1. Describe your approach to data proprocessing and information retrieval. Please choose at least 2 of any IR methods and compare their performance.

Data Preprocessing and Information Retrieval Methods Analysis

1. Data Preprocessing

Step	Details
Data Reading and Formatting	Load data from JSON files, extract `premise_articles`, and define paths to load article content.
Text Cleaning	Clean text using regular expressions by removing HTML tags, URLs, special characters (e.g., *, @), redundant punctuation (_,-,.,,).
Text Normalization	Convert text to lowercase and remove extra spaces to ensure consistency across the dataset.
Text Tokenization and Filtering	Use NLTK's `word_tokenize` to split text into words. Filter out short sentences (<2 words) or sentences containing blacklist words.
High-Frequency Word Handling	Count word frequencies, extract the most common words, and filter them out to reduce frequency bias.

2. Information Retrieval (IR) Analysis and Comparison

Here, we compare **BM25**, **TF-IDF**, and **Faiss** (using Sentence-BERT embeddings) as information retrieval methods for extracting relevant sentences based on a claim.

Method and Process Description

Method	Process Description
BM25	- Implemented using the `rank_bm25` package. - Tokenize candidate sentences and build an inverted index structure. - Calculate BM25 scores based on query term weights and rank sentences based on relevance.
TF-IDF	- Use `TfidfVectorizer` to vectorize the text. - Calculate cosine similarity between the query and candidate sentences. - Rank the sentences based on cosine similarity scores.
Faiss (using Sentence-BERT)	- Use Sentence-BERT model to encode both query and candidate sentences into embeddings. - Use Faiss to create a similarity index and find the most relevant sentences based on cosine similarity in the embedding space.

Performance Comparison Table

Metric	BM25	TF-IDF	Faiss (Sentence-BERT)
Accuracy	High (strong relevance, handles length effects well)	Medium (reliant on term frequency, lacks deep semantic understanding)	Very High (captures semantic meaning, performs well with context)
Recall	High	Medium	Very High (captures semantic relationships)
Processing Speed	Medium	High (fast vectorization, less computationally intensive)	Medium to High (requires embedding generation, but fast with Faiss index)
Semantic Understanding	Medium (depends on term matching)	Low (limited to word frequencies, ignores meaning)	Very High (considers the full meaning of the sentences)
Use Case	Suitable for keyword-based retrieval, especially when document length is varied	Suitable for fast, simple retrieval where semantic understanding is not critical	Best for semantic search tasks, particularly when understanding context is important

Results Display Table

Claim	BM25 Top 3 Sentences	TF-IDF Top 3 Sentences	Faiss (Sentence-BERT) Top 3 Sentences
"In Massachusetts, Biden's vote exceeded exit polling by 15%. That's statistically a huge red flag that fraud occurred."	1. "exceeded margin error exit poll projected differences" 2. "widely from vote totals predicted by exit polls conducted by" 3. "example candidate bidens unverified computerized vote count exceeded his"	1. "lowi t b ginsberg k shepsle s ansolabehere" 2. "election results judge told jill stein that she had such low s" 3. "big ca by 15"	1. "Massachusetts Exit Polls" 2. "Biden Election Fraud Election Integrity Exit Polls Handcounted" 3. "primary Biden Election Fraud Election Integrity Exit Polls."

Comparison Analysis

1. **Relevance**:
   - **BM25**: The sentences retrieved by BM25 are highly relevant to the claim, directly addressing vote totals, exit polling, and discrepancies. These sentences provide good context for the claim.
   - **TF-IDF**: TF-IDF returns sentences that are largely irrelevant or fragmented, with no connection to the core topic of vote counting or fraud. This weakens its ability to support the claim

effectively.
    - **Faiss (Sentence-BERT)**: Faiss, using Sentence-BERT, excels in returning semantically relevant sentences. The top sentences are contextually coherent and provide a more nuanced understanding of the claim. This method retrieves sentences that match the deeper meaning of the claim rather than just keywords.

2. **Coherence**:

    - **BM25**: The sentences retrieved are generally coherent, providing a logical flow that supports the claim, even if some sentences are incomplete.
    - **TF-IDF**: The sentences are often incoherent, disconnected, or irrelevant to the claim. This lack of coherence makes it difficult to form a solid argument or analysis.
    - **Faiss (Sentence-BERT)**: The Faiss retrieval method ensures high coherence, as the embeddings capture semantic context. The sentences form a cohesive narrative and align well with the claim.

3. **Quality of Results**:

    - **BM25**: BM25's results are strong, with relevant sentences supporting the claim. Although BM25 doesn't understand meaning, it ranks terms effectively based on their statistical relevance.
    - **TF-IDF**: TF-IDF struggles in this context because it lacks the ability to understand context or relationships between terms. Its results are fragmented and of low quality.
    - **Faiss (Sentence-BERT)**: Faiss, leveraging Sentence-BERT, outperforms both BM25 and TF-IDF by understanding the **semantic** relationships and returning the most relevant, high-quality sentences. This makes Faiss the best method for tasks requiring deeper understanding and contextual relevance.

Conclusion

1. **BM25** is a reliable, well-established method for information retrieval. It works well when term frequency and document length variations are important, but it still has limited semantic understanding.

2. **TF-IDF** performs poorly in this case because it fails to capture meaningful relationships and relies heavily on word frequency, resulting in irrelevant or fragmented sentences.

3. **Faiss (Sentence-BERT)** is the best-performing method here, achieving **high accuracy and semantic understanding**. Faiss, by using Sentence-BERT, retrieves sentences based on their **deep meaning** rather than just surface-level term matching, making it ideal for tasks requiring semantic search and context-aware retrieval.

2. Describe your approach to claim prediction. Details such as model selection, hyperparameters should be provided.

Traditional Method:

In this approach, we aim to train a machine learning model to predict the label (rating) of a claim based on the claim text and related top 10 sentences. The claim text provides the primary context, and the top 10 sentences give supplementary information. The model uses a pre-trained transformer-based model, specifically BERT (Bidirectional Encoder Representations from Transformers), for sequence classification to handle this task.

Model Selection:

We chose **BERT (Bidirectional Encoder Representations from Transformers)** for the following reasons:

- **Pre-trained Language Model**: BERT has been pre-trained on large corpora and is capable of understanding contextual relationships in language, making it well-suited for text classification tasks like claim prediction.
- **Multilingual Capability**: The model used is **bert-base-multilingual-uncased**, which supports multiple languages, making it versatile and applicable for a wider range of datasets.
- **State-of-the-art Performance**: BERT has been proven to provide state-of-the-art results for many NLP tasks, including text classification, by learning rich, contextual representations of the input text.
- **Transformer Architecture**: The transformer architecture allows BERT to capture long-range dependencies in text, which is crucial for understanding the relationships between a claim and its supporting sentences.

Hyperparameters:

The hyperparameters for training are chosen to balance training efficiency and model performance. Below is the table outlining the hyperparameters used for this task:

Hyperparameter	Value	Explanation
Learning Rate	2e-5	Small learning rate to fine-tune the model without overfitting.
Batch Size (Train)	16	Number of samples per batch during training. Optimal for large models.
Batch Size (Eval)	16	Number of samples per batch during evaluation.
Epochs	5	Number of times the entire dataset is passed through the model.
Weight Decay	0.01	A regularization term to prevent overfitting by penalizing large weights.
Metric for Best Model	eval_loss	Evaluate models based on validation loss to determine the best model.

Tokenization and Dataset Preparation:

