FactCheck Editor: Multilingual Text Editor with End-to-End fact-checking

Vinay Setty*
vsetty@acm.org
University of Stavanger
Stavanger, Norway

ABSTRACT

We introduce FactCheck Editor, an advanced text editor designed to automate fact-checking and correct factual inaccuracies. Given the widespread issue of misinformation, often a result of unintentional mistakes by content creators, our tool aims to address this challenge. It supports over 90 languages and utilizes transformer models to assist humans in the labor-intensive process of fact verification. This demonstration showcases a complete workflow that detects text claims in need of verification, generates relevant search engine queries, and retrieves appropriate documents from the web. It employs Natural Language Inference (NLI) to predict the veracity of claims and uses LLMs to summarize the evidence and suggest textual revisions to correct any errors in the text. Additionally, the effectiveness of models used in claim detection and veracity assessment is evaluated across multiple languages.

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**.

KEYWORDS

Multilingual Fact-checking; Factuality

ACM Reference Format:
Vinay Setty. 2024. FactCheck Editor: Multilingual Text Editor with End-to-End fact-checking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657663

1 INTRODUCTION

In the era of digital information, the proliferation of misinformation has emerged as a formidable challenge, impacting societies, politics, and public opinions. This is often a result of unintentional mistakes by content creators, has necessitated the development of tools that can effectively identify and correct factual errors [7, 18].

Most newsrooms rely on content management systems for news production, offering basic formatting and composition tools. After journalists write an article, it is typically proofread and fact-checked manually, using web searches and searching internal archives. Current automation extends only to grammar checkers like Grammarly and advanced tools like Writer.com, which automate writing styles. This paper presents FactCheck Editor, an innovative text editor capable of identifying factual inaccuracies and suggesting corrections in over 90 languages. FactCheck Editor could potentially assist humans writers in content creation in sectors like news and media by helping editors detect factual errors early. However, end-to-end multilingual fact-checking presents unresolved challenges for both academia and the industry [12].

To make this problem tangible, our approach is threefold: First, we address the problem of detecting check-worthy claims, a task that involves understanding the context, relevance, and potential impact of each statement. Second, we address the task of generating and executing search engine queries, for gathering relevant information from the web. Finally, this information is then utilized by a Natural Language Inference (NLI) model, for veracity prediction. Furthermore, we use LLMs, to generate justification summaries and also suggest precise textual amendments for error rectification.

We also present preliminary evaluation results which show that a smaller transformer model, fine-tuned using datasets in small number of languages, can outperform large language models (LLMs) such as GPT-3.5-Turbo and Mistral-7b for both claim detection and veracity prediction tasks. On the other hand, LLMs excel at generative tasks such as summarization and suggesting claim corrections.

2 RELATED WORK

Automated fact-checking has become popular in research recently. However, there is relatively low adaption in the industry. There are existing tools such as browser plugin proposed by Botnevik et al. [2] which can fact-check already written text. Wang et al. [16] propose a tool for annotating the factual mistakes made by LLMs. There are also tools for detecting hallucinations and factual mistakes made by LLMs, such as FactTool [5] and FAVA [10]. While these are very sophisticated solutions, they do not focus on end-to-end fact-checking in a multilingual setting.

Majority of the fact-checking literature is focused on English language [7, 15]. There are datasets for multilingual fact-checking [8, 11]. There is also recent survey on multilingual claim detection Panchendrarajan and Zubiaga [12]. However, there is still a need for research regarding end-to-end multilingual fact-checking.

3 SYSTEM OVERVIEW

This work aims to provide a user-friendly web-based editor to compose textual articles with fact-checking feature. FactCheck Editor identifies check-worthy claims in the written article by the user and verify those claims using evidence gathered from open web and

*Also with Factiverse AI.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657663

