Fact Checking with LLMs

Peyton Taylor
University of Hawai‘i at Hilo
peytonct@hawaii.edu

Abstract

This paper aims to act as a proof of concept that the fact-checking task of determining the truth of a given claim is plausible and actionable using existing LLM (Large Language Models). The barrier to entry for such a task may be simpler than methods performed by existing models that rely on a large pipeline of complicated NLP (Natural Language Processing) functions. We use a pre-existing model and focus its domain range on up-to-date data. We compare 2 types of the same model: LLAMA:3.2:1b (Touvron et al., 2023), and a modified version of the Thorne et al. (2018) system that will include Llama. We found that the model by itself had a bias to not definitively print labels for inputted claims, in fact, it not abstain from selecting the REFUTES label when given a choice between SUPPORTS, REFUTES, or NOT ENOUGH INFO. By providing supporting evidence with the claim, the model would output all 3 labels.

1 Introduction

Fact-checking is the process of determining if a given claim, oral or text, is factually backed by evidence. This task is often performed individually, as needed, or on-line sites such as `PolitiFact.com` or `FactCheck.org`. Although not common, some journalists outside of the aforementioned sites are also starting to conduct fact checking and include if they did so in their articles. The process involves identifying an existing claim to verify, obtaining supporting evidence or the lack thereof, and labeling the claim based on the retrieved evidence. Furthermore, automated fact-checking is the end goal for fact-checking tasks, which requires determining whether or not a piece of information is a check-worthy claim, and providing on-the-fly fact-checking. For example, ”COVID-19 first appeared in China in 2019” versus ”a dog is a dog”. The check-worthy claim being the former. Due to the nature of on-the-fly claims being difficult to collect in large scale, although some live claims were evaluated, for evaluation purposes we will focus on altering an existing Llama model to make predictions on the veracity of preexisting claims.

Current models are often trained on a wide range of information and may not be able to produce fact-specific claim labels. The common approach is to utilize separate models for the claim, evidence, and verdict stages in a pipeline fashion. An older system described by Thorne et al. (2018) is as follows:

- **Document Retrieval:** Identify existing literature to support a claim by retrieving relevant documents from Wikipedia. In Thorne et al. (2018)’s work, the document retrieval component from DrQA (Chen et al., 2017), produces k nearest documents for a given query using co-sine similarity between unigram and bigram TF-IDF vectors.

- **Sentence Selection:** Select potential evidence from the retrieved documents. Thorne et al. (2018) used a modified DrQA component of (Chen et al., 2017) to select the top relevant sentences, with bigram TF-IDF binning.

- **Textual Entailment:** Produce labeled claims paired with evidence (labels consist of SUPPORTS, REFUTES, or NOT ENOUGH INFO). Thorne et al. (2018) used a decomposed attention model (Parikh et al., 2016) for this task.

Newer existing pipelines implore the use of somewhat varying techniques, but much of the idea is the same. Some newer models use neural networks based on sequence or graph modeling in order to learn contextual meaning surrounding social media activity. They can aid in capturing patterns such as rumorousness around the way information is propagated through such channels (Zubiaga et al., 2016). Recent research utilized domain-specific features in neural models (Zhang et al., 2021). Challenges also arise in the task of determining whether or not a given claim is checkable or check-worthy as the volume of claims produced in real world scenarios is vast (e.g., all social media posts from a given day) and prohibits the retrieval and evidence for such Guo et al. (2022). There doesn't seem to be a clear consensus on the most effective model currently. My approach is different in that for our model, we take an already well generalized, trained model and focus its ability to label the veracity of claims specifically. We also aim to bridge the gap between creating extensive, complicated models by feeding potential evidence for a claim into a preexisting model and having it determine the supporting evidence itself.

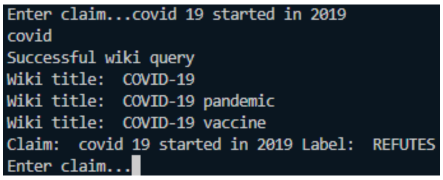
2 Dataset

The human-annotated fact-checked dataset FEVER, introduced by Thorne et al. (2018) consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. The claims are classified as Supported, Refuted or NotEnoughInfo. For the first two classes, the annotators also recorded the sentence(s) forming the necessary evidence for their judgment. The claims were created by annotators extracting claims from Wikipedia and altering them in a variety of ways, some of which would alter the meaning entirely. Claims were then verified by annotators through Wikipedia.

3 Models

We start with the base untrained Llama model and analyze its ability to label claims from FEVER. The untrained model refused to make specific labels for claims, and thus defaulted to REFUTES for every attempt. We did experiment with training Llama using (Developers, 2024) on FEVER to enhance its domain of fact labeling. We tested if using fine-tuning on a single but large dataset dedicated to fact-checking performs better than the base model. We also experimented using RAG and a vector based database, utilizing the same dataset. Results were not impressive in any of the aforementioned attempt, and thus omitted.

??.



