Advanced NLP: Final Project Submission

Anmol Goel
2021701045

Ravi Shankar Mishra
2021701044

1 Introduction

Automating fact verification is an important research problem due to the rise in misleading and fake content being generated recently. Misleading information has potential to affect real world events like elections, riots, mobilization, etc. Verifying and fact-checking claims thus becomes an integral task to ensure the smooth functioning of society. Since fact-checking is a time consuming task, it is imperative to develop automated methods for automatic fact checking and verification.

2 Dataset

FEVEROUS Aly et al. (2021) is a multimodal dataset relying on textual and structural information as part of evidence for the claims made.

The dataset is manually annotated and verified. Annotator agreement and screening was extensively used to ensure high quality of the data. The data statistics across splits are illustrated in Table 1.

	Train	**Dev**	**Test**
Supported	41835	3908	3372
Refuted	27215	3481	2973
NotEnoughInfo (NEI)	2241	501	1500
Total	71291	7890	7845

Table 1: Number of samples in each split of the FEVEROUS dataset

The data is quite rich in its modalities and variations. Not all claims in the dataset have multimodal evidence as illustrated by Table 2.

	Train	**Dev**	**Test**
E_{sent}	31607	3745	3589
E_{cells}	25020	2738	2816
$E_{sent+cells}$	20865	2468	2062

Table 2: Claims requiring evidence of only sentence, only cell or both

2.1 Data Statistics

Distribution of Multimodal Evidence

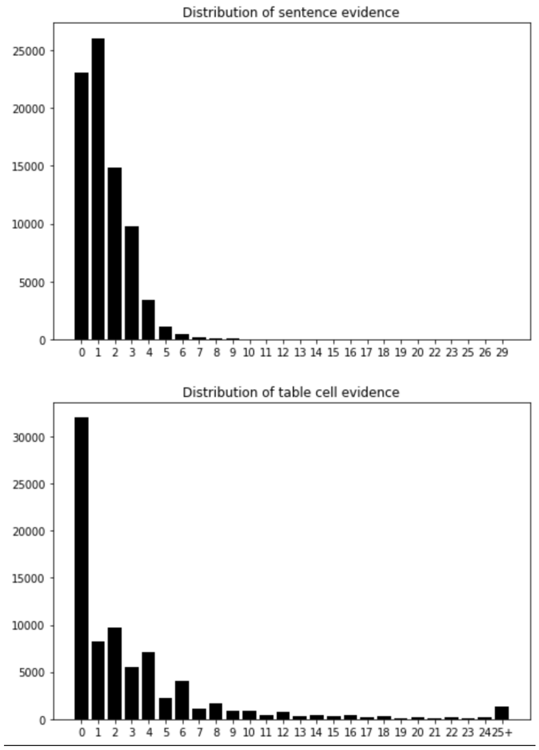


Figure 1: Evidence distribution

Distribution of Wiki article lengths

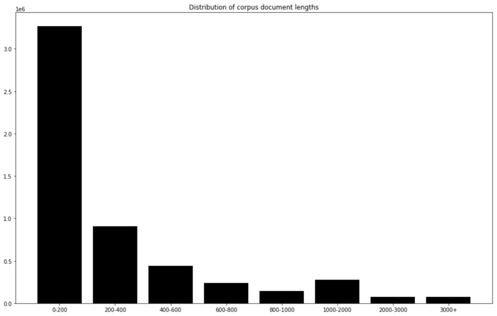


Figure 2: Distribution of corpora length

Distribution of table dimensions (rows and columns)