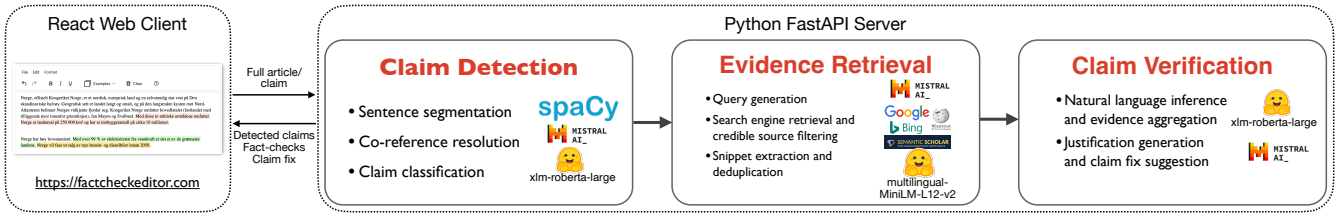

Figure 1: System Architecture of FactCheck Editor

previous fact-checks. Figure 1, the architecture of FactCheck Editor, with a web-based front-end implemented in React framework and a backend server. The frontend includes a text editor implemented using the TinyMCE text editor[1]. The backend, exposes REST APIs to interact with the machine learning (ML) models. The ML models used in the backend are grouped into (a) Check-worthy claim detection, (b) Evidence retrieval, and (c) Veracity prediction.

3.1 Check-worthy Claim Detection

The goal of this stage is to identify and enrich claims which warrant verification.

Sentence segmentation and Co-reference resolution: The initial step in processing an article is sentence segmentation, which involves breaking down the text into individual sentences. We primarily use models from Spacy[2] for this task due to their efficiency and accuracy. We use an LLM (Mistral-7b), for co-reference resolutions, which helps in identifying the pronouns and linking them back to the appropriate named entities they represent.

Claim classification: This step is often the first step in a manual fact-checking pipeline. The objective is to determine whether a sentence contains a claim that warrants verification. We approach this as a binary classification task, where the goal is to classify each claim as either 'check-worthy' or not. To accomplish this, we leverage established datasets ClaimBuster [9] and CLEF CheckThat! Lab [1]. These datasets are in English, therefore, we translate them to a handful of languages (Norwegian, German, and Danish) to fine-tune a multilingual classifier (**XLM-Roberta-Large**). Surprisingly, this limited training is able to transfer the knowledge to other languages, which the model didn't have any training data for. We also employ LLMs with **two-shot chain-of-thought (CoT) reasoning** prompts for comparison (See Section 4.1).

3.2 Evidence Retrieval

The goal of this stage is to retrieve highly relevant documents to verify the claims in the previous step. It is also important to retrieve both supporting and refuting documents for the claim.

Query generation: This process involves generating effective questions or search queries to find relevant documents. Typically, original claims are used as search queries. However, this approach often fails to retrieve relevant documents, particularly for claims containing incorrect information. To address this, we draw inspiration from existing works [4, 6, 14]. We utilize **Mistra-7b**[3] to create more effective questions and queries that are well-suited for search engines. The prompts used can be found on our GitHub repo[4].

Retrieval from search engines: We search across diverse platforms for queries and questions, including Wikipedia, Google Fact-check Explorer for previous fact-checks, Google and Bing search engines, and the semantic web for scholarly articles, ensuring a broad and comprehensive source coverage. To maintain source credibility, we filter out domains which are known to spread misinformation[5] and only consider scholarly articles from Semantic Scholar with at least 10 citations. To support multilingual search, we specify the language option in the respective search APIs.

Deduplication and Snippet Extraction. To streamline search results from various sources, we implement deduplication by combining URL, title, and content, using approximate matching to filter out duplicates. Additionally, we refine relevance by extracting the top three paragraphs most related to the claim, determined by cosine similarity scores from sentence embeddings, focusing on the most pertinent information. We use **Multilingual-MiniLM-L12-v2** for computing sentence embeddings.

3.3 Veracity Prediction

After having retrieved the evidence, and pre-processing them, the final step in the fact-checking pipeline is veracity prediction and justification generation. This is a crucial step in fact-checking.

Natural Language Inference (NLI):. The NLI task involves classifying whether a piece of evidence supports, refutes or unrelated to a claim [3]. In our process, we ensure that only relevant documents are considered as evidence, significantly reducing the number of unrelated documents. Therefore, we simplify it by modeling it as a binary classification (supports or refutes). We fine-tune XLM-Roberta-Large using FEVER [15], MNLI [17] and X-Fact [8] datasets. We also compare its performance against LLMs from OpenAI and Mistral in Section 4.2. For each claim, there are usually multiple evidence snippets, and the NLI prediction applies to each claim-evidence pair. To synthesize these individual predictions, we use the majority voting technique used in the literature [13, 14].