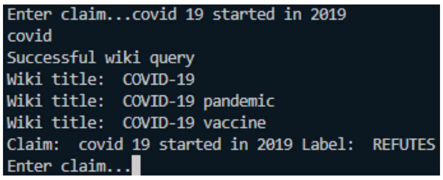
```
Enter claim...covid 19 started in 2019
covid
Successful wiki query
Wiki title:  COVID-19
Wiki title:  COVID-19 pandemic
Wiki title:  COVID-19 vaccine
Claim:  covid 19 started in 2019 Label:  REFUTES
Enter claim...
```

Figure 1: An example on-the-fly claim given by the authors.

4 Related Work

We implement a modified version of the Thorne et al. (2018) system. We took an existing claim from FEVER, extract keywords (such as nouns, proper nouns, etc.), and then perform a query search on the Wikipedia API. Dense vector embeddings and semantic similarity are generated for the input claim and prior retrieved evidence using a sentence transformer (Reimers and Gurevych, 2020), derived from a model developed by Wang et al. (2020). The now augmented model only selects the top 3 similar Wikipedia articles to our claim in order to avoid unnecessary noise. We then re-query the Wikipedia API with our top 3 document titles, and extract parsed data in descending query title order to form our 3 separate documents. The given documents were scored by a similarity index between varying degrees of information: full Wikipedia articles, individual paragraphs, and individual sentences. Parameters such as how many pages to score similarity between the claim were experimented with and taking 3 of the top-most Wikipedia pages to be initially scored, then 1 articles resulted in the best results for the ’article model’. 5 pages and 3 paragraphs worked best for the ’paragraph model’. And the same of 5 pages, 3 paragraphs worked best for the ’sentences model’. We feed the claim along with the evidence (documents) retrieved into Llama in a process akin to RAG (Retrieval Augmented-Generation) and prompt-engineering. The given prompt is ” ”role”: ”user”, ”content”: f”Given the following documents determine if the following claim is true. Documents: documents Claim: claim Only reply with SUPPORTS, REFUTES, or NOT ENOUGH INFO and nothing else. Choose the label with the best evidence.” The model then outputs a label for the claim based on the given documents.

??.



```
Enter claim...covid 19 started in 2019
covid
Successful wiki query
Wiki title:  COVID-19
Wiki title:  COVID-19 pandemic
Wiki title:  COVID-19 vaccine
Claim:  covid 19 started in 2019 Label:  REFUTES
Enter claim...
```

Figure 2: Another on-the-fly claim given by the authors that struggled to produce accurate results.

5 Experiments and Analysis

Our model that separated evidence into sentences had a particularly higher recall score of 0.7 for NOT ENOUGH INFO, and a very low recall score of 0.03 for REFUTES. Its precision for SUPPORTS is relatively the same as the other models, and its REFUTES precision score of 0.27 is higher than the others, but not greatly.

Although it had a better time at selecting NOT ENOUGH INFO, its scores indicate that sentences greatly struggled at providing the correct evidence appropriate for Llama to make an accurate prediction. Separating evidence into paragraphs yielded a more fairly averaged system specifically with a slightly lower precision score for REFUTES of 0.16, but a higher recall score for the same label, 0.12. The other scores were relatively similar. Our article model had a notably larger recall score of 0.32 for REFUTES, but lower score of 0.14 for SUPPORTS of the same score type.

Label	Precision	Recall	F1-Score
NOT ENOUGH INFO	0.27	0.70	0.39
SUPPORTS	0.57	0.31	0.40
REFUTES	0.27	0.03	0.06
Accuracy			0.36
Macro Avg	0.37	0.35	0.28
Weighted Avg	0.44	0.36	0.33

Table 1: Classification metrics for sentences as documents.

Label	Precision	Recall	F1-Score
NOT ENOUGH INFO	0.27	0.60	0.37
SUPPORTS	0.59	0.28	0.38
REFUTES	0.16	0.12	0.14
Accuracy			0.33
Macro Avg	0.34	0.33	0.29
Weighted Avg	0.42	0.33	0.33

Table 2: Classification metrics for paragraphs as documents.

The articles model arguably had the most evenly distributed f1-score over each label. Each model seems to have over-predicted SUPPORTS labels, indicating Llama frequently chose this label based on the provided evidence. No scores were great or consistent overall, and with the no-contest results of the Llama model using no supporting evidence for claims, this indicates the overall lack of efficacy in utilizing current language models to make specific predictions on the veracity of claims.

Label	Precision	Recall	F1-Score
NOT ENOUGH INFO	0.26	0.59	0.36
SUPPORTS	0.56	0.14	0.22
REFUTES	0.22	0.32	0.26
Accuracy			0.29
Macro Avg	0.35	0.35	0.28
Weighted Avg	0.42	0.29	0.27

Table 3: Classification metrics for articles as documents.

The results clearly indicate that although the provided methods emboldened Llama to specifically choose a label, other methods are further required post-claim/evidence-processing in order to accurately label claims. We would like to implement a custom BiLSTM model with a classification head in future work. Although the BiLSTM model wouldn't be trained on on-the-fly data (which is ultimately the end goal for predictions of this task), being able to adjust the pipeline or parameters, and having a model specifically trained for the task of fact-checking, may produce better results.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *CoRR*, abs/1704.00051.

Unsloth Developers. 2024. https://github.com/unslothai/unsloth. Accessed: 2024-12-13.

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

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *Preprint*, arXiv:arXiv:2010.08275. The model all-MiniLM-L6-v2 is available as part of the Sentence Transformers framework.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.