula foi rejeitada com a aplicação do teste de Friedman (valores disponíveis no Apêndice B). Na Figura 4.6, estão dispostos os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05. Nela, verificamos que o DistilBERT obteve diferença estatística em todos os classificadores, enquanto LLAMA 2 obteve para todos, com exceção do Naive Bayes e KNN, e o TF-IDF obteve diferença significativa na Regressão Logística, SVM, AdaBoost e XGBoost.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.6) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título da base Politifact (Acurácia). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Desempenho similar também foi observado com os resultados obtidos com métrica do F1-Score, os quais podem ser observados na Tabela 4.4. Os melhores resultados por algoritmo foram com a inclusão do rótulo do viés, sendo o LLAMA 2 o melhor extrator de características. Na Figura 4.7, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

Tabela (4.4) Resultados da F1-Score para o Título da Politifact. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	0,65 (3,50)	72,48 (5,37)	74,03 (4,53)	82,78 (5,21)	81,13 (4,90)	83,95 (3,70)	**89,43 (4,14)**
SVM	77,90 (5,72)	69,68 (5,18)	77,63 (6,15)	81,59 (5,01)	82,42 (4,81)	83,51 (4,72)	**89,37 (4,18)**
Naive Bayes	49,38 (8,09)	60,67 (7,82)	72,84 (6,77)	80,63 (4,04)	64,34 (8,19)	83,06 (5,09)	**83,69 (4,03)**
KNN	74,75 (7,69)	63,06 (7,80)	70,26 (5,65)	76,59 (5,66)	64,23 (7,86)	81,77 (5,55)	**82,40 (4,84)**
Random Forest	74,97 (6,31)	8,26 (9,86)	73,16 (6,78)	72,97 (5,70)	41,50 (21,62)	**81,52 (5,98)**	**81,52 (5,59)**
AdaBoost	75,84 (4,61)	68,47 (5,14)	70,74 (6,31)	72,47 (5,68)	78,58 (4,98)	78,93 (4,90)	**83,14 (5,12)**
XGBoost	77,36 (4,43)	64,19 (14,47)	73,78 (5,15)	73,61 (3,61)	76,95 (5,45)	82,30 (4,50)	**85,22 (4,67)**

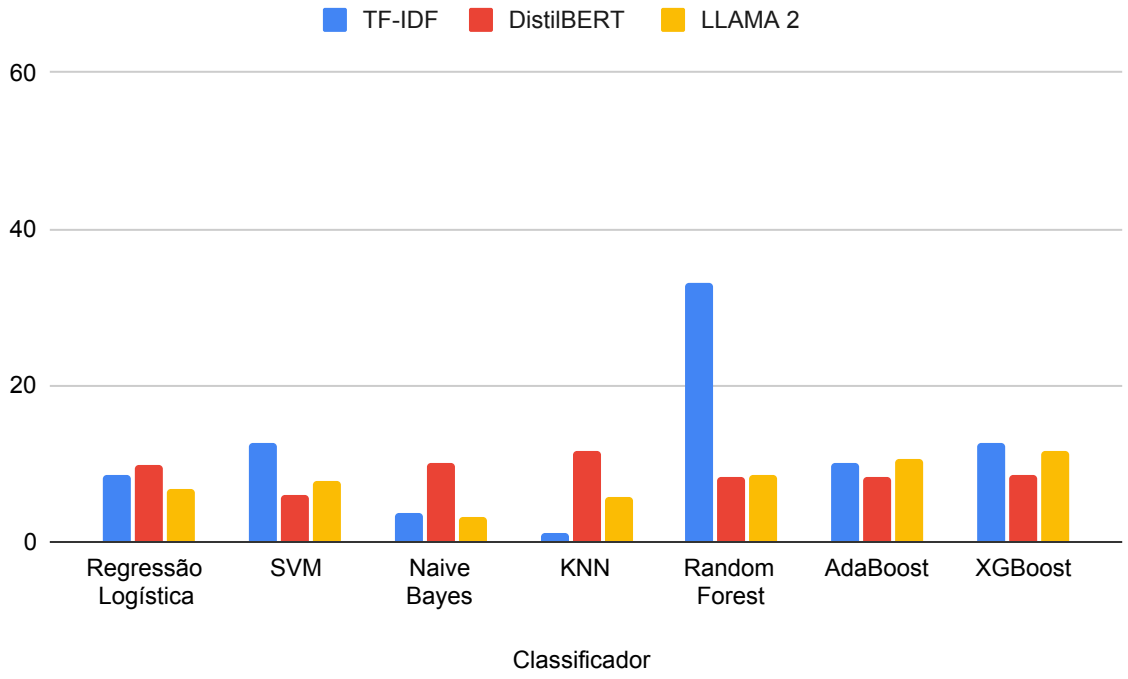
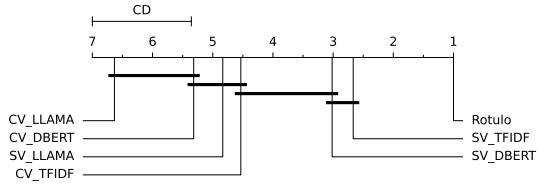


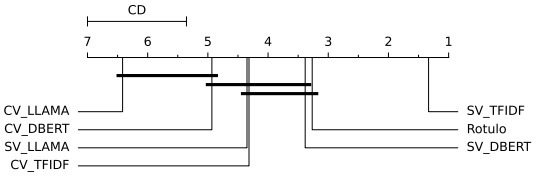
Figura (4.7) Percentuais de Ganhos com a Rotulação na Politifact - Título (F1-Score)

Novamente, todos os resultados do DistilBERT e do LLAMA 2 com o uso do rótulo e do texto foram melhores que os resultados envolvendo apenas o rótulo. O TF-IDF obteve o maior percentual de melhora, com cerca de 33,24% (no caso do Random Forest), enquanto o maior valor obtido foi o LLAMA 2 com 89,43% (no caso da Regressão Logística).

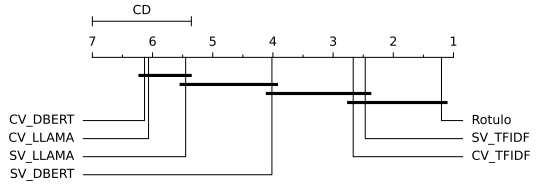
Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada, estando os valores ilustrados no Apêndice B. Na Figura 4.8, estão ilustrados os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. O LLAMA 2 apresentou diferença significativa entre os cenários com e sem rotulação em todos os algoritmos, com exceção de Naive Bayes e KNN, enquanto que, apenas nos casos de SVM e XGBoost, o DistilBERT não apresentou diferença significativa na rotulação. TF-IDF apresentou diferença significativa em Regressão Logística, SVM, AdaBoost e XGBoost.



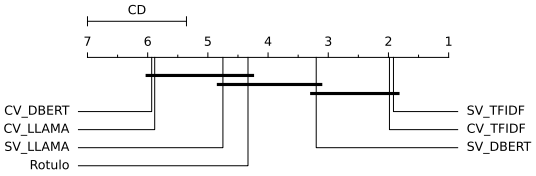
(a) Regressão Logística



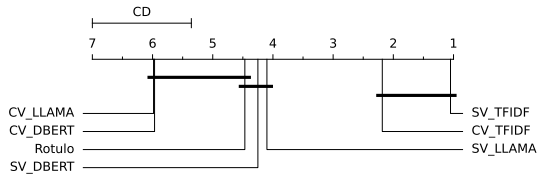
(b) SVM



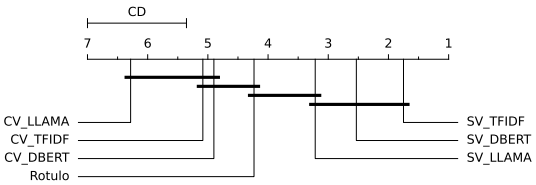
(c) Naive Bayes



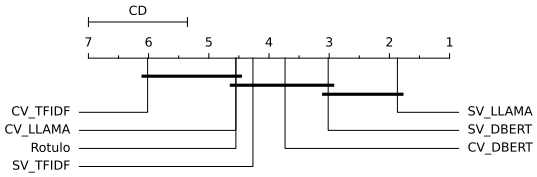
(d) KNN



(e) Random Forest



(f) AdaBoost



(g) XGBoost

Figura (4.8) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título da base Politifact (F1-Score). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Analisando os resultados da acurácia para o vetor de Título + Texto, dispostos na Tabela 4.5, todos os classificadores atingiram o maior valor com a rotulação. Na Figura 4.9, estão dispostos os percentuais de ganhos de cada classificador. Em quase todos, houve variação positiva, com exceção do Naive Bayes, KNN e Random Fores com TF-IDF. Nestes três casos, há uma variação negativa inferior a 1%, a qual não é significativa estaticamente. Destaque para o SVM que obteve o melhor resultado, e o LLAMA 2 que obteve os maiores percentuais de ganhos.

Tabela (4.5) Resultados da Acurácia para o Título + Texto da Politifact. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	55,61 (0,51)	85,85 (5,00)	81,33 (3,41)	86,47 (3,47)	87,76 (4,01)	85,52 (3,41)	**90,00 (3,90)**
SVM	82,14 (4,45)	87,09 (3,22)	81,71 (4,22)	84,52 (3,15)	87,61 (4,27)	84,76 (4,30)	**<u>90,66 (3,23)</u>**
Naive Bayes	66,66 (4,27)	79,28 (4,88)	74,38 (6,86)	83,57 (4,45)	79,04 (4,27)	80,23 (3,15)	**85,33 (4,34)**
KNN	79,33 (5,40)	70,42 (5,32)	75,33 (4,20)	79,23 (4,70)	69,71 (4,41)	80,04 (3,99)	**84,71 (3,17)**
Random Forest	80,61 (4,40)	77,99 (4,46)	76,57 (4,94)	81,00 (3,72)	77,95 (5,44)	81,85 (4,71)	**85,52 (3,20)**
AdaBoost	81,04 (3,47)	76,80 (4,77)	75,42 (4,24)	79,80 (4,16)	86,09 (3,81)	80,95 (3,69)	**88,47 (3,61)**
XGBoost	82,00 (2,98)	77,57 (5,00)	78,38 (4,01)	81,38 (3,78)	83,90 (3,90)	81,47 (4,26)	**87,28 (2,88)**

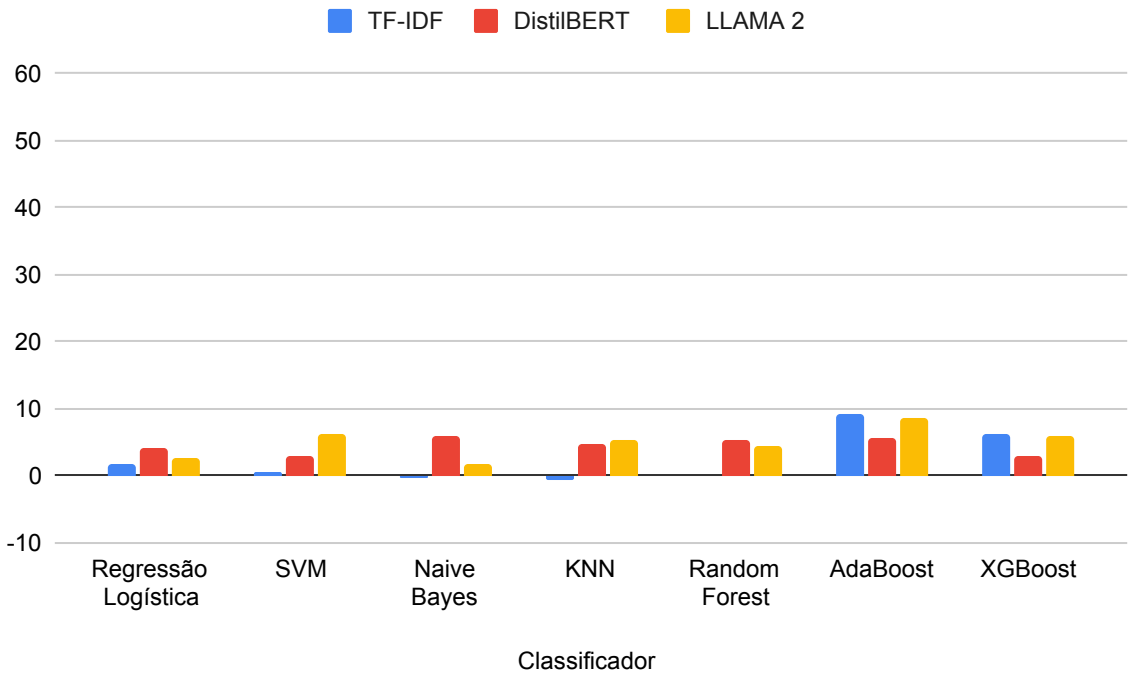
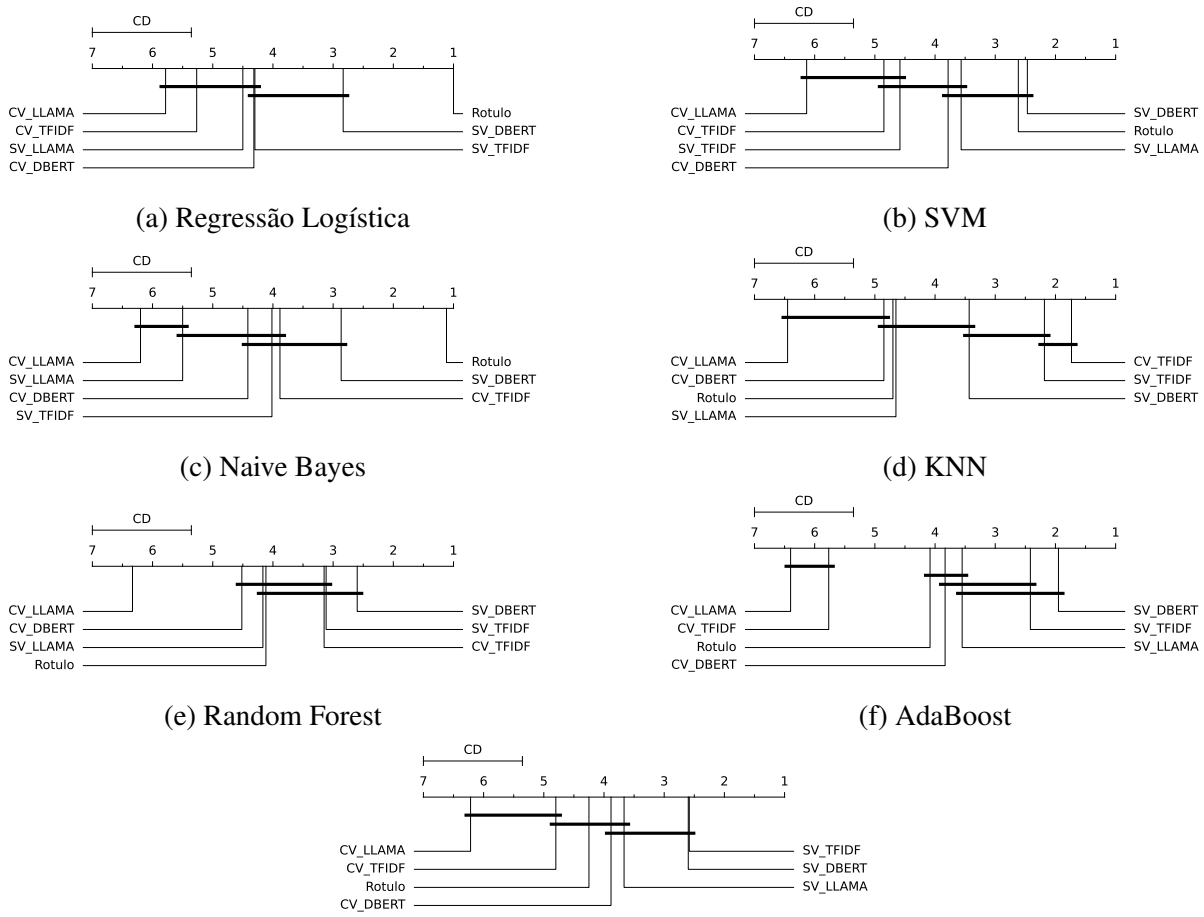


Figura (4.9) Percentuais de Ganhos com a Rotulação na Politifact - Título + Texto (Acurácia)

Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada, estando os valores ilustrados no Apêndice B. Na Figura 4.10, estão ilustrados os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. Com a adoção da rotulação, o LLAMA 2 apresentou diferença estatística para o SVM, KNN, Random Forest, AdaBoost e XGBoost, enquanto o DistilBERT foi diferente nos casos de Random Forest e AdaBoost e o TF-IDF nos casos de AdaBoost e XGBoost.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.10) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título + Texto da base Politifact (Acurácia). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Para a métrica do F1-Score, os resultados estão dispostos na Tabela 4.6. Novamente, todos os melhores valores por classificador foram do LLAMA 2 com a adoção da rotulação, tendo como melhor classificador o SVM. Na Figura 4.11, é possível verificar os ganhos percentuais de cada classificador, sendo todos positivos, com exceção do Naive Bayes e do KNN com TF-IDF. Nestes dois casos, há uma variação negativa inferior a 2%, a qual não é significativa estaticamente.

