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

EVALUATION STRATEGY

Training Setup:

The dataset was split into:

- 70% Training
- 15% Validation
- 15% Testing

Feature Representation:

- Each claim is converted into numerical form using TF-IDF (Term Frequency–Inverse Document Frequency).
- TF-IDF captures important words and short phrases that help differentiate across: Supported, Refuted, NEI
- This representation is well-suited for traditional ML models and fast to compute.

Evaluation Strategy:

Model performance is measured using two key metrics:

- Accuracy – Measures overall correctness of predictions.
- Macro-F1 – Gives equal weight to all three classes, ensuring fair evaluation even when class distribution is imbalanced.

BASELINE MODELS

Baseline Models Implemented:

We trained the following classical machine learning models using TF-IDF features:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest Classifier

RESULTS:

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

KEY OBSERVATIONS:

- SVM outperforms all baseline models, achieving the highest accuracy and macro-F1 due to its ability to separate high-dimensional TF-IDF features effectively.
- Random Forest performs better than expected, suggesting the dataset contains non-linear word patterns that tree-based models can capture.

TRANSFORMER MODELS

Transformer Models Implemented:

We trained the following transformer-based models after preprocessing and tokenizing the FEVER dataset:

- DistilBERT
- BERT
- RoBERTa

RESULTS:

Transformer Model Performance

Model	Accuracy	Macro-F1
DistilBERT	0.8685	0.8669
BERT	0.8711	0.8695
RoBERTa	0.8707	0.8687

KEY OBSERVATIONS:

- All three Transformer models perform almost identically (~0.87 accuracy), showing that the FEVER task is handled very well by pretrained models.
- DistilBERT matches BERT and RoBERTa, despite being smaller and faster, indicating that model size does not significantly impact performance here.

THANK YOU