Justification Generation and Claim Fix Suggestion: With several evidence snippets available for a claim, each presenting different

[1] https://www.tiny.cloud
[2] https://spacy.io/usage/models/
[3] https://ollama.ai/library/mistral
[4] https://github.com/factiverse/factcheck-editor/tree/main/code/prompts
[5] https://en.wikipedia.org/wiki/List_of_fake_news_websites

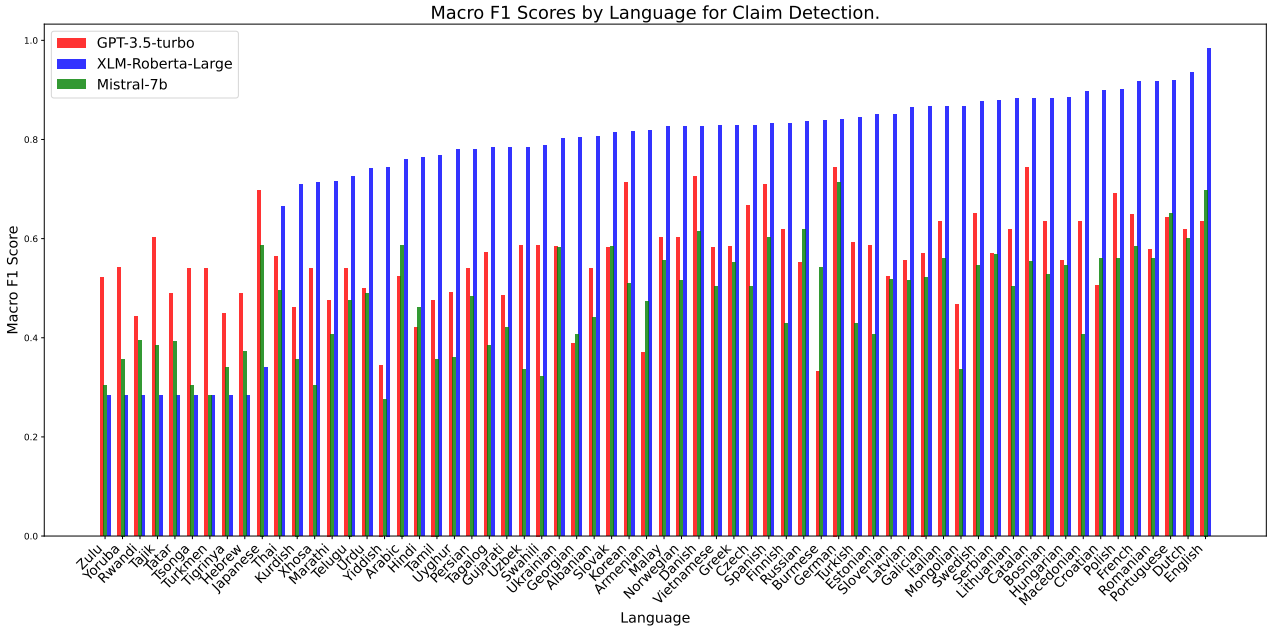

Figure 2: Evaluation of claim detection for 118 languages using XLM-RoBERTa-Large, GPT-3.5-Turbo and Mistral-7b.

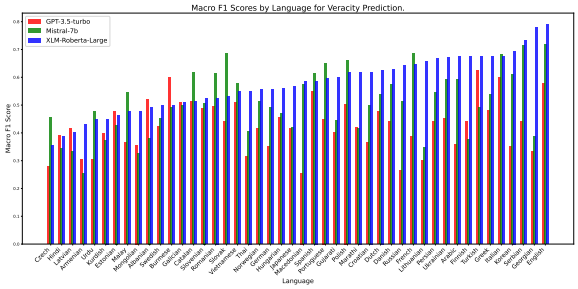

Figure 3: Evaluation of veracity prediction for 46 languages.

arguments that support or refute it, the information can become overwhelming for users. To enhance accessibility, we summarize the evidence in relation to the claim and its predicted veracity label, offering a concise and coherent overview that simplifies understanding the basis for the claim's classification. Similarly, in the way tools like Grammarly suggest corrections for typos and grammar, we introduce a method for suggesting fixes to inaccuracies in claims based on the evidence found. Utilizing Mistral-7b and crafted prompts, this feature not only identifies potential errors in claims but also offers suggestions for correction, thereby improving the accuracy and reliability of the information. We omit the qualitative evaluation of these suggestions as future work.