Tabela (4.6) Resultados do F1-Score para o Texto + Título da Politifact. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	0,65 (3,50)	84,93 (4,97)	78,72 (3,87)	84,54 (3,90)	86,87 (3,91)	83,41 (4,24)	**88,61 (4,43)**
SVM	77,90 (5,72)	86,16 (3,29)	79,45 (4,71)	82,69 (3,47)	86,72 (4,44)	82,61 (5,01)	**89,26 (3,88)**
Naive Bayes	49,38 (8,09)	73,68 (6,93)	70,76 (8,71)	81,76 (4,87)	73,58 (6,12)	77,94 (4,18)	**82,63 (5,32)**
KNN	74,75 (7,69)	53,81 (10,03)	67,41 (5,86)	74,19 (6,61)	52,47 (8,84)	74,75 (5,79)	**80,75 (4,67)**
Random Forest	74,97 (6,31)	77,13 (4,43)	71,96 (6,15)	75,53 (5,73)	77,21 (5,54)	78,82 (5,78)	**82,48 (4,02)**
AdaBoost	75,84 (4,61)	75,05 (4,86)	70,98 (5,03)	76,83 (4,96)	83,74 (4,76)	77,95 (4,20)	**86,59 (4,30)**
XGBoost	77,36 (4,43)	75,03 (6,11)	74,25 (5,41)	78,29 (4,74)	81,16 (4,69)	78,49 (5,23)	**84,89 (3,50)**

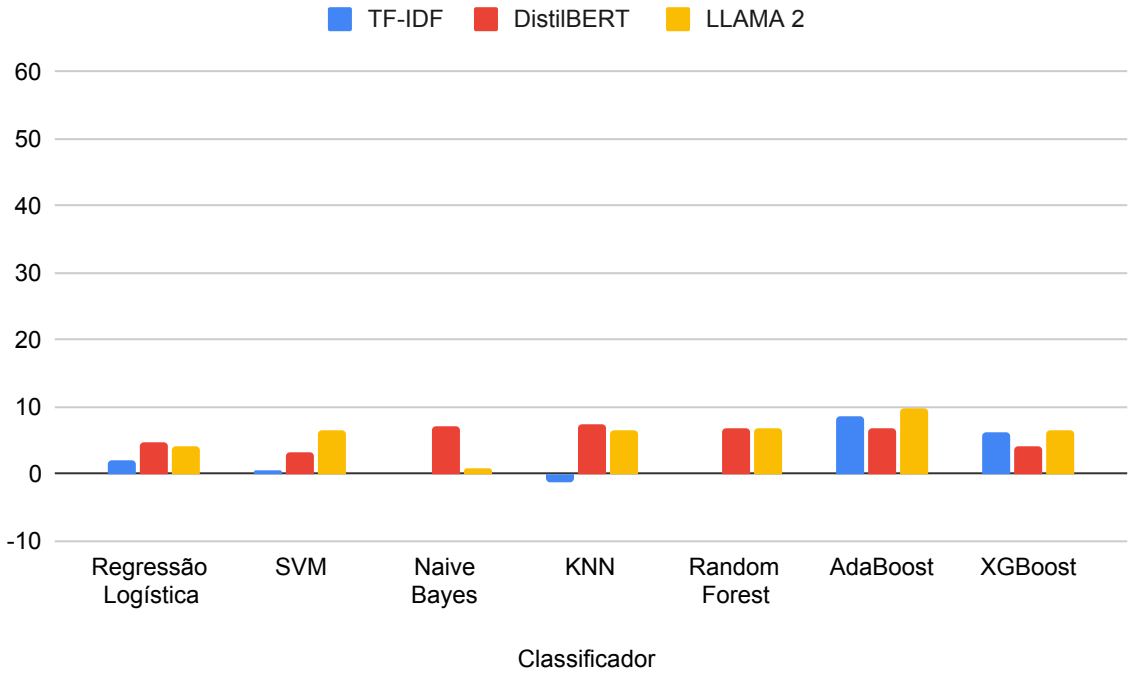
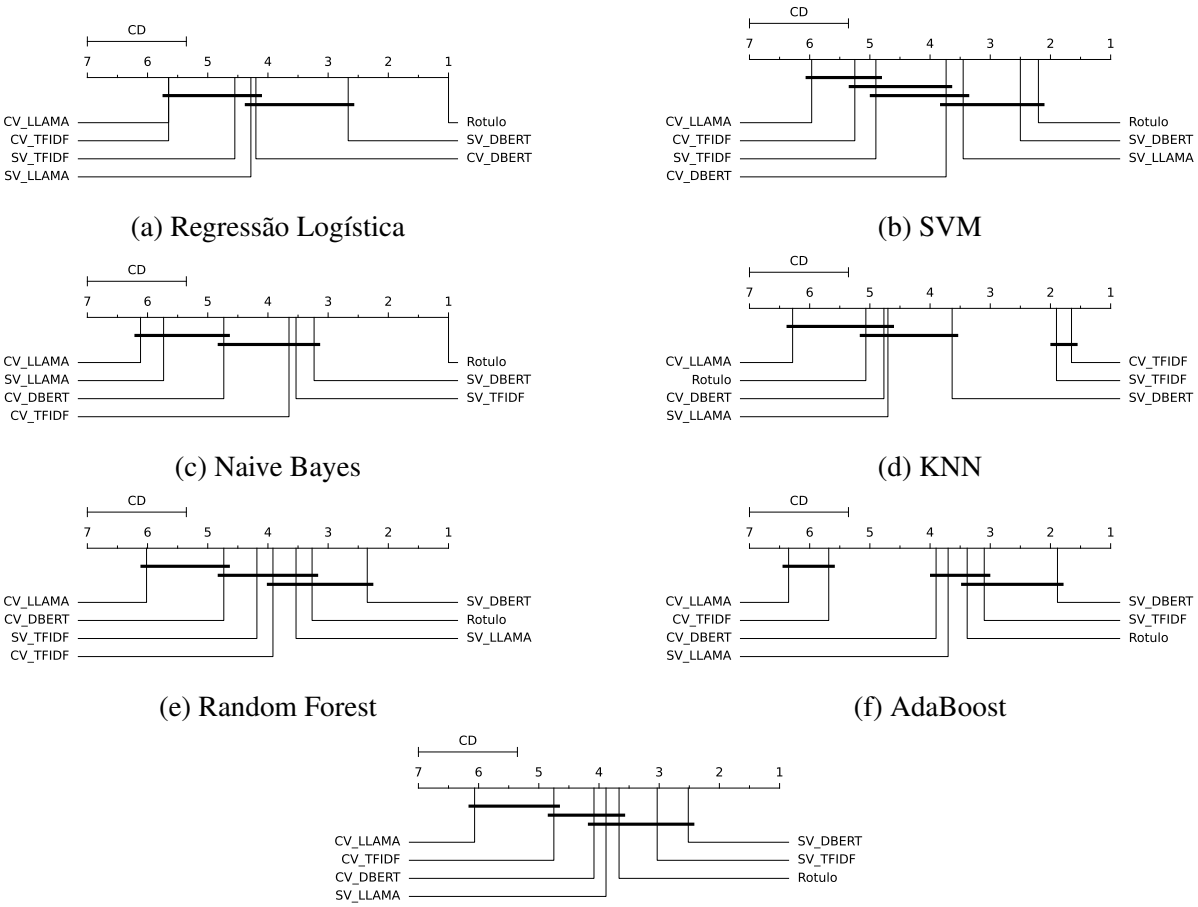


Figura (4.11) Percentuais de Ganhos com a Rotulação na Politifact - Título + Texto (F1-Score)

Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada, estando os valores ilustrados no Apêndice B. Na Figura 4.12, estão ilustrados os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. Com a adoção da rotulação, LLAMA 2 apresentou diferença estatística para o SVM, Random Forest, AdaBoost e XGBoost, enquanto o DistilBERT foi diferente nos casos de Random Forest e AdaBoost e o TF-IDF nos casos de AdaBoost e XGBoost.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.12) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título + Texto da base Politifact (F1). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Constata-se que a rotulação apresentou melhoras percentuais nas duas métricas para todos os extratores de características e todos os algoritmos de aprendizagem de máquina, tanto a nível de texto, quanto a nível de título, para a base FakeNewsNet - Politifact. Tais diferenças se mostraram significativa em grande parte dos cenários avaliados, indicando que é uma abordagem consistente. Destacam-se o TF-IDF e o LLAMA 2 com a rotulação, que obtiveram os melhores resultados para esta base, além do SVM como classificador. No caso do vetor de título + texto, também foram apresentaram bons percentuais de otimização na maioria dos cenários avaliados.

Conforme discutido no Capítulo 3, esta base foi bastante reduzida, estando com 349 instâncias ao final, com cerca de 55% de notícias verdadeiras e 45% de falsas. Dessa forma, os resultados apontam que a metodologia pode ser uma importante alternativa para bases que sofreram grande redução no processo de limpeza ou de bases de menor tamanho. O desbalanceamento da base não parece ter sido determinante para os resultados, tendo em vista que não há uma diferença tão grande entre as classes. No que tange os rótulos, a concentração de notícias verdadeiras no rótulo *left-center* e de notícias falsas no rótulo *fake-news* pode ter contribuído para a classificação, mas pode não ser determinante, conforme será discutido na Seção 4.2.

4.2 KaggleFN

Tabela (4.7) Resultados da Acurácia para o Texto da KaggleFN. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	65,13 (1,65)	74,07 (2,18)	69,73 (1,78)	69,58 (1,78)	**96,29 (0,93)**	88,59 (1,36)	93,82 (0,94)
SVM	93,01 (1,12)	75,15 (2,28)	72,53 (1,53)	70,96 (2,07)	85,29 (1,68)	89,59 (1,43)	**93,59 (1,38)**
Naive Bayes	74,31 (1,57)	67,23 (2,01)	61,17 (1,24)	64,10 (2,37)	68,45 (1,98)	74,50 (2,32)	**92,96 (1,15)**
KNN	92,81 (3,95)	42,73 (1,30)	67,20 (2,36)	68,10 (2,22)	70,25 (1,78)	75,89 (1,66)	**93,40 (1,01)**
Random Forest	93,11 (1,23)	60,14 (0,46)	67,15 (1,54)	69,34 (1,55)	60,32 (0,50)	77,78 (1,90)	**93,68 (1,12)**
AdaBoost	93,01 (1,46)	71,89 (2,09)	65,42 (2,44)	67,42 (1,92)	**96,81 (0,77)**	81,99 (1,94)	93,50 (1,37)
XGBoost	93,33 (1,28)	75,20 (2,46)	70,08 (2,15)	70,16 (2,41)	**97,10 (0,83)**	83,51 (2,28)	93,27 (0,97)

Na Tabela 4.7 estão dispostos os resultados da acurácia aplicados ao texto da KaggleFN, bem como também ao rótulo. Os melhores resultados por algoritmo estão destacados em negrito, sendo todos eles com a inclusão do viés. Na Figura 4.13, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

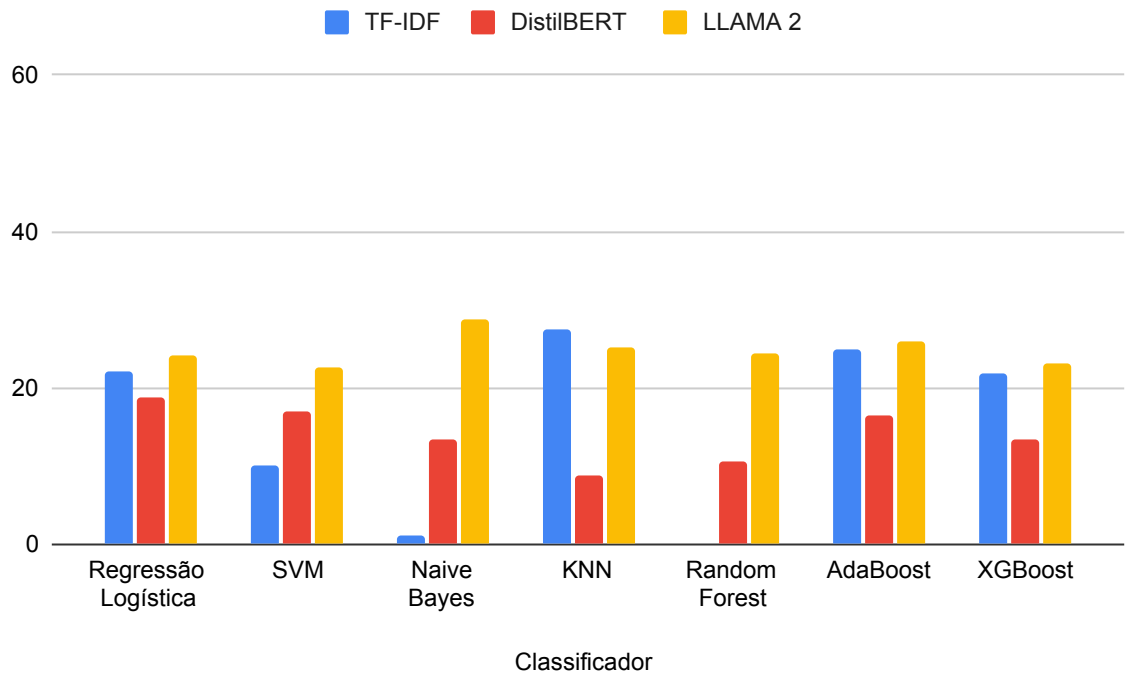
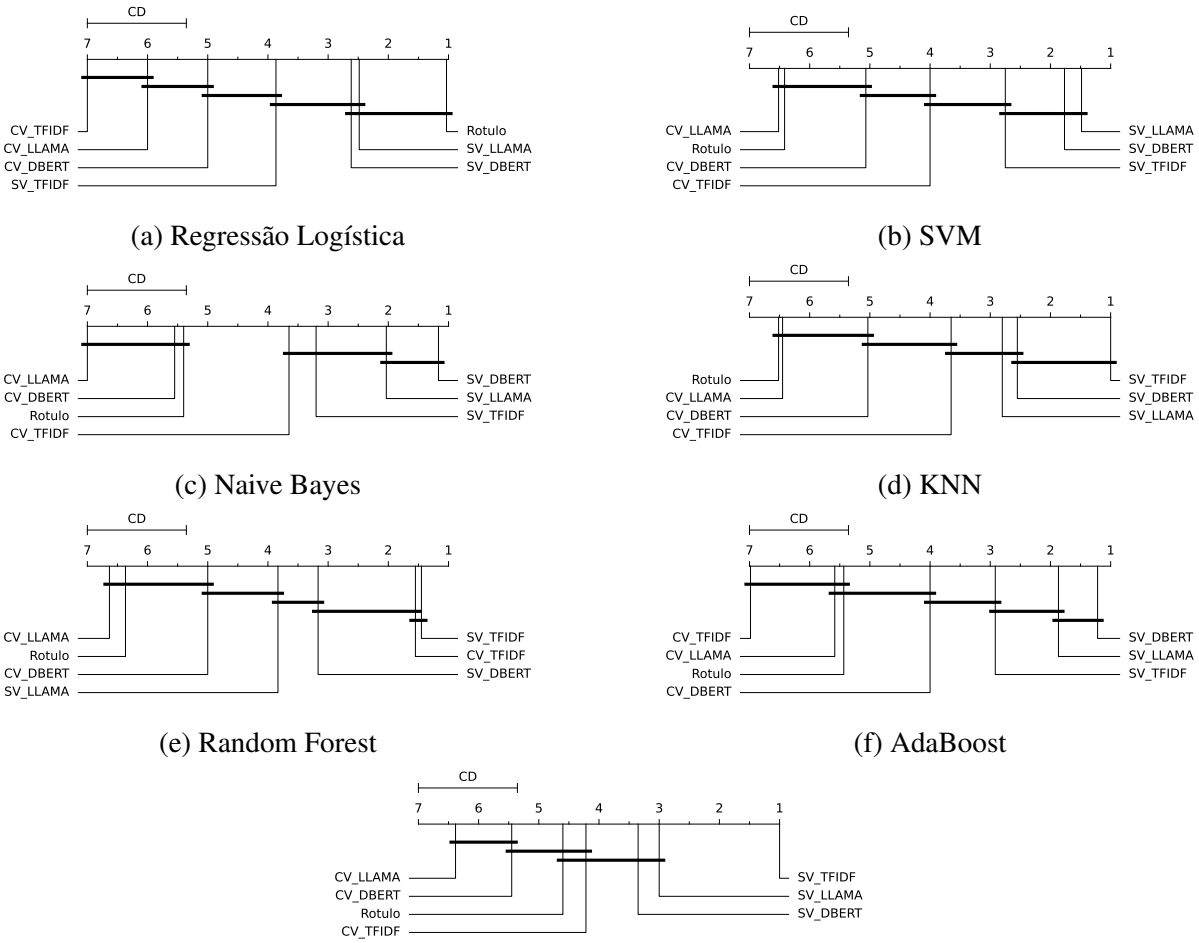


Figura (4.13) Percentuais de Ganhos com a Rotulação na KaggleFN - Texto (Acurácia)

LLAMA 2 obteve grandes percentuais de melhora, de forma que o resultado menos expressivo com a rotulação, de 92,96% (no caso do Naive Bayes), foi muito melhor que o resultado

mais expressivo sem a rotulação, de 70,96% (no caso do do SVM). Isso também aconteceu com o DistilBERT, tendo como o menor valor com a rotulação 74,50% (no caso do Naive Bayes) e como maior valor sem a rotulação 72,53% (no caso do SVM). Por sua vez, o TF-IDF com a rotulação obteve o melhor resultado, de 97,10% (caso do XGBoost). Apesar dos resultados com Regressão Logística e Naive Bayes, o cenário com apenas o rótulo obteve valores altos, indicando o poder de classificação da variável do viés.

