Automated Claim Verification Using Transformer Models and the FEVER Dataset

Mihika Rao, Nina Chinnam

05/09

Motivation

- **Problem:**
  - Misinformation spread rapidly
  - Fact checking is slow
- **Need:**
  - Automated systems
  - Can LLMs discern truth from false?
- **Our Goal:**
  - Build an automated pipeline that:
    - Extracts verifiable claims from text
    - Retrieves relevant evidence
    - Classifies each claim as **Supported**, **Contradicted**, or **Not Verifiable**

Background

- **Fact-Checking in NLP:**
  - Fact-checking needs claim understanding, evidence, retrieval, and classification.
- **FEVER Dataset:**
  - Fact Extraction and VERification
  - 185,000+ human-generated claims
  - benchmark for evaluating fact-checking models
- **Pretrained Language Models:**
  - FLAN-T5 and BART-MNLI:
    - Zero-shot NLI
    - Extracting factual evidence from text

Related Work

- **Traditional Fact-Checking Approaches:**
  - Rule-based systems (e.g., symbolic logic, KB matching) limited by rigidity and scalability (e.g., Vlachos & Riedel, 2014)
- **Neural Approaches with FEVER:**
  - FEVER baseline: sentence retrieval + Recognizing Textual Entailment (RTE) (Thorne et al., 2018)
  - DeFactNLP: used neural entailment models on FEVER claims
- **Transformer Based Pipelines:**
  - BERT for fact-checking improved sentence-level verification (Liu et al., 2019)

Claim / Target Task

- **Problem Definition:**
  - Input: A natural language glaim
    - e.g., “The capital of Australia is Sydney.”
- **Goal:** Automatically classify the claim based on retrieved evidence as:
  - Supported ✓
  - Contradicted ✗
  - Not verifiable ?
- **Task Components:**
  - Claim Understanding
  - Evidence Retrieval
  - Veracity Classification

Intuitive figure showing WHY

1 Input Text

"The Eiffel Tower is in Berlin and was built in 1887"

2 Claim Extraction

- Use LLM to extract factual claims
- spaCy, transformers, etc.

"The Eiffel Tower is in Berlin"

"It was built in 1887"

3 Information Retrieval

1. Generate search query
2. Use API (WIkipedia)
3. Retrieve top-k snippets

Evidence
The Eiffel tower is in Paris
It was built between 1887-1889

4 Claim Verification

- Feed claim + evidence into the language model
- Prompt-based LLM Reasoning

```
prompt = f"""
Claim: "{claim}"
Evidence: "{evidence}"
Based on the evidence, is the claim Supported, Contradicted, or Not Verifiable?
Answer with one of the three labels and provide a one-line justification.
"""
```

5 Output Display

- Produce a final output with claims color coded

"The Eiffel Tower is in Berlin and was built in 1900"

Proposed Solution

- **Claim Extraction (Flan-T5):**
  - Text2text model prompts the generation of discrete factual claims from a user-submitted paragraph
- **Evidence Retrieval (Wikipedia):**
  - Each claim is used to query Wiki via keyword/entity extraction
  - Returns a context snippet from most relevant Wiki page
- **Claim Verification (BART-MNLI):**
  - Each claim and evidence pair is passed to a zero-shot classifier
  - “This claim is {Supported/Contradicted/Not Verifiable} based on the evidence:..”
- **Output:**
  - Each claim gets assigned a label

Implementation

- **Claim Extraction:**
  - Uses Flan-T5 via text2text-generation to extract factual claims from input text
  - Each generated line is further split into atomic sentences using spaCy
- **Evidence Retrieval:**
  - Extracts best search term using entity recognition
  - Uses Wikipedia to search for matching pages
  - Returns the first 1000 characters of relevant article content
- **Claim Verification:**
  - Uses facebook/bart-large-mnli model for zero-shot classification
  - Forms natural language hypotheses using label-specific template
  - Returns the label with highest entailment probability

Implementation

- **Frontend:**
  - Built with Gradio Blocks UI
  - Users enter a paragraph -> extracted claims populate a dropdown
  - Selected claim triggers evidence retrieval and verification output

