Automated Fact Checking

Matheus C. Lindino

01 de Dezembro de 2022

Abstract

Este trabalho explora dois sistemas distintos para a tarefa de verificação de alegações científicas no domínio biomédico. Os dois sistemas tem três principais módulos: (1) recuperação de documentos relevantes, (2) seleção de sentenças e (3) classificação dos documentos caso apoiam ou refutam a premissa. O primeiro modelo é baseado no BERT, utilizando o *token* $[CLS]$ para classificar, enquanto o segundo utiliza o modelo T5, prevendo *tokens* específicos em cada módulo. Os dois sistemas foram avaliados no conjunto de dados *SciFact*, um *dataset* que requer que os modelos não apenas identificam a veracidade das alegações, mas também devem fornecer sentenças relevantes de corpus científicos, para apoiar a decisão. Esses dois sistemas tiveram resultados comparativos com os encontrados na literatura.

1 Introdução

Nas últimos anos, foi cunhado o termo *Fake News* para caracterizar notícias falsas publicadas por veículos de comunicação, muitas vezes em canais não oficiais. Esse movimento acabou se tornando uma poderosa ferramenta para a desinformação, uma vez que tem um grande poder viral, espalhando-se rapidamente. O fraco discernimento da verdade está associado à falta de raciocínio cuidadoso e conhecimento relevante do assunto. Além disso, há uma desconexão substancial entre o que as pessoas acreditam e o que compartilham nas mídias sociais, o que acaba impulsionando a transmissão dessas notícias [13].

Por sua vez, com a vinda da pandemia do COVID-19 em 2020, foi possível quantificar o estrago da *Fake News* na sociedade. De acordo com a Organização Pan-Americana de Saúde (OPAS), em abril do mesmo ano, as pesquisas na internet sobre a doença cresceram entre 50% e 70% nos públicos das diversas faixas etárias em todo o mundo [4]. Esse aumento nas buscas, muitas vezes ligadas a desinformação, fez com que as organizações internacionais alertassem sobre a desinfodemia, entendida como um processo sistemático de desinformação com o objetivo de deslegitimar a ciência médica voltada para a Covid-19, com impacto na vida de um grande número de pessoas – segundo a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (Unesco) [3].

Assim, a checagem de fatos consiste em uma tarefa de avaliar se reivindicações realizadas em diferentes meios de comunicação são verdadeiras. Esta é uma tarefa essencial no jornalismo, no qual tem o objetivo de filtrar as notícias falsas, diminuindo assim a disseminação de informações questionáveis [10]. Todavia, por causa da massiva quantidade de informações circulando no meio digital, se tornou impossível realizar a verificação de cada notícia manualmente. Portanto, técnicas de Processamento de Linguagem Natural (PLN) podem auxiliar a automatização do processo de checagem de fatos, classificando as informações, mas também identificando artigos que colaboram ou refutam tal informação.

Pesquisas voltadas a automatizar o processo de checagem de fatos geralmente se concentram em três objetivos: identificação das reivindicações, verificação de evidências apropriadas e produção de um veredito [18]. Na Figura 1 mostra o esse fluxograma simples, no qual consiste em:

- *Claim Detection:* Etapa que identifica se uma premissa requer verificação. Algumas alegações são rumores ou afirmações de conhecimento popular (e.g. a água é molhada) e devem ser subtraídas para a análise do modelo.
- *Evidence Retrieval:* Etapa que visa encontrar informações relevantes que auxiliam na identificação de veracidade da premissa. As informações podem ser textos longos ou apenas uma sentença.
- *Claim Verification:* Etapa de veredito da amostra. Nessa etapa, o modelo deve ser capaz de identificar a veracidade da premissa, além de justificar a decisão.

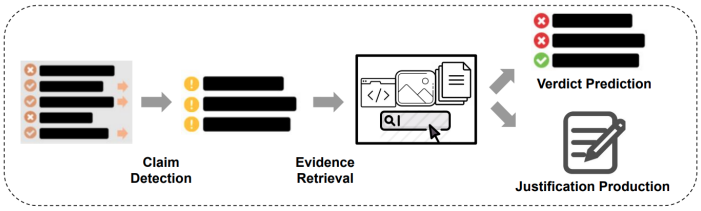

Figure 1: Uma estrutura de processamento de linguagem natural para verificação automatizada de fatos. Fonte: [6].