No Apêndice B, estão dispostos os valores obtidos no teste de Friedman, nos quais, em todos os casos, a hipótese nula foi rejeitada. Na Figura 4.14, estão dispostos os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05. Nela, é possível identificar que o DistilBERT e LLAMA 2 obtiveram diferença significativa entre o cenário com e sem rotulação em todos os algoritmos, enquanto o TF-IDF obteve diferença significativa nos algoritmos de Regressão Logística, AdaBoost e XGBoost.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.14) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Texto da base KaggleFN (Acurácia). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Desempenho similar também foi observado com os resultados obtidos com métrica do F1-Score, os quais podem ser observados na Tabela 4.8. Novamente, os melhores resultados por algoritmo foram com a inclusão do rótulo do viés. Na Figura 4.15, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

Tabela (4.8) Resultados da F1-Score para o Texto da KaggleFN. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	73,98 (1,16)	79,37 (1,77)	76,14 (1,57)	74,97 (1,59)	**95,50 (1,10)**	90,52 (1,11)	94,60 (0,87)
SVM	93,84 (1,05)	80,34 (1,81)	78,27 (1,31)	77,44 (1,58)	80,65 (2,46)	91,18 (1,22)	**94,38 (1,28)**
Naive Bayes	81,99 (0,94)	72,32 (1,98)	72,80 (1,79)	68,73 (2,46)	73,34 (2,40)	80,01 (1,82)	**93,80 (1,05)**
KNN	93,85 (2,67)	9,90 (3,99)	74,95 (2,03)	72,88 (2,08)	60,62 (3,10)	81,84 (1,20)	**94,23 (0,91)**
Random Forest	93,94 (1,13)	74,99 (0,28)	77,85 (0,91)	78,59 (0,97)	75,08 (0,31)	83,55 (1,27)	**94,47 (1,03)**
AdaBoost	93,83 (1,38)	80,05 (1,50)	74,09 (3,11)	74,27 (1,75)	**97,29 (0,67)**	85,09 (1,62)	94,31 (1,26)
XGBoost	92,29 (1,37)	63,04 (4,18)	55,84 (4,00)	57,38 (3,81)	**96,47 (0,99)**	78,18 (3,14)	92,16 (1,07)

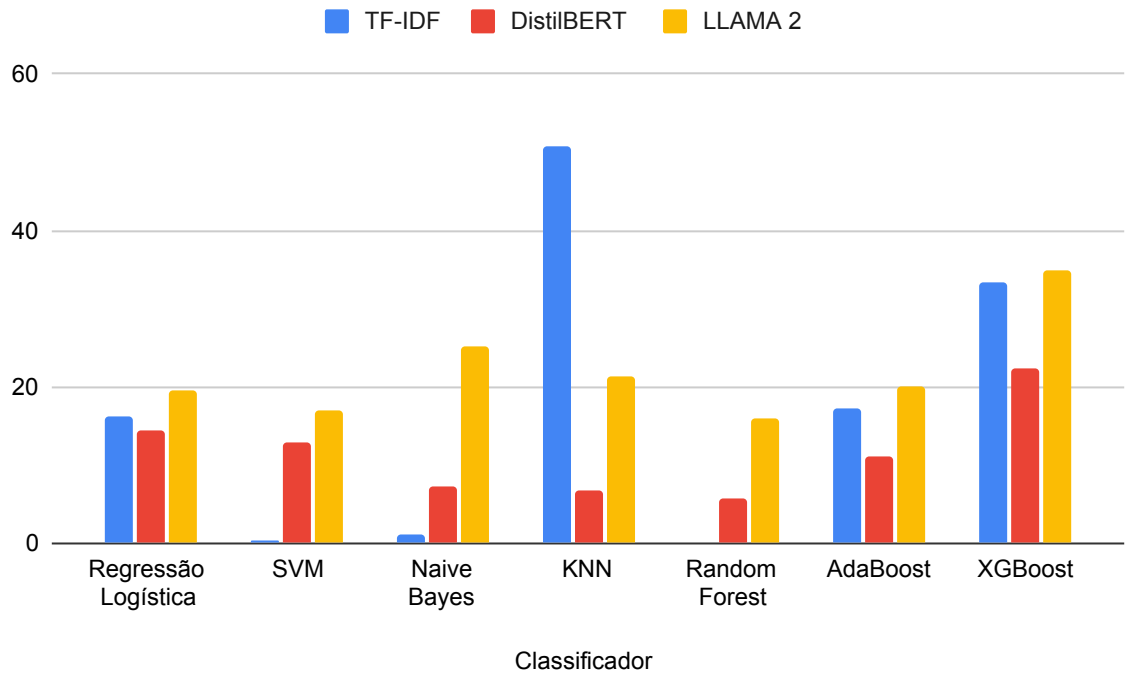
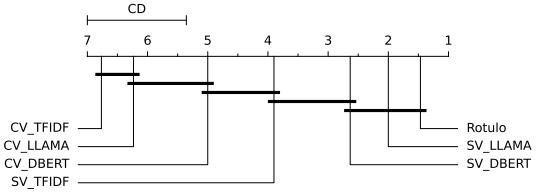


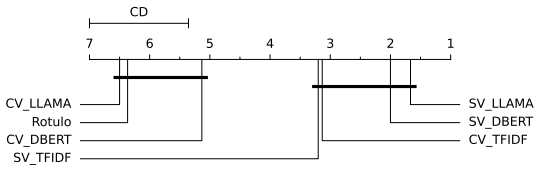
Figura (4.15) Percentuais de Ganhos com a Rotulação na KaggleFN - Texto (F1-Score)

Os ganhos percentuais chegaram a cerca de 51%, obtendo resultados expressivos. Novamente, o resultado do LLAMA 2 menos expressivo com a rotulação, de 92,16% (no caso do XGBoost), foi melhor que o mais expressivo sem a rotulação, de 78,59% (no caso do Random Forest). DistilBERT e TF-IDF também apresentaram resultados substanciais, com este último

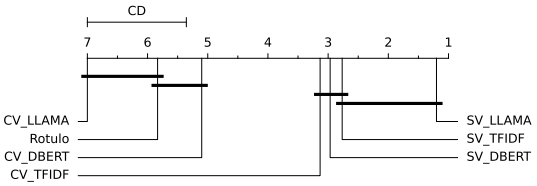
atingindo 97,29%. Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada (valores disponíveis no Apêndice B). Na Figura 4.16, estão ilustrados os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. Novamente, maior parte dos cenários houve diferença significativa entre a rotulação e a não rotulação, com destaque para o LLAMA 2, que em todos casos obteve tal diferença.



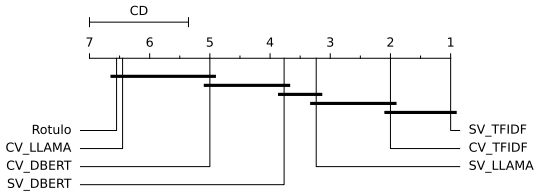
(a) Regressão Logística



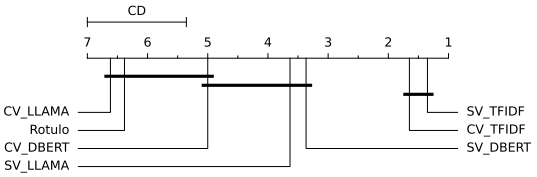
(b) SVM



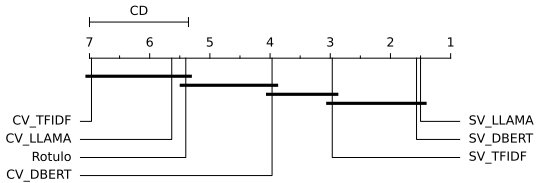
(c) Naive Bayes



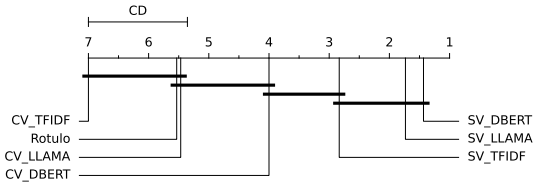
(d) KNN



(e) Random Forest



(f) AdaBoost



(g) XGBoost

Figura (4.16) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Texto da base KaggleFN (F1-Score). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Tabela (4.9) Resultados da Acurácia para o Título da KaggleFN. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	65,13 (1,65)	72,16 (2,42)	66,74 (2,24)	66,98 (1,74)	**93,15 (1,32)**	92,39 (1,32)	93,09 (1,41)
SVM	93,01 (1,12)	71,50 (1,87)	71,22 (2,12)	68,27 (1,78)	93,37 (1,13)	92,62 (1,03)	**93,45 (1,19)**
Naive Bayes	74,31 (1,57)	69,19 (1,91)	59,75 (2,24)	62,16 (2,54)	87,02 (2,52)	83,89 (1,68)	**92,16 (1,28)**
KNN	92,81 (3,95)	65,86 (1,87)	64,32 (2,71)	64,21 (2,48)	80,38 (2,06)	88,35 (1,57)	**93,49 (1,12)**
Random Forest	93,11 (1,23)	60,19 (0,28)	68,33 (1,88)	67,54 (1,57)	60,67 (1,24)	89,14 (1,70)	**93,35 (1,19)**
AdaBoost	93,01 (1,46)	68,66 (2,22)	63,16 (1,93)	64,87 (1,80)	93,37 (1,12)	88,14 (1,53)	**93,46 (1,14)**
XGBoost	93,33 (1,28)	69,74 (1,65)	68,11 (2,02)	67,60 (2,38)	**93,74 (1,13)**	90,58 (1,39)	93,17 (1,08)

Passando para a análise do Título da KaggleFN, na Tabela 4.9, estão dispostos os resultados da acurácia. Os melhores resultados por algoritmo estão destacados em negrito, sendo todos eles com a inclusão do viés. Na Figura 4.17, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

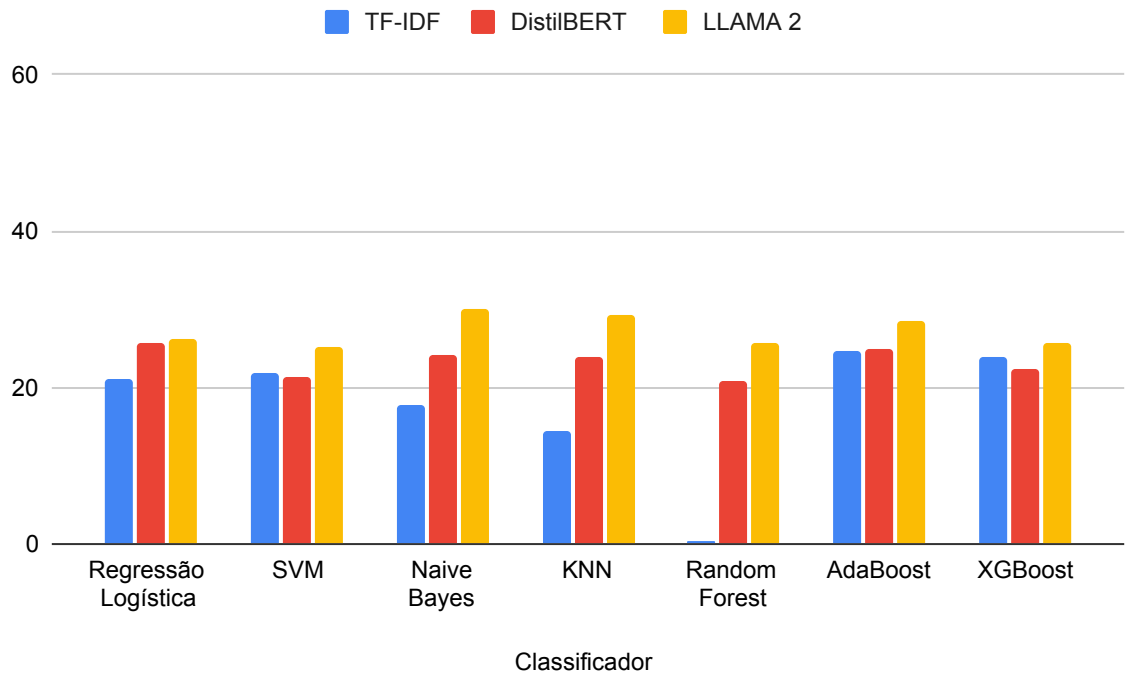
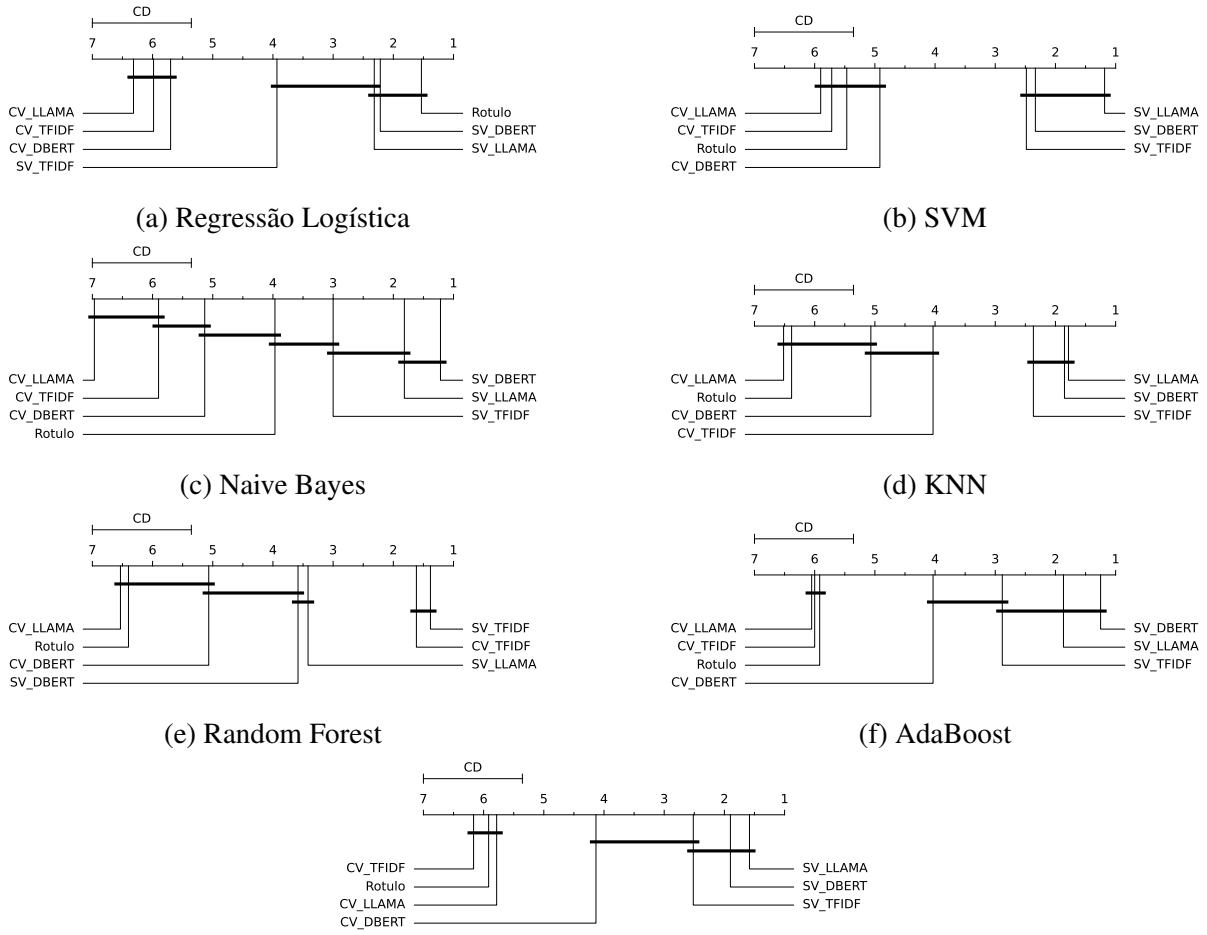


Figura (4.17) Percentuais de Ganhos com a Rotulação na KaggleFN - Título (Acurácia)

Tanto DistilBERT, quanto LLAMA 2, obtiveram ganhos altos de desempenhos após a rotulação, de forma que os valores menos expressivos do cenário com rotulação, de 83,89% e 92,16% (ambos no caso do Naive Bayes) respectivamente, foram melhores que o valores mais expressivos sem a rotulação, de 71,22% e 68,27% (ambos no caso do SVM). O TF-IDF também apresentou bons resultados, obtendo 93,74% (no caso do AdaBoost). Os resultados do cenário com apenas rótulo também apresentaram altos valores.

No Apêndice B, estão dispostos os valores obtidos no teste de Friedman, nos quais, em todos os casos, a hipótese nula foi rejeitada. Na Figura 4.6, estão dispostos os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05. LLAMA 2 apresentou diferença significativa dentre o cenário com e sem rotulação em todos os algoritmos de classificação, enquanto o DistilBERT e o TF-IDF apresentaram diferença em todos os algoritmos, com exceção do Random Forest.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.18) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título da base KaggleFN (Acurácia). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Desempenho similar também foi observado com os resultados obtidos com métrica do F1-Score, os quais podem ser observados na Tabela 4.10. Novamente, os melhores resultados por algoritmo foram com a inclusão do rótulo do viés. Na Figura 4.19, estão dispostos os percentuais de ganho com a rotulação para cada classificador, em que é perceptível uma melhora em todos os cenários.

