FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information

Project Report

Raveesh Vyas 2022114002
Ketaki Shetye 2022114013
Gargi Shroff 2022114009

1 Introduction

Fact verification has attracted a lot of interest in recent times as it is one of the key methods for detecting misinformation. In order to address the problem of potentially mislaeading false claims, the need for automated fact verification is on rise. This has resulted not only development of models, but also creation of appropriate datasets of scale, quality, and complexity to evaluate models for fact extraction and verification. While there has been a lot of progress in building fact verification systems, these methods have mostly ignored the valuable information found in structured formats like tables.

The project tries to cover this gap by implementing baseline approaches outlined in the study (Aly et al., 2021) for evidence retrieval and verdict prediction. The goal is to determine the veracity of a claim c by:

i) retrieving a set of evidence pieces E, which can be either a sentence or a table cell, and

ii) assigning a label y belonging to {Supports; Refutes; Not Enough Info}.

The source of evidence is derived from the English Wikipedia (excluding pages and sections flagged to require additional references or citations) `https://doi.org/10.5281/zenodo.4911508` and consists of sentences and tables. The evidence retrieval model is developed using a combination of entity matching and TF-IDF (Robertson, 2004) to extract the most relevant sentences and tables. This is followed by a cell extraction model that returns relevant cells from tables by linearizing them and treating the extraction as a sequence labeling task. Further, a RoBERTa classifier (Liu et al., 2019) pre-trained on multiple NLI datasets predicts the veracity of the claim as 'supported', 'refuted', or 'lacks enough information' using the retrieved evidence and its context. Moreover, we detail our efforts to track and minimize the biases present in the dataset and could be exploited by models, e.g. being able to predict the label without using evidence. etc. We implement a baseline for verifying claims against text and tables which predicts both the correct evidence and verdict.

2 Appropriate baselines from the literature study

2.1 Literature review

With the development of various baselines to assess the performance of models on fact-checking datasets, fact verification has seen significant advancements. These baselines serve as benchmarks, evaluating the ability of systems to retrieve evidence and predict verdict. Existing work primarily focuses on datasets like FEVER(Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018) which focus on textual evidences. These have laid the foundation for large-scale fact-checking systems but are limited in scope when it comes to structured data such as tables.
FEVEROUS(Aly et al., 2021) dataset addresses this limitation by integrating both structured (tables) and unstructured (textual) data. It consists of 87,026 verified claims, each annotated with evidence in the form of sentences or table cells from Wikipedia. The integration of structured evidence introduces complexities not addressed in earlier baselines. For instance, datasets like TabFact (Chen et al., 2019) and SEM-TAB-FACTS (Wang, Mahajan, Danilevsky, & Rosenthal, 2021) rely on pre-provided tables, bypassing retrieval challenges. The FEVEROUS baseline combines entity matching and TF-IDF retrieval techniques to extract relevant sentences and tables. For structured evidence, table linearization is employed, treating cell extraction as a sequence labeling task. These

methods are augmented by RoBERTa-based classifiers, fine-tuned for predicting verdicts based on retrieved evidence (Aly et al., 2021).
The FEVEROUS baseline also tackles challenges such as multi-hop reasoning and numerical operations, which are underexplored in prior works. By addressing these aspects, the FEVEROUS baseline provides a comprehensive framework for evaluating models' reasoning and retrieval capabilities. In implementing these baselines, our project builds upon the methodologies used for FEVEROUS while exploring improvements in retrieval accuracy and reasoning. This implementation also aims to identify areas where current baselines fall short, providing insights for future model development in multi-modal fact verification

3 Dataset Characteristics

As presented in the study (Aly et al., 2021), the FEVEROUS dataset and benchmark (Fact Extraction and VERification Over Unstructured and Structured information) introduces a novel approach for verifying claims using Wikipedia pages. Each claim is categorized as supported, refuted, or having insufficient information, with corresponding evidence drawn from Wikipedia in the form of sentences and/or table cells which is used for the verdict prediction.

3.1 FEVEROUS Dataset

The FEVEROUS dataset focuses on large-scale fact verification, combining sentences, tables, and the integration of both. It is publicly available online at `https://fever.ai/dataset/feverous.html`, and is used in for the implemented approach. Figure 1 provides two example instances from the dataset, showcasing the complexity of the dataset. As shown, the evidence for a claim can come from a single table cell, a sentence, or a combination of both, extracted from different articles.

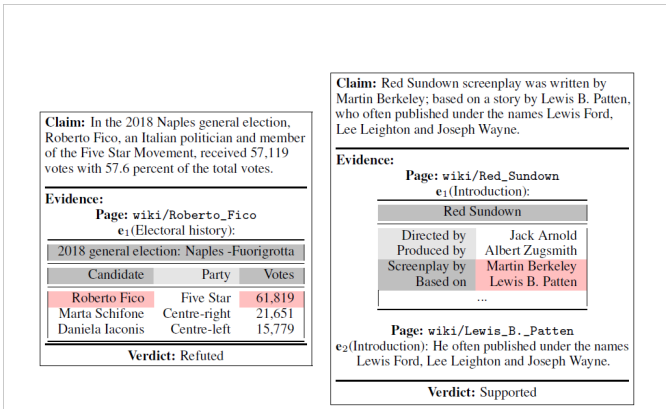

Figure 1: FEVEROUS sample instances. Evidence in tables is highlighted in red. Each piece of evidence e_i has associated context. Image source: (Aly et al., 2021).

The FEVEROUS dataset is designed for fact extraction and verification tasks. The dataset combines various types of evidence (sentences and tables) for verifying claims and labels them as either supported, refuted, or lacking enough information (NEI). Below are some key quantitative characteristics of the dataset:

3.2 Dataset Splits and Evaluations

For the purposes of this project, the FEVEROUS dataset is split into training, development (dev), and test sets. The training and development datasets were obtained directly from the FEVEROUS website `https://fever.ai/dataset/feverous.html`. A total of 71,291 claims were used for training. The test set, however, was constructed by implementing a retriever model using claims from the dev set. The retrieved evidences and the gold labels from dev set were used for testing.

The quantitative characteristics of the dataset, split into different classes for each set, are shown in the table below. It includes the number of claims classified into three categories: **SUPPORTS**, **REFUTES**, and **NEI** (Not Enough Information). Additionally, a balanced NEI label set was used for training the RoBERTa+NLI+NEI model.

Statistic	FEVEROUS
Total Claims	87,026
Avg. Claim Length	25.3
Avg. Evidence	1.4 sentences, 3.3 cells (0.8 tables)
Evidence Sets by Type	34,963 sentences, 28,760 tables, 24,667 combined
Size of Evidence Source	95.6M sentences, 11.8M tables
Veracity Labels	49,115 Supported, 33,669 Refuted, 4,242 NEI

Table 1: FEVEROUS Dataset Statistics. Table source: (Aly et al., 2021).

	SUPPORTS	REFUTES	NEI	Total
Train	41,835	27,215	2,241	71,291
Dev	3,908	3,481	501	6,890
Test	3,789	3,341	466	7,596
Train (NEI Balanced)	37,563	22,990	10,738	71,291

Table 2: Dataset Split and Class Distribution. The table includes a balanced NEI label set used for training the RoBERTa+NLI+NEI model.

This table summarizes the distribution of claims across the different categories in the training, development, and test sets, and also reflects the changes made during the creation of the "Train (NEI Balanced)" set, where the NEI labels were balanced to improve model training performance.

4 Methodology

The proposed solution is divided into two subtasks, one being retrieval of the evidence, and the other being verdict prediction (refuted, supported, or not enough information) using the claim and evidence. The pipeline(2) explains the basic pipeline of the approach implemented. Firstly, the evidence is retrieved from the Wikipedia pages for the claims in the FEVEROUS dataset. The retrieved evidences contains both the sentences and tables relevant to the claim. The claim and retrieved evidence together using RoBERTa classsifier is used to predict the verdict for the claim.

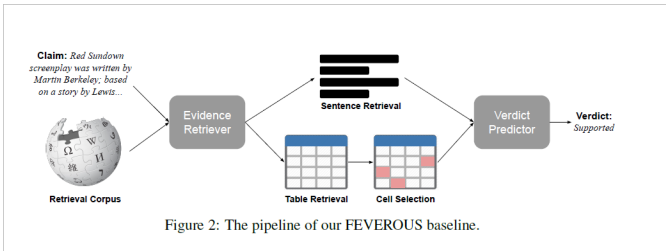

Figure 2: The pipeline of our FEVEROUS baseline.

Figure 2: The pipeline of our FEVEROUS baseline. Image source: (Aly et al., 2021).

4.1 Evidence retrieval

For evidence retrieval we use a model which is a combination of entity matching and TF-IDF values. The retrieval process can be broken down into the following parts

4.1.1 Page retrieval

Firstly, we find all relevant wikipedia pages by extracting the named entities in the claim, and matching the pages. All of these pages are then scored by the TF-IDF score of the first sentence in the article. We extract the top k pages using this, where k has been set to 5

4.1.2 Sentence retrieval

To extract the unstructured data from the articles, we extract the relevant sentences from the retrieved pages. All the sentences in these pages are scored by their TF-IDF score with the claim, and the top l (=5) sentences are extracted.

4.1.3 Table retrieval

For the structured data, we retrieve the relevant tables from the articles. The top q (=3) tables are chosen using TF-IDF scores of their contents. We also extract the top cells from these tables by their TF-IDF scores.

4.2 Verdict Prediction

To predict the verdict for claims, we use a RoBERTa-based classifier with a linear layer on top. RoBERTa, a pretrained transformer-based model that excels at understanding the contextual relationships in text, makes it highly effective for extracting meaningful features that are crucial for classification.

For each piece of evidence, its content is concatenated with the associated context, forming a single input sequence. This input is fed into the RoBERTa model to predict the verdict. The FEVEROUS training dataset is utilized for training and fine-tuning the model.

4.2.1 Experiment 1: Fine-tuning Base RoBERTa

In the first experiment, the RoBERTa classifier model (Liu et al., 2019) is fine-tuned on the provided training data, which consists of 71,092 samples. The fine-tuned model is then used to predict verdicts for the claims in the test set.

However, as the base RoBERTa model is not pretrained on Natural Language Inference (NLI) tasks, it does not inherently understand the relationships between claims and evidence.

4.2.2 Experiment 2: Fine-tuning NLI-Pretrained RoBERTa

To address the limitations observed in Experiment 1, we leverage an NLI-pretrained RoBERTa model. This model has been pretrained on SNLI and MultiNLI datasets, enabling it to capture semantic relationships between sentence pairs effectively.

We fine-tune the NLI-pretrained RoBERTa model on the FEVEROUS training data to predict verdicts for claims.

To make further improvements to the model, we address the imbalance in the training data, where the "NEI" (Not Enough Information) label is significantly underrepresented.

4.2.3 Experiment 3: Balancing the Training Data

As shown in Table 2, there are only 2,241 "NEI" labelled claims compared to 41,835 "SUPPORT" and 27,215 "REFUTES" labelled claims. This imbalance impacts the model's ability to handle "NEI" cases effectively. To improve the representation of "NEI" labelled claims, we employ the following data augmentation strategies:

- **Partial Evidence Removal:** For claims annotated with both sentence and table cell evidence, we create new samples by removing either the sentence or the cells, thereby converting them into "NEI" cases.
- **Irrelevant Evidence Replacement:** We generate additional "NEI" labelled samples by replacing the original evidence with irrelevant evidence from the dataset.

Using these techniques, we increase the number of "NEI" labelled claims to 10,738, resulting in a more balanced dataset. The NLI-pretrained RoBERTa model is then fine-tuned on this augmented dataset to predict the verdicts more accurately.

4.2.4 Hyperparameters

The hyperparameters used have been kept constant for all three experiments and are as follows:

The experiments demonstrate the importance of leveraging an NLI-pretrained model and addressing label imbalances in the training data for the verdict prediction of the claims using the retrieved evidence.

