Verificação de Fatos com Transformers: Um Estudo com DistilBERT no Benchmark FEVER

Beneilton Martins Leite[1], **Fael Faray de Paiva**[1], **Anselmo Cardoso de Paiva**[1]

[1]Núcleo de Computação Aplicada – Univesidade Federal do Maranhão (UFMA)
Caixa Postal: 65.085-580 – São Luís, MA – Brasil

{beneilton.martins,fael.paiva,anselmo.paiva}@nca.ufma.br

Abstract. *Automated fact-checking is an essential task in the digital age to combat misinformation. This work investigates claim verification using the FEVER dataset with a supervised transformer-based model. Trained on provided claims and evidence, our model achieves 91.98%, showing that high performance is possible even with lightweight architectures. The results highlight that compact models such as DistilBERT can achieve performance comparable to larger ones, underscoring the potential of efficient solutions for large-scale fact-checking systems.*

Resumo. *A verificação automatizada de fatos é uma tarefa essencial na era digital para combater a desinformação. Este trabalho investiga a classificação da veracidade de afirmações no dataset FEVER, adotando um modelo supervisionado baseado em transformers. O modelo alcança 91,98% de acurácia, demonstrando que é possível obter alto desempenho com arquiteturas leves. Os resultados evidenciam que modelos compactos, como o DistilBERT, podem alcançar desempenho comparável a modelos maiores, reforçando o potencial de soluções eficientes para verificação de fatos em larga escala.*

1. Introdução

O crescimento exponencial das redes sociais e da disseminação de conteúdo digital nos últimos anos transformou o panorama informacional contemporâneo, trazendo à tona desafios críticos relacionados à desinformação e à propagação de afirmações falsas ou enganosas [Guo et al. 2022]. Nesse contexto, a verificação automática de fatos tornou-se um problema fundamental no processamento de linguagem natural (PLN), uma vez que o volume massivo de informações em circulação em plataformas digitais torna impraticável a verificação manual de todas as reivindicações. Assim, o desenvolvimento de ferramentas automatizadas capazes de identificar, recuperar evidências e classificar a veracidade de afirmações é essencial para mitigar os efeitos adversos da desinformação na sociedade.

Como resposta a essa demanda, o benchmark FEVER (Fact Extraction and VERification) emergiu como referência fundamental para treinar e avaliar sistemas de verificação de fatos [Thorne et al. 2018]. O FEVER estrutura o problema de verificação como um desafio composto por três subtarefas interdependentes: (i) a recuperação de documentos e sentenças relevantes que contenham evidências potenciais; (ii) a verificação da consistência lógica entre as evidências recuperadas e a afirmação submetida; e (iii) a classificação da veracidade da reivindicação em uma das três categorias: SUPPORTS (Sustentada), REFUTES (Refutada) ou NOT ENOUGH INFO (Informação Insuficiente). Essa estruturação

modular permitiu o desenvolvimento de sistemas mais eficazes e a identificação de gargalos específicos em pipelines de verificação.

Paralelamente, a eficiência computacional tornou-se uma consideração crítica no desenvolvimento de sistemas de verificação em larga escala. Embora modelos de linguagem grandes, como o BERT, demonstrem alto desempenho em diversas tarefas de PLN, seus requisitos computacionais elevados limitam sua aplicabilidade em contextos com restrições de recursos, como dispositivos móveis e servidores com orçamentos computacionais limitados. Nesse cenário, o DistilBERT, um modelo obtido por destilação de conhecimento, apresenta-se como uma alternativa promissora, mantendo 97% do desempenho do BERT enquanto reduz seu tamanho em 40% e aumenta a velocidade de inferência em 60% [Sanh et al. 2019]. Essa otimização torna viável a implementação de sistemas de verificação eficientes e escaláveis sem comprometer significativamente a qualidade das predições.