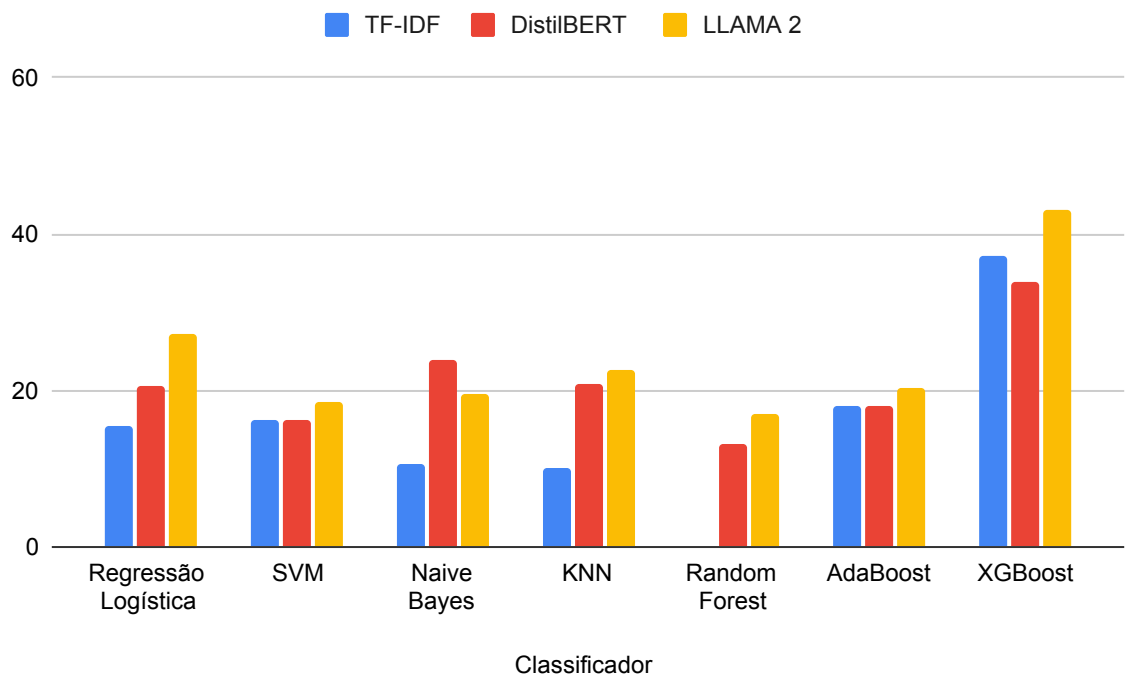


Figura (4.19) Percentuais de Ganhos com a Rotulação na KaggleFN - Título (F1-Score)

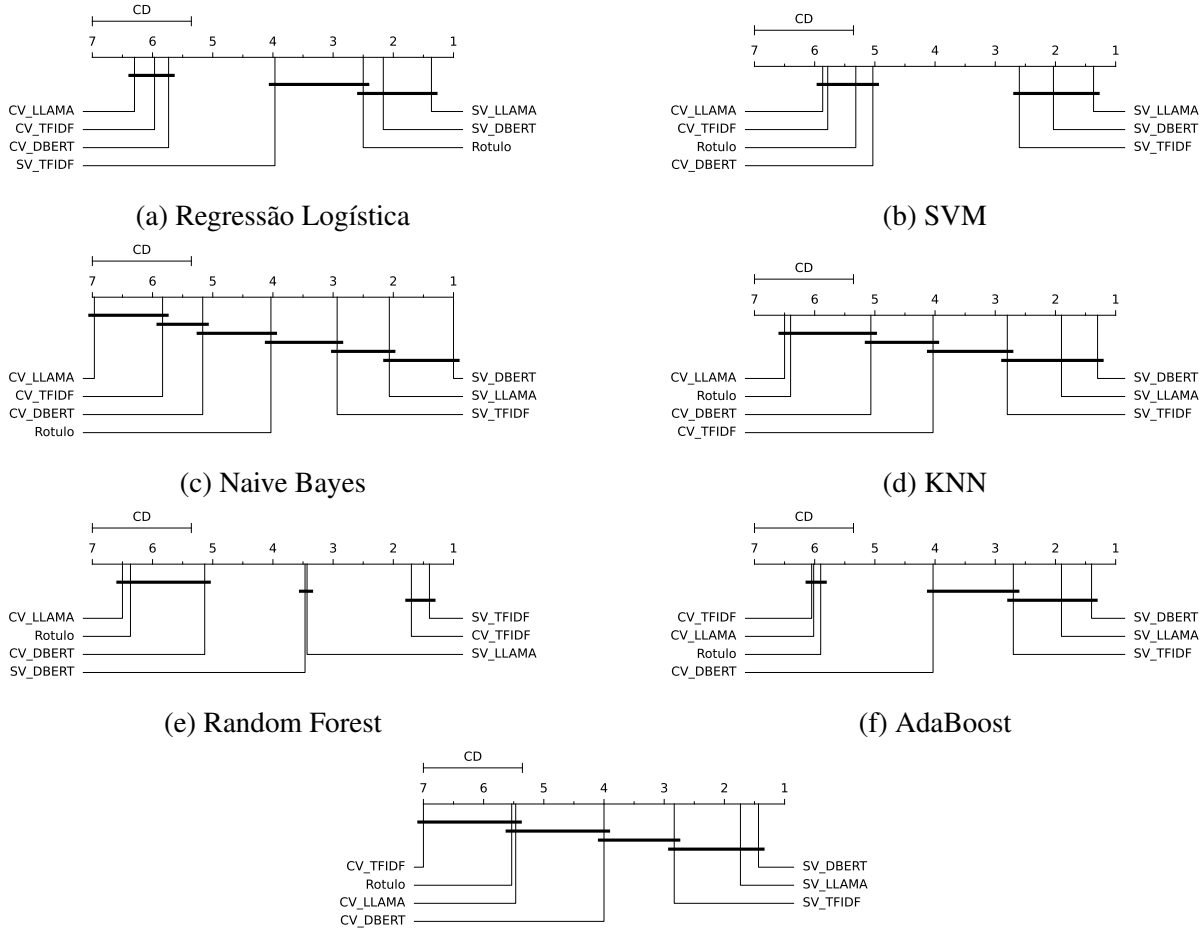
Tabela (4.10) Resultados da F1-Score para o Título da KaggleFN. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	73,98 (1,16)	78,54 (1,90)	72,96 (1,99)	66,79 (7,57)	**94,06 (1,17)**	93,51 (1,13)	93,94 (1,31)
SVM	93,84 (1,05)	78,09 (1,89)	77,27 (1,68)	75,71 (1,78)	94,22 (1,04)	93,60 (0,95)	**94,25 (1,10)**
Naive Bayes	81,99 (0,94)	77,38 (2,68)	62,11 (2,77)	73,38 (1,89)	87,88 (2,65)	85,99 (1,59)	**93,04 (1,21)**
KNN	93,85 (2,67)	74,14 (1,46)	69,55 (2,51)	71,62 (1,91)	84,30 (1,59)	90,42 (1,30)	**94,30 (1,04)**
Random Forest	93,94 (1,13)	75,09 (0,14)	77,64 (1,40)	77,24 (1,14)	75,31 (0,58)	90,93 (1,40)	**94,16 (1,09)**
AdaBoost	93,83 (1,38)	76,16 (1,79)	72,05 (2,77)	74,01 (2,75)	94,23 (1,03)	90,11 (1,27)	**94,30 (1,04)**
XGBoost	92,29 (1,37)	55,38 (3,04)	54,54 (4,04)	48,98 (5,39)	**92,56 (1,28)**	88,52 (1,73)	92,04 (1,17)

Tanto DistilBERT, quanto LLAMA 2, obtiveram ganhos altos de desempenhos após a rotulação, de forma que os valores menos expressivos do cenário com rotulação, de 83,89% (no caso do Naive Bayes) e 92,04% (no caso do XGBoost) respectivamente, foram melhores que o valores mais expressivos sem a rotulação, de 77,64% e 77,24% (ambos no caso do Random Forest). O TF-IDF também apresentou bons resultados, obtendo 94,23% (no caso do AdaBoost). Os resultados do cenário com apenas rótulo também apresentaram altos valores.

Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada (valores disponíveis no Apêndice B). Na Figura 4.20, estão ilustrados os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. Todos os cenários do LLAMA 2 e do DistilBERT apresentaram diferenças significativas entre

os cenários com e sem rotulação e o TF-IDF apenas os casos de KNN e Random Forest não apresentaram diferença significativa.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.20) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título da base KaggleFN (F1-Score). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Analisando os resultados da acurácia para o vetor de Título + Texto para a KaggleFN, dispostos na Tabela 4.11, cinco dos sete classificadores avaliados obtiveram os melhores valores a partir da rotulação, sendo eles Regressão Logística, Naive Bayes, Random Forest, AdaBoost e XGBoost. Na Figura 4.21, estão dispostos os ganhos percentuais da rotulação, sendo quase todos positivos, com exceção do Naive Bayes e do Random Forest com o TF-IDF. Nestes dois casos, há uma variação negativa inferior à 1%, a qual não é significativa estatisticamente.

Tabela (4.11) Resultados da Acurácia para Título + Texto da KaggleFN. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	65,13 (1,65)	78,13 (1,83)	70,01 (1,83)	67,23 (2,07)	**95,95 (0,91)**	89,81 (1,80)	93,13 (0,98)
SVM	**93,01 (1,12)**	77,39 (1,52)	73,60 (2,50)	68,57 (1,51)	86,22 (1,43)	90,50 (1,43)	92,81 (1,02)
Naive Bayes	74,31 (1,57)	72,36 (1,80)	62,17 (1,55)	$60{,}05\ (2{,}22x10^{-16})$	71,86 (1,80)	76,08 (1,77)	**91,65 (1,02)**
KNN	**92,81 (3,95)**	70,86 (1,97)	67,48 (2,13)	63,19 (2,22)	71,78 (3,00)	77,95 (1,87)	92,56 (3,69)
Random Forest	93,11 (1,23)	60,17 (0,46)	66,28 (1,29)	66,71 (1,63)	60,07 (0,42)	78,39 (2,22)	**93,62 (1,03)**
AdaBoost	93,01 (1,46)	74,30 (2,00)	64,24 (2,10)	64,39 (2,24)	**96,84 (0,89)**	82,10 (1,56)	93,28 (1,36)
XGBoost	93,33 (1,28)	78,99 (1,89)	70,10 (1,62)	67,06 (2,57)	**96,84 (0,72)**	82,80 (1,93)	93,20 (1,12)

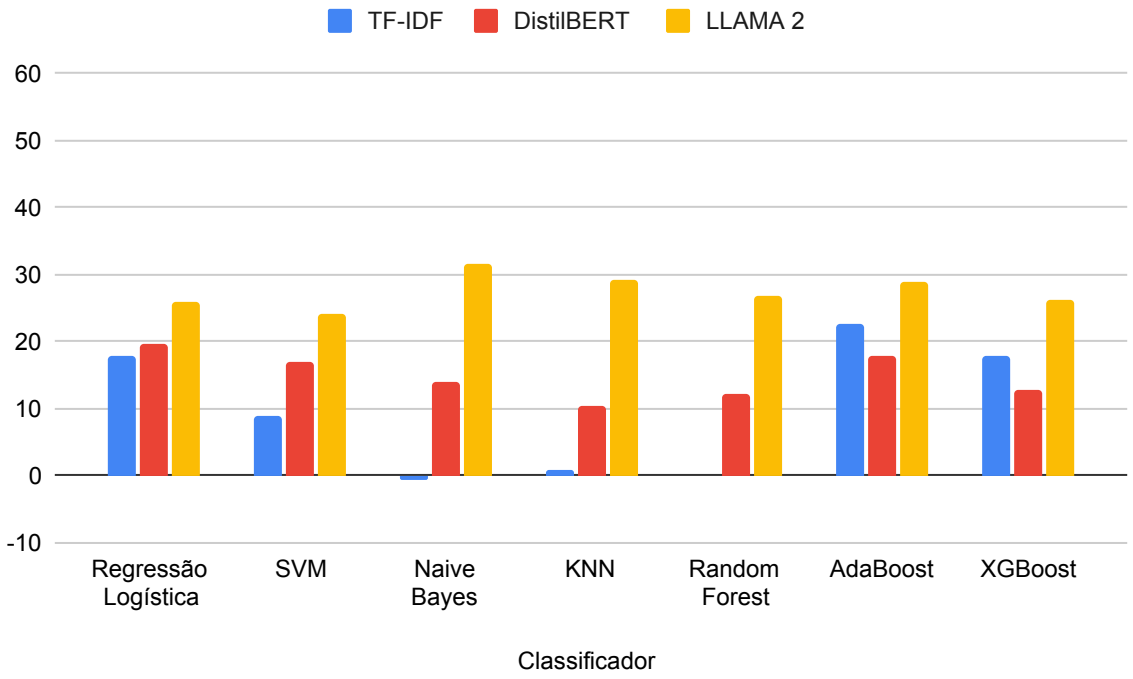
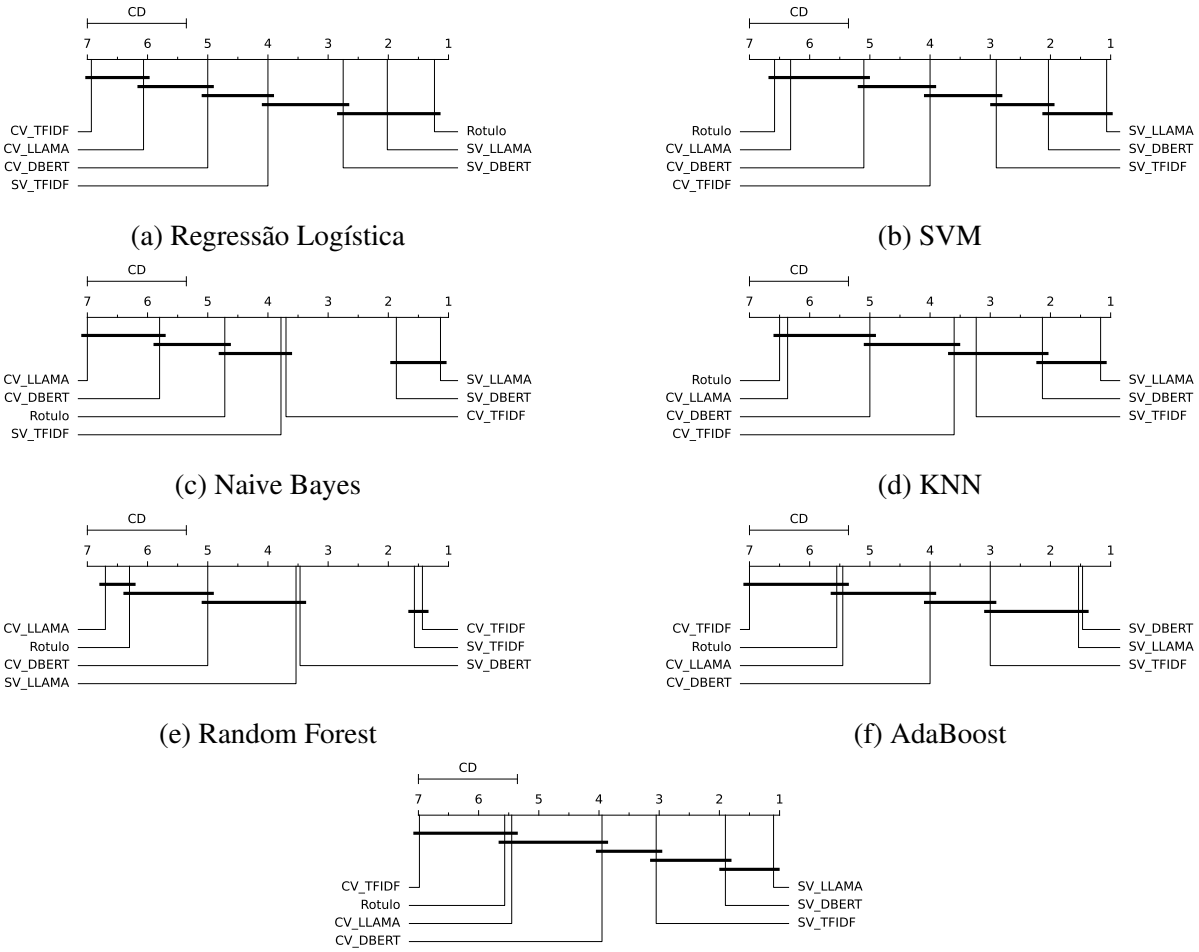


Figura (4.21) Percentuais de Ganhos com a Rotulação na KaggleFN - Título + Texto (Acurácia)

Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada (valores disponíveis no Apêndice B). Na Figura 4.22, estão ilustrados os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. Todos os cenários do LLAMA 2 apresentaram diferenças significativas entre os cenários com e sem rotulação, o DistilBERT apenas o caso de Random Forest não apresentou diferença significativa e o TF-IDF nos casos de SVM, Naive Bayes, KNN e Random Forest não apresentaram diferença significativa.



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.22) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título + Texto da base KaggleFN (Acurácia). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

O mesmo padrão ocorre nos resultados do F1-Score para o vetor de Título + Texto para a KaggleFN, dispostos na Tabela 4.12. Cinco dos sete classificadores avaliados obtiveram os melhores valores a partir da rotulação, sendo eles Regressão Logística, Naive Bayes, Random Forest, AdaBoost e XGBoost. Na Figura 4.24, estão dispostos os ganhos percentuais da rotulação, sendo quase todos positivos, com exceção do Naive Bayes e do Random Forest com o TF-IDF. Nestes dois casos, há uma variação negativa inferior à 1%, a qual não é significativa estatisticamente.

