Group Members:

Zeeshan Ali (197)
Shayan Hassan Abbasi (167)
Muhammad Sheraz Ajmal (130)

Natural Language Processing

Semester Project

Final Report

Submitted To: Dr. Arif ur Rahman

Project Title: Lateral Reading

Abstract:

We chose this project because the tasks are straightforward and require less domain-specific complexity compared to other projects enlisted on [TREC]. The detection and evaluation of misinformation remain significant challenges due to the complexity of defining truth. ***Lateral reading*** offers a systematic approach to assess the trustworthiness of information by formulating critical questions about document sources and evidence, followed by seeking answers through external searches. This track aims to develop innovative technologies that support and promote the practice of lateral reading for verifying information.

The track is divided into two core subtasks:

1. Question Generation: Given 100 articles from the manually created dataset, the objective is to produce 5 contextually relevant and fact-checking questions for each article.
2. Document Retrieval: Using the manually created dataset and the generated questions, the goal is to retrieve and rank the top 3 documents or passages that provide evidence-based answers to the questions.

The proposed solutions will enhance the lateral reading process by integrating natural language processing and information retrieval techniques to aid in misinformation detection and evaluation.

Introduction:

In today's digital world, the continuous evolving digital landscape has made it hard and challenging to tell which information is trustworthy and which might be misleading i.e. misinformation and biased narratives. In response, the concept of Lateral Reading has emerged as a robust methodology to critically evaluate the trustworthiness of information sources. Unlike traditional linear reading, lateral reading emphasizes cross-referencing content with external sources, checking the reliability of information by looking beyond the original source to verify claims and facts and uncover any hidden or potential biases.

This project taken from [TREC] focuses on using Natural Language Processing and Information Retrieval techniques to enhance lateral reading practices and make it easier and faster. Specifically, it aims to automate critical components of this process

i.e. generating fact-checking questions from articles and retrieving evidence-based responses from extensive datasets. By doing this, our goal is to help people detect misinformation more effectively on a larger scale.

By utilizing manually created dataset, transformer-based models and dense retrieval techniques, this project bridges the gap between manual evaluation and automated verification systems. This initiative offers a practical way to help readers fact-check information and make better decisions about what they trust online.

Requirements:

1. **Primary Goal:** Develop a system that supports and encourages lateral reading for evaluating the trustworthiness of information.
2. **Aims:** Following are the things that we want to achieve from this project:
   - To enhance misinformation detection process by using NLP & IR technologies.
   - To facilitate fact-checking practices using automated tools.
3. **Objectives:** Following actions and steps will be taken to accomplish our aims.
   - Design and implement an NLP-based system to generate meaningful fact-checking questions from given articles.
   - Build an efficient document retrieval system that identifies and ranks evidence from a large dataset.
   - Evaluate the system's effectiveness in generating relevant questions and retrieving accurate supporting documents.

Methodology:

For this project, we have designed a system that automates the generation of fact-checking questions and retrieves evidence-based documents. Hence the methodology consists two key components:

1. Question Generation:

We used a pre-trained transformer-based model **(valhalla/t5-base-qg-hl)** to generate contextually relevant questions. For each selected article, the system extracts the first sentence to form a context. The model then generates multiple fact-checking questions using beam search ensuring diversity and relevance.

2. Document Retrieval:

To retrieve and rank supporting documents for generated questions, we employed the **all-MiniLM-L6-v2** model from the Sentence Transformers

library. This model encodes both the questions and articles into embeddings and cosine similarity is used to measure relevance. The system ranks the documents and returns the top three most relevant results.

These steps ensure the system effectively supports the lateral reading process by automating the generation of critical questions and identifying reliable sources for verification.

Tools & Techniques:

1. Tools: Following are the software, libraries and frameworks used to implement specific functionalities:

- **Programming Language:** Python (due to its extensive support for NLP and IR libraries) and HTML (for creating the website).
- **Dataset:** Manually created by collecting 100+ articles.
- **Frameworks and Libraries:**
  - **NLP:** Hugging Face Transformers (for text-to-text generation), NLTK (for sentence tokenization) and RE (for preprocessing tasks).
  - **Retrieval Models:** Sentence-BERT (dense embeddings using the all-MiniLM-L6-v2 model).

2. Techniques: Following are the methods and approaches applied to solve tasks or problems to accomplish the goal:

- **Natural Language Processing:**
  - Uses a transformer-based model **(valhalla/t5-base-qg-hl)** to generate fact-checking questions.
  - Tokenize articles into sentences using **NLTK** and selects the first sentence to provide context for question generation.
  - Text normalization and preprocessing are conducted to ensure consistent inputs for modeling.

- **Information Retrieval:**
  - Dense retrieval techniques are implemented using **Sentence-BERT** to encode both the questions and articles into embeddings.
  - **Cosine similarity** is computed to rank articles based on their relevance to generated questions.

- **Evaluation Metrics:**
  - **Question Quality**: Assessed manually based on clarity, relevance and diversity. Also, google forms & ratings.
  - **Retrieval Effectiveness:** Evaluated using **cosine similarity scores** to determine the relevance of retrieved documents.

Project Phases:

This project is structured into three key phases to systematically develop and evaluate the system:

1. Data Preparation:

- **Dataset Collection:**
  The ClueWeb22B-English dataset was supposed to be chosen but due to it's paid version, we created dataset manually by collecting 100+ articles from different topics.