Diversos trabalhos na literatura seguem esse fluxograma, dividindo o problema em pequenas atividades. Zhang et al. propõem um sistema denominado de ARSJOINT, no qual utiliza o modelo RoBerta [8] para aprender conjuntamente três tarefas (*abstract retrieval*, *rationale selection* e *stance prediction*) com uma estrutura *Machine Reading Comprehension* (MRC). Também, para minimizar o erro de propagação entre as tarefas, eles propõem um termo de regularização entre os *scores* de atenção e os modelos responsáveis pela tarefa *abstract retrieval* e *rationale selection* [21].

Wadden et al. (2022) apresenta o modelo MULTIVERS (do inglês - *Multi-task Verification for Science*) no qual dado uma alegação e evidências científicas, esse modelo cria um *encoding* compartilhado de toda a reivindicação/*abstract*, utilizando o Longformer [2] para processar sequências longas [20]. Em seguida, o MULTIVERS prevê um rótulo para cada premissa em dois níveis: 1) *abstract-level* e *sentence-level*, reforçando a consistência entre as saídas das duas tarefas durante a decodificação. Por fim, Rana et al. (2022) propõem o sistema denominado RERRFACT, no qual consiste em vários classificadores binários para cada atividade (todos baseados no Bert – RoBerta e BioBert) [16].

Desta forma, o presente trabalho visa construir um sistema completo de verificação automática de factualidade, explorando as três etapas principais: *abstract retrieval*, *sentence selection* e *label prediction*. O conjunto de dados utilizado para a obtenção dos resultados foi o *SciFact* [19].

2 Metodologia

A metodologia desse trabalho foi semelhante a arquitetura *VerT5erini* [14]. A Figura 2 apresenta os principais módulos desse sistema.

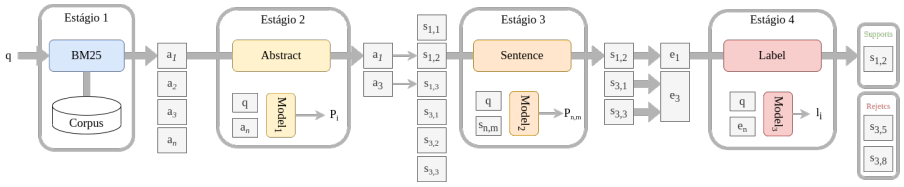

Figure 2: Ilustração do sistema completo de verificação de factualidade. Inspirado em [14].

O sistema possuí três principais componentes:

1. **Abstract Retrieval:** Dado uma afirmação científica c, retorna os top-k resumos mais relevantes do corpus $\mathcal{C}$ (Engloba estágio 1 e 2);
2. **Sentence Selection:** Dado uma afirmação científica c e um dos top-k resumos relevantes a, seleciona sentenças mais relevantes ($\mathcal{S}(c, a)$) do resumo a;
3. **Label Prediction:** Dado uma afirmação científica c e um conjunto de sentenças $\mathcal{S}(c, a)$, identifica o rótulo adequado $y(c, a)$.

2.1 Abstract Retrieval

Dado uma afirmação científica c e um corpus $\mathcal{C}$ de resumos de artigos científicos, esse módulo tem como objetivo recuperar os top-k resumos mais relevantes. Para tanto, foi utilizado dois estágios: (1) utilizando algoritmos tradicionais de busca de documentos e (2) utilizando técnicas de PLN.

O primeiro estágio envolve classificar os documentos do corpus, utilizando o *score* do BM25 [17]. Essa função de pontuação utiliza a técnica de *bag-of-words* para consultar a quantidade de termos que a premissa tem em comum com cada documento. Como há variações dessa função, foi implementado utilizando duas bibliotecas diferentes, a fim de mensurar a qualidade das duas: a primeira utiliza o *kit* de ferramentas *Pyserini* [7] e a outra utiliza o *Rank-BM25* [1]. A saída desta etapa é uma lista de k resumos.