Tabela (4.12) Resultados do F1-Score para Título + Texto da KaggleFN. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	73,98 (1,16)	82,95 (1,35)	76,10 (1,62)	73,19 (1,91)	**96,56 (0,80)**	91,51 (1,52)	93,99 (0,91)
SVM	**93,84 (1,05)**	82,22 (1,26)	79,25 (2,08)	75,67 (1,03)	88,78 (1,21)	92,05 (1,20)	93,67 (0,94)
Naive Bayes	81,99 (0,94)	76,41 (1,97)	72,85 (2,38)	$75{,}04\ (1,11x10^{-16})$	76,18 (1,64)	80,90 (1,31)	**92,59 (0,97)**
KNN	**93,85 (2,67)**	76,46 (1,61)	74,16 (1,66)	70,88 (1,97)	77,10 (2,60)	83,35 (1,32)	93,61 (2,49)
Random Forest	93,94 (1,13)	75,01 (0,28)	77,37 (0,78)	77,35 (1,12)	74,93 (0,26)	84,03 (1,54)	**94,41 (0,96)**
AdaBoost	93,83 (1,38)	81,18 (1,87)	73,98 (2,29)	74,78 (1,92)	**97,31 (0,77)**	85,25 (1,36)	94,12 (1,23)
XGBoost	92,29 (1,37)	69,60 (2,98)	55,02 (2,37)	50,45 (7,06)	**96,16 (0,85)**	77,36 (2,89)	92,09 (1,21)

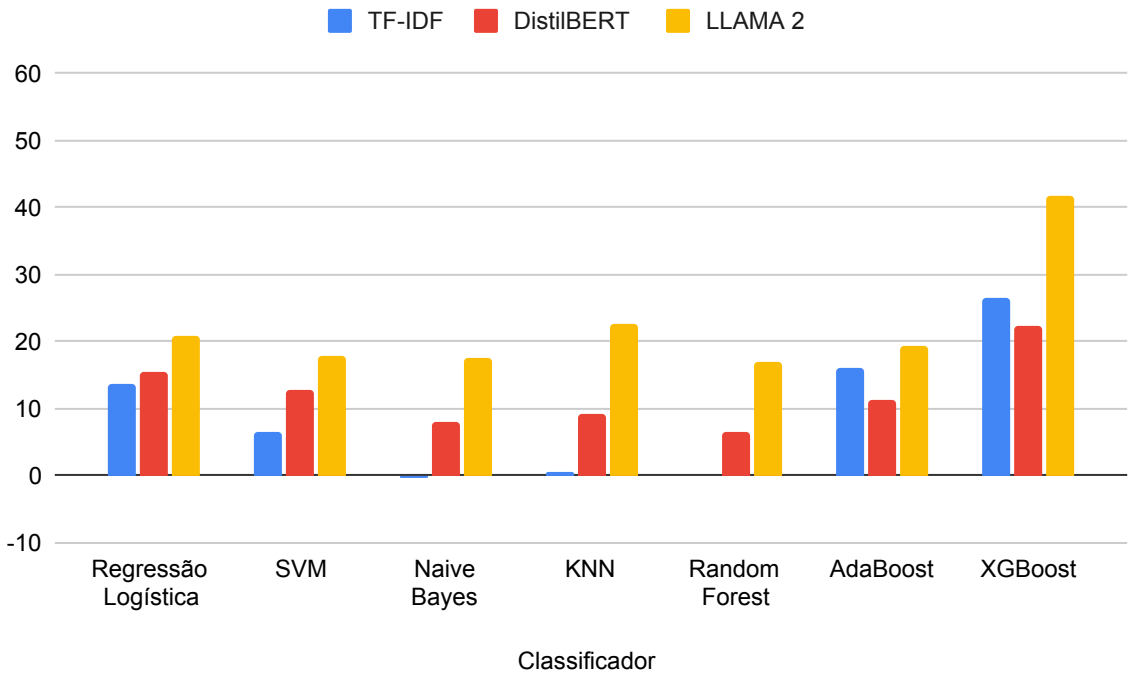
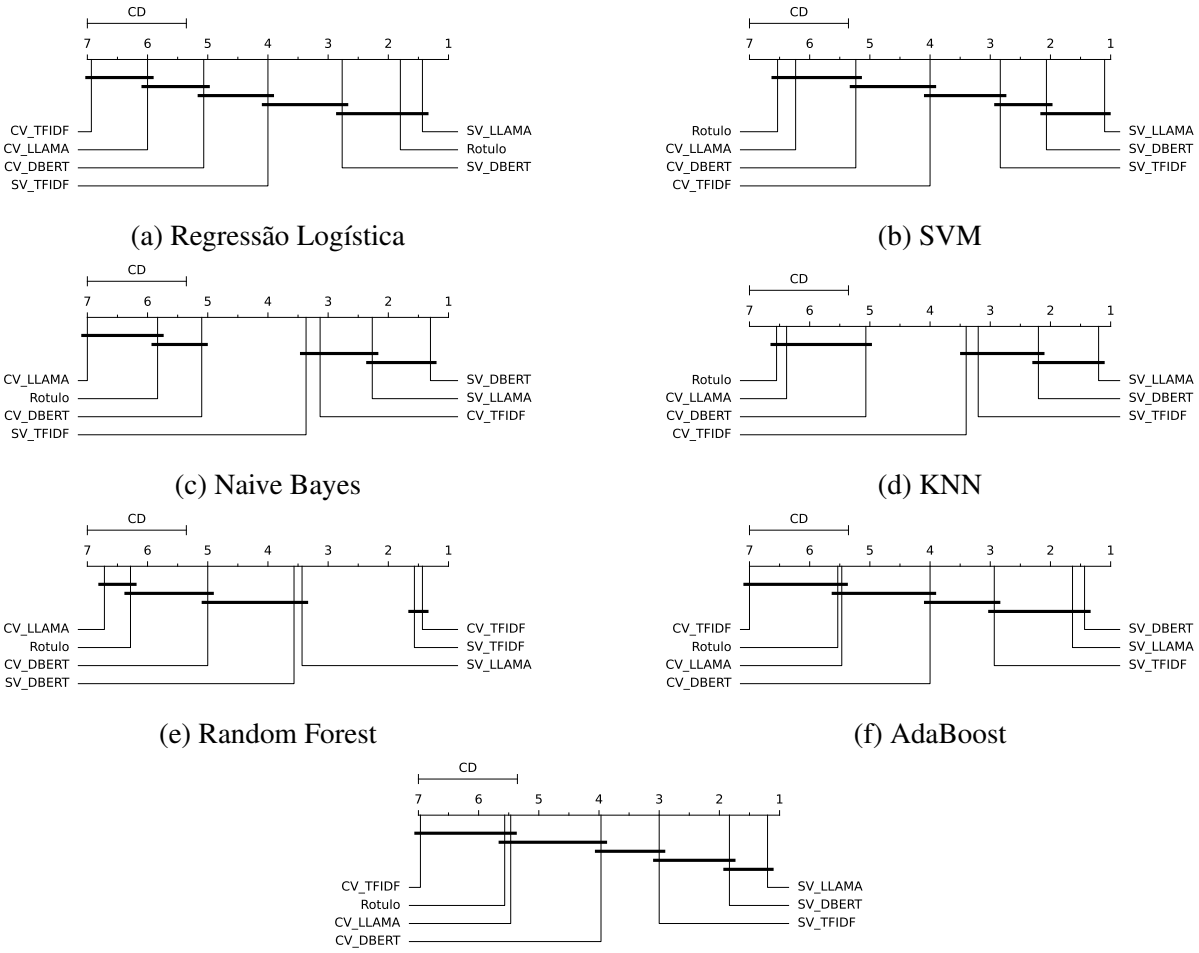


Figura (4.23) Percentuais de Ganhos com a Rotulação na KaggleFN - Título + Texto (F1-Score)

Avaliando os testes de Friedman realizados, em todos, a hipótese nula foi rejeitada (valores disponíveis no Apêndice B). Na Figura 4.22, estão ilustrados os Diagramas de Distâncias Críticas aplicados o pós-teste de Nemenyi com significância de 0,05 para a métrica F1-Score. O LLAMA 2 apresentou diferenças significativas entre os cenários com e sem rotulação para a Regressão Logísitca, AdaBoost e XGBoost, o DistilBERT apenas o caso de Random Forest não apresentou diferença significativa e o TF-IDF nos casos de SVM, Naive Bayes, KNN e Random Forest não apresentaram diferença significativa.



(a) Regressão Logística (b) SVM

(c) Naive Bayes (d) KNN

(e) Random Forest (f) AdaBoost

(g) XGBoost

Figura (4.24) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para Título + Texto da base KaggleFN (F1). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

É perceptível que há uma melhora nas duas métricas a partir do processo de rotulação na base KaggleFN, em todos os extratores de características e algoritmos de classificação, tanto a nível de texto, quanto a nível de título. Destaca-se o LLAMA 2 como o extrator que apresentou os melhores índices. No caso do vetor de título + texto, também foram apresentaram bons percentuais de otimização na maioria dos cenários avaliados.

Nesse sentido, os resultados obtidos aqui corroboram para os achados na base FakeNewsNet - Politifact. Levando em conta que a KaggleFN ficou com 1.787 instâncias, um tamanho cinco vez maior que a Politifact, há um indicativo de que o tamanho da base não é determinante, podendo a metodologia ser aplicada a bases com um número mais elevado de instâncias. O desbalanceamento da base também não demonstra ter impacto nos resultados, tendo em vista que a base ficou com cerca de 40% de notícias verdadeiras e 60% de notícias falsas, uma proporção levemente maior que a Politifact.

No que tange a distribuição dos rótulos, houve uma predominância do *conspirancy-pseudo-*

science para notícias falsas e *fake-news* para notícias verdadeiras, este último contrastando com a Politifact. Esse fator pode ser um indicativo de que o rótulo por si só não é o que determina a classe, mas a combinação dos elementos léxicos com a proporção do viés das instâncias. Esse apontamento é também um fator a ser considerado quando analisado o cenário de apenas rótulo. Apesar dos valores altos encontrados, esse cenário não foi consistente em todos os classificadores e métricas, enquanto a metodologia proposta de junção de rótulo e texto conseguiu manter um padrão elevado para os casos avaliados na duas bases.

4.3 LIAR

Tabela (4.13) Resultados da Acurácia para LIAR. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	64,15 (0,00)	64,92 (0,70)	64,96 (0,47)	60,70 (1,04)	65,43 (0,58)	**65,76 (0,72)**	63,29 (0,65)
SVM	59,08 (2,00)	61,80 (1,18)	**64,15 (0,00)**	60,09 (7,52)	61,93 (1,00)	**64,15 (0,00)**	59,04 (5,30)
Naive Bayes	64,15 (0,00)	55,61 (1,02)	64,61 (0,60)	64,15 (0,00)	56,04 (1,49)	**64,88 (0,54)**	64,15 (0,00)
KNN	58,09 (8,37)	62,00 (0,63)	62,90 (0,79)	61,31 (0,79)	62,16 (0,73)	**63,57 (0,74)**	61,06 (2,47)
Random Forest	64,15 (0,00)	64,15 (0,00)	64,68 (0,28)	64,15 (0,13)	64,15 (0,00)	**64,95 (0,38)**	64,34 (0,29)
AdaBoost	64,15 (0,00)	64,09 (0,16)	**64,83 (0,47)**	64,14 (0,06)	64,18 (0,47)	64,73 (0,51)	64,35 (0,24)
XGBoost	64,15 (0,00)	64,10 (0,34)	64,60 (0,62)	63,99 (0,28)	**65,05 (0,55)**	64,84 (0,60)	64,08 (0,28)

Na Tabela 4.13, estão descritos os resultados da acurácia obtidos com a base LIAR. A maioria dos algoritmos apresentou melhora com a inclusão do viés, bem como foram os que apresentaram os melhores índices, em especial o DistilBERT como extrator de *features*. Na Figura 4.25, estão ilustrados os percentuais de ganhos com a rotulação, sendo quase todos positivos, com exceção de SVM e KNN com LLAMA 2 e AdaBoost com DistilBERT. Observando os Diagramas de Distâncias Críticas, presentes na Figura 4.26, apesar do DistilBERT sem rótulo do viés ter obtido melhor resultado em relação às demais variações do AdaBoost, não há diferença significativa entre ele e o DistilBERT com inclusão do viés.

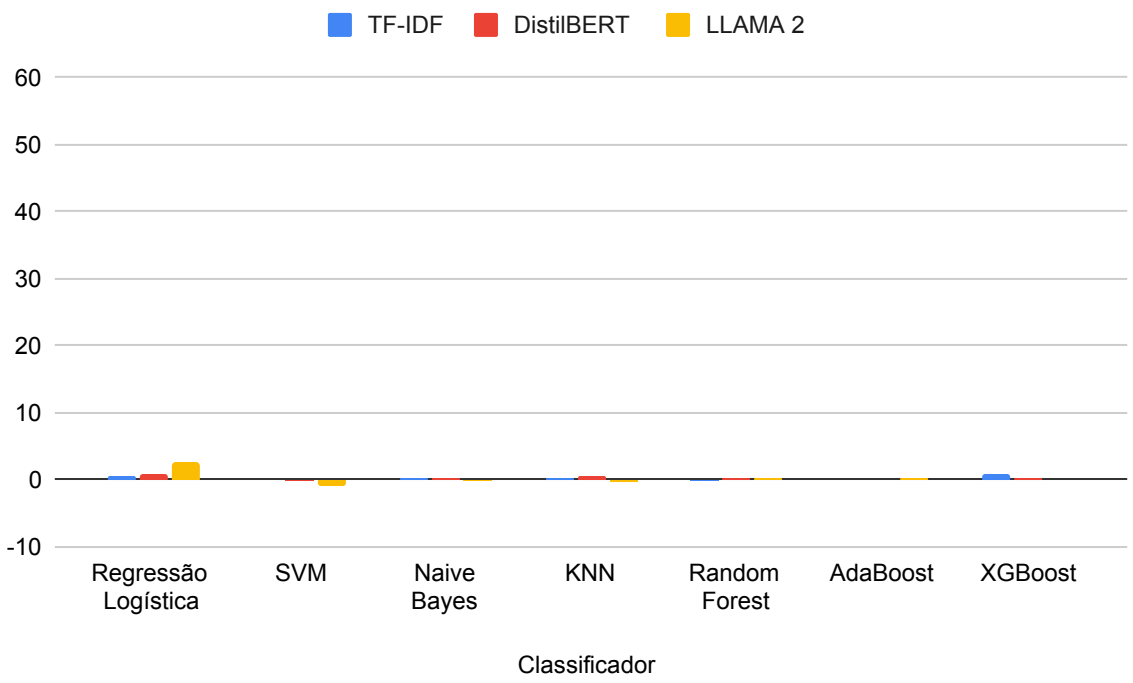


Figura (4.25) Percentuais de Ganhos com a Rotulação na LIAR (Acurácia)

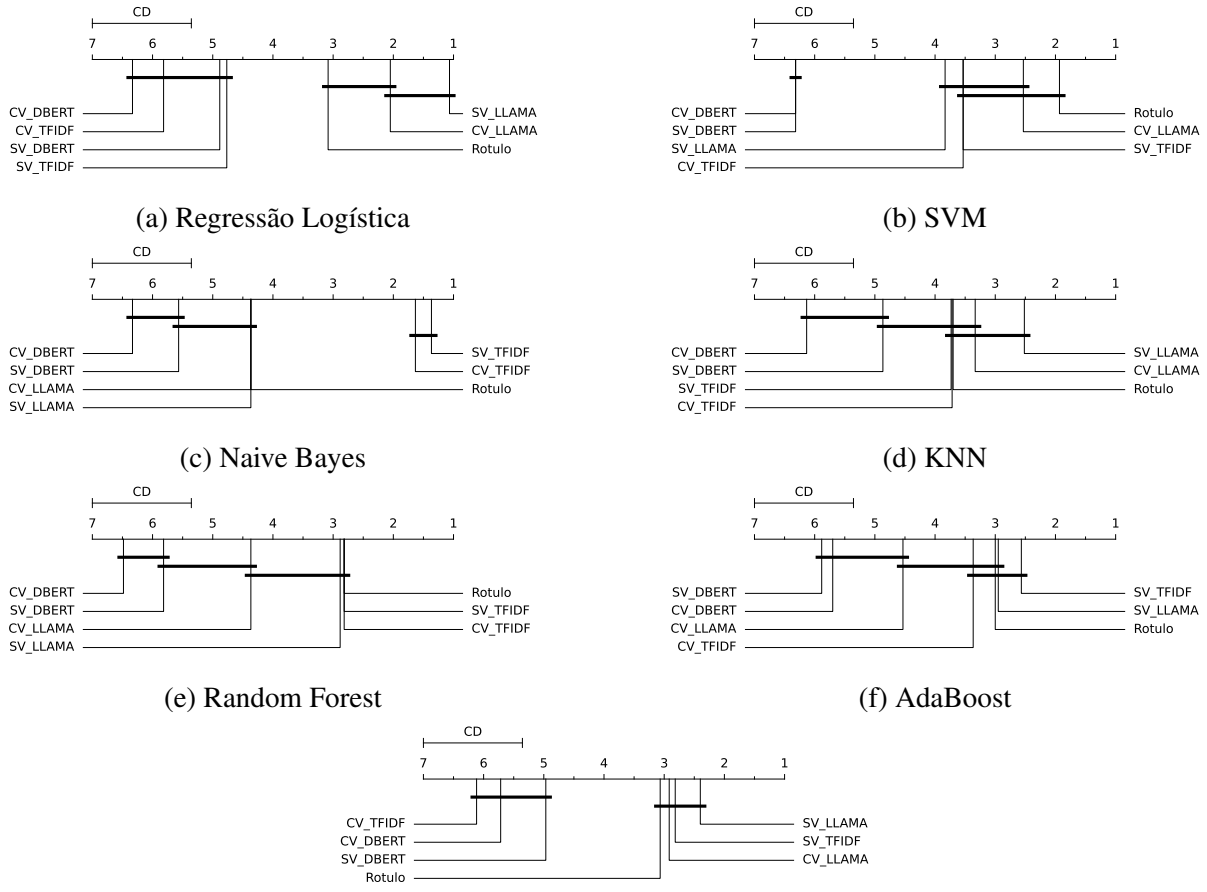


Figura (4.26) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para LIAR (Acurácia). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Tabela (4.14) Resultados da F1-Score para LIAR. Os valores representam a média das 30 execuções. Valores entre parênteses representam o desvio padrão. Valores em negrito destacam o melhor resultado por classificador, enquanto a célula com o valor sublinhado representa o melhor resultado geral.

