FactHunt
Retrieval-driven Fact Verification for Accurate Conclusions

Mayank Kumar 19CS30029
Shrinivas Khiste 19CS30043
Ishan Goel 19CS30052
Ashish Gupta 19IE10010

Problem Statement

The task of fact verification involves determining the veracity of textual hypotheses based on provided evidence. In the context of tabular data, where tables and their captions serve as evidence, the problem lies in effectively selecting relevant rows and columns to support or refute the given hypotheses. This project aims to address this challenge by evaluating different techniques for evidence selection and investigating the impact of using various language models for fact verification from tabular data.

1 Introduction

The process of factual verification plays a crucial role in determining the accuracy and reliability of information presented in various forms of textual content. It involves the rigorous task of assessing whether a given statement or hypothesis holds true based on the available evidence. With the exponential growth of data and the vast amount of information available, the need for efficient and reliable methods to verify the accuracy of textual claims has become increasingly important.

In recent years, researchers and practitioners have focused on developing automated systems capable of fact-checking and verifying textual claims. This project report delves into the realm of factual verification from tabular data, where tables, along with captions, are presented as evidence, and hypotheses are framed as natural language statements.

We leverage a comprehensive and large-scale dataset known as TabFact, which comprises a vast collection of Wikipedia tables that serve as evidence, and human-annotated natural language statements that act as hypotheses. The dataset's diverse range of topics and the incorporation of real-world information make it an ideal resource for training and evaluating fact verification models.

Let us have a look at an example of TabFact using the table from 2-16570286-3.html.csv. The example provided consists of a table 1 and several statements related to the 1948 Ashes series, a cricket tournament.

player	team	matches	wickets	average	best bowling
ray lindwall	australia	5	27	19.62	6 / 20
bill johnston	australia	5	27	23.33	5 / 36
alec bedser	england	5	18	38.22	4 / 81
keith miller	australia	5	13	23.15	4 / 125
ernie toshack	australia	4	11	33.09	5 / 40
norman yardley	england	5	9	22.66	2 / 32
jim laker	england	3	9	52.44	4 / 138

Table 1: 1948 ashes series

Entailed Statements are:
the bowler with 13 wicket appear in more match than the bowler with 11 wicket (Ray Lindwall, with 13 wickets, has played in more matches than Ernie Toshack, who had 11 wickets)
none of the england player take as many wicket as bill johnston (Johnston's tally of 27 wickets is higher than any of the listed England players)
jim laker play in fewer match than any of the australian player (Laker participated in only three matches, which is fewer than any Australian player mentioned)

Refuted Statements are: *alec bedser have the best bowling average of any england player* (Bedser's bowling average of 38.22 is not the best among the listed England players)
all of the england player take as many wicket as bill johnston (Johnston's wicket tally of 27 is higher than any England player listed)
jim laker play in more match than any of the australian player (Laker participated in only three matches, which is fewer than any Australian player mentioned)

By exploring this dataset and applying state-of-the-art techniques in natural language processing, we aim to develop a robust and accurate system for verifying factual claims using tabular data. In the subsequent sections of this report, we will discuss the methodology employed, the dataset used, the experimental setup, and the results obtained.

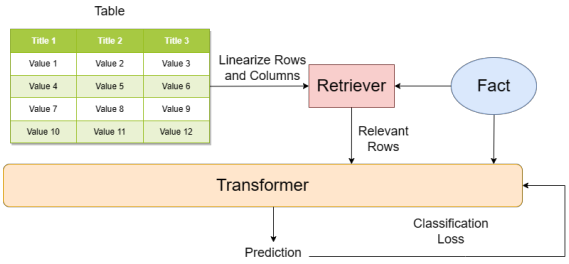

Figure 1: Model Architecture

2 Dataset

The dataset used in this project is TabFact, a large-scale and meticulously annotated dataset specifically designed for fact verification from tabular data. TabFact consists of 117,854 manually annotated statements, each associated with one of two possible relations: ENTAILED or REFUTED. These statements are evaluated based on the evidence presented in a collection of 16,573 distinct Wikipedia tables.

3 Methodology

Our general pipeline is shown in Figure 1. We first linearise the rows and columns of the table to sentences as done in the original paper. The linearised rows are sent to a retriever module that selects relevant rows based on the fact, and are combined with facts and sent to the Transformer model that predicts whether the fact is entailed or refuteds. This project employs a series of techniques to tackle the task of factual verification from tabular data.

1. **Contriever:** This technique adopts unsupervised dense information retrieval with contrastive learning, which enables the selection of relevant rows and columns to be sent to the fact verification model by leveraging contrastive learning.

2. **TF-IDF:** In the context of fact verification from tabular data, TF-IDF is utilized to select relevant rows and columns based on the textual information they contain. This is done by assigning higher weights to terms that are more discriminative.