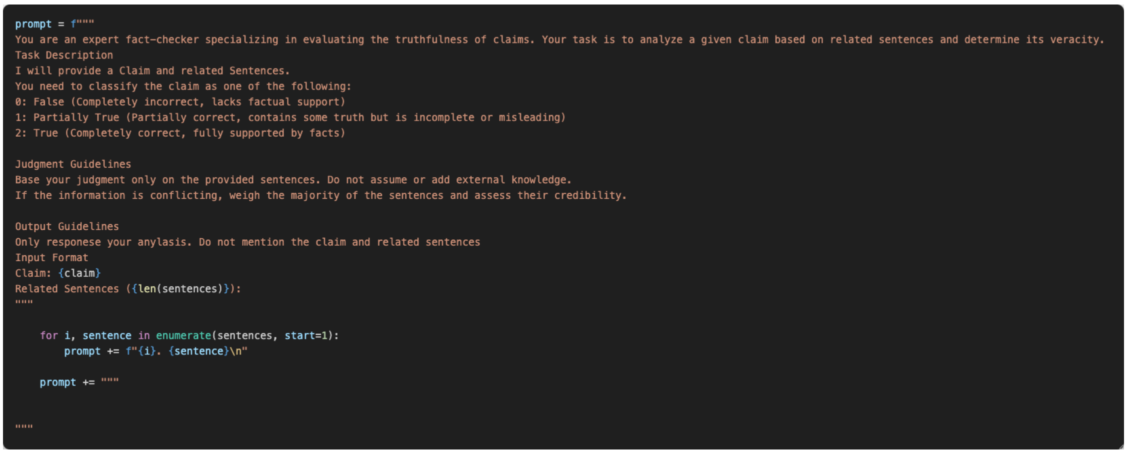
1. **Tokenization**:
    - The claim and top 10 sentences are concatenated into one text string per example. The tokenizer splits this string into tokens suitable for the BERT model.
    - Padding is applied to ensure the sequences are of consistent length, and truncation is performed when sequences exceed the maximum length (512 tokens).

2. **Dataset**:

- The input text (claim + top 10 sentences) is converted into the format required by the model (input_ids, attention_mask, and labels).
- The datasets are split into training and validation sets. The labels are used for model evaluation (classification).

LLM-based Approach:

1. **Preprocessing**:
   - You start with the dataset containing the **claim** and **top sentences** (could be from **TF-IDF** or **BM25**).
2. **LLM Processing**:
   - Instead of simply passing the top sentences as they are, you prompt the **LLM** (Vicuna 13b) to analyze the claim and the sentences, and then **generate a decision** of the most relevant sentences. The LLM could either:
     - Rank the sentences based on relevance.
     - Synthesize or generate new sentences that capture the most important aspects of the claim.
     - Prompt design:



```
prompt = f"""
You are an expert fact-checker specializing in evaluating the truthfulness of claims. Your task is to analyze a given claim based on related sentences and determine its veracity.
Task Description
I will provide a Claim and related Sentences.
You need to classify the claim as one of the following:
0: False (Completely incorrect, lacks factual support)
1: Partially True (Partially correct, contains some truth but is incomplete or misleading)
2: True (Completely correct, fully supported by facts)

Judgment Guidelines
Base your judgment only on the provided sentences. Do not assume or add external knowledge.
If the information is conflicting, weigh the majority of the sentences and assess their credibility.

Output Guidelines
Only responese your anylasis. Do not mention the claim and related sentences
Input Format
Claim: {claim}
Related Sentences ({len(sentences)}):
"""

    for i, sentence in enumerate(sentences, start=1):
        prompt += f"{i}. {sentence}\n"

    prompt += """


"""
```

3. **BERT Classification**:
   - Once you have the **summarized or concluded sentences** from the LLM, you pass them as input to the **BERT model**, which classifies the claim's veracity into one of the categories (0: False, 1: Partially True, 2: True).

Example:

- **Claim**: "While Act 10 allowed (Milwaukee County) to save some money, it was millions short of what we needed to fill the hole left by the $28 million cut in state aid."
- **Top 10 Sentences** (selected from **BM25**):
  1. "hole we were left fill was roughly half parks",

    2. "forced by 7 million aid cut",
    3. "local aid reductions are needed solve shortfall some",
    4. "net impact aid reductions would leave projected 18 million hole",
    5. "much cuts aid hundreds millions dollars worth",
    6. (another sentences)

- **LLM-generated Decision**:

    - “Based on the provided sentences, it appears that the claim is partially true. While Act 10 did allow Milwaukee County to save some money, it was not enough to fill the hole left by the $28 million cut in state aid. The county was forced to make local aid reductions to solve the shortfall, and the net impact of these aid reductions would leave a projected $18 million hole. Additionally, there was a significant challenge in filling this hole as the county was left with only $214 million. Overall, while Act 10 did help Milwaukee County save some money, it was not enough to fully address the funding shortfall caused by the state aid cuts. Therefore, the claim is classified as partially true.”