O segundo estágio tem como objetivo reclassificar os resumos obtidos na etapa anterior, com o intuito de afunilar os documentos. Para tanto foi utilizado dois modelos: (1) *MonoBERT* e (2) *MonoT5*. O *MonoBERT* [11] é um modelo baseado no BERT [5], no qual foi realizado um *fine-tunning* para classificar caso o resumo a é relevante ou não para a afirmação científica c. Para tanto, o *token* $[CLS]$ é utilizado como entrada para uma única camada neural para obter a probabilidade do documento ser relevante. Assim, os resumos top-k relevantes são aqueles com as k maiores probabilidades.

Já o *MonoT5* [12] é baseado no T5 [15], um modelo *seq2seq* no qual todas as tarefas tem como entrada e saída um texto. Portanto, foi necessário adaptar o texto de entrada para $Query : c\ Document : a\ Relevant :$, no qual c é a premissa e a é o resumo. Assim é realizado o *fine-tunning* desse modelo para prever as palavras *true* ou *false* para indicar se o documento é relevante ou não para a afirmação c. Diferente do *MonoBERT*, para a inferência desse modelo, é necessário calcular a probabilidade para cada par afirmação-resumo aplicando uma *softmax* apenas no *logits* dos *tokens true* e *false*. Os resumos top-k são aqueles com a maior probabilidade atribuída ao *token true*. Vale ressaltar que esses dois modelos foram ajustados para o *dataset* MS MARCO [1].

2.2 Sentence Selection

Nesse módulo, o objetivo é selecionar as sentenças $\mathcal{S}(c, a)$ de cada resumo a dos top-k documentos selecionados na etapa anterior. Para isso, foi utilizado os mesmo modelos da etapa *abstract-retrieval*. Para utilizar o *MonoT5*, a sequência de entrada foi modificada para: $Query : c\ Document : s\ Relevant :$, no qual c é a premissa e s é uma sentença do resumo a.

Para ajustar os dois modelos (*MonoBERT* e *MonoT5*), foi realizado um *fine-tunnig* no conjunto do *SciFact*, utilizando as sentenças ouro como positivas e sentenças aleatórias dos documentos retirados na etapa anterior como negativas. Para a inferência, como o modelo *MonoT5* é *seq2seq*, foi calculado a probabilidade da sentença ser relevante com base nos *logits* dos *tokens true* e *false*. As sentenças utilizadas foram aquelas com 0,99 de probabilidade do *token true*. No *MonoBERT* é necessário apenas visualizar a saída da rede neural para obter $\mathcal{S}(c, a)$.

[1] https://github.com/dorianbrown/rank_bm25

2.3 Label Prediction

Dado uma afirmação científica c, um resumo a e as sentenças $\mathcal{S}(c, a)$, esse módulo tem como objetivo classificar $y(c, a) \in \{Supports, Refutes, NoInfo\}$. A sequência de entrada foi alterada para: $hypothesis : c\ sentence_1 :\ s_1\ ...\ sentence_z :\ s_z$, no qual $s_1\ ...\ s_z$ são as sentenças de $\mathcal{S}(c, a)$. Nessa etapa, foi utilizado os modelos base *Bert* e *T5*, uma vez que a tarefa dos modelos *MonoBERT* e *MonoT5* são completamente diferentes. A saída do BERT são os *indexes* de cada classe, enquanto para o T5 são as sequências *true*, *false* e *weak*, respectivamente.

Para a criação desse conjunto de dados, os exemplos de treinamento de *Support* e *Refutes* são evidências ouro já rotuladas manualmente. Já para a classe *NoInfo*, exemplos aleatórios de uma ou mais sentenças foram utilizadas, a partir dos top-k documentos adquiridos.

3 Conjunto de Dados

Como mencionado na seção 1, o conjunto de dados utilizado nesse trabalho foi o *SciFact* [19]. Esse *dataset* consiste em 1.409 afirmações científicas verificados em um corpus com 5.183 resumos de artigos científicos. Cada resumo que confirmam ou refutam as reivindicações, forma anotados com justificativas. A distribuição dos rótulos está apresentada na Tabela 1. Nota-se que o conjunto de dados é relativamente pequeno, com um desbalanceamento significativo das classes. Isso acaba mostrando a importância de utilizar técnicas de *zero-shot* ou *few-shot*.

Table 1: Distribuição dos rótulos do SciFact.

