Team Number: 7
Team Members:
1. Gireeshee Pendela (gp590)
2. Veera Jeeshitha Kolla (vk536)

Multi-Hop Claim Verification

Description:

Our project works toward developing a system that can automatically identify whether a claim made in writing is true or false through the use of many pieces of text evidence. This is called multi-hop claim verification and is an important problem in NLP because it requires reasoning across more than one document or sentence to make a final decision.

For instance, if the claim is “Marie Curie won two Nobel Prizes in different sciences,” then the model has to be able to find different texts talking about Marie Curie, her Nobel Prizes, and the fields she won them for. Then it will connect this information and make a decision on whether the claim is true (Supported), false (Refuted), or if there isn’t enough information (Not Enough Info).

We will use transformer-based models like BERT or RoBERTa, which understand text in context. The baseline we plan to create first is a simple TF-IDF-based classifier, and then fine-tune a pre-trained transformer model for better results.

FEVER or HotpotQA are some of the available datasets that already have claims and supporting evidence for training and testing. We will divide the data into training, validation, and test sets in order to measure the model's performance.

Here, to check the quality of our model, we will make use of metrics such as accuracy, precision, recall, and F1-score. These show how frequently a model makes correct predictions and deals with challenging examples.

Our general objective in this work is to study how NLP models can make human-like reasoning by combining information from different sources before checking a claim.

NLP Task:

The broad NLP task under consideration for this project is text classification. It takes a given claim and classifies it into one of three categories - Supported, Refuted, or Not Enough Info using multiple pieces of textual evidence.

Since the model needs to read and reason across multiple sentences or documents to determine if a claim is true or not, this task also contains elements of multi-hop reasoning or fact verification, a specialist form of classification in NLP.

Dataset:

In this project, we will be working with **fact verification datasets** that support reasoning over multiple pieces of textual evidence. These datasets provide labeled claims along with corresponding evidence sentences, enabling the training and evaluation of models that can perform multi-hop reasoning and claim verification.

1. FEVER (Fact Extraction and VERification)

Link:https://huggingface.co/datasets/fever/fever
In this work, we will be using the **FEVER** dataset, which includes approximately **185,000 claims** generated from Wikipedia articles and labeled as **Supported**, **Refuted**, or **Not Enough Info**. Evidence sentences that support or refute a claim are also included, making it perfect for training models that need to reason over multiple pieces of text.

2. HotpotQA (Multi-Hop Question Answering)

Link:https://huggingface.co/datasets/hotpotqa/hotpot_qa
HotpotQA is a large-scale dataset designed for multi-hop question answering, where answering a question requires reasoning over multiple documents. While HotpotQA does not include Supported/Refuted labels, it can be adapted for reasoning experiments to test multi-hop capability.

Prospective Models:

We would like to start with a simple baseline model using TF-IDF features and a logistic regression or SVM classifier in order to establish a reference performance. After that, we will move to more advanced models based on pre-trained transformers such as BERT, RoBERTa, or DeBERTa.

These transformer models are well-suited for understanding context and relationships between different sentences, which is exactly what multi-hop reasoning requires. We will fine-tune these pre-trained models on our dataset so that they learn to connect multiple pieces of evidence and make accurate decisions in claim verification.

Plan for Training Models:

First, we will preprocess the dataset by cleaning the text, removing unnecessary characters, and tokenizing for model input. Then, we divided the data into training, validation, and test sets in a 70:15:15 ratio.

For the baseline models, which include logistic regression, SVM, and random forest, we train them from scratch using either TF-IDF or embedding-based features. The deep learning and transformer models, namely BiLSTM, GRU, BERT, and RoBERTa, are pre-trained and fine-tuned. Fine-tuning will primarily be done on the FEVER dataset, with optional evaluation on adapted HotpotQA samples.

Training will be performed with early stopping for multiple epochs to avoid overfitting. Different batch sizes and learning rates will be experimented upon. Hyperparameters will be tuned using the validation dataset, while the final evaluation will be done using the test set.

Metric for Measuring Quality of Model:

- Primary: precision, recall, F1-score (macro-averaged for class balance), accuracy as a secondary summary.
- Class-wise report: per-class precision/recall/F1 for Supported, Refuted, Not Enough Info.
- Confusion matrix to observe the common confusions, such as Refuted vs Not Enough Info.
- If the data is imbalanced: macro F1 and weighted F1.
- Validation protocol: report mean ± std over 3–5 runs with fixed splits/seeds.
- Final selection: choose the model with best macro F1 on validation, then report test results once.