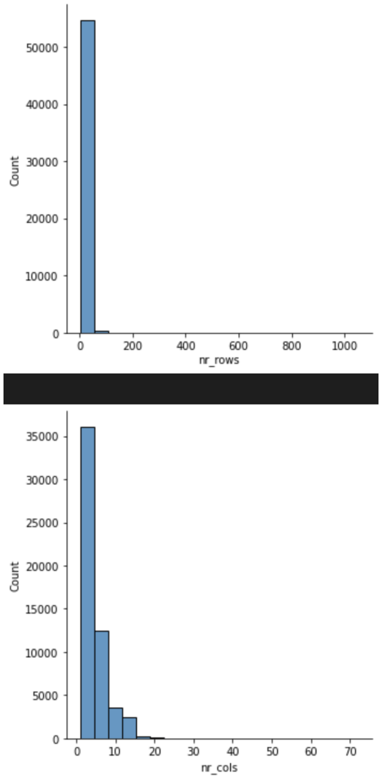


Figure 3: Distribution of table evidence rows and columns

2.2 Evaluation

The authors propose the FEVEROUS score to evaluate models on the given task. For any given claim, the model is correct in its prediction only if at least one complete gold evidence set E is a subset of the predicted evidence $\hat{E}$ and the predicted label is correct.

$$score(y, \hat{y}, \mathbf{E}, \hat{E}) = \begin{cases} 1 & \forall E \in \mathbf{E} : E \subseteq \hat{E} \wedge \hat{y} = y, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Additionally, traditional metrics like Precision, Recall, F-score and label accuracy can also be applied to the task. Precision measurements will likely lead to penalization of correct evidence missed by the annotators.

3 Proposed Model

- Baseline implemented based on the code included in the official GitHub repository: `https://github.com/Raldir/FEVEROUS`
- We have split the implementation of the task into the following components for our approach:
  - **Document Retrieval**: Similar documents to a given evidence are retrieved using TF-IDF. We plan to experiment with embedding based retrieval approaches for the final submission.
  - **Sentence Extraction**: Similarly, relevant sentences are extracted based on TF-IDF.

- **Table Extraction**: The pretrained TaPaS Herzig et al. (2020) model is used to extract tables based on the dense representations of the [CLS] token.
- **Table Cell Extraction**: The TaPaSForQuestionAnswering module from the HuggingFace library is used to extract relevant cells given a query.
- **Claim Veracity Prediction** A neural network consisting of 3 dense layers (2 ReLUs and 1 Softmax) is used. The claim is embedded using text representations based on RoBERTa and the table is embedded using TaPaS

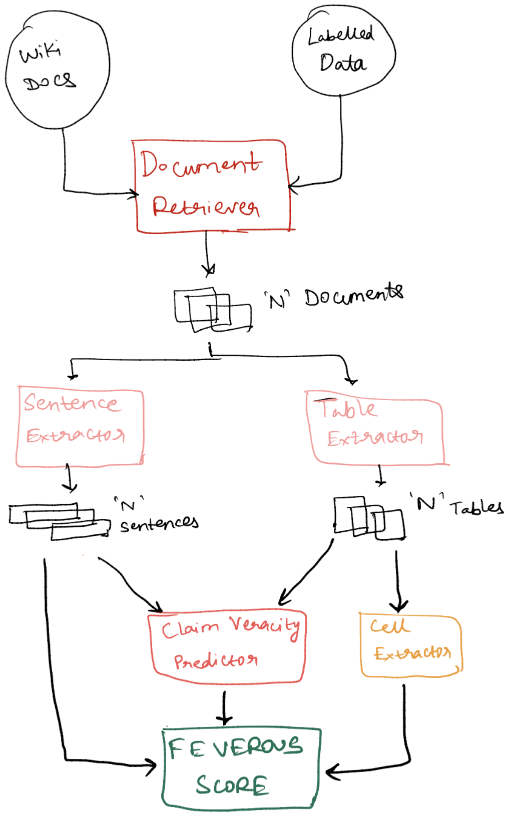


Figure 4: Flow chart of the proposed approach

4 Results

Method	Task	Metric	Performance
Baseline	Document Retrieval	Accuracy	77.24%
	Claim Verification	FEVEROUS	0.19
TaPaS+RoBERTa	Document Retrieval	Accuracy	80.2%
	Sentence Extraction	Precision@5	23.89
	Table Extraction	Precision@5	0.73
	Table Cell Extraction	Precision	9.03
	Claim Veracity Prediction	FEVEROUS	0.14