Set	Supports	Refutes	NoInfo	Total
Treino	332	173	304	809
Validação	124	64	112	300
Teste	100	100	100	300

Para construir o *SciFact*, os autores sub amostraram o corpus S2ORC, um corpus disponível publicamente com 81,1 milhões de artigos acadêmicos em inglês [9]. Para garantir que os documentos retirados desse corpus tenham uma alta qualidade, foi retirado artigos apenas de periódicos conceituados (e.g. Nature, Cell, JAMA e BMJ) e artigos que possuem pelo menos 10 citações. Além disso, para ampliar o corpus, foram identificados cinco artigos citados no mesmo artigo, adicionado-os como resumos distratores. Esses resumos geralmente discutem tópicos semelhantes aos documentos de evidência, aumentando a dificuldade de recuperação de resumos.

No *SciFact*, os sistemas são fornecidos com uma afirmação científica c e um corpus de resumos $\mathcal{A}$. Todos os resumos $a \in \mathcal{A}$ são rotulado como $y(c, a) \in \{Supports, Refutes, NoInfo\}$ em relação a afirmação c. O sistema deve ser encarregado de retornar:

- Um conjunto de resumos de evidências $\mathcal{E}(c)$;

- O rótulo $y(c, a)$ que mapeia a afirmação c e o resumo a;
- Um conjunto de sentenças $\mathcal{S}(c, a)$ que justificam o rótulo $y(c, a)$.

Assim, é possível avaliar a tarefa em dois níveis de granularidade: 1) *abstract-level* e 2) *sentence-level*. No *abstract-level*, é avaliado a capacidade do modelo de identificar os resumos que apoiam ou refutam a afirmação. Para tanto, dado uma afirmação c, uma evidência predita é classificado como *correctly labeled* caso o resumo a é uma evidência ouro e o rótulo predito está correto. É classificado como *correctly rationalized* se, além disso, contém uma sentença ouro. Estas avaliações são referidas como $Abstract_{Label-Only}$ e $Abstract_{Label+Rationale}$, respectivamente.

Já *sentence-level*, é avaliado o desempenho do modelo a identificar as sentenças que justificam as previsões realizads no *abstract-level*. Uma sentença é predita como *correctly selected* se (1) é uma sentença ouro, (2) todas as sentenças ouro estão preditas corretamente e (3) não é predita como $NoInfo$. É classificado como *correctly labelled,* se, além disso, o rótulo está correto. Estas avaliações são referidas como $Sentence_{Selection-Only}$ e $Sentence_{Selection+Label}$, respectivamente.

4 Resultados Experimentais

Essa seção tem como objetivo apresentar e discutir os resultados de cada módulo independente e o sistema como um todo. Vale ressaltar que todos os testes foram realizados utilizando o Jupyter Notebook em um servidor privado com uma placa de vídeo Nvidia RTX 5000, com 16GiB de mémoria.

4.1 Abstract Retrieval

A fim de avaliar os modelos, três métricas são bastante utilizadas para a tarefa de retirada de documentos: *precision*, *recall* e *F1 score*. Cada métrica pode ser vista com mais detalhes abaixo:

Precision = N. de documentos relevantes retirados / N. total de documentos retirados

Recall = N. de documentos relevantes retirados / N. total dos documentos relevantes

F1 Score = 2 * Precision * Recall / Precision + Recall

Vale ressaltar que, para obter os resultados dessa seção, foi utilizado apenas o conjunto de validação, uma vez que o conjunto de treinamento foi utilizado para realizar *fine-tunning* e o conjunto de teste não possuí rótulos para avaliação.

4.1.1 Estágio 1

No primeiro estágio, foi comparado o algoritmo BM25 de duas bibliotecas distintas. A Tabela 2 apresenta os resultados de *recall*, modificando o k para cada abordagem.

Table 2: Comparação entre diferentes abordagens do BM25

Abordagem	R@1	R@2	R@3	R@5	R@10	R@25	R@50	R@100
Rank BM25	55,27	65,3	68,8	74,78	80,36	86,4	88,39	90,65
Pyserini	**70,52**	**80,85**	**84,97**	**88,47**	**92,48**	**94,85**	**95,08**	**96,44**

Nota-se que a abordagem utilizado o kit de ferramentas *Pyserini* obteve os melhores resultados, quando comparado com o Rank BM25. Provavelmente, essa diferença é pela forma de tokenização que os algoritmos utilizam. Portanto, foi utilizado o BM25 do *Pyserini* para os próximos módulos.