O objetivo geral deste trabalho é contribuir para o avanço da verificação automática de fatos, com foco específico na etapa de verificação de consistência, que constitui um componente crítico do pipeline de verificação proposto pelo FEVER. Como objetivo específico, busca-se investigar a eficácia do DistilBERT, um modelo compacto baseado em arquitetura Transformer, na tarefa de classificar a relação entre uma afirmação (claim) e um conjunto de evidências extraídas. O modelo proposto recebe como entrada a concatenação da afirmação com a evidência e retorna a predição da classe apropriada, operando como classificador de rótulos ternários. Espera-se que esta pesquisa demonstre a viabilidade de empregar modelos destilados para a verificação de fatos, abrindo caminhos para a implementação de sistemas robustos e eficientes em ambientes onde recursos computacionais são limitados.

2. Trabalhos Relacionados

Desde o lançamento do modelo BERT original por Devlin et al. (2019), diversas melhorias e variantes vêm sendo desenvolvidas, impactando diretamente o desempenho de sistemas de verificação de fatos.

Trabalhos clássicos como [Nie et al. 2019] e [Hanselowski et al. 2018] utilizaram versões do BERT-base ou BERT-large pré-treinadas até 2019, época em que técnicas de fine-tuning e regularização ainda estavam em estágios iniciais.

O modelo DistilBERT [Sanh et al. 2019] oferece uma versão mais compacta e eficiente do BERT, mantendo cerca de 97% do desempenho do BERT-base, com menor custo computacional. Além disso, avanços no pré-treinamento e na regularização tornaram essas versões destiladas ainda mais eficazes.

Essa tendência é reforçada por linhas de pesquisa recentes e complementares. O trabalho mais recente das redes *GhostNet* [Han et al. 2020], *GhostNetV3* [Liu et al. 2024], aprimora a eficiência do treinamento e mantém desempenho próximo ao de redes muito maiores em tarefas de visão computacional usando de redes maiores para aplicar *Knowledge Distillation* assim como o modelo usado no presente artigo. Já o *Tiny Recursive Model* [Jolicoeur-Martineau 2025] apresenta uma abordagem de redes extremamente pequenas com conectividade recursiva, mostrando que estruturas compactas podem generalizar bem e até superar modelos maiores em determinadas tarefas.

Esses trabalhos reforçam a ideia central deste estudo: modelos menores e otimiza-

dos, quando bem projetados, podem manter ou até superar o desempenho de arquiteturas de grande porte, equilibrando eficiência e precisão.

3. Materiais e Métodos

Neste trabalho, abordamos a tarefa de verificação de afirmações do benchmark FEVER, focando exclusivamente na etapa de classificação de veracidade, assumindo que as evidências corretas já foram previamente fornecidas.

O objetivo foi treinar um modelo capaz de determinar se uma evidência textual suporta ou refuta uma determinada afirmação (claim). Além de simplificar a tarefa para fins de avaliação controlada, essa abordagem também visa desenvolver um componente leve e eficiente que possa, futuramente, ser integrado ao sistema RefChecker [Hu et al. 2024], onde as evidências seriam obtidas por um módulo externo de recuperação.

3.1. Conjunto de Dados

Cada instância do conjunto de dados é composta por uma claim (afirmação textual cuja veracidade deve ser verificada), uma evidence (sentença extraída da Wikipedia e previamente associada à claim) e um rótulo que indica a relação entre ambas. Os rótulos possíveis são: SUPPORTS, quando a evidência apoia a afirmação; REFUTES, quando a contradiz; e NOT ENOUGH INFO (NEI), quando não há informação suficiente para concluir.