4 EXPERIMENTAL EVALUATION

4.1 Claim Detection

Dataset: Fact-checking full articles is different from most existing datasets such as political debates, therefore, we annotate a smaller scale news dataset. The dataset statistics are described in Table 1. Since this dataset is in English, we translated this dataset into 118 languages using the Google translate API.

Table 1: Dataset distribution.

Split	Not Check-worthy	Check-worthy	True Claims	False Claims	Total
Train	609	548	332	196	1,076
Dev	38	25	15	10	63
Test	62	38	26	12	100

Table 2: Claim detection and veracity prediction results presented as mean Micro and Macro-F1 scores for all languages.

Model	Claim Detection		NLI	
	Ma.-F1	Mi.-F1	Ma.-F1	Mi.-F1
GPT-3.5-Turbo	0.562	0.567	0.427	0.461
Mistral-7b	0.477	0.510	0.509	0.557
XLM-RoBERTa-Large	**0.743**	**0.768**	**0.575**	**0.594**

Results: We compare XLM-RoBERTa-Large, GPT-3.5-Turbo and Mistral-7b using this dataset. In Figure 2, the F1-Macro is shown for all languages. Surprisingly, the fine-tuned XLM-RoBERTa-Large outperforms both GPT-3.5-Turbo and Mistral-7b in most languages. Since the model was trained mainly on English, not surprisingly it is the best performing language. For some languages, we see that XLM-RoBERTa-Large is the worst performing model. On closer inspection, these are the languages not supported (not included in the pre-training step). Mistral-7b seems to be the worst performing