Classificador	Apenas Rótulo	Sem Viés			Com Viés		
		TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
Regressão Logística	0,00 (0,00)	27,66 (2,35)	25,12 (6,52)	**34,88 (1,70)**	31,08 (1,78)	31,66 (2,27)	28,79 (1,60)
SVM	**42,21 (14,15)**	41,60 (1,91)	0,00 (0,00)	16,75 (14,13)	41,03 (3,40)	0,00 (0,00)	19,84 (9,97)
Naive Bayes	0,00 (0,00)	53,52 (0,90)	17,89 (4,19)	0,00 (0,00)	**53,94 (1,04)**	19,01 (3,63)	0,00 (0,00)
KNN	23,82 (22,54)	35,11 (1,35)	30,76 (1,63)	34,02 (1,69)	**35,68 (1,74)**	30,11 (1,84)	33,29 (4,29)
Random Forest	0,00 (0,00)	0,00 (0,00)	10,08 (1,38)	0,75 (0,79)	0,00 (0,00)	**10,84 (1,45)**	5,92 (3,00)
AdaBoost	0,00 (0,00)	1,39 (3,32)	12,85 (2,34)	0,27 (0,36)	8,94 (8,78)	**13,73 (3,11)**	4,60 (2,04)
XGBoost	0,00 (0,00)	7,76 (6,66)	19,06 (10,74)	2,12 (3,04)	19,14 (3,49)	**21,00 (11,13)**	6,65 (6,72)

Na Tabela 4.14, estão dispostos os resultados da F1-Score da LIAR. O TF-IDF e o DistilBERT com rotulação apresentaram os melhores resultados. Na Figura 4.27, estão dispostos os ganhos percentuais da rotulação, sendo quase todos positivos, com exceção de Regressão

Logística e KNN com LLAMA 2, KNN com DistilBERT e SVM com TF-IDF. Observando a Figura 4.28, é possível visualizar os Diagramas de Distâncias Críticas e constata-se que no caso do SVM, que não houve melhora no índice, a diferença não é significativa, com apenas a Regressão Logística apresentando perda no desempenho.

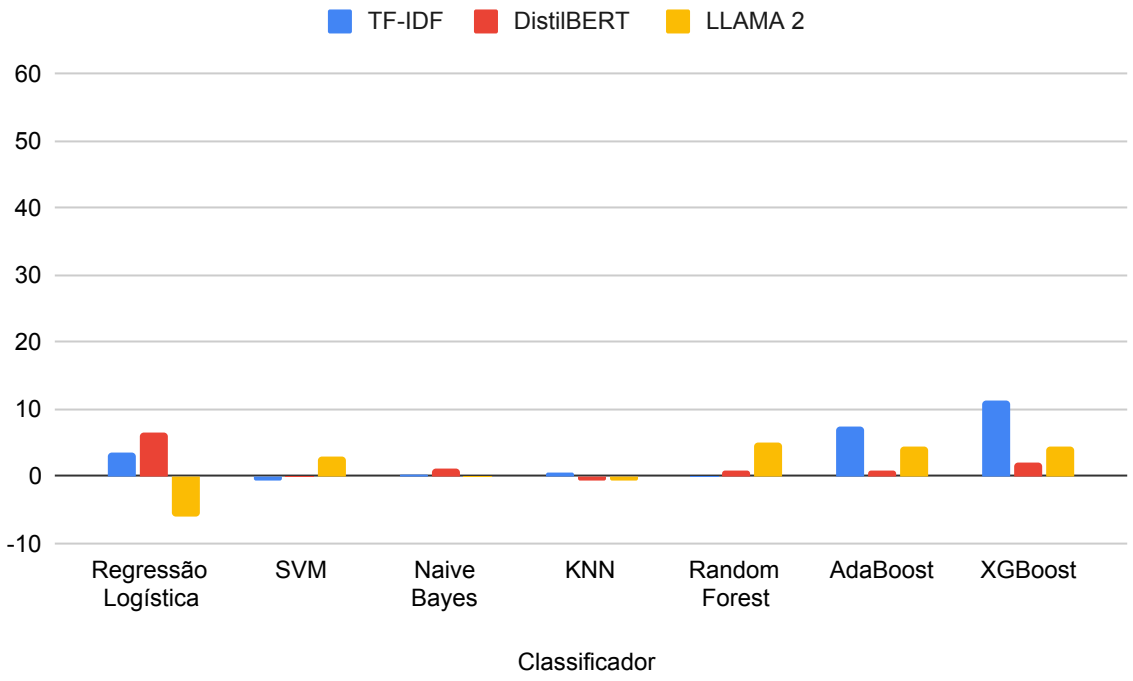
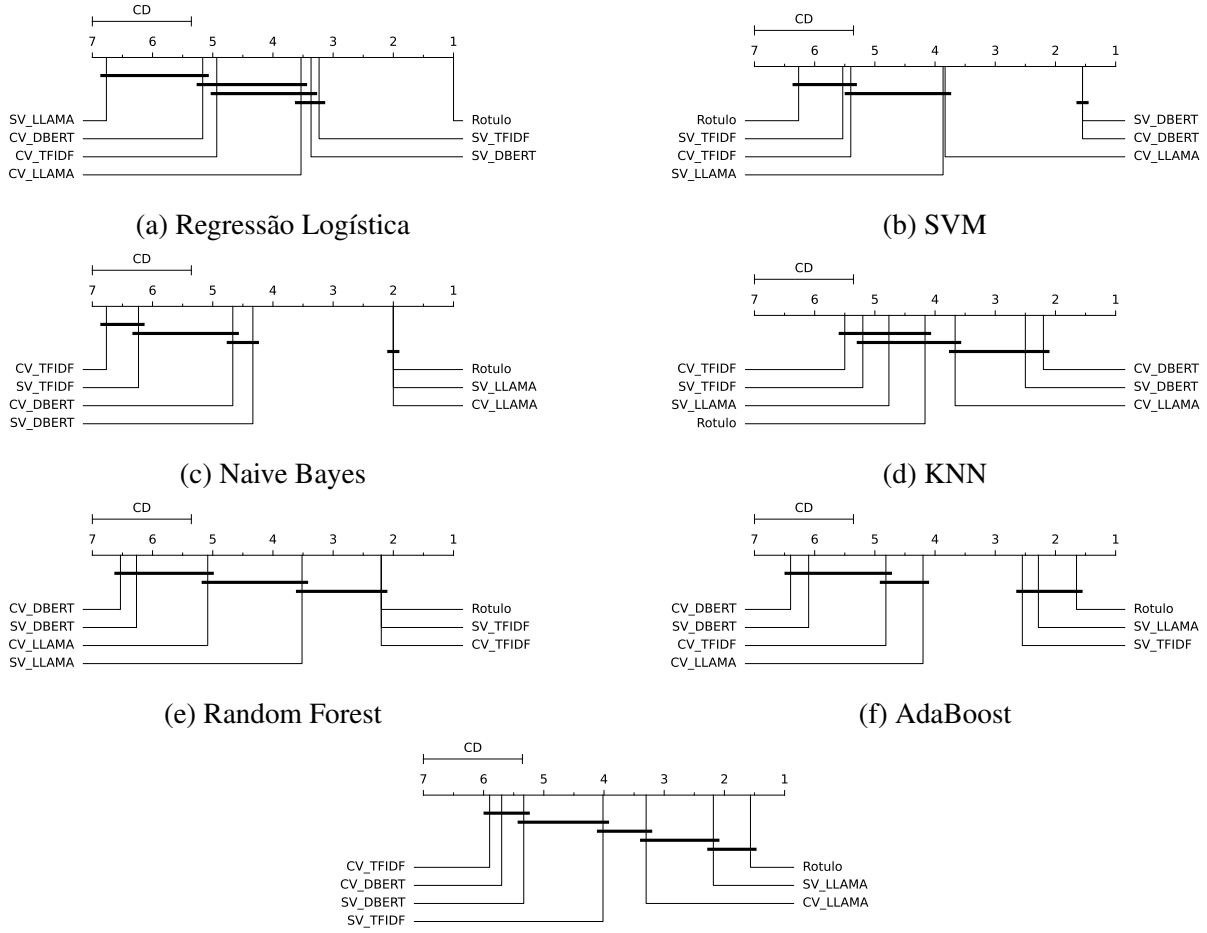


Figura (4.27) Percentuais de Ganhos com a Rotulação na LIAR (F1-Score)



(a) Regressão Logística

(b) SVM

(c) Naive Bayes

(d) KNN

(e) Random Forest

(f) AdaBoost

(g) XGBoost

Figura (4.28) Gráfico de distâncias críticas do teste post-hoc de Nemenyi para LIAR (F1-Score). Os prefixos "SV"e "CV"foram adicionados na legenda para descrever os cenários sem a rotulação do viés e com a rotulação do viés respectivamente.

Os resultados da base LIAR não são tão positivos quanto os das demais bases, no entanto ainda apresenta melhoras em determinados algoritmos e extratores, como o DistilBERT. Essa diferença pode ser explicada pelo processo de rotulação dela ser diferente das outras. Enquanto nas demais bases foi adotado a rotulação a partir da base do *Media Bias/Fact Check*, a LIAR utilizou uma rotulação da própria base, com indicadores de vieses diferentes. Além disso, deve-se considerar que houve adaptação nas classes da base, tornando ela um problema de classificação binário. Outro importante fator é ela ser uma base de declaração e não de notícias publicadas. Com isso, constata-se que o processo de rotulação pode ser mais efetivos a depender da base de rotulação usada e em bases de notícias em si, realizando a rotulação a nível de portal.

Há um fator a se considerar que é o aspecto de desbalanceamento da base, no qual cerca de 35% de notícias eram verdadeiras e 65% eram notícias falsas, dentre o total de 9.802 instâncias. A diferença percentual entre a quantidade das duas classes supera a das outras duas classes. É possível que essa discrepância tenha ocorrido em virtude do ajuste das classes junto da redução da base. Nesse sentido, mais experimentos são necessários para averiguar os efeitos

do desbalanceamento.

4.4 Lições Aprendidas

4.4.1 A inclusão do viés gera ganho de desempenho na classificação de notícias falsas?

Conforme demonstrado nas seções anteriores, há sim ganhos significativos de desempenho com a inclusão do viés. Na Tabela 4.15, estão dispostas as médias de ganho com a rotulação por base. Para a base Politifact, a média de ganhos da rotulação aplicada ao Texto é de 5,60% (acurácia) e 6,68% (F1-Score); a rotulação aplicada ao Título é de 7,93% (acurácia) e 9,48% (F1-Score); e a rotulação aplicada ao Título + Texto é de 3,98% (acurácia) e 4,65% (F1-Score). Para a base KaggleFN, a média de ganhos da rotulação aplicada ao Texto é de 18,15% (acurácia) e 16,81% (F1-Score); a rotulação aplicada ao Título é de 22,78% (acurácia) e 20,14% (F1-Score); e a rotulação aplicada ao Título + Texto é de 17,34% (acurácia) e 14,57% (F1-Score). Para a base LIAR, a média de ganhos da rotulação foi de 0,29% (acurácia) e 2,07% (F1-Score).

Tabela (4.15) Média dos Ganhos Percentuais da Rotulação por Base.

Base	Campo	Acurácia	F1-Score
FakeNewsNet - Politifact	Texto	5,60	6,68
	Título	7,93	9,48
	Título+Texto	3,98	4,65
KaggleFN	Texto	18,15	16,81
	Título	22,78	20,14
	Título+Texto	17,34	14,57
LIAR	Texto	0,29	2,07

Analisando os trabalhos originais das bases usadas, é perceptível uma melhora também. Considerando o melhor resultado encontrado, os algoritmos Regressão Logística, SVM e Naive Bayes tiveram um melhora de, respectivamente, 26,7%, 32,95% e 24,34% na acurácia para PolictFact da FakeNewsNet em relação ao trabalho original [SMW$^+$20]. Para a base KaggleFN, considerando o melhor resultado encontrado, os algoritmos de Regressão Logística, SVM, Naive Bayes, Random Forest e AdaBoost obtiveram um ganho de, respectivamente, 19,29%, 25,59%, 46,96%, 11,68% e 1,45% na acurácia em relação ao trabalho original [BAB$^+$20]. Para a base LIAR, não é possível realizar uma comparação direta, pois suas classes foram reorganizadas para esta pesquisa em duas, enquanto no trabalho original eram seis classes [Wan17].

Diante dos resultados exposto, a metodologia apresenta bons índices de otimização de desempenho nas duas métricas avaliadas. Destaca-se também que o campo de título obteve maior percentual médio de ganho em relação ao texto da notícia.

4.4.2 Qual melhor campo para o treinamento do classificador?

Na Tabela 4.16, estão dispostas as médias de cada campo utilizado, com e sem rotulação. Percebe-se que, para a base Politifact, a junção de Título + Texto com a rotulação obteve um desempenho médio superior aos demais campos, enquanto, para a base KaggleFN, o campo de Título com rotulação obteve um desempenho médio superior. A base LIAR só possui um campo para avaliar, o de Texto, obtendo um desempenho médio superior a partir da rotulação. Os resultados apontam que o uso do título por modelos de classificação de notícias falsas pode ter bons resultados, seja de maneira isolada ou de maneira conjunta.

Tabela (4.16) Média dos Resultados por Campo Utilizado. Em negrito, está destacado o melhor resultado por base.

Base	Métrica	Apenas Rótulo	Sem Viés			Com Viés		
			Texto	Título	Título + Texto	Texto	Título	Título + Texto
FakeNewsNet - Politifact	Acurácia	75,34	76,75	75,59	79,71	82,36	83,52	**83,75**
	F1-Score	61,55	72,24	69,51	75,86	78,92	78,99	**80,52**
KaggleFN	Acurácia	86,38	67,58	66,69	68,72	85,74	**89,47**	86,06
	F1-Score	89,10	70,2	70,88	73,8	87,01	**91,03**	88,37
LIAR	Acurácia	62,56	63,13	**-**	-	**63,43**	-	-
	F1-Score	9,43	17,69	**-**	-	**19,77**	-	-

4.4.3 Qual melhor extrator de características?

Na Figura 4.17, estão dispostas as médias de cada extrator de características utilizado, com e sem rotulação. Tanto na Politifact, quanto a KaggleFN, obtiveram um melhor desempenho médio com o uso do LLAMA 2 com a rotulação, em ambas as métricas avaliadas. No caso da base LIAR, o DistilBERT com rotulação obteve melhor desempenho na acurácia e o TF-IDF com rotulação obteve melhor valor do F1-Score.

Tabela (4.17) Média dos Resultados por Extrator de Característica. Em negrito, está destacado o melhor resultado por base.

Base	Métricas	Apenas Rótulo	Sem Viés			Com Viés		
			TF-IDF	DistilBERT	LLAMA 2	TF-IDF	DistilBERT	LLAMA 2
FakeNewsNet - Politifact	Acurácia	75,34	75,65	77,56	78,87	80,91	82,69	**86,04**
	F1-Score	61,55	69,38	73,43	74,83	75,46	79,69	**83,30**
KaggleFN	Acurácia	86,39	69,33	67,09	66,59	83,61	84,50	**93,17**
	F1-Score	89,10	72,32	71,66	70,92	86,12	86,57	**93,74**
LIAR	Acurácia	62,56	62,38	64,39	62,65	62,71	**64,70**	62,90
	F1-Score	9,43	23,86	16,54	12,68	**27,12**	18,05	14,16

4.4.4 Qual melhor classificador?

Na Figura 4.18, estão dispostas as médias dos classificadores. Tanto para a Politifact, quanto para a KaggleFN, o SVM obteve o melhor desempenho médio, em ambas as métricas. No caso da LIAR, o XGBoost obteve melhor acurácia e o KNN melhor F1-Score.

Tabela (4.18) Média dos Resultados por Classificador. Em negrito, está destacado o melhor resultado por base.

Base	Métricas	Regressão Logísitica	SVM	Naive Bayes	KNN	Random Forest	AdaBoost	XGBoost
FakeNewsNet - Politifact	Acurácia	80,41	**84,13**	77,51	76,89	78,35	79,72	80,06
	F1-Score	70,68	**81,72**	71,69	69,05	71,22	76,57	76,70
KaggleFN	Acurácia	79,34	**83,13**	73,32	76,02	74,87	81,09	82,92
	F1-Score	83,20	**86,12**	79,08	77,90	82,85	85,43	75,84
LIAR	Acurácia	64,17	61,46	61,94	61,58	64,37	64,35	**64,40**
	F1-Score	25,60	23,06	20,62	**31,83**	3,94	5,97	10,82

4.4.5 Por que o cenário de "Apenas Rótulo"obteve valores tão altos?

Uma questão que pode gerar questionamentos é acerca dos resultados do cenário de "Apenas Rótulo", pois é uma configuração de experimento que utilizou apenas uma variável no treinamento dos modelos. Apesar disso, ela obteve valores expressivos. Para entender melhor esse cenário, foi selecionada uma das 10 Árvores de Decisão geradas pelo algoritmo Random Forest para o cenário de "Apenas Rótulo", ilustradas nas Figuras 4.29, para a base FakeNewsNet - Politifact, e 4.30, para a base KaggleFN.

