CS 6120: Multi-Hop Fact Verification System

Roshan Chouhan – chouhan.ros@northeastern.edu
Daniel Xiong – xiong.d@northeastern.edu
Ashwin Satish – satish.ash@northeastern.edu
https://github.com/rchauhan1001/NLP-Final-Project/

Abstract

This project presents a multi-hop fact verification system designed to combat misinformation by verifying claims against multiple Wikipedia sources. The system implements a three-stage pipeline: BM25-based retrieval from 5.5 million Wikipedia articles, dense retrieval reranking for semantic understanding, and RoBERTa-Large-MNLI for final verification.Key challenges included computational constraints preventing iterative multi-hop retrieval and low initial evidence coverage. Results show that retrieval quality is the primary bottleneck in fact verification pipelines, with semantic understanding providing measurable improvements over keyword-based approaches.

Introduction

Misinformation spreads faster than verified facts, particularly regarding public figures and breaking news. Traditional fact-checkers verify claims against single sources, but real-world claims require connecting evidence across multiple documents. For example, rumors about Lewis Hamilton retiring from Formula 1 spread through media without factual basis, requiring multi-source verification to separate facts from speculation.

This project addresses automated fact verification requiring multi-hop reasoning across documents. We developed an end-to-end pipeline that processes textual claims, retrieves relevant evidence from 5.5 million Wikipedia articles, connects evidence across sources, and classifies claims as SUPPORTED or REFUTED with confidence scores.

The pipeline consists of three stages: BM25 retrieval for fast initial candidate selection, dense retrieval reranking for semantic refinement, and RoBERTa-Large-MNLI for natural language inference. The system outputs predictions with retrieved document references for transparency and interpretability.

1 Background/Related Work

Multi-hop fact verification requires connecting information across multiple documents, unlike single-hop verification using isolated evidence. The HoVER (Hop-over) benchmark specifically addresses this challenge through claims requiring reasoning across 2-4 Wikipedia articles.

Information Retrieval Approaches: Sparse retrieval methods like BM25 use TF-IDF scoring with length normalization for fast keyword matching but miss paraphrases and semantic relationships. Dense retrieval uses neural embeddings from transformer models to encode text into vectors, enabling semantic similarity matching beyond exact word overlap. While computationally intensive, dense methods better handle paraphrases and contextual meaning.

Natural Language Inference: RoBERTa-Large-MNLI represents state-of-the-art for textual entailment. Pre-trained on Multi-Genre Natural Language Inference data, it determines relationships between text pairs, making it suitable for matching claims against evidence. The model provides confidence scores rather than binary outputs, enabling threshold-based classification tuning.

Prior work on HoVER shows iterative retrieval methods achieving ~75% accuracy through multiple retrieval hops, while single-stage BM25 baselines achieve ~52.6% accuracy.

2 Data

HoVER Dataset: The HoVER (Hop-over Wikipedia-based Verification) dataset contains 18,171 training claims, 4,000 development claims

(2,000 SUPPORTED, 2,000 NOT_SUPPORTED), and 4,000 test claims without labels. Claims are specifically designed to require multi-hop reasoning across multiple evidence sources.

Wikipedia Corpus: Evidence source consists of 5.5 million preprocessed Wikipedia articles from the HotpotQA dump, totaling approximately 30GB raw text. This corpus provides comprehensive factual knowledge across diverse domains.

Preprocessing: Due to the 30GB dataset size, we implemented streaming processing using the ijson library to avoid memory overflow. Articles were preprocessed to create an inverted index mapping terms to documents. The indexing process compressed the corpus from ~30GB to an ~8GB indexed database for efficient retrieval.

3 Methods

System Architecture: Three-stage pipeline: (1) BM25 retrieval from indexed Wikipedia, (2) Dense retrieval reranking, (3) RoBERTa-Large-MNLI verification.

Stage 1 - BM25 Retrieval: BM25 algorithm uses probabilistic ranking based on term frequency (TF), inverse document frequency (IDF), and length normalization. Creates inverted index mapping terms to documents for fast lookup. Retrieves top 100 candidate documents per claim for broad coverage with millisecond-speed performance.

Stage 2 - Dense Retrieval Reranking: Transformer-based encoder converts claims and documents into dense vector embeddings. Computes semantic similarity between claim and each of top 100 BM25 candidates. Reranks based on neural similarity rather than keyword overlap. Outputs top 10 most semantically relevant documents, capturing paraphrases and contextual meaning missed by BM25.

Stage 3 - RoBERTa – Large - MNLI Verification: From each of top 10 documents, extracts top 5 relevant sentences (up to 50 claim-sentence pairs per claim). Passes each pair through RoBERTa-MNLI to obtain entailment scores (SUPPORTS/REFUTES/NEUTRAL). Aggregates confidence scores across all evidence sentences. Applies optimized threshold (0.05) to determine final SUPPORTED or REFUTED label.

Baseline: BM25-only retrieval with direct RoBERTa verification (no dense reranking) serves as baseline for comparison.