Para o treinamento supervisionado, cada par claim + evidência é tratado como uma instância independente. Nesta versão inicial do trabalho, a classe NEI foi removida para focar na verificação com evidências explícitas. O conjunto de dados FEVER possui aproximadamente 185.445 claims, rotuladas em uma das três classes: SUPPORTS, REFUTES ou NOT ENOUGH INFO (NEI). Para este trabalho, utilizamos apenas os exemplos com evidências válidas, ou seja, aqueles rotulados como SUPPORTS ou REFUTES resultando em:

Conjunto	SUPPORTS	REFUTES	Total	Observação
Treinamento (TRAIN)	79.551	29.584	109.135	Dataset desbalanceado
Validação (DEV)	6.615	6.614	13.229	Dataset balanceado

Tabela 1. Distribuição das classes SUPPORTS e REFUTES nos conjuntos de treinamento e de validação.

Ou seja, enquanto o conjunto de treinamento está desbalanceado (SUPPORTS: aproximadamente 72.9%, REFUTES: aproximadamente 27.1%), o conjunto de validação foi propositalmente balanceado para permitir uma avaliação mais justa e imparcial do modelo.

3.2. Pré-processamento

Foi utilizado o modelo DistilBERT, uma versão compacta do BERT, com entrada no formato `[CLS] afirmação [SEP] evidências [SEP]`. Essa estrutura permite que o modelo processe a relação entre a afirmação e a evidência em conjunto. O truncamento foi aplicado para respeitar o limite de 512 tokens do modelo.

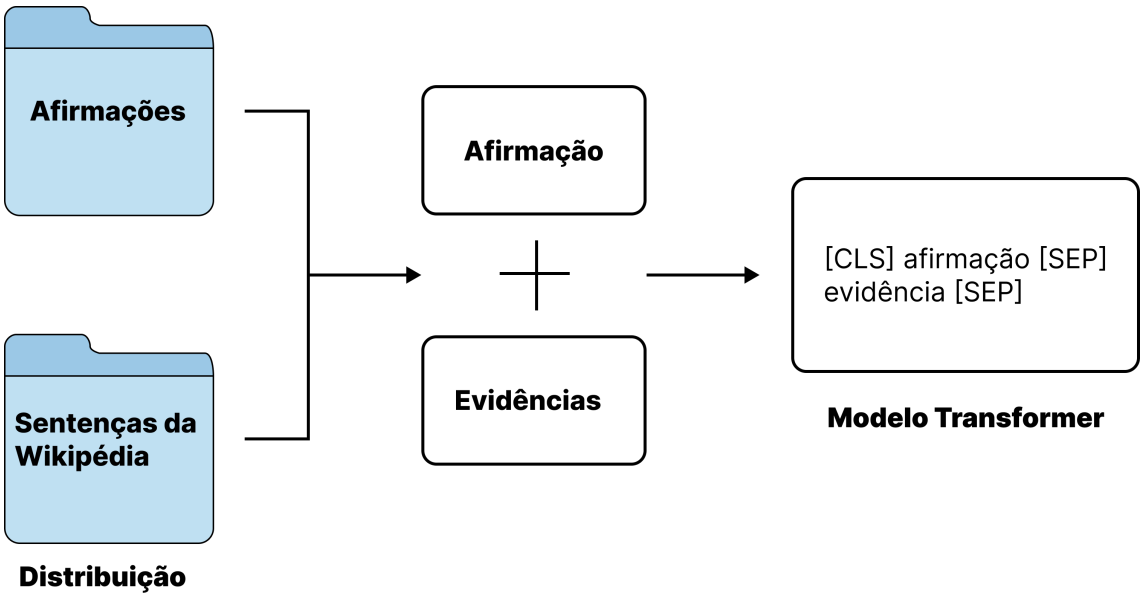

Figura 1. Esquema de entrada para o modelo transformer.

As afirmações *(claims)* e suas respectivas evidências *(evidences)*, extraídas da Wikipedia, são combinadas em uma única sequência formatada como "[CLS] afirmação [SEP] evidência [SEP]" e alimentadas no modelo para classificação.