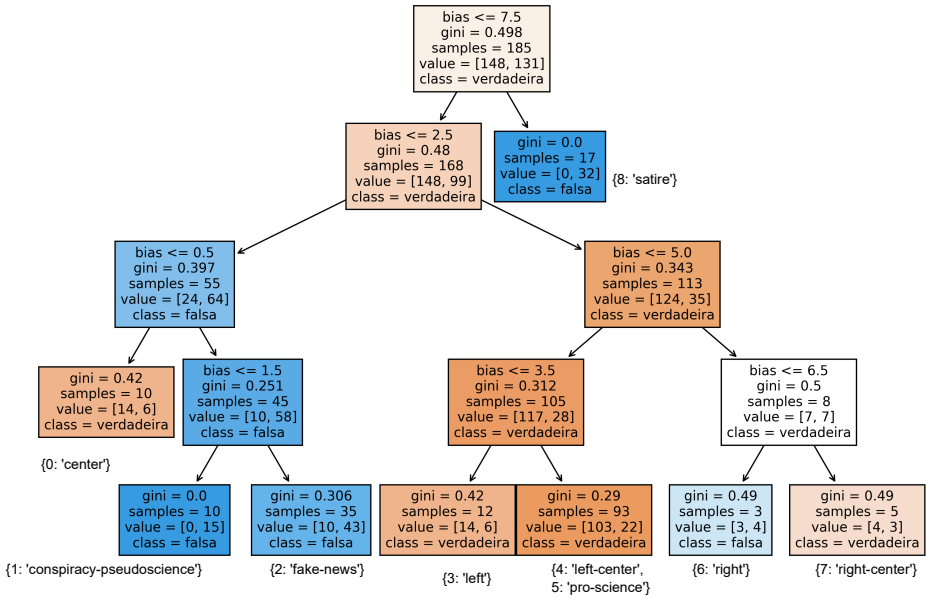


Figura (4.29) Árvore de Decisão do Random Forest aplicado à base FakeNewsNet - Politifact no cenário de "Apenas Rótulo".

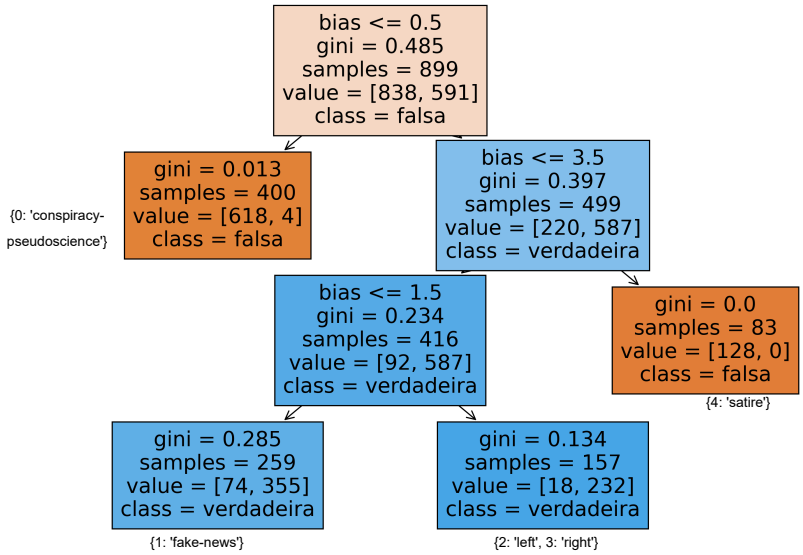


Figura (4.30) Árvore de Decisão do Random Forest aplicado à base KaggleFN no cenário de "Apenas Rótulo".

Em cada folha, dentro do retângulo, está apontada a classe inferida, enquanto fora do retângulo, está descrito o rótulo que gerou aquela classificação, de forma que é possível perceber que a árvore de decisão infere se a classe é verdadeira ou falsa em virtude da concentração em cada rótulo. Para a base Politifact, usando como exemplo o rótulo *"left-center"*, tem-se uma quantidade de 143 notícias verdadeiras para 28 notícias falsas desse rótulo. Dessa forma, com o classificador assumindo que toda notícia de *"left-center"* é inferida como verdadeira, ele acerta 83,62% das instâncias, errando em 16,38% das instâncias desse rótulo, o qual corresponde à 49,13% de toda a base. Já para a base KaggleFN, adotando como exemplo o rótulo *"conspiracy-pseudoscience"*, há uma quantidade de 796 de notícias falsas para apenas 5 de notícias falsas. Com o classificador assumindo que toda notícia de *"conspiracy-pseudoscience"* é falsa, ele acerta 99,37% das instâncias, errando em menos de 1% das instâncias desse rótulo, o qual corresponde a 44,82% de toda a base.

Conforme ilustrado na Seção 3.2, a distribuição de vieses nas duas bases em questões possui uma discrepância entre as notícias verdadeiras e falsas, de forma que a maioria dos rótulos possui uma quantidade de instâncias de uma classe muito maior que a outra. Essa proporção de instâncias de cada classe por rótulo pode explicar os resultados elevados mesmo com apenas uma variável. No entanto, em cenários em que haja uma distribuição mais igualitária entre os rótulos, esses resultados de "Apenas Rótulo"tenderão ter uma redução na acurácia. Isso pode ser constatado em outros classificadores que não conseguem separar tão bem as classes a partir de uma única variável, como a Regressão Logística. Nas Figuras 4.31 e 4.32, estão dispostoas as Fronteiras de Decisão extraídas do algoritmo Regressão Logística aplicado às bases Politifact

e KaggleFN, respectivamente, no cenário de "Apenas Rótulo". Nelas, os círculos dispostos no gráfico são os dados de treino, enquanto os "X"são os dados de teste, estando na maioria das vezes sobrepostos na Figura. A cor vermelha indica a classe das notícias falsas, enquanto a azul representa as das notícias verdadeiras. A área colorida no gráfico indica a escolha do classificador, seguindo o mesmo padrão de cores.

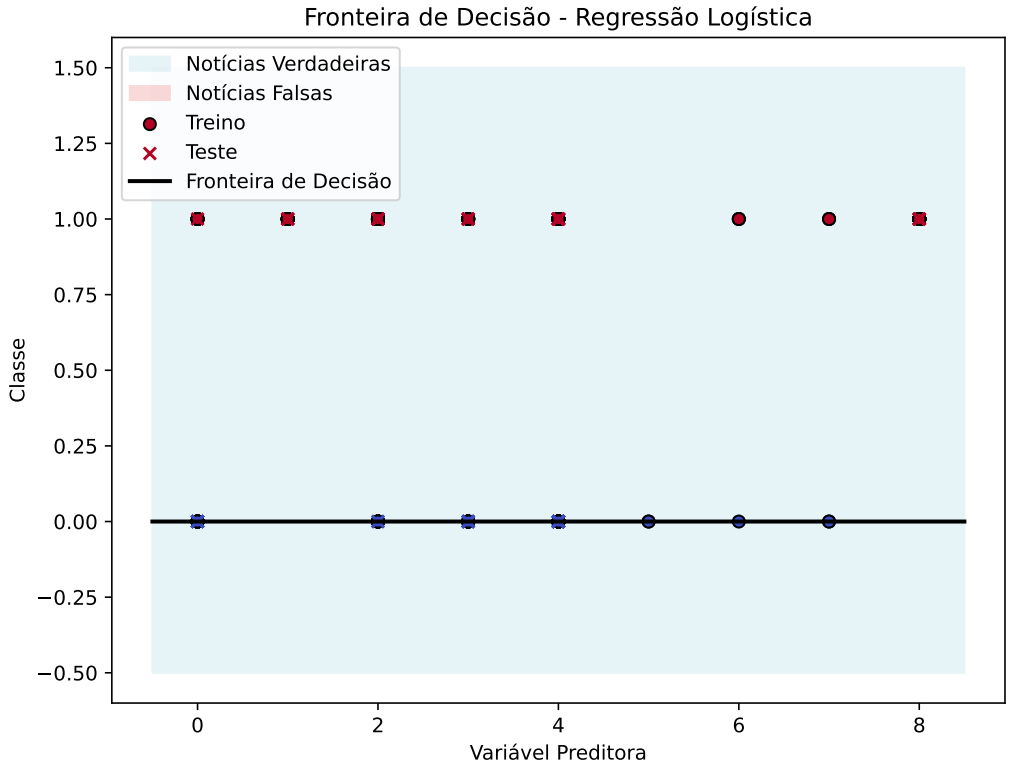


Figura (4.31) Fronteira de Decisão da Regressão Logística aplicada à base FakeNewsNet - Politifact no cenário de "Apenas Rótulo".

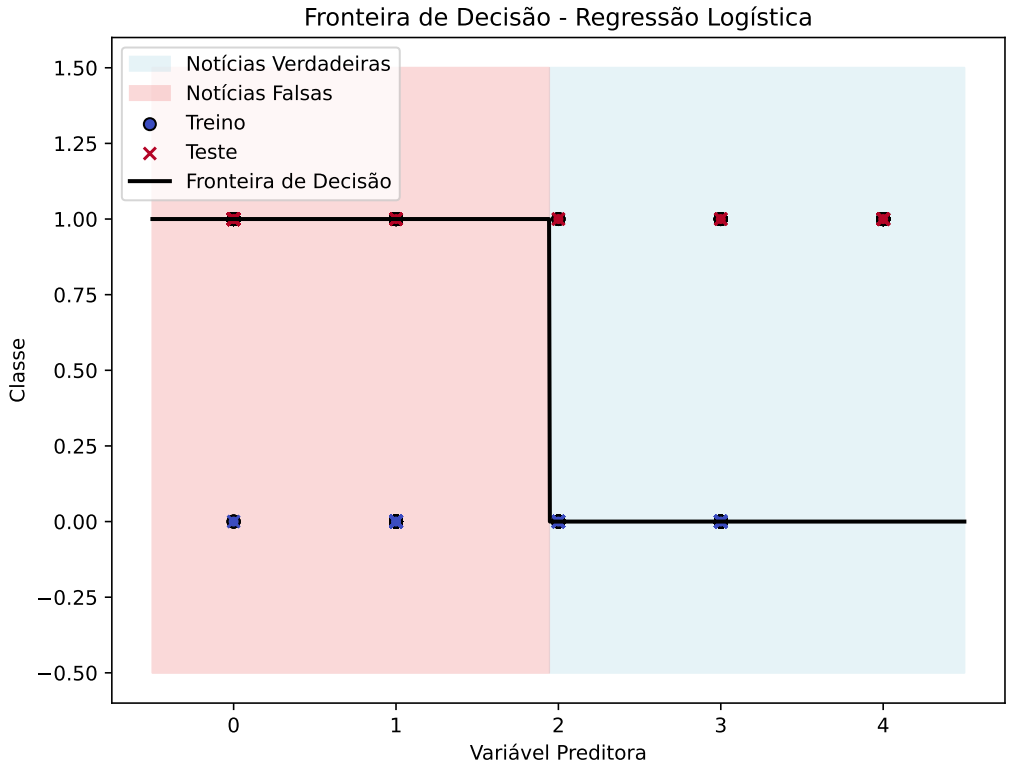


Figura (4.32) Fronteira de Decisão da Regressão Logística aplicada à base KaggleFN no cenário de "Apenas Rótulo".

Ao analisar a base Politifact, é possível constatar que o modelo traça a Fronteira de Decisão no eixo da classe de notícias verdadeiras, sendo esse o valor mínimo dos dados para o eixo Y, representado pelo valor zero, enquanto o valor um representa a classe das notícias falsas. Com isso, o modelo não conseguiu chegar a um *threshold* para separar as classes, de forma que todas as notícias avaliadas serão classificadas como verdadeiras. Isso explica o resultado inferior deste algoritmo no cenário de "Apenas Rótulo", pois enquanto outros algoritmo, como SVM, chegaram a obter cerca de 82% na acurácia, a Regressão Logística só obteve cerca de 56% nessa base. Apesar de traçar uma curva melhor adaptada, chegando a um *threshold*, na Regressão Logística na base KaggleFN também possui dificuldade de separar corretamente as classes neste cenário e apresenta resultados abaixo dos demais classificadores, obtendo cerca de 65% na acurácia, enquanto outros algoritmos, como SVM, obtiveram cerca de 93%.

Nessa perspectiva, é possível verificar que há uma importante capacidade de classificação na variável do rótulo, obtendo altos valores nas métricas avaliadas. No entanto, apenas ela pode não ser tão estável quanto a proposta metodológica de sua concatenação com o texto, pois a distribuição das classes nos vieses impacta significativamente nesses resultados, além de não ser tão promissora em determinados algoritmos, enquanto a rotulação no texto se mostrou eficiente na grande maioria dos cenários avaliados. Essa maior estabilidade nos resultados contribui para explicar a maior média da rotulação nas Subseções anteriores.

CAPÍTULO 5

Conclusão

A classificação de notícias falsas tem tomado cada vez mais importância perante os riscos da desinformação. Há, ainda, conceitos pouco explorados na literatura, em especial do impacto do viés político em classificadores de notícias falsas. Nesse sentido, este trabalho apresentou uma metodologia para auxiliar no processo de classificação, indicando que a análise de viés pode contribuir para construção de modelos mais eficientes, em especial quando trata-se de notícias com a temática política, as quais foram o objeto dos experimentos realizados. Nessa perspectiva, a maior parte dos cenários avaliados apresentou melhoras em relação à metodologia tradicional que usa apenas o texto, com a maioria possuindo diferença significativa. Assim, destaca-se que houve melhora de até de 29,28% na acurácia, no caso do KNN com LLAMA 2 para a base Politifact, e 50,72% no *F1-Score*, no caso do KNN com TF-IDF para a base KaggleFN.

Dentre os três cenários avaliados, a adoção de campo textual e do viés político, além de obter os melhores resultados em 92,85% dos casos, apresentou maior consistência quando são comparados os diferentes experimentos, isto é, o padrão dos resultados manteve-se ao alternar entre classificadores e extratores de características. O cenário de Apenas Rótulo não apresentou bons resultados para todos os classificadores, como no caso da Regressão Logística, e o cenário de Apenas Texto (Sem Viés) obteve um desempenho médio inferior. Enquanto isso, o cenário de Texto com Viés obteve altos índices nos três extratores de *features* e nos sete classificadores avaliados, demonstrando versatilidade e praticidade ao possuir maior aderência nos experimentos. Dentre os resultados, destacam-se o SVM e o LLAMA 2, obtendo um desempenho médio na acurácia de até 84,13% (para a base Politifact) e 93,17% (para a base KaggleFN) e no *F1-Score* de 86,12% (para a base KaggleFN) e 93,74% (para a base KaggleFN), respectivamente.

Um dos diferenciais desta abordagem é ter sido verificada seu uso tanto para texto, quanto título da notícia, possibilitando seu uso em contextos em que os informação está parcial, isto é, estando a base com apenas um desses dois campos, além de ter sido avaliada a junção desses dois tipos de campos. Atrelado a isso, o uso de um ambiente de fácil reprodução, como o Google Colab, permite a melhor reprodutibilidade do experimento, além da adoção da metodologia por outros segmentos, como jornalistas, os quais têm adotado cada vez mais soluções automatizadas para realizar o processo de *Fact Checking* [DL23] [Ünv23].

Os resultados apontam algumas considerações importantes, como o fato dos maiores aumentos nas métricas avaliadas terem ocorrido nos cenários envolvendo as manchetes. Isso pode ser explicado por dois fatores: o primeiro é pelo tamanho reduzido do título, resultando em vetores e camadas menores em relação a todo texto, de forma que a adição da informação do viés termine por ser mais significativa proporcionalmente. Para ilustrar esse fator, considerando o TF-IDF, o vetor gerado para o título da Politifact possui 1.491 de colunas, enquanto

o vetor para o texto possui 21.260, ou seja, o vetor de título corresponde a apenas 7,01% do tamanho do de texto. O mesmo ocorre na KaggleFN, em que o vetor de título possui 4.973 colunas e o vetor de texto possui 39.744, ou seja, 12,51% do tamanho. Já o segundo consiste no fator *Clickbait*, o qual resulta em formatação de manchetes mais sensacionalistas, de maneira enviesada, de modo que a adição do viés do portal em questão ajude os modelos encontrar melhor os padrões em questão.

Considerando as bases avaliadas, as duas de notícias publicadas, Politifact e KaggleFN, obtiveram melhores resultados, enquanto a base de declarações, LIAR, obteve resultados menos expressivos, com o desempenho médio abaixo em comparativo às demais bases. Isso pode indicar que a metodologia possui maior aderência em notícias veiculadas em portais do que em textos extraídos de declarações ou que a abordagem da rotulação adotada para a base LIAR é menos eficiente que a adotada para as demais.

Como limitações, é importante considerar a redução das bases, em especial na base Politifact, de forma em que houve uma perda significativa na quantidade de dados disponíveis para o treino. Outro ponto também a se considerar foi o número reduzido de bases de vieses disponíveis, impossibilitando análises mais diversificadas. Há a se considerar, ainda, mais formas de anexar o viés, como a adoção de máscaras visuais para os rótulos. Nesse sentido, mais experimentos são importantes para ajustar a metodologia, a fim de torná-la mais generalizável possível.

Por fim, como trabalhos futuros, é possível destacar: a incorporação de uma etapa de classificação para os vieses dos portais, evitando assim a redução das bases; aplicar tal metodologia em processos de *chat bots*, a fim de verificar emissão de falas enviesadas por parte do usuário ou do robô; e ampliar o número de classificadores e extratores de *features* avaliados.

