

Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models

Vinay Setty

Factiverse AI

Norway

vinay@factiverse.ai

ABSTRACT

In this paper, we explore the challenges associated with establishing an end-to-end fact-checking pipeline in a real-world context, covering over 90 languages. Our real-world experimental benchmarks demonstrate that fine-tuning Transformer models specifically for fact-checking tasks, such as claim detection and veracity prediction, provide superior performance over large language models (LLMs) like GPT-4, GPT-3.5-Turbo, and Mistral-7b. However, we illustrate that LLMs excel in generative tasks such as question decomposition for evidence retrieval. Through extensive evaluation, we show the efficacy of fine-tuned models for fact-checking in a multilingual setting and complex claims that include numerical quantities.

CCS CONCEPTS

- Information systems → Retrieval tasks and goals.

KEYWORDS

Multilingual; Fact-checking

ACM Reference Format:

Vinay Setty. 2024. Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models. In *Proceedings of Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

The spread of online misinformation poses a significant challenge globally, impacting societies, political landscapes, and public opinions. This problem is compounded by the complexity of fact-checking, a task that is challenging even for humans¹. Research communities specializing in natural language processing and information retrieval have made significant strides in automating aspects of this task, addressing some challenges.

However, several challenges remain that hinder the widespread adoption of these technologies in the industry. These challenges include but not limited to multilinguality, verifying numerical claims,

¹<https://www.niemanlab.org/2024/01/asking-people-to-do-the-research-on-fake-news-stories-makes-them-seem-more-believable-not-less/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '24, July 14–18, 2024, Washington, D.C., USA

© 2024 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

and the temporal aspects of facts [7, 11]. The emergence of large language models (LLMs), such as GPT-4, Gemini and Mistral, has raised expectations that they might address these issues effectively. There are many recent works that propose use of LLMs for end-to-end fact-checking pipeline [5, 18]. Particularly, LLMs, being inherently multilingual, are anticipated to excel in processing and understanding non-English languages, especially those considered low-resource. Furthermore, advanced models like GPT-4 have shown increased capabilities in numerical reasoning, which could potentially enhance the accuracy of fact-checking numerical data.

Despite expectations, this paper demonstrates that, for certain fact-checking tasks like identifying check-worthy claims and assessing their veracity, specialized Transformer models perform better than LLMs in practical situations. Nonetheless, it's worth mentioning that LLMs are superior in generative tasks, such as decomposing claims into queries and posing questions, which helps in the more effective gathering of evidence. This paper specifically addresses fact-checking within multilingual contexts and for claims involving numerical quantities. We show that multilingual Transformer models, such as XLM-RoBERTa, fine-tuned for select languages, can surpass LLMs in effectiveness across more than 90 languages. Similarly, Transformer models specialized in numerical question answering demonstrate superior performance.

An additional obstacle encountered by fact-checking tools in the industry pertains to privacy concerns. Journalists and fact-checkers, the primary users of these technologies, hesitate to transmit sensitive and private information to third-party servers that host LLMs. This paper explores the effectiveness of fine-tuned models and smaller, self-hostable LLMs in diverse fact-checking situations.

1.1 Related Work

Recently, the use of machine learning models for fact-checking has gained attention in the academic research, yet its implementation in the industry remains limited. Factiverse has a few commercial tools capable of fact-checking textual content² [2, 10]. Meanwhile, large language model (LLM) chat applications like ChatGPT and Gemini have started providing references with their outputs. Similarly, there are writer tools such as Originality AI³ which use LLMs to do the fact-checking. However, these solutions still can still generate erroneous text, and they rely on a limited range of sources for their references. Factiverse has also released a custom GPT to fact-check ChatGPT generated text⁴. There are also tools such as Google Search-Augmented Factuality Evaluator (SAFE) [19], SelfCheckGPT [9] FactTool [18] and FAVA [5] to identify factual

²<https://editor.factiverse.ai>

³<https://originality.ai>

⁴<https://gpt.factiverse.ai>

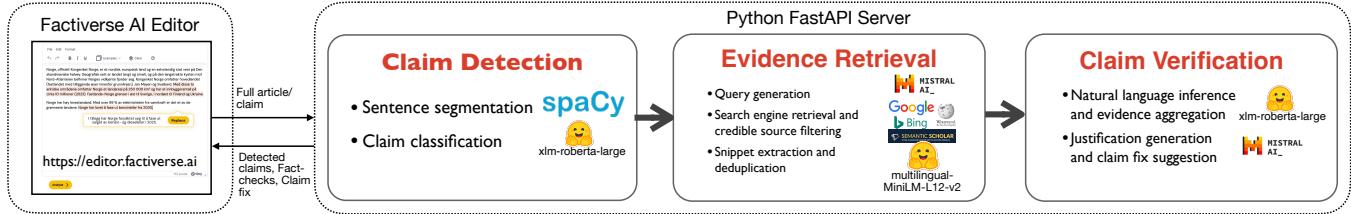


Figure 1: System Architecture of the Fact-Checking Pipeline at Factiverse

inaccuracies in LLM outputs using search results and prompts. Despite their complexity, these tools have not been adapted for real-world applications.

2 FACT-CHECKING PIPELINE

We follow a three-stage fact-checking pipeline (shown in Figure 1) that is widely used in the literature [7]. We have a front-end React app, which can be used to compose articles and fact-check them⁵. The frontend communicates with the backend via the REST APIs.

The first stage identifies check-worthy sentences (also referred to as ‘claims’ in the rest of the paper), in the given text, either written by users or generated using LLMs. The second stage gathers evidence from the open web and previous fact-checks for a given claim. Finally, the claim is verified using the gathered evidence in the third stage. The pipeline is hosted using a Python FastAPI backend that is hosted on a Kubernetes cluster with auto-scaling on the Google Cloud Platform. There are several ML models used in the backend targeted for (a) Check-worthy claim detection, (b) Evidence search, and (c) Veracity prediction. We now explain the sub-tasks within these steps of the pipeline.

2.1 Check-worthy Claim Detection

The objective of this stage is to enrich the sentences within the text that merit fact-checking. We first segment sentences using spaCy models. While there is no fixed definition, it is generally agreed upon which sentences require verification: (1) they invite the public to scrutinize their accuracy and authenticity, and (2) they exclude subjective statements such as opinions, beliefs, or queries [12].

We fine-tune the *XLM-RoBERTa-Large* model⁶ [6], using the data from ClaimBuster [8] and CLEF CheckThat Lab! [1] along with a dataset collected from Factiverse production system (see Table 1) to classify sentences into ‘Check-worthy’ and ‘Not check-worthy’.

2.2 Evidence Search

The objective of the second stage is to retrieve highly relevant articles essential for fact-checking the verification of the check-worthy claims detected in the previous phase. Since using a claim verbatim as a search query may not yield sufficiently relevant results, we need to decompose it into questions to diversify the search results. In this regard, we employ Large Language Models (LLMs), to generate targeted questions to verify the claim. The prompt

used for this purpose can be found in⁷. We then leverage a diverse range of search engines including Google, Bing, You.com, Wikipedia, Semantic Scholar⁸ (contains 212M scholarly articles), and Factiverse’s fact-checking collection⁹, which comprises 280K fact-checks updated in real-time. We then deduplicate the search results by matching URLs, titles, and content with approximate similarity. Furthermore, to enhance the relevancy of the data, we select the top three paragraphs that are most similar to the claim using a multilingual sentence encoder [15].

2.3 Veracity Prediction

The final step in the fact-checking pipeline is predicting the veracity of the claim based on the evidence. To ensure the evidence is coming from credible sources, we exclude websites known to spread misinformation¹⁰. We first predict the stance of the evidence for the claim (natural language inference or NLI). I.e., if the claim is supported or refuted by the evidence. To simplify, we omit claims which are half true or false, similar to existing works [16]. For this step, we also fine-tune *XLM-RoBERTa-Large* model using a combination of FEVER data and real-world fact-checks from Factisearch. We then aggregate the stance predictions for individual evidence snippets by majority voting following existing works [13]. However, we acknowledge that there is no consensus on how to aggregate multiple evidence snippets and account for the credibility of these sources. We further summarize the evidence snippets, using an LLM, which acts as a justification for the veracity prediction. In addition, we also generate a correction for the refuted claims based on the justification summary using an LLM.

3 EXPERIMENTAL SETUP

Dataset: We use a dataset sample from production deployment at Factiverse¹¹. Since the original data was only in English, we translated the claims into 114 languages using the Google Translate API. An overview of the dataset is shown in Table 1. Contrary to existing fact-checking datasets from fact-checking organizations, this sample has ‘True’ class, the majority class. This is because Factiverse applications focus on fact-checking full articles, while manual fact-checkers tend to pick ‘False’ claims to verify. For numerical

⁵<https://editor.factiverse.ai>

⁶<https://huggingface.co/FacebookAI/xlm-roberta-large>

⁷<https://github.com/vinaysetty/factcheck-editor/blob/main/code/prompts/prompts.py>

⁸<https://www.semanticscholar.org>

⁹<https://factisearch.ai>

¹⁰https://en.wikipedia.org/wiki/List_of_fake_news_websites

¹¹<https://github.com/vinaysetty/factcheck-editor/tree/main/data>

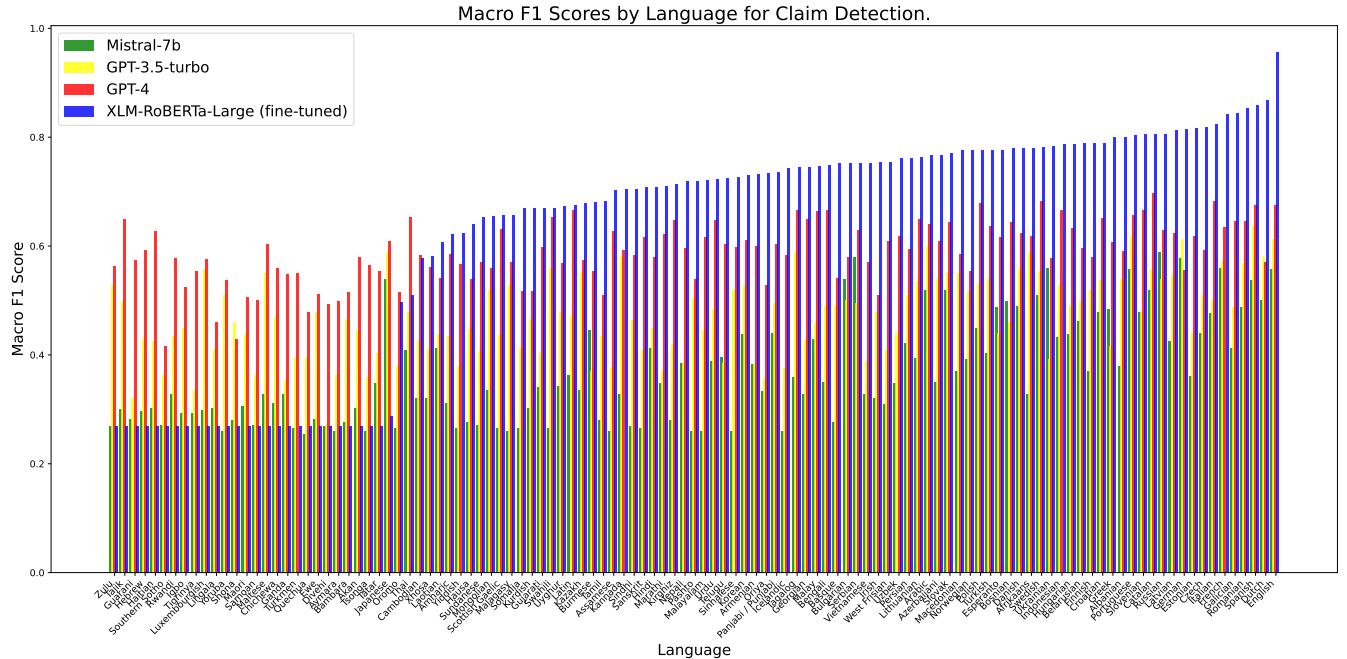


Figure 2: Evaluation of claim detection for 114 languages using Factiverse model, GPT-3.5-Turbo, GPT-4 and Mistral-7b.

claims, we sample 100 claims from a collection of past fact-checks from Factisearch¹² with numerical quantities¹³.

Models: For claim detection and veracity prediction tasks, we fine-tune XLM-RoBERTa-Large [6]. For verifying numerical claims, we also use FinQA-RoBERTa-Large [20] fine-tuned using the financial QA dataset, FinQA [4]. We use GPT-4 and GPT-3.5-Turbo and Mistral-7b¹⁴ as LLMs. To adapt LLMs to perform fact-checking, we do prompt engineering to draft a prompt to predict both check-worthiness of a claim and for veracity prediction. To make the comparison fair, the same prompts are used for all LLMs¹⁵. We set the temperature to ‘0.2’, set a random seed and repeat all experiments 3 times and report the mean scores.