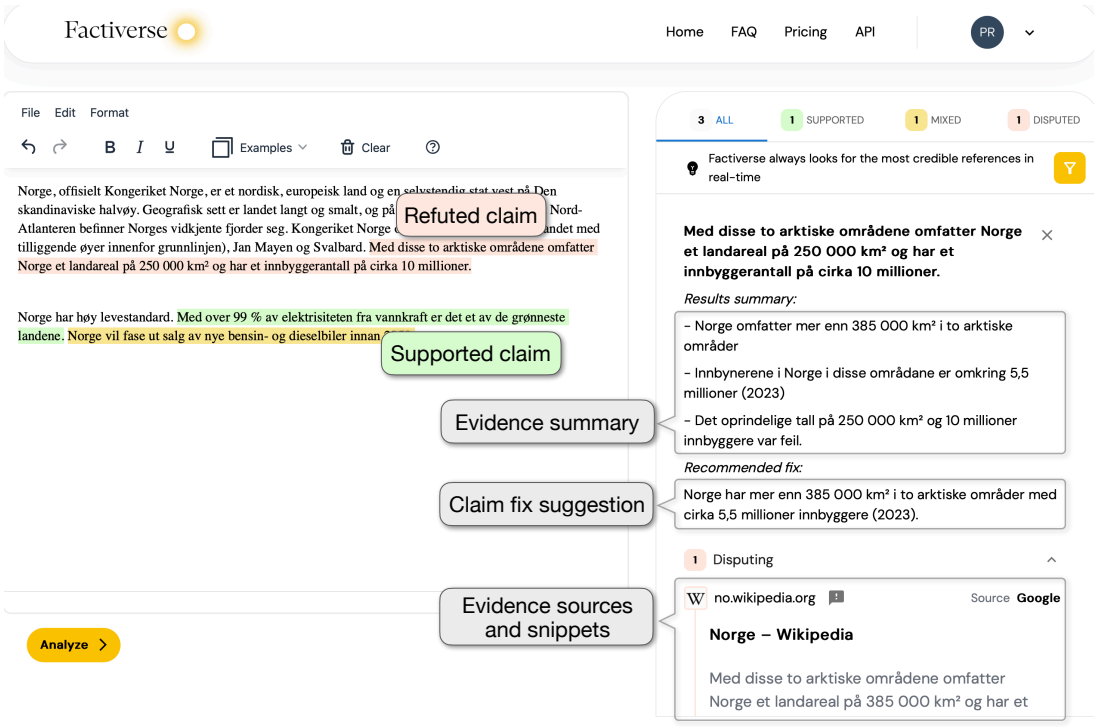

Figure 4: FactCheck Editor demonstration

model, it seems to be because Mistral, struggles to follow instructions in the prompt. We observed a similar pattern in Micro-F1 scores, therefore, we omit those results due to lack of space. This suggests that for claim detection, fine-tuning a multilingual transformer model is promising rather than, few-shot chain-of-thought reasoning prompts with LLMs in a multilingual setting. Table 2 shows the mean Macro-F1 and Micro-F1 scores for all languages evaluated. *This shows that a fine-tuned XLM-RoBERTa-Large can outperform LLMs in multilingual setting for claim detection.*

4.2 Veracity Prediction

Dataset: We use the same data from the claim detection dataset for the NLI and veracity prediction tasks. The distribution of True and False claims are shown in Table 1.

Results: As shown in Figure 3, fine-tuned XLM-RoBERTa-Large outperforms GPT-3.5-Turbo and Mistral-7b for most languages. It is interesting to see that Mistral performs better than GPT-3.5-Turbo despite being much a smaller LLM. Mistral-7b seems to be the best model for some European languages, such as French and Portuguese. Since for some languages, we couldn't find any evidence snippets for any of the claims, they are omitted. The overall results are shown in Table 2 with similar observations to claim detection.

5 IMPLEMENTATION AND DEMONSTRATION

We use a docker container to deploy the backend on a public cloud provider. The front-end is also hosted on the same provider. We use Ollama framework for self-hosting the LLMs.

Figure 4 shows an example involving an article written in Norwegian that contains factual inaccuracies. For instance, the claim "Norge et landareal på 250 000 km^2 og har et innbyggerantall på cirka 10 millioner" (which translates to "Norway has a land area of 250,000 km^2 and a population of approximately 10 million") is flagged as incorrect, with a suggestion to replace "250 000" with "385 000" and '10 million" with "5.5 million" based on the evidence found. The editor also marks claims in red and 'green to indicate disputed and supported claims, respectively, based on the evidence. Additionally, the right-hand pane displays evidence snippets along with a summary of the generated justification. The demonstration can be accessed live[6] and the evaluation code is shared[7].

6 CONCLUSION

In this paper, we demonstrated a multilingual text editor designed for identifying factual errors in written text. We also conduct preliminary experiments, which show that fine-tuning transformer models are more effective for claim detection and veracity prediction in multilingual setting with over 90 languages, warranting further research on end-to-end multilingual fact-checking.

7 ACKNOWLEDGEMENTS

This work is in part funded by the Research Council of Norway project EXPLAIN (grant number 337133). We also acknowledge Tobias Tykvart and Domante Stirbytė from Factiverse for the implementation of the FactCheck Editor front-end.

[6]https://factcheckeditor.com
[7]https://github.com/factiverse/factcheck-editor

REFERENCES

[1] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In *Findings of the Association for Computational Linguistics: EMNLP 2021 (EMNLP '21)*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 611–649.

[2] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. BRENDA: Browser Extension for Fake News Detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 2117–2120.

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. arXiv:1508.05326 [cs]

[4] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating Literal and Implied Subquestions to Fact-check Complex Claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 3495–3516.

[5] I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. arXiv:2307.13528 [cs.CL]

[6] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating Fact Checking Briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 7147–7161.

[7] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[8] Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Online) *(ACL)*. 675–682.

[9] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proc. VLDB Endow.* 10, 12 (2017), 1945–1948.

[10] Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained Hallucination Detection and Editing for Language Models. arXiv:2401.06855 [cs]

[11] Dan S. Nielsen and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, 3141–3153.

[12] Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal* 7 (2024), 100066.

[13] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. 1003–1012.

[14] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems* 36 (2024).

[15] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. arXiv:1803.05355 [cs]

[16] Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. arXiv:2311.09000 [cs.CL]

[17] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. arXiv:1704.05426 [cs]

[18] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *Comput. Surveys* 53, 5 (2020), 1–40.