Glossário

Backfire Effects Efeito em que uma pessoa reforça determinada crença ao entrar em contato com uma evidência que a contraria. 12

Belief Perseverance Perseverança de crença, resistência em mudar sua opinião mesmo após a constatação de uma nova informação que contradiga o que já se acredita previamente. 11

Clickbait Estratégias de construção da notícia para atribuir cliques, muitas vezes se utilizando de manchetes sensacionalistas. 3, 10, 22, 72

Fact Checking Instituições jornalísticas voltadas para checagem de dados e informações já veiculadas, seja por agentes ou pela própria imprensa, a fim de validá-las ou não. 22, 23, 71

APÊNDICE A

Parâmetros do DistilBERT e LLAMA 2

Os itens abaixo correspondem aos parâmetros adotados no métodos de extração de características DistilBERT e LLAMA2 para o *tokenizer.batch_encode_plus*:

- news["lemmas"].values.tolist();
- add_special_tokens = True;
- padding = 'max_length';
- max_length = x; [1]
- truncation = True;
- return_tensors = 'pt',
- return_attention_mask = True.

[1] A variável x no parâmetro *max_length* foi substituído por 64 em nas bases da FakeNewsNet e KaggleFN e por 32 na LIAR no caso do DistilBERT e 16 na base da FakeNewsNet e por 4 na KaggleFN e LIAR no caso do LLAMA 2.

APÊNDICE B

Resultados para o Teste de Friedman

Resultados do Teste de Friedman aplicados para validação estatística nos experimentos.

Tabela (B.1) Resultados do Teste de Friedman.

Base	Métricas	Regressão Logística	SVM	Naive Bayes	KNN	Random Forest	AdaBoost	XGBoost
Politifact - Texto	Acurácia	$1,28x10^{-27}$	$1,40x10^{-18}$	$1,00x10^{-18}$	$4,22x10^{-18}$	$4,30x10^{-12}$	$2,04x10^{-22}$	$1,10x10^{-13}$
	F1-Score	$1,00x10^{-27}$	$1,66x10^{-20}$	$2,40x10^{-17}$	$3,14x10^{-20}$	$5,30x10^{-12}$	$1,49x10^{-20}$	$2,08x10^{-12}$
Politifact - Título	Acurácia	$2,74x10^{-28}$	$1,10x10^{-19}$	$4,41x10^{-27}$	$2,54x10^{-22}$	$6,02x10^{-25}$	$3,63x10^{-23}$	$6,05x10^{-24}$
	F1-Score	$1,86x10^{-9}$	$1,15x10^{-18}$	$2,93x10^{-29}$	$2,34x10^{-21}$	$1,71x10^{-25}$	$8,50x10^{-19}$	$1,54x10^{-20}$
Politifact - Título + Texto	Acurácia	$3,49x10^{-20}$	$1,40x10^{-12}$	$1,53x10^{-21}$	$5,93x10^{-21}$	$2,55x10^{-11}$	$1,93x10^{-20}$	$4,83x10^{-12}$
	F1-Score	$7,16x10^{-21}$	$7,75x10^{-15}$	$1,15x10^{-22}$	$5,08x10^{-22}$	$1,52x10^{-09}$	$1,85x10^{-17}$	$1,60x10^{-09}$
KaggleFN - Texto	Acurácia	$6,73x10^{-35}$	$7,68x10^{-34}$	$1,40x10^{-33}$	$1,20x10^{-33}$	$1,05x10^{-34}$	$6,80x10^{-35}$	$1,74x10^{-34}$
	F1-Score	$2,20x10^{-33}$	$1,05x10^{-30}$	$9,12x10^{-32}$	$4,70x10^{-35}$	$2,43x10^{-34}$	$2,22x10^{-34}$	$2,85x10^{-34}$
KaggleFN - Título	Acurácia	$2,78x10^{-31}$	$2,09x10^{-31}$	$3,85x10^{-35}$	$1,18x10^{-32}$	$2,67x10^{-34}$	$1,35x10^{-32}$	$6,07x10^{-31}$
	F1-Score	$1,04x10^{-31}$	$2,35x10^{-28}$	$2,23x10^{-35}$	$7,32x10^{-34}$	$2,37x10^{-33}$	$9,00x10^{-32}$	$2,07x10^{-31}$
KaggleFN - Título + Texto	Acurácia	$6,88x10^{-35}$	$8,34x10^{-35}$	$3,95x10^{-33}$	$2,68x10^{-32}$	$1,90x10^{-34}$	$6,60x10^{-35}$	$3,61x10^{-35}$
	F1-Score	$2,07x10^{-34}$	$4,62x10^{-34}$	$3,95x10^{-32}$	$1,13x10^{-32}$	$2,29x10^{-34}$	$1,34x10^{-34}$	$9,42x10^{-35}$
LIAR	Acurácia	$5,72x10^{-30}$	$1,14x10^{-22}$	$7,00x10^{-29}$	$1,32x10^{-09}$	$4,66x10^{-21}$	$1,95x10^{-14}$	$1,69x10^{-18}$
	F1-Score	$1,92x10^{-25}$	$4,84x10^{-28}$	$5,12x10^{-35}$	$7,99x10^{-12}$	$6,17x10^{-33}$	$1,33x10^{-28}$	$1,87x10^{-23}$

ANEXO I

Número de Artigos Sobre Fake News ao Longo dos Anos

Na Figura I.1, estão listados as publicações na literatura acerca de *fake news* ao longo dos anos. Gráfico retirado do trabalho de Domenico et al. [DDSIN21].

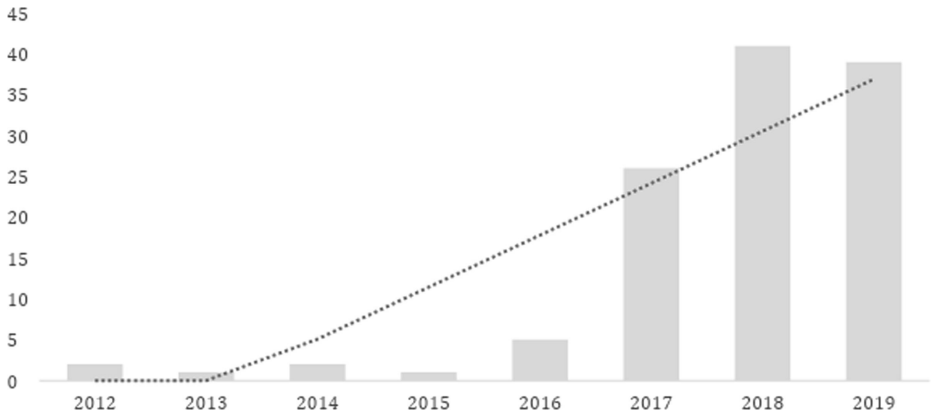


Figura (I.1) Número de Artigos Sobre Fake News ao Longo dos Anos (Domenico et al., 2021).

Referências Bibliográficas

[AG17] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[All23] Allsides | balanced news and media bias ratings. unbiased news doesn't exist, 2023.

[AM21] Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

[Ans21] Nicolas M Anspach. Trumping the equality norm? presidential tweets and revealed racial attitudes. *New Media & Society*, 23(9):2691–2707, 2021.

[Arr15] Gabriel Domingos de Arruda. *Análise de viés em notícias na língua portuguesa*. PhD thesis, Universidade de São Paulo, 2015.

[ASLA20] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. *arXiv preprint arXiv:2004.05773*, 2020.

[ATS17] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference*, pages 127–138, Vancouver, BC, Canada, 2017. Springer.

[BAB+20] A Bharadwaj, Brinda Ashar, P Barbhaya, R Bhatia, and Z Shaikh. Source based fake news classification using machine learning. *International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)*, pages 2320–6710, 2020.

[BCD+15] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093, 2015.

[BCGPN21] João Pedro Baptista, Elisete Correia, Anabela Gradim, and Valeriano Piñeiro-Naval. The influence of political ideology on fake news belief: The portuguese case. *Publications*, 9(2):23, 2021.

[BFG21] Sven Bernecker, Amy K. Flowerree, and Thomas Grundmann. *The Epistemology of Fake News*. Oxford University Press, 06 2021.

[BFJ+12] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[BMC19] Luís Borges, Bruno Martins, and Pável Calado. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–26, 2019.

[BPY16] Daniel Bradley, Christos Pantzalis, and Xiaojing Yuan. The influence of political bias in state pension funds. *Journal of Financial Economics*, 119(1):69–91, 2016.

[BR20] Jennifer L Bonnet and Judith E Rosenbaum. "fake news," misinformation, and political bias: Teaching news literacy in the 21st century. *Communication teacher*, 34(2):103–108, 2020.

[BRP20] Bence Bago, David G Rand, and Gordon Pennycook. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general*, 149(8):1608, 2020.

[BSP17] Sean Baird, Doug Sibley, and Yuxi Pan. Talos targets disinformation with fake news challenge victory. *Fake News Challenge*, 2017.

[BSR+23] Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natália da Silva Perez, Natacha Klein Käfer, and Isabelle Augenstein. Measuring intersectional biases in historical documents. *arXiv preprint arXiv:2305.12376*, 2023.

[CGBM21] Dustin P. Calvillo, Ryan J.B. Garcia, Kiana Bertrand, and Tommi A. Mayers. Personality factors and self-reported political news consumption predict susceptibility to political fake news. *Personality and Individual Differences*, 174:110666, 2021.

[Com24] The Political Compass. The political compass. https://politicalcompass.org/, 2024.

[COU23] Sara Bandeira COUTINHO. Seleção de ensemble heterogêneo para a detecção de fake news. Master's thesis, Universidade Federal de Pernambuco, 2023.

[CPYM21] Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. Neutral bots probe political bias on social media. *Nature communications*, 12(1):5580, 2021.

[CSS18] Santanu Chakrabarti, Lucile Stengel, and Sapna Solanki. Duty, identity, credibility: Fake news and the ordinary citizen in india. *BBC News*, 2018.

[DBL+14] Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. Pheme: computing veracity: the fourth challenge of big social data. In *European Semantic Web Conference ESWC*, 2014.

[DBL+17] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*, 2017.

[DCFG21] Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518, 2021.

[DDSIN21] Giandomenico Di Domenico, Jason Sit, Alessio Ishizaka, and Daniel Nunan. Fake news, social media and marketing: A systematic review. *Journal of Business Research*, 124:329–341, 2021.

[Dem06] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[DL23] Laurence Dierickx and Carl-Gustav Lindén. Journalism and fact-checking technologies: Understanding user needs. *communication+ 1*, 10(1), 2023.

[dSdAV19] Mayara Karla Dantas da Silva, Maria Elizabeth Baltar Carneiro de Albuquerque, and Maria do Socorro Furtado Veloso. Representação da informação noticiosa pelas agências de fact-checking: do acesso à informação ao excesso de informação. *Revista Brasileira de Biblioteconomia e Documentação*, 15(2):410–426, 2019.

[DV18] Caroline Delmazo and Jonas CL Valente. Fake news nas redes sociais online: propagação e reações à desinformação em busca de cliques. *Media & Jornalismo*, 18(32):155–169, 2018.

[ER15] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

[FCC24] Faramarz Farhangian, Rafael MO Cruz, and George DC Cavalcanti. Fake news detection: Taxonomy and comparative study. *Information Fusion*, 103:102140, 2024.

[FMA+23] Frederik Federspiel, Ruth Mitchell, Asha Asokan, Carlos Umana, and David McCoy. Threats by artificial intelligence to human health and human existence. *BMJ global health*, 8(5):e010435, 2023.

[FNR17] Daniel J Flynn, Brendan Nyhan, and Jason Reifler. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150, 2017.

[FV16] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American*

chapter of the association for computational linguistics: Human language technologies. ACL, 2016.

[G2024] G20 dewg maceio ministerial declaration, Sep 2024.

[Gaw21] Bertram Gawronski. Partisan bias in the identification of fake news. *Trends in Cognitive Sciences*, 25(9):723–724, 2021.

[GBL23] Sandra González-Bailón and Yphtach Lelkes. Do social media undermine social cohesion? a critical review. *Social Issues and Policy Review*, 17(1):155–180, 2023.

[GDM19] Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 77–84, 2019.

[GGI24] Eitan Goldman, Nandini Gupta, and Ryan Israelsen. Political polarization in financial news. *Journal of Financial Economics*, 155:103816, 2024.

[GJF+19] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

[GKL+19] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, United States, June 2019. Association for Computational Linguistics.

[GRR18] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. Stance detection in fake news a combined feature representation. In *Proceedings of the first workshop on fact extraction and VERification (FEVER)*, pages 66–71, 2018.

[H+13] Lee Howell et al. Digital wildfires in a hyperconnected world. *World Economic Forum Report*, 3:15–94, 2013.

[HANA21] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242*, 2021.

[HKO+22] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022.

[HLW22] D Hopkins, Y Lelkes, and S Wolken. Which news goes viral? measuring identity threats and engagement on social media. In *American Political Science Association annual meetings, Montreal, Canada*, 2022.

[IP18] Cherilyn Ireton and Julie Posetti. *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing, 2018.

[KEM+19] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22:188–227, 2019.

[KGN21] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788, 2021.

[Koo04] Ruud Koopmans. Movements and media: Selection processes and evolutionary dynamics in the public sphere. *Theory and society*, 33(3-4):367–391, 2004.

[KSG21] Ayush Kaushal, Avirup Saha, and Niloy Ganguly. twt–wt: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889, 2021.

[LBB+18] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[LFBL20] Lucas Lisboa, João Victor Ferro, José Rubens Brito, and Roberta Lopes. A disseminação da desinformação promovida por líderes estatais na pandemia da covid-19. In *Anais do I Workshop sobre as Implicações da Computação na Sociedade*, pages 114–121, Porto Alegre, RS, Brasil, 2020. SBC.

[LSOLH23] Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, 7(1):74–101, 2023.

[Mar23] Susie Marino. What happens in an internet minute? [2024 statistics], Dec 2023.

[MBG+18] Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. Automatic stance detection using end-to-end memory networks. *arXiv preprint arXiv:1804.07581*, 2018.

[MPNR24] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: Measuring chatgpt political bias. *Public Choice*, 198(1):3–23, 2024.

[NSG17] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. Online political discourse in the trump era. *arXiv preprint arXiv:1711.05303*, 2017.

[OBV+21] MATHIAS OSMUNDSEN, ALEXANDER BOR, PETER BJERREGAARD VAHLSTRUP, ANJA BECHMANN, and MICHAEL BANG PETERSEN. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015, 2021.

[Pat17] Thomas E Patterson. News coverage of donald trump's first 100 days. *HKS Faculty Research Working Paper Series*, 2017.

[PB20] Julie Posetti and Kalina Bontcheva. Disinfodemic: deciphering covid-19 disinformation. policy brief 1. *Paris: United Nations Educational, Scientific and Cultural Organization*, 2020.

[PKCS12] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. A computational framework for media bias mitigation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2):1–32, 2012.

[RKA14] Mazhar Iqbal Rana, Shehzad Khalid, and Muhammad Usman Akbar. News classification based on their headlines: A review. In *17th IEEE International Multi Topic Conference 2014*, pages 211–216. IEEE, 2014.

[RMCS22] Mozer de Miranda Ramos, Rodrigo de Oliveira Machado, and Elder Cerqueira-Santos. "it's true! i saw it on whatsapp": Social media, covid-19, and political-ideological orientation in brazil. *Trends in Psychology*, 30(3):570–590, 2022.

[RRP19] Robert M Ross, David G Rand, and Gordon Pennycook. Beyond "fake news": The role of analytic thinking in the detection of inaccuracy and partisan bias in news headlines. *PsyArXiv*, 2019.

[RSB+19] Filipe N Ribeiro, Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabricio Benevenuto, Oana Goga, Krishna P Gummadi, and Elissa M Redmiles. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 140–149, 2019.

[SAB+18] Rose Marie Santini, Larissa Agostini, Carlos Eduardo Barros, Danilo Carvalho, Rafael Centeno de Rezende, Debora G Salles, Kenzo Seto, Camyla Terra, and Giulia Tucci. Software power as soft power. a literature review on computational propaganda effects in public opinion and political process. *Partecipazione e Conflitto*, 11(2):332–360, 2018.

[San21] Rose Marie Santini. *A Indústria da Desinformação: Fábrica de Mentiras, AdTechs e as Novas Formas de Resistência*, pages 122–138. Intercom, 2021.

[SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[SDP22] PYKL Srinivas, Amitava Das, and Viswanath Pulabaigari. Fake spreader is narcissist; real spreader is machiavellian prediction of fake news diffusion using psycho-sociological facets. *Expert systems with applications*, 207:117952, 2022.

[SMW+20] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[SOU20] Luís Fred Gonçalves de SOUSA. Uso de aprendizado supervisionado multivisão para atribuição automática de autoria de textos. Master's thesis, Universidade Federal de Pernambuco, 2020.

[SSS+16] Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate, Oct 2016.

[SST21] Rose Marie Santini, Debora Salles, and Giulia Tucci. When machine behavior targets future voters: the use of social bots to test narratives for political campaigns in brazil. *International Journal of Communication*, 15(0):1220–1223, 2021.

[SSW+17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[STCL13] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1679–1684, 2013.

[SWL17] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 8, 2017.

[SWL19] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 312–320, 2019.

[SZW+19] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439, 2019.

[TJLL18] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital Journalism*, 6(2):137–153, 2018.

[TMS+23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Ünv23] Akın Ünver. Emerging technologies and automated fact-checking: Tools, techniques and algorithms. *Techniques and Algorithms (August 29, 2023)*, 2023.

[VdLPR20] Sander Van der Linden, Costas Panagopoulos, and Jon Roozenbeek. You are fake news: Political bias in perceptions of fake news. *Media, Culture & Society*, 42(3):460–470, 2020.

[VM20] Federico Vegetti and Moreno Mancosu. The impact of political sophistication and motivated reasoning on misinformation. *Political Communication*, 37(5):678–695, 2020.

[VRA18] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[VZ22] Dave Van Zandt. Media bias/fact check. https://mediabiasfactcheck.com/, 2022.

[Wan17] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[ZG20] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.

[ZZ20] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis).