ExPrompt: Augmenting Prompts Using Examples as Modern Baseline for Stance Classification

Umair Qudus
Data Science Group, Department of Computer Science,
Paderborn University
Paderborn, Germany
umair.qudus@uni-paderborn.de

Michael Röder
Data Science Group, Department of Computer Science,
Paderborn University
Paderborn, Germany
michael.roeder@uni-paderborn.de

Daniel Vollmers
Data Science Group, Department of Computer Science,
Paderborn University
Paderborn, Germany
daniel.vollmers@uni-paderborn.de

Axel-Cyrille Ngonga Ngomo
Data Science Group, Department of Computer Science,
Paderborn University
Paderborn, Germany
axel.ngonga@uni-paderborn.de

Abstract

Detecting the veracity of a statement automatically is a challenge the world is grappling with due to the vast amount of data spread across the web. Verifying a given claim typically entails validating it within the framework of supporting evidence like a retrieved piece of text. Classifying the stance of the text with respect to the claim is called stance classification. Despite advancements in automated fact-checking, most systems still rely on a substantial quantity of labeled training data, which can be costly. In this work, we avoid the costly training or fine-tuning of models by reusing pre-trained large language models together with few-shot in-context learning. Since we do not train any model, our approach ExPrompt is lightweight, demands fewer resources than other stance classification methods and can serve as a modern baseline for future developments. At the same time, our evaluation shows that our approach is able to outperform former state-of-the-art stance classification approaches regarding accuracy by at least 2 percent. Our scripts and data used in this paper are available at https://github.com/dice-group/ExPrompt.

CCS Concepts

• **Computing methodologies** → *Knowledge representation and reasoning*; Artificial intelligence; • **Information systems** → *Data cleaning*; Graph-based database models.

Keywords

Stance Classification; Few-shot in-context learning; Pre-trained large language models.

ACM Reference Format:
Umair Qudus, Michael Röder, Daniel Vollmers, and Axel-Cyrille Ngonga Ngomo. 2024. ExPrompt: Augmenting Prompts Using Examples as Modern Baseline for Stance Classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3627673.3679923

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679923

1 Introduction

With the rapid proliferation of misinformation on the internet—such as fake news, socio-political deception, and online rumors—journalists, broadcasters, political figures, and the general populace face challenges in keeping abreast of the latest factual information [1]. This difficulty is amplified during public emergencies such as the COVID-19 pandemic, where new discoveries are swiftly disseminated and decisions made on outdated or incomplete data can pose significant risks [29]. Consequently, there is a growing demand for automated tools to help users assess the accuracy of claims. Many existing approaches are based on deriving pieces of evidence for a given claim that either support or refute the claim. The task to classify such a piece as either supporting or refuting is known as stance classification [24].

Recent advancements in stance classification reveal significant challenges with supervised learning models, prone to dataset-specific biases [32]. Schuster et al. [24] demonstrate the effectiveness of a claim-only model, potentially exploiting dataset idiosyncrasies. Thorne et al. [27] highlighted Fever systems' vulnerability to adversarial conditions, causing performance drops. In natural language inference (NLI), neural models rely on surface-level cues over genuine comprehension. Zero-shot fact-checking methods [18, 32] have emerged but struggle with out-of-domain claims. To address these challenges, we propose a novel baseline approach using pre-trained large language models (LLMs) for stance classification, conserving resources and overcoming limitations posed by out-of-domain claims.

While previous efforts to extract knowledge from LLMs have primarily focused on open-domain question answering, to the best of our knowledge, this is the first study to explore the use of LLMs with in-context learning in this domain. The main contributions of this paper are as follows:

- We propose a modern baseline approach for stance classification on given claims and their respective pieces of evidence.
- We find that with the correct input prompts and in-context examples, our approach outperforms all previous works and

achieves state-of-the-art performance on Fever [26], SciFact [29], Climate-Fever [7], Fever-Symmetric [24], and Fever-Symmetric-Generated [24] datasets.

2 Related work

Recent advancements in stance classification have highlighted significant hurdles linked to supervised learning models, particularly regarding their vulnerability to biases specific to the training data [32]. For example, Schuster et al. [24] showcase the efficacy of a claim-only model, which assesses individual claims in solitude, independent of supporting evidence. The model's superiority to baseline systems suggests the potential exploitation of dataset peculiarities rather than a true grasp of linguistic nuances. Similarly, Thorne et al. [27] underscored the susceptibility of various Fever systems to adversarial conditions, where even minor perturbations resulted in significant performance drops. In the realm of natural language inference (NLI), prior research [11, 20] has unveiled neural models' vulnerability to superficial correlations present in the data, indicating a reliance on surface-level cues over genuine linguistic comprehension. These observations collectively hint at the presence of annotation artifacts within datasets, which could introduce biases and impact model effectiveness.

In response to these challenges, there is a growing need for approaches that are independent of specific datasets and do not require extensive training [21, 22]. To tackle these issues, zero-shot fact-checking approaches have emerged [18, 32]. However, these methods also have inherent limitations. For example, zero-shot approaches often struggle with out-of-domain claims if the training set differs extensively from the validation set, due to a lack of specific training on such instances. This limitation can lead to inaccurate or unreliable predictions, particularly with novel or unfamiliar topics.

In this paper, we aim to address the aforementioned challenges by proposing a novel baseline approach. Our method utilizes pre-trained LLMs for stance classification, offering the dual benefit of conserving resources and reducing bias. It also overcomes the limitations posed by out-of-domain or previously unseen claims due to the generic nature of LLMs, which have already been exposed to vast corpora of textual data.

3 Methodology

3.1 Problem statement

Stance classification is the task to decide whether a given claim is either supported by a given evidence, refuted by the evidence, or whether there is not enough information to make such a decision [8]. More formally, let C be the set of all claims, $\mathcal{E}$ the set of all evidence, and $\mathcal{S} = \{\textsc{Supports}, \textsc{Refutes}, \textsc{NotEnoughInfo}\}$ the set of stances. The goal of stance classification is to assign a stance $y_i \in \mathcal{S}$ to a given pair comprising a claim $c_i \in C$ and a piece of evidence $e_i \in \mathcal{E}$ [8, 27, 32]. We define stance classification as a single-label multi-class classification function f as follows:

$$f : C \times \mathcal{E} \rightarrow \mathcal{S}\,. \tag{1}$$

A Dataset $D = ((c_i, e_i), y_i)$ for stance classification comprises claim-evidence pairs and their stance. It is typically divided into training (D_T), validation (D_V) and test data (D_E).

3.2 ExPrompt

Instruction	You are an expert in stance detection. You only have three options (REFUTES, SUPPORTS, and NOT ENOUGH INFO) to detect stance from a textual evidence: refuting, supporting, or not finding enough information for the given claim. Only output must be one of these three options: (1) REFUTES, (2) SUPPORTS, or (3) NOT ENOUGH INFO.
Examples	Examples of each cases are the following: Example 1: (c_1, e_1) Answer: REFUTES. Example 2: (c_2, e_2) Answer: SUPPORTS. Example 3: (c_3, e_3) Answer: NOT ENOUGH INFO.
Instruction	You should not output more than the option, i.e., (1) REFUTES, (2) SUPPORTS, or (3) NOT ENOUGH INFO. The output should not contain explanations, notes, or numbers, and it should not begin with a number.
Task	The given claim is: c_t Given textual document is: e_t

Figure 1: The template for an LLM prompt.

Our approach ExPrompt uses few-shot in-context learning [3, 16], i.e., it relies on the idea that a pre-trained LLM can be used to tackle the stance classification task when it is queried with a fine-tuned prompt containing class examples. We use the template shown in Figure 1 to generate candidates for this prompt. The template starts with instructions, gives three examples—one per class—before it briefly repeats the instructions and gives the actual task, i.e., the current claim-evidence pair (c_t, e_t) that should be classified.

Choosing the examples for the classes that are used in the prompt can be done in various ways, e.g., a domain expert can choose the examples manually. However, since ExPrompt has the goal to be a baseline, we choose an automatic and basic approach by randomly sampling N example sets. Figure 2 gives an overview of this process. Let $D_j = (c_1, e_1), (c_2, e_2), (c_3, e_3)$ be the j-th set of examples comprising three pairs that have the classification labels Supports, Refutes, and NotEnoughInfo, respectively. These pairs are randomly sampled from the training data D_T. We insert the chosen pairs into our prompt template to generate the prompt p_j. We evaluate this prompt by measuring the performance of the LLM when it is queried with p_j and the claim-evidence pairs from the validation data D_V. We repeat this until we have generated N sets. Finally, we use the examples from the set with which the LLM achieved the best performance on the validation set.

An advantage of our approach in comparison to previous works that made use of LLMs is that we do not train the LLM itself and it can be used with a small number of iterations to find a set of examples. Hence, our approach needs less computational power and, thus, consumes less resources.

4 Evaluation

In this section, we describe the datasets and LLMs used in our experiments as well as competing approaches.

Table 1: Post-processing statistics comprising the number of claims in the datasets. The abbreviations are: S/SUPPORTS, R/REFUTES, NEI/NOTENOUGHINFO, and #/Number of.

Dataset Name	Year	Source	Text type	Train	Validation	Labels
FEVER [26]	2018	Wikipedia	Wiki pages	145,449	9,999	S/R/NEI
FEVER-Symmetric [24]	2019	Wikipedia	Wiki pages	956	956	S/R
FEVER-Symmetric-Gen. [24]	2019	Wikipedia	Wiki pages	285	285	S/R
SCIFACT [29]	2020	Scientific articles	Abstracts	1109	300	S/R/NEI
Climate-FEVER [7]	2021	Wikipedia	Wiki pages	7,675	1,535	S/R/NEI

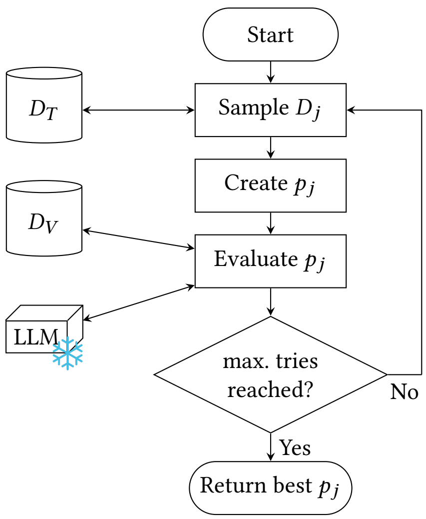

Figure 2: Overview of the proposed workflow to choose examples for the prompt. The snowflake means that the LLM is not changed by this process.

4.1 Datasets

In our evaluation, we utilize five benchmark datasets listed in Table 1. The FEVER dataset comprises 155,448 claims generated by modifying sentences from Wikipedia, which are subsequently verified against Wikipedia without access to their original sentences. The FEVER-Symmetric dataset [24] addresses biases identified in the original FEVER dataset by employing a regularization procedure to mitigate potential biases from giveaway phrases. The complete FEVER-Symmetric test set comprises 956 claim-evidence pairs. These pairs were created by manually generating a synthetic pair for each claim-evidence pair, maintaining the same relation (SUPPORTS or REFUTES) as the original FEVER dataset while expressing a contradictory fact. Following their creation, Schuster et al. selected two individuals to annotate a randomly chosen subset of 285 claim-evidence pairs (representing 30% of the total pairs in the FEVER-Symmetric test set) with labels indicating SUPPORTS, or REFUTES, dubbed FEVER-Symmetric-Generated. Their agreement with the dataset labels was observed in 94% of cases, resulting in a Cohen's κ of 0.88 [5]. We extracted all SUPPORTS and REFUTES claims, and their corresponding gold evidence sentences, from these two datasets for our evaluation. The Climate-FEVER dataset [7] is specifically designed to verify real-world climate change claims, excluding those disputed. The SCIFACT [29] dataset comprises scientific claims verified against a corpus of 5,183 abstracts. Each claim is annotated with rationales from abstracts that either support or refute it.

We exclude the FEVER 2.0 [27] dataset from our analysis because it is tailored for methods that leverage structured data, such as tables sourced from Wikipedia. Additionally, we exclude the AVeriTeC [23] dataset because it comprises question-answer pairs rather than evidence sentences, as it is primarily designed for question-answering tasks.

4.2 LLMs

In our evaluation, we use Mixtral-8x7B and Llama-3-70B as pre-trained LLMs. We describe both models in the following.[1]

4.2.1 Mixtral-8x7B. Mixtral [12] is a large language model that uses a sparse mixture of expert models. For each token, it uses 2 out of 8 experts, that are implemented as feed-forward networks. As a result, for each token, only a limited set of all model parameters is used, which allows faster inference time. It outperforms the Llama2 model [28] with 70B parameters, on tasks such as mathematics and code writing by using fewer parameters [12].

4.2.2 Llama-3-70B. Llama 3 is a publicly available large decoder-only language model, developed by Meta.[2] There are different versions available ranging from 7 billion up to 70 billion parameters. Compared to Llama 2, the Llama 3 model uses a tokenizer with 128K tokens and grouped query attention.[3]

4.3 Competitors

We compare our system with several approaches including the state of the art approaches for stance detection by reusing results available in various publication for the same datasets that we use. To the best of our knowledge, we include all available results reported for recent stance classification approaches on the selected datasets. Schuster et al.[24] present the accuracy results of three different classifiers: NSMN[17], ESIM [4], and BERT [31] on the FEVER-Symmetric and FEVER-Symmetric-Generated datasets. NSMN is derived from the ESIM model and further enhanced with additional features like contextual word embeddings [19]. Additionally, Schuster et al. train their own ESIM model using GloVe embeddings, leveraging code provided by [9]. The third classifier is based on a fine-tuned BERT model that has been trained for three epochs to

[1]These 2 LLMS are open-source and available in the Ollama framework. We use the latter for an efficient setup and fast inference. https://ollama.com/
[2]https://llama.meta.com/
[3]https://ai.meta.com/blog/meta-llama-3/

Table 2: Accuracy scores on test sets. Y is the abbreviation for Yuan et. al. [32].

Datasets	Supervised						Zero-Shot				Encoders							Ours	
	BEVERS [6]	Diggelmann et al. [7]	DREAM [33]	GEAR [34]	KGAT [15]	Random Guess	QACG [18]	Y-base [32]	Y-large [32]	Y-large + USchema [32]	RoBERTa-base [14]	RoBERTa-large [14]	SCIBERT [2]	BioMedRoBERTa [10]	NSMN [17]	ESIM [4]	BERT [31]	ExPrompt (Mixtral:8x7B)	ExPrompt (Llama3:70b)
Fever	80.2	77.7	76.9	71.6	72.8	33.3	–	38.6	60.4	61.3	36.1	58.1	–	–	–	–	–	80.9	**82.9**
Fever-Symmetric	75.9	–	–	–	–	50.0	77.1	56.6	79.8	79.8	51.7	78.9	–	–	81.8	80.8	86.2	**93.6**	93.5
Fever-Sym-Gen.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	58.7	55.9	58.3	**81.0**	**81.0**
SciFact	73.2	–	–	–	–	–	–	–	–	–	62.9	75.7	69.2	71.7	–	–	–	**87.3**	**81.1**
Climate-Fever	–	38.8	–	–	–	33.3	–	–	46.7	46.7	–	44.4	–	–	–	–	–	63.9	**70.2**

classify the relation by concatenating the claim and evidence using a delimiter token.[4]

We further report the zero-shot results from Yuan et al. [32], who also provide the results of a random guessing baseline and QACG [18]. Yuan et al. [32] also utilize the Wikidata5m dataset [30] for training a universal schema (dubbed "large + USchema") model. We also report the results of supervised approaches such as BEVERS [6] and Diggelmann et al. [7] as reported by Yuan et al. [32]. BEVERS uses a transformer model [25] for the stance detection task. However, Diggelmann et al. use an ALBERT (large-v2) model [13] with a three-way classifier applied to the [CLS] token of the concatenated claim and evidence sentences. We also report the results of top-3 approaches reported by Zhong et al. [33]: DREAM [33], KGAT [15], and GEAR [34]. DREAM and KGAT regard pieces of evidence as nodes in a graph and utilize a Kernel Graph Attention Network to aggregate information. GEAR uses BERT for claim-specific evidence representation and applies a graph network, treating each evidence sentence as a node. Additionally, we report the results for sentence encoder-based approaches, namely SCIBERT [2], BioMedRoBERTa [10], RoBERTa-base [14], and RoBERTa-large [14], on the SciFact dataset reported by Wadden et al. [29].

5 Results and Analysis

Table 2 shows the accuracy scores of the different approaches. ExPrompt significantly outperforms all other stance classification approaches with both LLMs.[5] The zero-shot-based approaches are out-performed on the Fever dataset by at 19.6%. BEVERS [6] is currently the state of the art on the Fever dataset and achieves a high accuracy on the Fever-Symmetric dataset without fine-tuning, which highlights the robustness of this model. Our approach outperforms this and the other supervised approaches by at least 0.7% in terms of accuracy, as seen in the Mixtral-based experiments on the Fever dataset. Yuan et al.[32], report that their approach using the large model and BERT [31] are the state of the art on the Climate-Fever and the Fever-Symmetric dataset, respectively. However, our proposed approach outperforms these approaches by at least 17.2% and 7.3% accuracy, respectively.

We use the reported results of sentence encoders on the SciFact dataset from Wadden et al. [29]. We observe that our approach outperforms all encoder-based approaches by at least 5.4% in accuracy. Our Mixtral-based approach achieves an accuracy of 87.3%, while the best encoder-based approach, RoBERTa-large, achieves an accuracy of 75.7% on the SciFact dataset. On the Climate-Fever dataset, we only have results for RoBERTa-large, which achieves 19.5% less than our Llama3-based approach and 25.8% less than our Mixtral-based approach. Additionally, we obtain results for RoBERTa-base, RoBERTa-large, NSMN, ESIM, and BERT on Fever-Symmetric, and for NSMN, ESIM, and BERT on Fever-Symmetric-Generated. However, we observe that our approaches outperform all these methods by at least 7.4% and 22.3% on both datasets, respectively.

6 Conclusion

In this paper, we introduce ExPrompt—a modern baseline approach for stance classification. Our results indicate that using LLMs with fine-tuned prompts and in-context learning outperforms all former state-of-the-art stance classification methods across all datasets used in our evaluation. Hence, ExPrompt can serve as a new contemporary baseline for future stance detection algorithms.

In future work, we plan to evaluate our approach on data that was not available on the web to ensure that the pre-trained LLMs haven't seen the evaluation dataset within the data that they have been trained on. We also plan to optimize the number of examples including an automatic guidance for their selection.

Acknowledgments

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie, No. 860801), the German Federal Ministry of Education and Research (BMBF) within the project NEBULA (13N16364), KIAM (02L19C115), and COLIDE (01IS21005D), the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW) within the project SAIL (NW21-059D).

[4] https://github.com/huggingface/pytorch-pretrained-BERT
[5] We performed a Wilcoxon signed-rank test with $\alpha = 0.05$.

References

[1] Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward "the holy grail": The continued quest to automate fact-checking. In *Computation+ Journalism Symposium,(September)*.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP*.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1657–1668. http://www.aclweb.org/anthology/P17-1152

[5] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[6] Mitchell DeHaven and Stephen Scott. 2023. Bevers: A general, simple, and performant framework for automatic fact verification. *arXiv preprint arXiv:2303.16974* (2023).

[7] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614* (2020).

[8] W. Ferreira and A. Vlachos. 2016. Emergent: a novel data-set for stance classification. https://eprints.whiterose.ac.uk/97416/ © 2016 ACL. This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited..

[9] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, Eunjeong L. Park, Masato Hagiwara, Dmitrijs Milajevs, and Liling Tan (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1–6. https://doi.org/10.18653/v1/W18-2501

[10] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks.. In *ACL*.

[11] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 107–112. https://doi.org/10.18653/v1/N18-2017

[12] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG]

[13] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *CoRR* abs/1909.11942 (2019). arXiv:1909.11942 http://arxiv.org/abs/1909.11942

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019).

[15] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2021. Fine-grained Fact Verification with Kernel Graph Attention Network. arXiv:1910.09796 [cs.CL]

[16] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? arXiv:2202.12837 [cs.CL]

[17] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Association for the Advancement of Artificial Intelligence*. https://arxiv.org/abs/1811.07039

[18] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot Fact Verification by Claim Generation. 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)*. Association for Computational Linguistics, Online, 476–483. https://doi.org/10.18653/v1/2021.acl-short.61

[19] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2227–2237. http://www.aclweb.org/anthology/N18-1202

[20] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana, 180–191. https://doi.org/10.18653/v1/S18-2023

[21] Umair Qudus, Michael Röder, Sabrina Kirrane, and Axel-Cyrille Ngonga Ngomo. 2023. TemporalFC: A Temporal Fact Checking Approach over Knowledge Graphs. In *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part I* (Athens, Greece). Springer-Verlag, Berlin, Heidelberg, 465–483. https://doi.org/10.1007/978-3-031-47240-4_25

[22] Umair Qudus, Michael Röder, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. 2022. HybridFC: A Hybrid Fact-Checking Approach for Knowledge Graphs. In *The Semantic Web – ISWC 2022*, Ulrike Sattler, Aidan Hogan, Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato (Eds.). Springer International Publishing, Cham, 462–480.

[23] Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=fKzSz0oyaI

[24] Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://arxiv.org/abs/1908.05267

[25] Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for Evidence Retrieval and Claim Verification. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 359–366.

[26] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*.

[27] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The FEVER2.0 Shared Task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*.

[28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[29] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[30] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics* 9 (2021), 176–194.

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[32] Zhangdie Yuan and Andreas Vlachos. 2023. Zero-Shot Fact-Checking with Semantic Triples and Knowledge Graphs. (2023). arXiv:2312.11785 [cs.CL]

[33] Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6170–6180. https://doi.org/10.18653/v1/2020.acl-main.549

[34] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 892–901. https://doi.org/10.18653/v1/P19-1085