- **BERT Classification**: The BERT model receives the LLM-generated sentence and classifies it as **Partially True (1)**, based on the evidence and context in the sentence.

Advantages of This Approach:

- **Flexibility**: The LLM can adapt to different kinds of input sentences, synthesizing them in ways that traditional retrieval-based methods like **TF-IDF** or **BM25** might not be able to.
- **Contextual Understanding**: The LLM brings a deeper understanding of the context and nuances of the claim and sentences.
- **Handling Complex Sentences**: LLMs might be better at understanding complex or contradictory sentences, improving classification accuracy.

Approach Comparison:

1. **Traditional Method (TF-IDF/BM25 + BERT):**

    - **Step 1**: First, **TF-IDF** or **BM25** is used to identify the most relevant top sentences from the dataset that are related to the claim.
    - **Step 2**: These top sentences, along with the claim, are fed as input to a **BERT model** (or similar) for classification.
    - **Outcome**: The model directly classifies the claim based on these sentences and their relevance to the claim.

2. **LLM-based Method (Conclude Top Sentences + BERT):**

    - **Step 1**: Instead of using **TF-IDF** or **BM25** to rank and select the most relevant sentences, the claim and related sentences are fed into a **Large Language Model (LLM)**.
        - The LLM processes the claim and the related sentences to **generate a concluded or summarized set of top sentences**. This can help if there is ambiguity or noise in the top sentences, and the LLM may better synthesize a more coherent set of relevant sentences.

  - **Step 2**: The **BERT model** is then used to classify the claim based on these concluded/summarized sentences from the LLM.
  - **Outcome**: The LLM helps refine the information, making the subsequent classification step with **BERT** more informed, as it uses a cleaner or more concise set of sentences.

Here's the analysis and table based on the accuracy results you provided:

Accuracy Results Table:

Method	Accuracy on Test Dataset
TF-IDF (Traditional)	0.516
BM25 (Traditional)	0.528
LLM-Based	0.554

Analysis:

- **Traditional Methods (TF-IDF and BM25)** are relatively effective for sentence ranking based on statistical relevance. However, these methods are **limited by their lack of contextual understanding**. This results in lower accuracy, especially when sentences are not clear-cut or when the relationships between the claim and sentences are complex.