4 Experiments

Training: No model training performed. Used pre-trained models: standard BM25 implementation for retrieval, pre-trained transformer for dense embeddings, and RoBERTa-Large-MNLI without fine-tuning. Constraint: Hardware limitations prevented fine-tuning on HoVER's 18,171 training examples.

Hyperparameters:

- BM25: Optimized k1 and b parameters for fact verification
- Dense Retrieval: Top 100→10 reranking
- Verification: Sentence aggregation (top 5 per document), threshold optimization on dev set
- Optimal threshold: 0.05 (determined through grid search on development set)

Infrastructure: Initial development on local machine, transitioned to Google Colab for GPU-accelerated dense retrieval. Single-stage retrieval pipeline due to computational constraints (versus iterative multi-hop in literature).

Evaluation Metrics:

- Retrieval: Recall@10, Recall@100, Coverage@10
- Verification: Accuracy, F1-Score, Precision-Recall curves
- Evaluated on 4,000-claim development set with balanced SUPPORTED/NOT_SUPPORTED distribution

Inference: For each claim: (1) BM25 retrieves top 100 documents, (2) Dense reranker selects top 10, (3) Extract top 5 sentences per document, (4) RoBERTa scores each claim-sentence pair, (5) Aggregate scores and apply threshold for final prediction.

4.1 Results

- Recall@10: 43.24% (+9.48pp improvement), Coverage@10:83.97%
- Verification: 52-52.5% accuracy, ~57% F1-score (+2pp)
- Threshold: 0.05 optimal (consistent)

- Improvements: Smoother precision-recall curves, more balanced predictions, relevant documents pushed to ranks 1-5

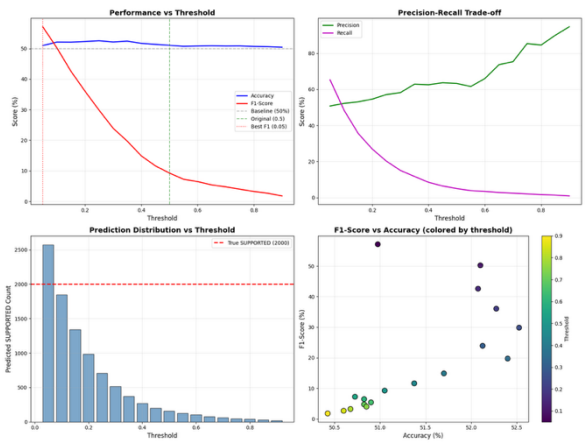

Figure 1 Evaluation Results

5 Conclusions

Successfully developed end-to-end multi-hop fact verification pipeline achieving 52.5% accuracy and 57% F1-score on HoVER dataset. Dense retrieval reranking improved Recall@10 by 9.48 percentage points, demonstrating semantic understanding's value over keyword matching.

Key Findings: Retrieval quality is the primary bottleneck—poor retrieval fundamentally limits verification regardless of NLI model quality. Dense retrieval's semantic understanding directly translated to better verification performance. Single-stage retrieval proves insufficient for true multi-hop reasoning requiring connected evidence chains across 2-4 documents.

Limitations: Computational constraints prevented iterative multi-hop retrieval (HoVER paper's approach). Only 20% of gold-labeled supporting documents found by our retrieval. Binary classification forced (no NOT_ENOUGH_INFO class). No fine-tuning on 18,171 training examples due to hardware limitations.

Future Work: Implement iterative two-stage retrieval to find connected evidence chains. Fine-tune models on HoVER training data for domain adaptation. Add three-class classification (SUPPORTED/REFUTED/NOT_ENOUGH_INFO). Improve sentence selection from retrieved documents. These improvements could potentially reach 65-75% accuracy range.

Impact: Demonstrates practical fact verification with constrained resources. Validates importance of retrieval in multi-hop reasoning tasks. Provides foundation for combating misinformation through transparent, evidence-based verification systems.

6 References

1. Thorne et al. (2018) - FEVER: Fact Extraction and VERification: Introduces the FEVER dataset with 185K claims and establishes the baseline pipeline for evidence retrieval and claim verification.
2. Jiang et al. (2020) - HoVer: A Dataset for Many-Hop Fact Extraction: Presents our primary dataset specifically designed for multi-hop reasoning, where claims require connecting 2-4 pieces of evidence.
3. Zhou et al. (2019) - GEAR: Graph-based Evidence Aggregating and Reasoning: Proposes graph attention networks to model evidence relationships and aggregate multiple pieces of evidence for better verification.
4. Zhao et al. (2020) - Transformer-XH: Multi-hop question answering:Demonstrates how transformer architectures can handle multi-hop reasoning by learning to compose information across multiple passages.
5. Guo et al. (2022) - Survey of Automated Fact-Checking: Comprehensive overview of current fact-checking methods, datasets, and challenges in the field.
6. Xiong et al. (2021) - Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval: About multi-hop dense retrieval which is related to the evidence retrieval stage