As classes textuais foram transformadas em inteiros por meio de codificação (LabelEncoder). O conjunto de validação (dev) apresenta, originalmente, proporções equivalentes entre as classes SUPPORTS e REFUTES, conforme definido pelos organizadores do benchmark FEVER. O conjunto de treinamento também manteve sua distribuição original, com predominância da classe SUPPORTS. A classe NOT ENOUGH INFO foi excluída desta fase inicial, de forma a focar exclusivamente em exemplos com evidência explícita.

3.3. Arquitetura do Modelo

O modelo empregado neste estudo foi o DistilBERT-base-uncased, uma versão compacta e eficiente da família BERT, escolhida por oferecer um bom equilíbrio entre desempenho e custo computacional. Sobre ele, foi acoplado um cabeçalho de classificação construído com a classe AutoModelForSequenceClassification, configurado originalmente com três unidades de saída, correspondentes às classes SUPPORTS, REFUTES e NOT ENOUGH INFO do benchmark FEVER.

Contudo, nesta etapa do trabalho, somente as classes SUPPORTS e REFUTES foram utilizadas efetivamente no treinamento, uma vez que o conjunto de dados empregado já excluía exemplos rotulados como NEI. Por esse motivo, embora a arquitetura mantivesse três logits para testes futuros, apenas dois deles contribuíam para esta etapa de otimização. Para converter esses logits em probabilidades, foi aplicada a função de ativação softmax na camada final, permitindo estimar a distribuição de probabilidade sobre as classes consideradas. Dessa forma, o modelo pôde aprender a distinguir de maneira robusta entre afirmações suportadas e refutadas com base nas evidências fornecidas.

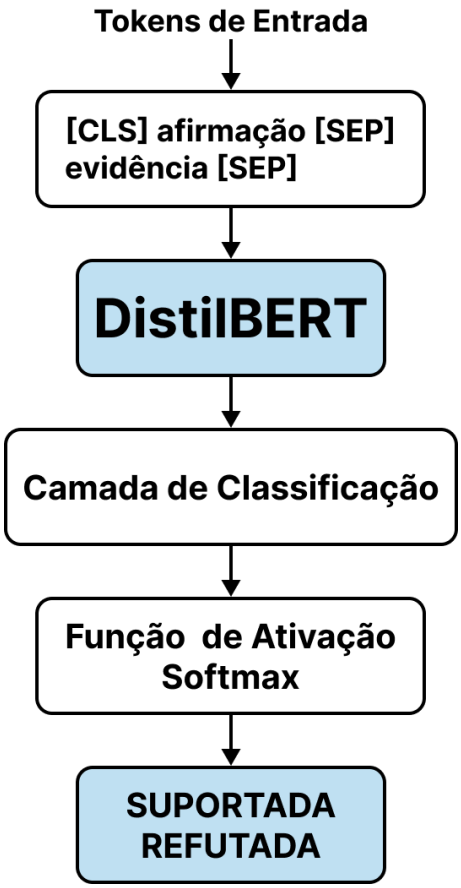

Figura 2. Arquitetura do modelo de verificação de fatos baseado em DistilBERT.

A entrada consiste nos tokens da afirmação (claim) e da evidência (evidence), separados pelos tokens especiais [CLS] e [SEP]. Esses tokens são processados pelo modelo DistilBERT, seguido por uma camada de classificação e uma função Softmax para determinar se a afirmação é SUPPORTS (suportada) ou REFUTES (refutada) pela evidência fornecida.

3.4. Configurações de Treinamento

O treinamento foi realizado com o framework PyTorch, em conjunto com a biblioteca Hugging Face Transformers. O otimizador adotado foi o AdamW, com taxa de aprendizado (learning rate) definida em 2×10^{-5}. O tamanho do lote (batch size) foi fixado em 16, e o número total de épocas de treinamento foi 3. A função de perda utilizada foi a CrossEntropyLoss, sem aplicação de pesos para balanceamento entre classes. A principal métrica de avaliação considerada foi a acurácia.