Hyperparameter	Value
Criterion	CrossEntropyLoss
Number of Warmup Steps	500
Total Training Steps	10000
Maximum Sequence Length	512
Number of Epochs	3
Learning Rate (lr)	1e-5
Batch Size	16
Optimizer	AdamW

Table 3: Hyperparameters kept constant throughout the experiments

5 Results

5.1 Retrieval scores

The following are the scores of the retrieved sentences compared to the gold evidence sets

	Precision	Recall	F1	Feverous score (evidence)
Sentences	0.4316	0.4227	0.4271	0.27
Tables	0.3144	0.3295	0.3217	

Table 4: Performance metrics of the retriever

5.2 Verdict Prediction

The performance metrics of the RoBERTa-based models across different experiments for verdict prediction highlight the overall and per-class performance on both the development and test datasets. Metrics such as F1-score, accuracy, precision, and recall are reported to provide a comprehensive evaluation of the models.
Table 5 provides the overall performance metrics—F1-score, accuracy, precision, and recall—achieved by the RoBERTa models across different experimental setups

	Dev			**Test**		
Metric	RoBERTa	RoBERTa + NLI	RoBERTa + NLI + NEI	RoBERTa	RoBERTa + NLI	RoBERTa + NLI + NEI
F1 Score	0.6046	0.6588	0.6748	0.4108	0.3950	0.3956
Accuracy	0.8501	0.85314	0.8562	0.5316	0.4942	0.5133
Precision	0.6950	0.7248	0.6762	0.4149	0.4163	0.4122
Recall	0.6136	0.6475	0.6319	0.4139	0.4149	0.4060

Table 5: Overall performance metrics (F1-score, accuracy, precision, and recall) on development and test datasets for RoBERTa, RoBERTa with NLI pretraining, and RoBERTa with NLI pretraining and NEI label balancing.

The detailed breakdown in Table 6 shows the model's ability to predict each verdict category accurately.

	Dev			**Test**		
Model	**Support**	**Refute**	**NEI**	**Support**	**Refute**	**NEI**
RoBERTa	0.8906	0.8642	0.0589	0.5475	0.5696	0.1153
RoBERTa + NLI	0.8896	0.8709	0.2158	0.4873	0.5789	0.1188
RoBERTa + NLI + NEI	0.8827	0.8598	0.2218	0.5880	0.5277	0.1911

Table 6: Verdict classification using gold evidence for the development dataset and retrieved evidence for the test dataset. NLI denotes pre-training on NLI corpora and NEI sampling. Scores are reported in per-class F1. The overall score is reported using macro-averaged F1.

5.3 Feverous scores

The feverous score of the model is calculated as a fraction of the claims for which the verdict is correctly classified, and at least 1 evidence set of the gold evidence sets is a subset of the extracted evidences. For each claim, we assign a score as:

$$Score(y, \hat{y}, \mathbb{E}, \hat{E}) = \begin{cases} 1 & \exists E \in \mathbb{E} : E \subseteq \hat{E} \wedge \hat{y} = y \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Here, $\hat{y}$ and $\hat{E}$ are the predicted label and evidence respectively, and y and $\mathbb{E}$ are the gold label and evidence sets respectively. Our metric for evaluation would be the score over all claims divided by the total number of claims. The following are the scores for retrieved data when compared with the gold data

	Baseline RoBERTa	**NLI pretrained RoBERTa**	**NLI + NEI balanced**
Sentence only	0.1326	0.1415	0.1428
Table only	0.1225	0.1247	0.1276
Combined	0.1597	0.1635	0.1670

Table 7: Feverous scores comparing when there is only sentences, tables or both combined used as evidence

6 Analysis

The performance of the RoBERTa-based models for verdict prediction is evaluated through a series of experiments, with results presented in Tables 5 and 6. These results demonstrate the effectiveness of various training strategies and the challenges associated with the task.

6.1 Retriever

The retriever model shows reasonable performance across both structured and unstructured data. Recall values of 0.4227 and 0.3295 show that the retrieved evidences match the gold evidence sets. We also calculate the feverous score for evidence only as the number of claims for which the retrieved evidence is a subset of at least one of the gold evidence sets. The feverous score of 0.27 shows that the retriever model extracts valid sentences and table entries.

6.2 Verdict Predictor

6.2.1 Overall Performance Across Experiments

Baseline RoBERTa: The baseline RoBERTa model demonstrates reasonable performance on the development dataset, achieving an F1-score of 0.6046 and an accuracy of 0.8501. These scores indicate that the model benefits from the pre-trained representations of RoBERTa. However, its performance on the test dataset is significantly weaker, with an F1-score of 0.4108 and an accuracy of 0.5316. This gap highlights the model's limited ability to generalize to unseen data.

Incorporating NLI Pretraining: Introducing pretraining on the NLI corpus leads to noticeable improvements on the development data, with the F1-score rising to 0.6588. This indicates that the additional NLI training enhances the model's ability to process logical relationships, such as entailment and contradiction, which are critical for verdict prediction. However, on the test data, the F1-score slightly drops to 0.3950, and accuracy decreases to 0.4942 compared to the baseline. This discrepancy suggests that while NLI pretraining improves the model's capacity for understanding semantic relationships, it does not fully address label imbalances present in the dataset as the the model is not able to recognise a NEI samples correctly.

Balancing NEI Labels: In the third experiment, balancing the underrepresented NEI class further enhances the F1-score on the development dataset to 0.6748, the highest among all models. This improvement indicates that addressing class imbalance helps the model better recognize and classify NEI instances. On the test dataset, the accuracy marginally increases to 0.5133, and the F1-score improves to 0.3956. While these gains are incremental, they show the importance of balanced data representation. However, the performance on the test data still lags, suggesting the need for advanced sampling techniques in order to improve generalization.

6.2.2 Per-Class Performance Analysis

Support Class: The Support class achieves high F1-scores on the development dataset across all models, exceeding 0.88. This strong performance indicates that the models can effectively identify supportive evidence on the development gold dataset as sufficient training examples are available for it to generalise. However, the F1-scores on the test dataset drop significantly, with the baseline RoBERTa model achieving 0.5475. While the NLI-pretrained model slightly reduces this drop, the results suggest that the test data may include some noise like irrelevent or incomplete evidence retrieved that the models struggle to handle.

Refute Class: Similar to the Support class, the Refute class shows strong performance on the development dataset, with F1-scores above 0.85 for all models. The inclusion of NLI pretraining improves the model's ability to handle refutational relationships, as seen in the development scores. However, the test performance remains relatively low, with only marginal improvements across experiments. This result highlights that while NLI pretraining aids the model in processing refutational evidence, it cannot predict for unseen or noisy data.

NEI Class: The NEI class is the most challenging for all models, with F1-scores significantly lower than those for the other classes due to imbalance in number of NEI samples in the training dataset. The baseline RoBERTa model achieves an F1-score of only 0.0589 on the development data and 0.1153 on the test data. Although, incorporating NLI pretraining improves these scores, balancing the NEI class further increases the F1-score on the development dataset. Even after increasing the NEI samples from 2,241 to 10,738 for the third experiment, the model performs decently well. It also highlights that with more advanced sampling of NEI instances, the model can predict the NEI samples more accurately.

7 Conclusion

In this report, we presented a the detailed implemented solution for solving the problem of evidence retrieval and verdict prediction. The proposed approach effectively combines entity matching and Term Frequency-Inverse Document Frequency (TF-IDF) techniques to identify relevant evidence from Wikipedia. Further, various RoBERTa models are used for verdict prediction for the claims. By leveraging both textual and tabular data, the solution aims to provide robust evidence retrieval and accurate verdict prediction.

References

Aly, R., Guo, Z., Schlichtkrull, M., Thorne, J., Vlachos, A., Christodoulopoulos, C., ... Mittal, A. (2021). *Feverous: Fact extraction and verification over unstructured and structured information.* Retrieved from `https://arxiv.org/abs/2106.05707`

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., ... Wang, W. Y. (2019). Tabfact: A large-scale dataset for table-based fact verification. *CoRR*, *abs/1909.02164*. Retrieved from `http://arxiv.org/abs/1909.02164`

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.* Retrieved from `https://arxiv.org/abs/1907.11692`

Robertson, S. (2004, 10). Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation - J DOC*, *60*, 503-520. doi: 10.1108/00220410410560582

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018, June). FEVER: a large-scale dataset for fact extraction and VERification. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 809–819). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N18-1074` doi: 10.18653/v1/N18-1074

Wang, N. X. R., Mahajan, D., Danilevsky, M., & Rosenthal, S. (2021, August). SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th international workshop on semantic evaluation (semeval-2021)* (pp. 317–326). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.semeval-1.39` doi: 10.18653/v1/2021.semeval-1.39