For question decomposition, in addition to LLMs, we also compare with T5-3b¹⁶[14] fine-tuned using the ClaimDecomp dataset [3]. Since T5-3b model only supports English, we perform question decomposition effectiveness analysis only for English claims.

Metrics: We use the Macro-F1 and Micro-F1 scores to compare the performance of the models since there is an imbalance in the classes (applies to both claim detection and veracity prediction tasks). For multilingual evaluation, we only show Macro-F1 but a similar trend was observed in Micro-F1.

The code used for the experiments is available on GitHub¹⁷.

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

4 EXPERIMENTAL RESULTS

4.1 Claim Detection

As shown in Figure 2, the fine-tuned XLM-RoBERTa-Large outperforms both OpenAI and Mistral models in most languages. Since the model was trained mostly in English, it is unsurprisingly the best-performing language. For some languages (towards the left side of the plot), we see that LLMs perform better. On closer inspection, these are the languages not yet supported by the model. Mistral-7b seems to be the worst-performing model overall. It seems to be because Mistral struggles to follow instructions in the prompt for text classification and hallucinates more than other models. Table 3 shows the average Macro-F1 and Micro-F1 scores for all four models. This suggests that a fine-tuned model is significantly better for claim detection compared to using LLMs with prompt engineering.

4.2 Evidence Retrieval

Here, we compare the effectiveness of different question decomposition methods for evidence retrieval. In this paper, we report the effect of these methods only on the final task of veracity prediction. We leave the evaluation of question quality and information retrieval measures for a future work.

¹²<https://factisearch.ai>

¹³https://github.com/vinaysetty/factcheck-editor/tree/main/data/veracity_prediction

¹⁴<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

¹⁵Prompts can be found here: <https://github.com/factiverse/factcheck-editor/blob/main/code/prompts/prompts.py>

¹⁶<https://huggingface.co/google-t5/t5-3b>

¹⁷<https://github.com/factiverse/factcheck-editor>

Table 2: Comparison of Effectiveness of Question Decomposition Methods for Veracity Prediction for English claims. Rows contain question decomposition methods and columns contain natural language inference (NLI) methods.

NLI Question Decomposition \	XLM-RoBERTa-Large		Mistral-7b		GPT-3.5-Turbo		GPT-4	
	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
T5-3b (fine-tuned)	0.632	0.639	0.398	0.444	0.438	0.444	0.469	0.473
Mistral-7b	0.642	0.647	0.410	0.412	0.454	0.471	0.463	0.470
GPT-4	0.719	0.722	0.550	0.556	0.494	0.500	0.494	0.500
GPT-3.5-Turbo	0.741	0.750	0.556	0.562	0.498	0.500	0.530	0.531

Table 3: Claim detection and veracity prediction results presented as mean Micro and Macro-F1 scores for all languages.

Model	Claim Detection		Veracity Prediction	
	Ma.-F1	Mi.-F1	Ma.-F1	Mi.-F1
Mistral-7b	0.477	0.510	0.509	0.557
GPT-3.5-Turbo	0.562	0.567	0.440	0.396
GPT-4	0.624	0.591	0.460	0.426
XLM-RoBERTa-Large	0.743	0.768	0.575	0.594

Table 4: Effectiveness of fine-tuned models compared to LLMs for numerical claims.

NLI Question Decomposition \	XLM-RoBERTa		FinQA-RoBERTa	
	Ma. F1	Mi. F1	Ma. F1	Mi. F1
T5-3B (fine-tuned)	0.579	0.611	0.642	0.750
GPT-4	0.642	0.700	0.641	0.750
GPT-3.5-Turbo	0.562	0.737	0.733	0.755
Mistral-7b	0.672	0.722	0.781	0.842

Table 2 shows that GPT-3.5-Turbo outperforms other question decomposition methods, including GPT-4, with the fine-tuned T5-3b model ranking lowest. This suggests LLMs are better at generating questions for evidence retrieval, though GPT-4’s longer questions may cause topic drift in search results.