O treinamento foi realizado com o conjunto desequilibrado, refletindo a distribuição real das evidências no corpus. Esta decisão buscou expor o modelo à distribuição natural dos dados que ele encontraria em cenários práticos de verificação de fatos, onde afirmações suportadas por evidências tendem a ser mais frequentes que as refutadas. No entanto, reconhece-se que este desbalanceamento pode induzir vieses no classificador, fazendo que ele tenda a favorecer a classe majoritária.

A validação, no entanto, utilizou um conjunto balanceado, o que ajuda a garantir

que a avaliação do modelo não seja enviesada pela classe majoritária. Esta abordagem híbrida, treinamento com dados desbalanceados e validação com dados balanceados, permite que o modelo aprenda a distribuição real dos dados enquanto possibilita uma avaliação justa de sua capacidade de generalização para ambas as classes. A matriz de confusão apresentada posteriormente confirma essa estratégia, revelando como o modelo lidou com o desafio do desbalanceamento durante o treinamento.

4. Resultados e Discussões

Os resultados obtidos pelo modelo proposto foram comparados aos de trabalhos relevantes da literatura que atuam na tarefa de verificação de afirmações com base no dataset FEVER. É importante destacar que, neste trabalho, assumimos que as evidências relevantes já foram fornecidas, com foco exclusivo na etapa de verificação da veracidade.

Durante o treinamento do modelo, foi possível observar uma redução consistente na training loss ao longo das épocas, o que indica que o modelo foi capaz de aprender padrões relevantes para a tarefa de classificação binária entre as classes SUPPORTS e REFUTES. Esse comportamento pode ser visualizado no gráfico a seguir, que ilustra a evolução da função de perda ao longo das três épocas de treinamento. A convergência da curva sugere que o modelo não sofreu de sobreajuste prematuro e que os hiperparâmetros adotados foram adequados ao cenário proposto.

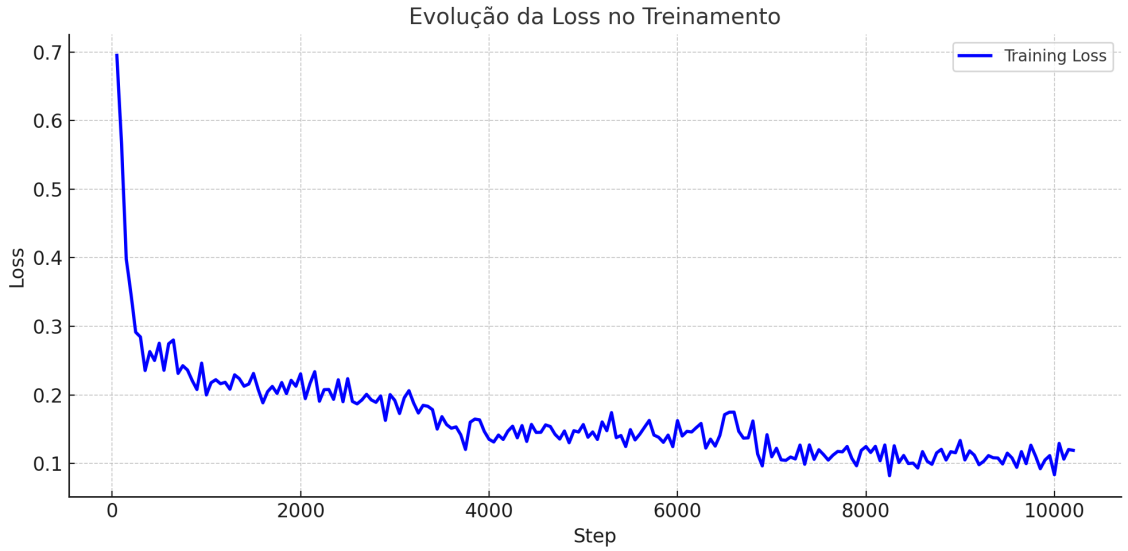

Figura 3. Evolução da *training loss* durante o treinamento do modelo DistilBERT.