3. **Latent Semantic Analysis:** LSA is a statistical approach that analyzes the relationships between terms and documents based on their co-occurrence patterns. By representing the textual content of the tables as a matrix, LSA identifies the underlying latent semantic structure within the data. This enables the identification of rows and columns that are semantically related to the hypotheses.

4. **BERT-Base:** BERT-Base, a transformer-based model, is utilized to assess the textual hypotheses. It employs masked language modeling to learn contextual representations and has demonstrated exceptional performance in various natural language processing tasks.

5. **RoBERTa:** RoBERTa, another transformer-based model, is employed to evaluate the textual hypotheses. It uses masked language modeling and next sentence prediction objectives to learn contextual representations of input text. Through fine-tuning on the task of fact verification, RoBERTa captures the semantic relationships between the hypotheses and the tabular evidence.

4 Experiments

In this section, we present the experimental setup conducted to evaluate the performance of various fact verification techniques on the TabFact dataset. We achieved an F1 score of 0.706 and 64.6% accuracy on the test set using Latent Program Analysis (LPA). The experiments were conducted using different configurations of row and column selection:

1. **All Rows**: The fact verification models were trained and evaluated using all available rows from the tables.

2. **Select Rows**: The models were trained and evaluated after selecting relevant rows using Contriever, TF-IDF, and LSA techniques.

3. **All Rows and Columns**: The models were trained and evaluated using both relevant rows and columns from the tables.

4. **Select Rows and Columns**: The models were trained and evaluated after selecting relevant rows and columns using Contriever, TF-IDF, and LSA techniques.

Evaluation Metrics: The performance of the fact verification models was assessed using **accuracy** (percentage of correctly predicted relations, entailed or refuted) and **F1 score** (balance between precision and recal) metrics.

		RoBERTa		BERT-Base	
		Accuracy	**F1 Score**	**Accuracy**	**F1 Score**
All Rows		65.12	0.7083	65.63	0.7102
Select Rows	Contriever	62.71	0.6813	63.63	0.6842
	TF-IDF	63.13	0.6768	62.73	0.6731
	LSA	63.02	0.6651	62.37	0.6636
All Rows and Columns		66.34	0.7171	66.11	0.7153
Select Rows and Columns	Contriever	67.30	0.7242	**67.53**	**0.7250**
	TF-IDF	66.78	0.7192	66.98	0.7189
	LSA	66.10	0.7143	66.58	0.7137

Table 2: Results

5 Evaluation & Analysis

In this section, we present a detailed evaluation and analysis of the fact verification models and techniques applied to the TabFact dataset. The performance of each technique is assessed based on accuracy and F1 score metrics. The results have been provided in tabular format in Table 2.

The first set of experiments involves using all rows from the tables for fact verification. BERT-Base achieved slightly better results than RoBERTa in this case. The results indicate that both models can effectively process the textual evidence in the rows to determine the validity of the hypotheses.

Next, we explore the impact of selecting relevant rows using different techniques. Contriever, TF-IDF, and Latent Semantic Analysis (LSA) are employed to identify informative rows for fact verification. When using the retrievers, the accuracy and F1 score drop. These outcomes suggest that the selection of relevant rows has a slight negative impact on the overall performance of the models, indicating that some useful information might be discarded during the selection process.

Furthermore, we investigate the effectiveness of incorporating column information along with rows for fact verification. The models are evaluated on the modified dataset, which includes both relevant rows and columns. The performance is slightly improved compared to using only rows, indicating that columns provide additional valuable information for the fact verification task. This is because many of the facts are based on information given in columns.

Finally, we assess the impact of selecting relevant rows and columns using Contriever, TF-IDF, and LSA techniques. These findings demonstrate that incorporating relevant rows and columns improves the performance of fact verification models.

6 Conclusion

In this project, we conducted a thorough investigation into the task of fact verification using tabular data, focusing on the selection of relevant rows and columns and the utilization of different language models. Through rigorous experimentation and analysis, we have obtained valuable insights and drawn significant conclusions.

Our evaluation indicate that incorporating both rows and columns leads to improved fact verification results compared to using only rows. While the selection of relevant rows and columns can have a slight negative impact on performance, the benefits of incorporating additional relevant information outweigh the potential drawbacks. The findings highlight the importance of considering both row and column data for effective fact verification from tabular evidence.

References

- Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., & Wang, W. Y. (2020). TabFact: A Large-scale Dataset for Table-based Fact Verification. In International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia.

- Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2022). Unsupervised Dense Information Retrieval with Contrastive Learning. Meta AI Research.

- Dumais, S. T. (2005). Latent Semantic Analysis. Annual Review of Information Science and Technology, 38, 188-230.