The fine-tuned XLM-RoBERTa-Large NLI model consistently outperforms others across all question generation methods. In contrast, the fine-tuned T5-3b emerges as the least effective question generation method. For this particular task, Large Language Models (LLMs) exhibit markedly superior performance compared to the fine-tuned T5-3b model. Furthermore, it appears that the larger the LLM, the better the performance. While self-hosted Mistral-7b does not seem to generate effective questions for evidence retrieval, it is worth exploring other open larger models in future.

4.3 Veracity Prediction

As shown in the Figure, 3, the fine-tuned XLM-RoBERTa-Large model outperforms the other models in 37 languages. GPT-4 is the best model only for three languages: Swedish, Albanian, and Georgian. Mistral-7b is the best model in 8 languages, and it is interesting to see that Mistral performs better than OpenAI models despite being a much smaller LLM. Unsurprisingly, Mistral-7b seems to be the best model for some European languages, such

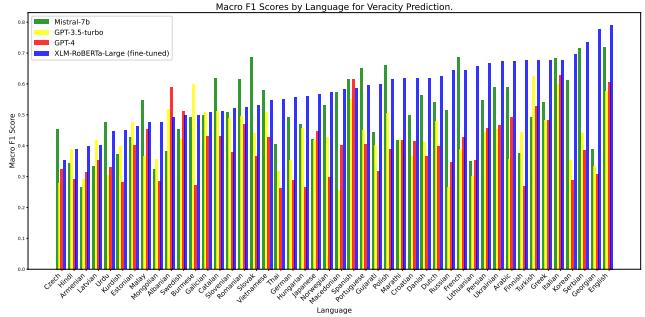


Figure 3: Evaluation of veracity prediction for 46 languages.

as French, Spanish, Catalan, and Portuguese. Due to the lack of evidence snippets for some languages, they were excluded. Veracity prediction results are summarized in Table 3.

Table 4 compares the fine-tuned XLM-RoBERTa-Large and the FinQA-RoBERTa-Large for numerical claims fine-tuned by [17]. Not surprisingly, FinQA-RoBERTa-Large performs better because it is tailored to perform reasoning over numerical data [20]. However, it is surprising to note that the question decomposition by Mistral-7b is better for numerical claims than OpenAI and T5 models. We believe this is because the numerical claims need different style of questions, which Mistral seems to perform better. We omit the comparison of LLMs for the NLI task due to lack of space, but the performance is similar to the results in Table 2.

5 CONCLUSION

In this paper, we explored the complexities involved in fact-checking systems and demonstrated that fine-tuned, smaller models surpass LLMs in multilingual contexts and when evaluating numerical claims. However, LLMs exhibit superior performance in generative tasks such as question decomposition. The outcomes suggest that both fine-tuning Transformer models and employing self-hosted LLMs could be effective strategies for enhancing fact-checking operations. Given the preliminary nature of this research and smaller scale of the dataset used, it becomes imperative to undertake further studies to confirm these findings. Future work will delve into evaluating the evidence retrieval step and larger scale experiments.

6 ACKNOWLEDGEMENTS

This work is in part funded by the Research Council of Norway project EXPLAIN (grant number 337133) with partners from Factiverse and TU Delft.

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-Homaied, 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] 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.
- [4] Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP '21)*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 3697–3711.
- [5] I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FactTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. arXiv:2307.13528 [cs.CL]
- [6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, 8440–8451.
- [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] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulakarni, 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.
- [9] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP '23)*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 9004–9017.
- [10] Rahul Mishra and Vinay Setty. 2019. SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19)*. Association for Computing Machinery, 197–204.
- [11] Andreas L Opdahl, Bjørnar Tessem, Duc-Tien Dang-Nguyen, Enrico Motta, Vinay Setty, Eivind Throndsen, Are Tverberg, and Christoph Trattner. 2023. Trustworthy journalism through AI. *Data & Knowledge Engineering* 146 (2023), 102182.
- [12] Rubaa Panchendarajan 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] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.
- [15] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP '19)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3982–3992.
- [16] 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).
- [17] Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims.

- arXiv:2403.17169 [cs.CL]
- [18] 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]
- [19] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. arXiv:2403.18802 [cs.CL]
- [20] Jiaxin Zhang and Yashar Moshfeghi. 2022. ELASTIC: numerical reasoning with adaptive symbolic compiler. *Advances in Neural Information Processing Systems* 35 (2022), 12647–12661.