A convergência estável observada na função de perda indica que o modelo assimilou efetivamente os padrões de relação entre afirmações e evidências. No entanto, para avaliar qualitativamente o tipo de aprendizado desenvolvido pelo classificador, é essencial analisar sua performance preditiva de forma mais granular. A matriz de confusão, apresentada na figura 4a seguir, permite identificar não apenas a acurácia global do modelo, mas também padrões específicos de acertos e erros entre as classes SUPPORTS e REFUTES.

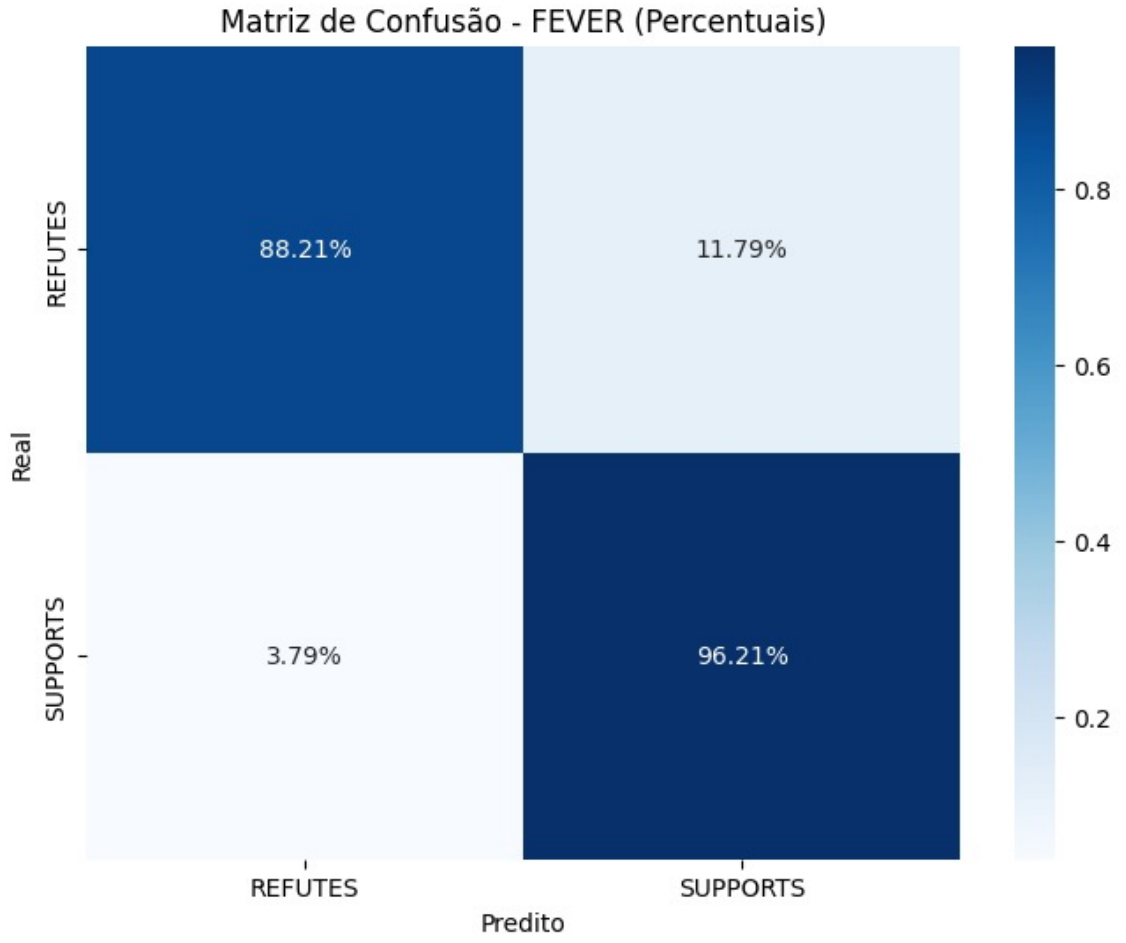

Figura 4. Matriz de confusão do modelo DistilBERT no conjunto de validação balanceado.