4.1.2 Estágio 2

No segundo estágio, foi retirado os top-100 resumos relevantes do BM25 no estágio anterior, e utilizado tanto o *MonoBERT* quanto o *MonoT5* para reorganizar os documentos (apenas inferência). A Tabela 3 apresenta os resultados desses dois modelos, em termo de *recall* com três e cinco documentos mais relevantes.

Table 3: Comparação entre modelos do estágio 2

Abordagem	R@3	R@5
Oráculo	97,6	100
BM25	84,97	88,47
monoBERT + BM25	88,40	91,97
monoT5 + BM25	**89,33**	**92,9**
BM25 [14]	79,90	84,69
T5 (MS MARCO) [14]	86,12	89,95
T5 (SciFact) [14]	86,60	89,40

Nota-se que o oráculo mostra que grande parte das afirmações científicas do conjunto de validação possui menos de três documentos relevantes, enquanto todos possuem menos de cinco resumos relevantes. Além disso, os modelos obtiveram os melhores resultados quando comparado com modelos encontrados na literatura. Vale ressaltar que o tanto o *MonoBERT* quanto o *MonoT5* não foram ajustados para o conjunto de dados *SciFact*, uma vez que o trabalho de Pradeep et. al (2020) mostra que não há ganhos significativos em realizar o *fine-tunning* [14].

4.2 Sentence Selection

A Tabela 4 apresenta os resultados de cada modelo para a tarefa de seleção de sentenças, em termos de *precision* (P), *recall* (R) e *F1 score* (F1). Nota-se que o melhor modelo foi o T5, com *F1 score* igual a $76,14$. O modelo *MonoT5* foi retreinado com o conjunto de dados do *SciFact* (como comentado na seção

2.2),, obtendo resultados similares ao artigo referência [14]. Vale ressaltar que foi realizado *over sampling* nas amostras positivas para balancear os dados.

Table 4: Comparação entre modelos para seleção de sentenças

Abordagem	P	R	F1
monoBERT	74,8	71,3	73,00
monoT5	78,19	72,54	75,26
RoBERTa-large [19]	73,71	70,40	72,07
T5 [14]	**79,29**	**73,22**	**76,14**

4.3 Label Prediction

A Tabela 5 apresenta os resultados dos modelos na etapa de predição de rótulo, em termos de *precision* (P), *recall* (R) e *F1 score* (F1). Vale ressaltar que, para os rótulos *Support* e *Refutes*, a entrada dos modelos são as sentenças ouro com seus respectivos rótulos. Para o *NoInfo* são sentenças que estão presentes no resumo, mas não são ouro, selecionando as sentenças mais parecidas a partir do BM25. Vale ressaltar que foi realizado *over sampling* nas amostras positivas para balancear os dados.

Table 5: Comparação entre modelos para a etapa de predição de rótulos

Abordagem	Rótulo	P	R	F1
BERT	Supports	80,12	75,51	77,74
	Refutes	66,84	76,56	71,37
	NoInfo	75,79	81,20	78,40
T5	Supports	91,14	84,21	87,53
	Refutes	81,13	**83,64**	82,36
	NoInfo	**97,54**	78,39	86,92
T5 [14]	Supports	**93,13**	**88,41**	**90,71**
	Refutes	**86,76**	83,10	**84,89**
	NoInfo	85,25	**92,86**	**88,89**

Neste caso, todos modelos foram ajustados com suas versões bases, ou seja, não foram utilizados os modelos *MonoBERT* e *MonoT5* para essa atividade. Invés disso, foi realizado *fine-tunning* no *BERT-large* com 340M de parâmetros e no *T5-large* com 770M de parâmetros. Os resultados foram inferiores ao T5 referência.

4.4 Full Pipeline

Para avaliar o sistema inteiro, foi calculado três métricas, sendo elas *precision* (P), *recall* (R) e *F1 score* (F1). Como citado na seção 3, o sistema inteiro é avaliado de quatro formas diferentes: sendo duas no *abstract-level* e duas no

sentence-level. A Tabela 6 e 7 aprensentam os resultado ao nível de resumo, avaliando apenas a o rótulo atribuído ao documento e, além disso, verificando se há sentenças ouro, respectivamente.

