MULTI-HOP CLAIM VERIFICATION

Team 7:
Gireeshee Pendela(gp590)
Veera Jeeshitha Kolla(vk536)

Problem Statement

- The goal of this project is to build an automated system that can verify the factual correctness of a given claim by analyzing textual information.
- Each claim must be classified into one of three categories:
    1. Supported: Evidence confirms the claim
    2. Refuted: Evidence contradicts the claim
    3. Not Enough Info(NEI): Evidence is insufficient or missing

NLP Task:

- It is a NLP text classification task
- The model reads the claim written in natural language
- It predicts whether the claim is supported by known facts

DATASET:

- FEVER - Large-scale fact verification dataset(~185k claims)
- Each claim labeled as supported/ refuted/ NEI
- Based on real evidence from Wikipedia
- Standard benchmark for claim verification research

APPROACH & PROGRESS

Training Setup:

- The FEVER dataset is split into 70% training, 15% validation, and 15% testing to build, tune, and evaluate the models reliably.

Feature Representation:

- Each claim is converted to numerical form using TF-IDF, capturing important words and short phrases that help differentiate between Supported, Refuted, and NEI labels.

Baseline Models Implemented:

- We trained multiple classical machine learning models including Logistic Regression, Support Vector Machine (SVM), and Random Forest, using TF-IDF features for claim classification.

Evaluation Strategy:

- Model quality is measured using Accuracy and Macro-F1, to ensure fair evaluation across all three classes.

Current Progress:

- Data preprocessing, training, and evaluation are successfully completed, and the system can now classify new claims with the baseline models and DistilBERT.

DISTILBERT FINE-TUNING APPROACH

Why DistilBERT?

- DistilBERT is a lighter, faster version of BERT that still retains strong language understanding. It is well-suited for claim verification because it can read the entire claim + evidence sequence and reason about the relationship between them.

Training Setup:

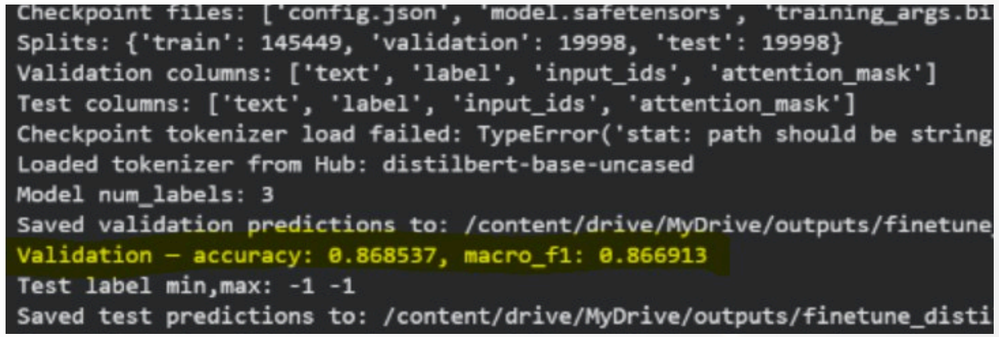
- Base model: distilbert-base-uncased
- 70% training, 15% validation, 15% testing
- Input format:
- “claim [SEP] evidence sentences”
- Sequence length capped at 128 tokens to reduce training time.
- Trained for 3 epochs using mixed-precision (fp16) on GPU with an effective batch size of 32.

Model Performance:

- DistilBERT significantly improves over classical baselines by using contextual embeddings and attention mechanisms.
- It achieved:
  - Accuracy: ~0.87
  - Macro-F1: ~0.87
- This shows strong balanced performance across all three claim labels.

RESULTS & NEXT STEPS:

Model	Accuracy	Macro-F1
Logistic Regression	0.694	0.688
SVM (Best)	**0.719**	**0.718**
Random Forest	0.711	0.700



KEY OBSERVATIONS:

- DistilBERT clearly outperformed all classical baselines, achieving around 0.87 accuracy and macro-F1, demonstrating the advantage of contextualized transformers.
- SVM was the strongest traditional model on TF-IDF features.
- Random Forest performed better than expected, possibly because the sampled dataset captured some nonlinear patterns.

NEXT STEPS:

- Fine-tune BERT or RoBERTa to compare their performance with DistilBERT and determine whether larger transformer models yield further gains.
- Evaluate all models using the same metrics (Accuracy + Macro-F1) to ensure a fair comparison across baselines and transformers.

THANK YOU