A análise da matriz de confusão revela um desempenho geral robusto, porém com um padrão assimétrico interessante entre as classes. O modelo demonstrou excelente capacidade de identificar afirmações SUPPORTS, alcançando 6.364 acertos (96,2% de recall) com apenas 251 falsos negativos (3,8%). Por outro lado, embora a performance na classe REFUTES também seja sólida com 5.834 acertos (88,2% de recall), observa-se uma taxa de erro significativamente maior nesta classe, com 780 falsos positivos (11,8%), casos onde afirmações refutadas foram erroneamente classificadas como suportadas.

Esta assimetria sugere que o modelo desenvolveu um viés conservador, tendendo a classificar afirmações ambíguas ou de difícil verificação como ”suportadas”. Esse comportamento pode ser atribuído à distribuição desbalanceada do conjunto de treinamento, onde a classe SUPPORTS é majoritária ($\approx 73\%$), influenciando o modelo a preferir esta classe em casos de maior incerteza. Os erros residuais concentram-se provavelmente em casos que exigem inferências linguísticas complexas, conhecimento de mundo externo ou onde a relação entre a afirmação e a evidência apresenta nuances semânticas difíceis de capturar.

Os resultados indicam que, mesmo com uma arquitetura mais leve, é possível alcançar altos níveis de acurácia. O balanceamento do conjunto de validação garante

que as métricas reflitam adequadamente o desempenho em ambas as classes. O uso de DistilBERT proporciona ganhos de eficiência computacional, tornando o modelo atraente para aplicações em larga escala. Alguns dos trabalhos listados abaixo incluem tanto a recuperação automática de evidências quanto a verificação, o que pode prejudicar suas métricas globais. Outros, assim como este estudo, realizam a verificação assumindo evidências previamente selecionadas, o que permite uma comparação mais direta dos classificadores em si.

A Tabela 2 resume a acurácia reportada por cada modelo. Observa-se que o modelo proposto, baseado em DistilBERT, alcança desempenho competitivo, com acurácia de 91,98%, superando a maioria dos trabalhos anteriores, mesmo sendo uma arquitetura mais leve.

Modelo	**Acurácia (%)**
Este trabalho (DistilBERT, sem recuperação)	91,98
[Thorne et al. 2018] (DA, com recuperação[1])	88,00
[Hanselowski et al. 2018] (ESIM, com recuperação[2])	68,49
[Nie et al. 2019] (vNSMN, com recuperação[3])	69,90
[Soleimani et al. 2020] (BERT(Large), com recuperação [4])	74,59
[Casillas et al. 2022] (BERT-Based, com recuperação)	88,73

Tabela 2. Comparação de acurácia com outros trabalhos na tarefa de verificação de afirmações utilizando o dataset FEVER.

Diversos trabalhos têm explorado modelos baseados em transformers para a tarefa de verificação de fatos. Por exemplo, [Liu et al. 2019] utilizou o modelo RoBERTa, alcançando alta acurácia, porém com maior custo computacional. Outros trabalhos, como [Thorne et al. 2018] e [Hanselowski et al. 2018], exploram tanto a recuperação automática de evidências quanto a classificação da veracidade, o que torna a tarefa mais complexa. Nosso trabalho foca na etapa de classificação assumindo evidências previamente fornecidas, similar a [Nie et al. 2019] (este, porém, também faz recuperação, mas fornece dados com as chamadas *Golden Evidences*, assumindo o caso em que todas as evidências são as melhores e mais corretas), porém, com a vantagem do uso de uma arquitetura mais leve.