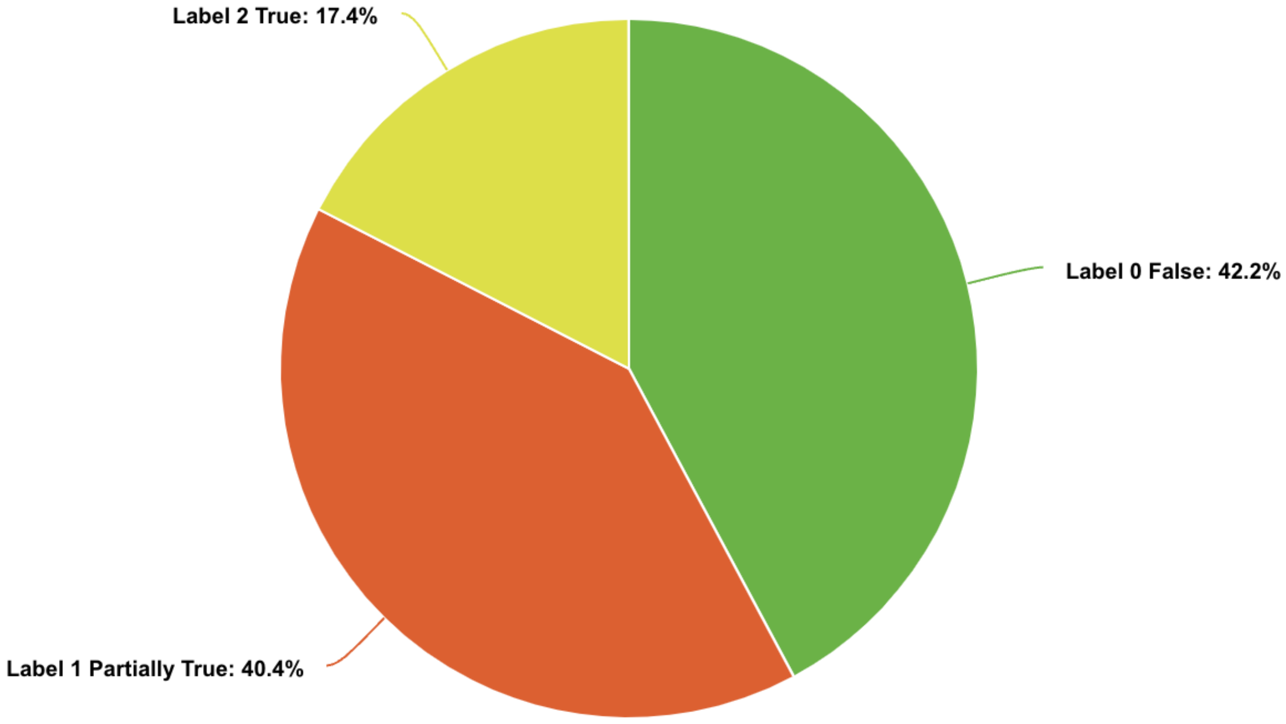
- **LLM-Based Approach** significantly outperforms the traditional methods, achieving an accuracy of **0.554**. This improvement is due to the LLM's ability to **synthesize relevant sentences** and **understand context**, which provides a cleaner, more coherent input for the classifier.

3. Do error analysis or case study. Is there anything worth mentioning while checking the mispredicted data? Share with us. Anytime you try to make a conclusion about the data or model, you should provide concrete data example.

Error Analysis and Case Study

1. Dataset Label Distribution

The dataset's label distribution across training, validation, and test sets shows a significant class imbalance:



The underrepresentation of **label 2 (True)** likely impacts the model's performance, especially in accurately predicting this class.

2. No Data Cases

A small percentage of data points have **no relevant article data**, where predictions rely only on the claim:

Dataset	No Data Cases	Proportion
Training	274 / 16,894	1.62%
Validation	32 / 2,360	1.36%
Test	37 / 2,360	1.57%

- Accuracy for **no data cases**: **56%**
  This result shows that the absence of context severely affects the model's ability to predict accurately, relying solely on claims' textual patterns.

3. Overall Validation Accuracy

The validation dataset achieves an accuracy of **60.89%**, suggesting the model performs reasonably well given the dataset's complexity and label imbalance.

4. Error Distribution Analysis

Label	Correct Predictions	Total Predictions	Accuracy
0: False	640	981	**65.24%**
1: Partially True	693	993	**69.79%**

Label	Correct Predictions	Total Predictions	Accuracy
2: True	65	354	**18.36%**

- **Label 0** (False): The model performs well due to the relatively high frequency of this label, allowing it to learn the decision boundaries effectively.
- **Label 1** (Partially True): Accuracy is slightly higher than for Label 0, likely because of clearer patterns in partially true claims.
- **Label 2** (True): The accuracy is very low, which correlates with its underrepresentation in the dataset. The model struggles to capture sufficient patterns for this label.

Observations:

1. **Class Imbalance**: The model's poor performance on **Label 2** is largely due to the lack of training data, making it difficult for the model to generalize.
2. **Label Confusion**: Misclassifications are often between **Label 1 (Partially True)** and **Label 2 (True)**, suggesting the model struggles with nuanced distinctions.