5 Zero Shot Pipeline

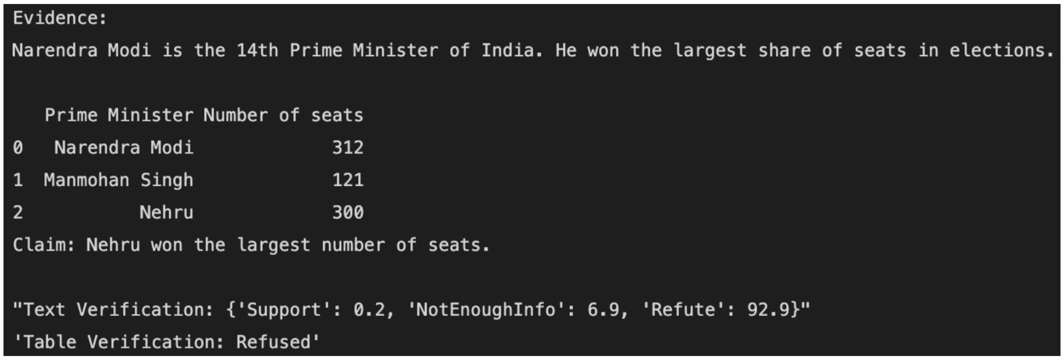
We experimented with zero shot capabilities of pretrained LLMs to check their performance on the fact verification task. In particular, we use:

- TAPEX: LLM pretrained on tabular data for fact verification based on Wikipedia tables
- mDeBERTa: Multilingual LLM finetuned on NLI for fact verification

We observe good performance of zero shot pipeline on the FEVEROUS task, considering no training or finetuning was done at all. This finding is corroborated by previous works in the FEVEROUS shared task.

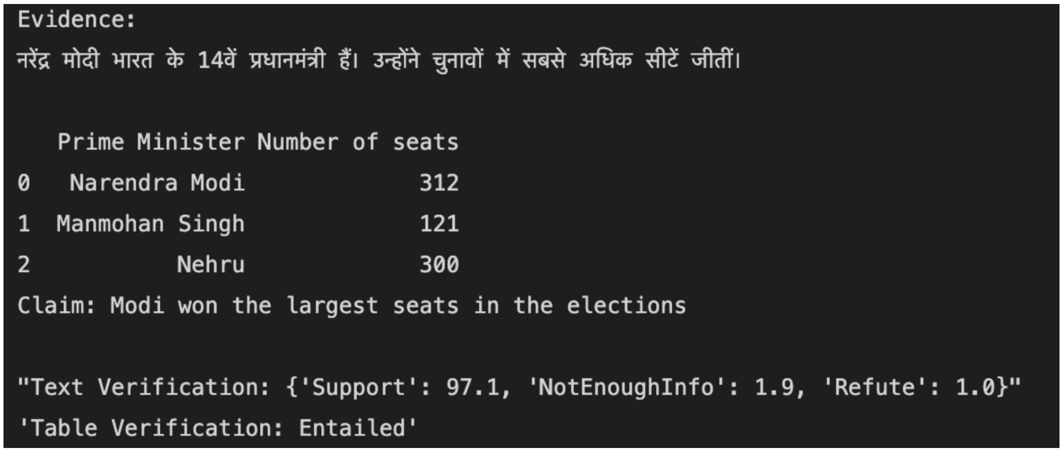
5.1 Sample Results

Output on test sample Output on multilingual sample



```
Evidence:
Narendra Modi is the 14th Prime Minister of India. He won the largest share of seats in elections.

   Prime Minister Number of seats
0   Narendra Modi            312
1  Manmohan Singh            121
2           Nehru            300
Claim: Nehru won the largest number of seats.

"Text Verification: {'Support': 0.2, 'NotEnoughInfo': 6.9, 'Refute': 92.9}"
'Table Verification: Refused'
```

Figure 5: Prediction on a sample taken from the dataset. Multimodal input of text and table is considered for the output.