- **Preprocessing:**
  Articles are cleaned and tokenized using natural language processing libraries **(nltk and re)** to ensure consistency. Tokenization helps segment the articles into meaningful units like sentences which are crucial for generating questions.

2. System Implementation:

- **Question Generation Module:**
  A transformer-based model **(valhalla/t5-base-qg-hl)** is integrated for generating fact-checking questions. The system processes the first sentence of each article to form a concise context. These sentences are fed into the model to produce five diverse and contextually relevant questions for each article using beam search.
  For Example:
  Article 1: "What is the tourism industry recovering from?"
  Article 2: "What is reshaping the concept of workplaces?"

- **Document Retrieval Module:**
  The **all-MiniLM-L6-v2** model from Sentence Transformers is used to encode the generated questions and all articles into high-dimensional embeddings. These embeddings allow the system to compute cosine similarity between questions and articles to determine relevance. The top three most relevant articles for each question are ranked and returned as results.

For Example:
For the question "What is the tourism industry recovering from?", the system ranked the top document with a cosine similarity score of 0.7978.

3. Testing and Evaluation:

- **Testing:**
  A random article is selected for testing the system. For the article, questions are generated and relevant documents are retrieved. The results are manually inspected to ensure the quality of the generated questions and the accuracy of the retrieved documents.

- **Evaluation:** *The system's performance is evaluated based on:*
  - **Question Quality:** Questions are checked for relevance, clarity and diversity.
  - **Document Relevance:** Retrieved documents are ranked according to their cosine similarity scores which are used to measure how well they matched the generated questions.

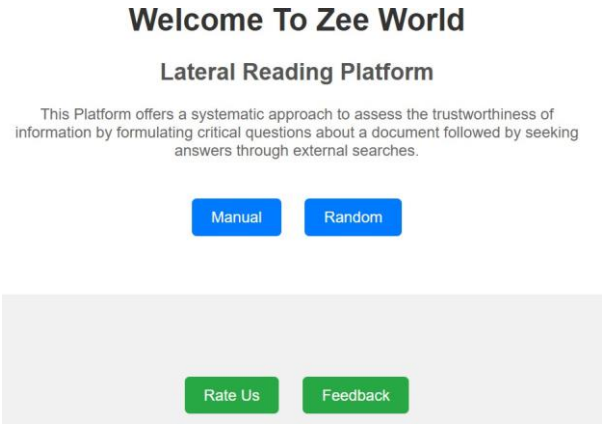
Results:

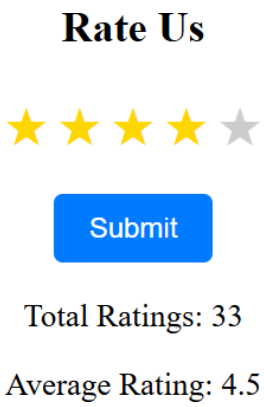
The system successfully generated five contextually relevant fact-checking questions for randomly selected 3 articles and retrieved and ranked the top three relevant documents based on cosine similarity scores using these questions.

Output:

```
Generating questions for Article 1...

Generated Questions:
1. What is the tourism industry recovering from?
2. How is the tourism industry recovering from global restrictions?
3. How is the tourism industry recovering?
4. What is the tourism industry recovering from global restrictions?
5. How is the tourism industry recovering after restrictions?

Retrieving top documents for the questions...

Top 3 Ranked Documents for Question: What is the tourism industry recovering from?
1. Score: 0.7978, Document: The tourism industry is recovering rapidly after global restrictions....
2. Score: 0.4764, Document: Virtual tourism is providing unique travel experiences....
3. Score: 0.4236, Document: Space tourism is on the brink of becoming a reality for civilians....

Generating questions for Article 2...

Generated Questions:
1. What is reshaping the concept of workplaces?
2. What is the role of remote work in workplaces?
3. What is the definition of a remote work environment?
4. What is the definition of remote work?
5. What is the role of remote work in the workplace?

Retrieving top documents for the questions...
```

Website: It consists of total 4 buttons. The ***manual*** button helps you to provide an article/fact on your own and it will serve you the purpose of lateral reading. The ***random*** button consists of 100+ articles stored in dataset which randomly selects any article and then performs lateral reading. The ***rate us*** button helps users to provide ratings out of 5 stars and it also stores total number of ratings along with average. The ***Feedback*** button redirects user to Google form which helps to provide detailed review.





Conclusion:

This project successfully implemented a system to enhance lateral reading practices using NLP and IR technologies. By automating question generation and document retrieval, it provides an efficient way to verify information and combat misinformation. The use of transformer models and semantic similarity techniques ensured high-quality outputs as evidenced by the generated questions and retrieved documents. This work highlights the potential of automated systems in supporting critical thinking and improving trust in digital information.

GitHub: [ZeeshawnAley/NLP_FINAL_PROJECT]

LinkedIn: [(22) Zeeshan Ali | LinkedIn]

References:

1. Text REtrieval Conference (TREC) Call to TREC 2024
2. 2024 | TREC - Home
3. The ClueWeb22 Dataset [Could not used due to paid version]
4. Sentence-BERT-base
5. valhalla/t5-base-qg-hl
6. Flask Tutorial - GeeksforGeeks