Multi-Hop Fact Checking of Political Claims

Introduction

Fact checking in the real world often requires analyzing multiple connected pieces of evidence. However, most existing datasets and models simplify this process by focusing on single-sentence evidence or artificially constructed claims. In response to this limitation, this paper introduces **PolitiHop**, a novel dataset specifically designed to support **multi-hop reasoning** in the context of **political claim verification**.

Dataset: PolitiHop

The authors present **PolitiHop**, a dataset containing **500 real political claims** extracted from **PolitiFact**. Each claim is annotated with:

- A **veracity label** (True, False, Half-True)
- Multiple **sets of evidence sentences**, selected from the full article text

Unlike prior datasets such as FEVER, which typically require only one sentence for verification, PolitiHop emphasizes **complex reasoning** by requiring multiple evidence sentences (often 2–3+ per claim).

Models and Methods

The authors evaluate several models:

- **Baselines**: Majority class, Random, TF-IDF
- **Single-hop model**: BERT
- **Multi-hop model**: Transformer-XH

Transformer-XH incorporates **graph attention layers** that simulate multiple reasoning "hops" between sentences, which makes it better suited for complex evidence reasoning.

Experiments and Findings

The experiments address three key questions:

1. Can multi-hop models perform well on PolitiHop?
2. Are they robust to overlapping named entities (NEs)?
3. Can reasoning skills transfer from other datasets?

Key Results:

- Transformer-XH generally outperforms BERT, especially in **evidence retrieval**.
- **Pre-training on LIAR-PLUS** (a related dataset from PolitiFact) followed by fine-tuning on PolitiHop produces the best results.
- **Pre-training on FEVER** (a Wikipedia-based dataset) does **not help** and may hurt performance due to domain mismatch.
- Transformer-XH shows **greater robustness** to misleading overlaps in named entities.

Error Analysis and Insights

- **Larger evidence sets** lead to more prediction errors.
- Fixed hop counts (e.g., 3-hop, 6-hop) do not consistently outperform others; dynamic hop selection might be more effective.
- **Attention analysis** shows that evidence sentences get higher attention scores than non-evidence ones, though the difference is modest.

Conclusion

This work is the first to focus on **multi-hop fact checking of real-world political claims**. It contributes:

- A challenging, explainable dataset (PolitiHop)
- Insights into the performance of multi-hop models
- Evidence that **domain-specific pre-training is critical**

While multi-hop models show promise, **accurate and explainable evidence retrieval** remains a key challenge in fact checking.