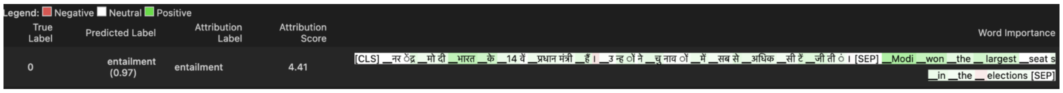


```
Evidence:
नरेंद्र मोदी भारत के 14वें प्रधानमंत्री हैं। उन्होंने चुनावों में सबसे अधिक सीटें जीतीं।

   Prime Minister Number of seats
0   Narendra Modi            312
1  Manmohan Singh            121
2           Nehru            300
Claim: Modi won the largest seats in the elections

"Text Verification: {'Support': 97.1, 'NotEnoughInfo': 1.9, 'Refute': 1.0}"
'Table Verification: Entailed'
```

Figure 6: Prediction on a multilingual sample taken from the dataset. Multimodal input of text and table is considered for the output.

5.2 Analysis

GradCAM visualization

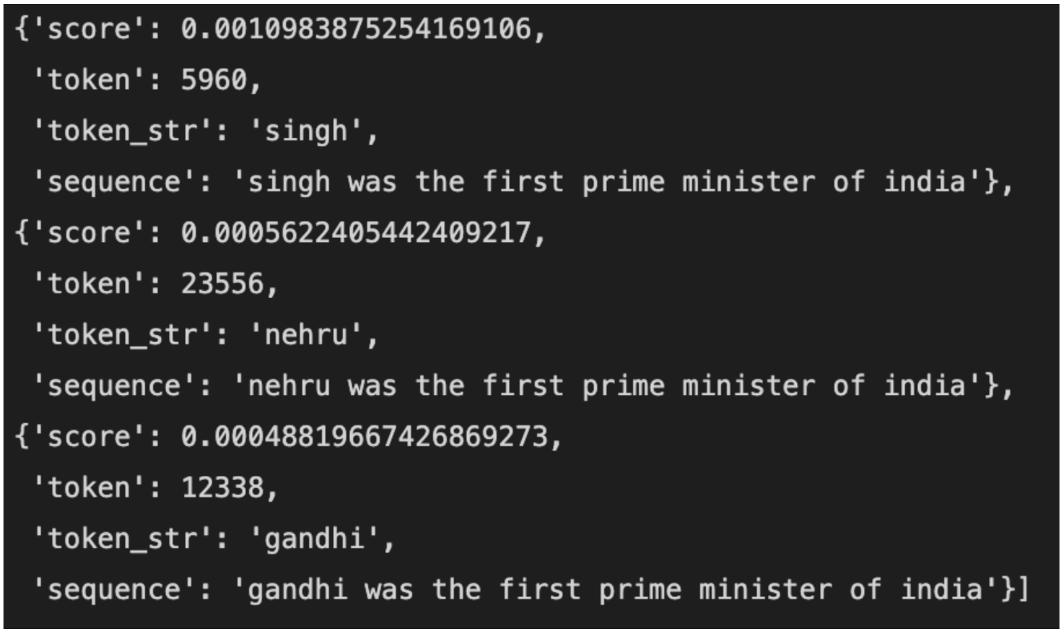


Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	entailment (0.97)	entailment	4.41	[CLS] __नर ेंद्र __मो दी __भारत __के __14 वें __प्रधान मंत्री __हैं । __उ न्ह ों ने __चु नाव ों __में __सब से __अधिक __सी टें __जी ती ं । [SEP] __Modi __won __the __largest __seat s __in __the __ elections [SEP]

Figure 7: GradCAM visualization of a sample evidence/claim pair to aid interpretability.

Probing World Knowledge in Pretrained LLMs



```
{'score': 0.0010983875254169106,
 'token': 5960,
 'token_str': 'singh',
 'sequence': 'singh was the first prime minister of india'},
{'score': 0.0005622405442409217,
 'token': 23556,
 'token_str': 'nehru',
 'sequence': 'nehru was the first prime minister of india'},
{'score': 0.00048819667426869273,
 'token': 12338,
 'token_str': 'gandhi',
 'sequence': 'gandhi was the first prime minister of india'}]
```

Figure 8: Probing for facts in pretrained MLMs. This shows that the pretraining helped model to learn some information about the world.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*, 2021.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*, 2020.