Data Summary

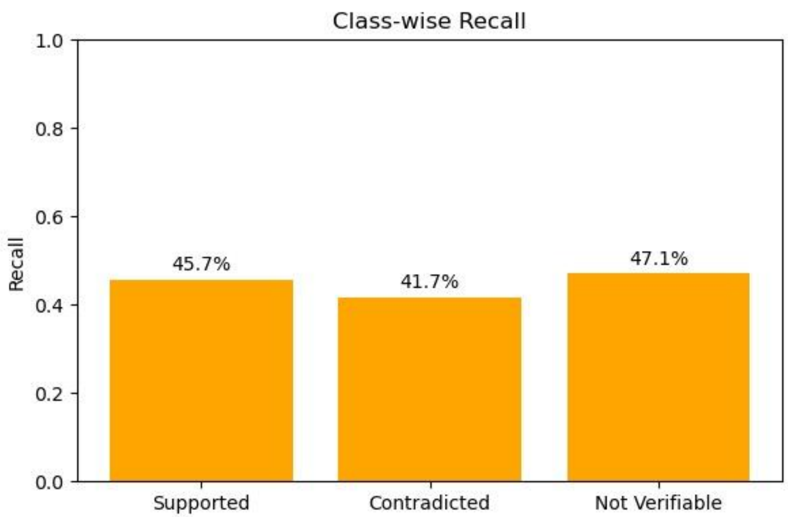
- **FEVER** = Fact Extraction and VERification
- Created for large-scale evaluation of fact-checking systems
- Built on Wikipedia as the sole source of evidence
- Statistics:
  - Total Claims: ~185,000
  - Train Set: ~145,000 claims
  - Dev Set: ~20,000 claims
  - Avg. Evidence Sentences: ~1-5/claim
- Labels:
  - Supports: Claim is fully supported by evidence
  - Refutes: Claim is clearly contradicted
  - Not Enough Information

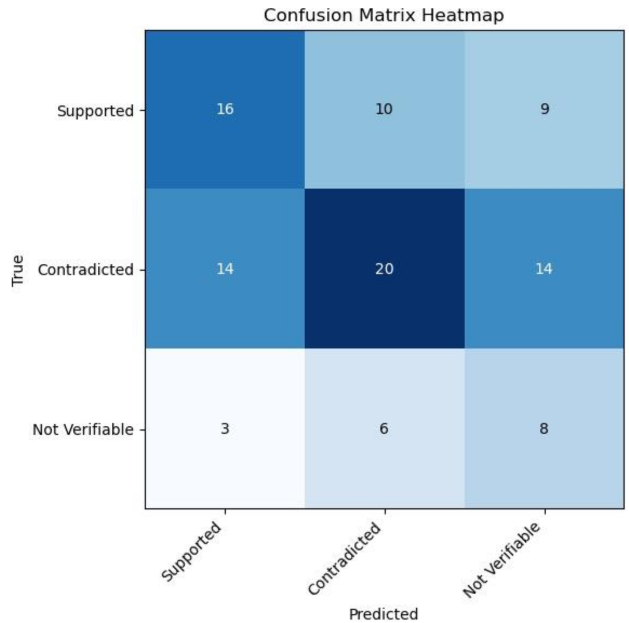
"Barack Obama was the first American president to be born in Kenya"

Label: REFUTES

Evidence: [
  [
    ["Barack_Obama", 0, "Barack Obama was born on August 4, 1961, at Kapiʻolani Medical Center for Women and Children in Honolulu, Hawaii."]
  ]
]

Experimental Results





Gold \ Pred	Supported	Contradicted	Not Verifiable	Total
Supported (35)	16	10	9	35
Contradicted (48)	14	20	14	48
Not Verifiable (17)	3	6	8	17
Total Pred	33	36	31	100

Experimental Analysis

- **Key Metrics:**
  - Overall accuracy: 44% (44/100)
  - Class-wise recall:
    - Supported: 16/35 -> 45.7%
    - Contradicted: 20/48 -> 41.7%
    - Not Verifiable: 8/17 -> 47.1%
- **Main Failure Modes:**
  - Over-abstention: 31% of all predictions are “Not Verifiable,” even when evidence is clear
  - Contradiction detection weak: only 42% recall on false claims
  - Support under-detection: ~46% recall on true claims

Conclusion and Future Work

- Future Work:
  - Model & Pipeline Improvements
    - Integrate dense retrievers for more accuracy evidence retrieval than simple keyword-based search
    - Replace zero-shot NLI with fine-tuned verifiers trained directly on FEVER for improved claim classification.
  - Real-World Adaptation:
    - Apply the pipeline to real-world misinformation, such as social media posts or news headlines
    - Add explanation generation to improve the user trust and interpretability.

References

Vlachos, A., & Riedel, S. (2014). *Fact Checking: Task definition and dataset construction*. ACL.

Thorne, J., et al. (2018). *FEVER: a large-scale dataset for Fact Extraction and VERification*. NAACL.

Yoneda, T., et al. (2018). *UCL Machine Reading Group: Four Factor Framework for Fact Verification*. FEVER Workshop.

Liu, Y., et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv.

Lewis, M., et al. (2020). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation*. ACL.

Yin, W., et al. (2019). *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach*. EMNLP.

Wadden, D., et al. (2020). *Fact or Fiction: Verifying Scientific Claims*. EMNLP.

Rashkin, H., et al. (2021). *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. ACL.

Raffel, C., et al. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. JMLR.