Outros trabalhos recentes também utilizam o FEVER com evidências previamente fornecidas, concentrando-se exclusivamente na etapa de classificação. Por exemplo, o estudo de [Casillas et al. 2022] propõe uma arquitetura baseada em BERT, com atenção interpretável, voltada a afirmações sobre COVID-19, mas avalia a capacidade de generalização do seu modelo usando o FEVER como *benchmark*. Nesse caso, as evidências já são fornecidas no par entrada (afirmação + evidência), sem envolvimento da etapa de recuperação automática. Essa abordagem é similar à adotada neste trabalho, reforçando a viabilidade do uso de arquiteturas mais enxutas em cenários em que as evidências são previamente selecionadas.

5. Conclusão

Neste trabalho, apresentamos um modelo baseado em DistilBERT para a tarefa de verificação de afirmações utilizando o dataset FEVER. O foco foi na etapa de classificação

da veracidade com base em afirmações e evidências fornecidas. Os resultados obtidos, com acurácia de 91,98%, destacam a eficácia do modelo e sua competitividade em relação à literatura existente.

A acurácia de 91,98% obtida em validação reflete a efetiva capacidade do modelo em distinguir entre afirmações suportadas e refutadas, uma vez que a avaliação foi realizada em um conjunto balanceado entre essas duas classes. Esse balanceamento, definido pelos criadores do benchmark FEVER, assegura que a métrica de acurácia seja representativa e não distorcida por uma distribuição enviesada de rótulos.

Apesar dos resultados promissores apresentados, este trabalho ainda não aborda plenamente todos os aspectos da verificação automática de fatos. Uma importante direção para trabalhos futuros é a inclusão da terceira classe Not Enough Information (NEI) no modelo, o que permitirá uma classificação ternária mais realista e alinhada ao desafio proposto pelo benchmark FEVER. A incorporação dessa classe permitirá ao sistema distinguir melhor entre afirmações para as quais há evidências claras e aquelas para as quais a informação disponível é insuficiente, aprimorando a precisão e a utilidade prática do modelo.

Além disso, a adaptação do modelo para uso como verificador (checker) no benchmark RefChecker [Hu et al. 2024] representa uma oportunidade significativa para ampliar a aplicabilidade do sistema. Essa integração permitirá avaliar o desempenho do modelo em um cenário diferente, com dados e requisitos específicos do RefChecker, contribuindo para a validação e generalização da abordagem em contextos mais amplos. A combinação dessas duas melhorias, a inclusão da classe NEI e a adaptação para o RefChecker, tem o potencial de tornar o sistema mais robusto e versátil para aplicações reais na área de verificação de fatos.

Referências

Casillas, R., Gómez-Adorno, H., Lomas-Barrie, V., and Ramos-Flores, O. (2022). Automatic fact checking using an interpretable bert-based architecture on covid-19 claims. *Applied Sciences*, 12(20):10644.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. (2020). Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589.

Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., and Gurevych, I. (2018). UKP-athene: Multi-sentence textual entailment for claim verification. In Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A., editors,

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Hu, X., Ru, D., Qiu, L., Guo, Q., Zhang, T., Xu, Y., Luo, Y., Liu, P., Zhang, Y., and Zhang, Z. (2024). Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*.

Jolicoeur-Martineau, A. (2025). Less is more: Recursive reasoning with tiny networks. *arXiv preprint arXiv:2510.04871*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Hao, Z., Han, K., Tang, Y., and Wang, Y. (2024). Ghostnetv3: Exploring the training strategies for compact models. *arXiv preprint arXiv:2404.11202*.

Nie, Y., Chen, H., and Bansal, M. (2019). Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Soleimani, A., Monz, C., and Worring, M. (2020). Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, page 359–366, Berlin, Heidelberg. Springer-Verlag.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.