Vale ressaltar que o oráculo apresentado nas Tabelas [6-9] apresenta sempre a evidência ouro para o sistema. Também, os modelos utilizados foram os mesmos apresentados no R@3 na Tabela 3.

Table 6: *Abstract-level evaluation* - apenas com o rótulo

Label Only			
Abordagem	P	R	F1
Oráculo	92,70	78,95	85,27
VeriSci	55,31	47,37	51,03
VerT5erini	**70,88**	**61,72**	**65,98**
BERTBased	24,61	26,78	25,65
T5Based	40,15	42,12	41,11

Table 7: *Abstract-level evaluation* - com sentenças

Label + Rationale			
Abordagem	P	R	F1
Oráculo	88,76	75,60	81,65
VeriSci	52,51	44,98	48,45
VerT5erini	**61,72**	**61,72**	**61,72**
BERTBased	23,05	20,12	21,48
T5Based	30,01	29,51	29,81

Table 8: *Sentence-selection evaluation* - apenas sentenças

Selection Only			
Abordagem	P	R	F1
Oráculo	83,54	72,13	77,42
VeriSci	52,46	43,72	47,69
VerT5erini	**64,81**	**57,37**	**60,87**
BERTBased	30,11	31,20	30,64
T5Based	40,12	42,10	41,09

Nota-se que os sistemas *BERTBased* e *T5Based* obtiveram inferiores aos encontrados na literatura. O sistema referência [14] teve os melhores resultados em todos os âmbitos.

Table 9: *Sentence-selection evaluation* - com rótulo

Selection + Label			
Abordagem	P	R	F1
Oráculo	78,16	67,49	72,43
VeriSci	46,89	39,07	42,62
VerT5erini	**60,80**	**53,83**	**57,10**
BERTBased	16,46	15,95	16,20
T5Based	20,12	21,01	20,55

5 Conclusão

Esse trabalho teve como objetivo apresentar um sistema para verificação de alegação científica que explora um dois modelos distintos para realizar três módulos: (1) reivindicação de documentos relevantes, (2) seleção de sentenças e (3) classificação das premissas. Tais sistemas são importantes nesta era desinfodemia, principalmente na era da pandemia de COVID-19. Os resultados obtidos se assemelham com os encontrados na literatura, mostrando que ainda há melhorias para serem feitas.

6 Trabalhos Futuros

Para trabalhos futuros deve pensar em três âmbitos:

- **Utilizar versões dos modelos maiores:** Tanto o *BERT* quanto o *T5* utilizados nesse trabalho forma as suas versões básicas. Isso porque houve limitação de *hardware* para realizar o treinamento de redes maiores.
- **Testar modelos genéricos como GPT-2, GPT-J:** Utilizando técnicas de *few-shot* acretida-se que esses modelos maiores podem ter resultados mais interessantes.
- **Testar em diferentes conjuntos de dados:** Outros conjuntos de dados devem ser testados, tanto no domínio biomédico (como o SciFact), quanto em outros domínios.

References

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[3] Kalina BONTCHEVA and Julie POSETTI. Desinfodemia-decifrar a desinformação sobre a covid-19. *Resumo de políticas*, 1, 2020.

[4] Organização Pan-Americana da Saúde. Entenda a infodemia ea desinformação na luta contra a covid-19, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[7] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362, 2021.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[9] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*, 2019.

[10] Alexios Mantzarlis. Fact-checking 101. *Journalism, fake news & disinformation: Handbook for journalism education and training*, pages 85–100, 2018.

[11] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[12] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.

[13] Gordon Pennycook and David G. Rand. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402, 2021.

[14] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. Scientific claim verification with vert5erini. *arXiv preprint arXiv:2010.11930*, 2020.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[16] Ashish Rana, Deepanshu Khanna, Tirthankar Ghosal, Muskaan Singh, Harpreet Singh, and Prashant Singh Rana. Rerrfact: Reduced evidence retrieval representations for scientific claim verification. *arXiv preprint arXiv:2202.02646*, 2022.

[17] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[18] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[19] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

[20] David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, 2022.

[21] Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. Abstract, rationale, stance: A joint model for scientific claim verification. *arXiv preprint arXiv:2110.15116*, 2021.