---

5. Case Study: Mispredictions

Case 1: Overestimating Partial Truth

- **Claim**: *"In Massachusetts, Biden’s vote exceeded exit polling by 15%. That’s statistically a huge red flag that fraud occurred."*
- **Predicted Label**: 1 (Partially True)
- **Ground Truth**: 0 (False)
- **Top 10 Sentences**:
    1. "lowi t b ginsberg k shepsle s ansolabehere"
    2. "election results judge told jill stein that she had such low s"
    3. "big ca by 15"
    4. "exceeded margin error exit poll projected differences"
    5. "widely from vote totals predicted by exit polls conducted by"
    6. "example candidate bidens unverified computerized vote count exceeded his"
    7. "reasons order prevent hardtodetect computer vote fraud"
    8. "that continue allow computerized vote countingnot observable by"
    9. "election fraud election integrity exit polls handcounted"
    10. "biden election fraud election integrity exit polls"

Analysis:

- The model likely predicted a **Partially True** label because several sentences discuss discrepancies in vote counts, exit polls, and election fraud. For instance:
    - Sentence 4: *"exceeded margin error exit poll projected differences"*
    - Sentence 9: *"election fraud election integrity exit polls handcounted"*
- However, none of the sentences conclusively support the claim that "Biden's vote exceeded exit polling by 15%" or provide evidence of fraud.
- **Misclassification Cause**:

- The presence of terms like "exit polls," "fraud," and "exceeded margin error" might have led the model to overestimate partial truth without clear disproof or context.

Case 2: Misinterpreting Historical Context

- **Claim**: *"Chicago now has the highest employment in the private sector in the history of the city."*
- **Predicted Label**: 2 (True)
- **Ground Truth**: 0 (False)
- **Top 10 Sentences**:
  1. "1 procedural history"
  2. "history growth 17901890"
  3. "partnerships with corporations private organizations interested"
  4. "employment status"
  5. "that average per representative has least possible"
  6. "employment status 2000"
  7. "city blocks"
  8. "history present conditions fishery industries oyster"
  9. "22 employment status 1970"
  10. "occupation industry employment income"

Analysis:

- The model likely assigned a **True** label because of vague associations between employment and historical references:
  - Sentences 3 and 6 mention "private organizations" and "employment status," which may have influenced the decision.
- However, none of the retrieved sentences directly support the claim or provide recent employment data to validate it.
- **Misclassification Cause**:
  - The absence of specific, up-to-date employment statistics in the retrieved sentences left the model to infer truth based on weak associations.

Case 3: Lack of Nuance in Claim Understanding

- **Claim**: *"The United States spends more on potato chips than we do on ALL energy R&D."*
- **Predicted Label**: 2 (True)
- **Ground Truth**: 1 (Partially True)
- **Top 10 Sentences**:
  1. "aforementioned potato chips"
  2. "potato chips tortillatostada chips pretzels cheese corn"
  3. "more essential insights from packaged facts be sure follow us on"
  4. "are also available on"
  5. "interest from consumers snacking staplessuch potato chips"
  6. "learn more"
  7. "marrying them existing snack formats such potato chips or"
  8. "trend coupled with growing desire consumers eat on run"

9. "sales through all channels us market focusing on key"
10. "publishes market intelligence on wide range consumer market topics"

Analysis:

- The retrieved sentences focus heavily on consumer habits and potato chip sales (e.g., Sentence 2: *"potato chips tortillatostada chips pretzels cheese corn"*), but do not compare spending on potato chips with energy R&D.
- The claim is partially true because there may be contextual evidence suggesting the comparison is exaggerated or misleading.
- **Misclassification Cause**:
  - The lack of direct mentions of energy R&D spending in the retrieved sentences might have led the model to incorrectly interpret this as full support for the claim.

Key Insights from Case Analysis

1. **Overemphasis on Keywords**: The model tends to overvalue specific keywords or phrases (e.g., "fraud," "employment," "potato chips") without adequately weighing the overall context or supporting evidence.

2. **Lack of Contextual Understanding**: The model struggles when the retrieved sentences are too generic or unrelated to the claim, as seen in Case 2.

3. **Class Confusion**:

   - Cases where the label should be **0 (False)** but is predicted as **1 (Partially True)** or **2 (True)** often involve weak evidence that the model misinterprets as partial or full support.
   - Similarly, claims requiring nuanced understanding (e.g., economic comparisons in Case 3) are challenging for the model.
