PUBHEALTH Fact-Checking System

Overview

Project builds a fact-checking pipeline to verify the veracity of public health claims. Using the **PUBHEALTH dataset** and state-of-the-art natural language processing (NLP) models, we fine-tune a pre-trained model to predict the truthfulness of a given claim. The pipeline is designed to be modular and robust, including data ingestion, preprocessing, model training, deployment, and monitoring steps.

Process Overview

1. **Data Ingestion**: We downloaded the PUBHEALTH dataset using Hugging Face's datasets library, which provides labeled examples of public health claims with veracity labels (true, false, unproven, mixture).
2. **Data Processing**: Raw data is cleaned and preprocessed by:
    - Removing duplicates and irrelevant columns.
    - Encoding labels to numerical values.
    - Combining claim and explanation text into a single feature.

This processed data is split into training and test sets to prepare for model fine-tuning.

3. **Model Selection**: The model chosen is distilbert-base-uncased-finetuned-health_facts, which balances performance and computational efficiency. DistilBERT is a distilled version of BERT, optimized for faster inference without significant loss in accuracy, making it suitable for this large-scale text classification task.
4. **Training**: The chosen model is fine-tuned on the training set. Hyperparameters like batch size, learning rate, and the number of epochs is configurable. During training, accuracy and loss are monitored to evaluate the model's performance.
5. **Deployment**: The trained model is deployed as a FastAPI service in a Docker container, allowing external applications to make predictions through an API endpoint.
6. **Monitoring and Updates**: We employ Weights & Biases (WandB) to track model performance and resource utilization over time. WandB provides insights into metrics like accuracy, F1 score, and helps monitor potential drifts, enabling efficient retraining and optimization.

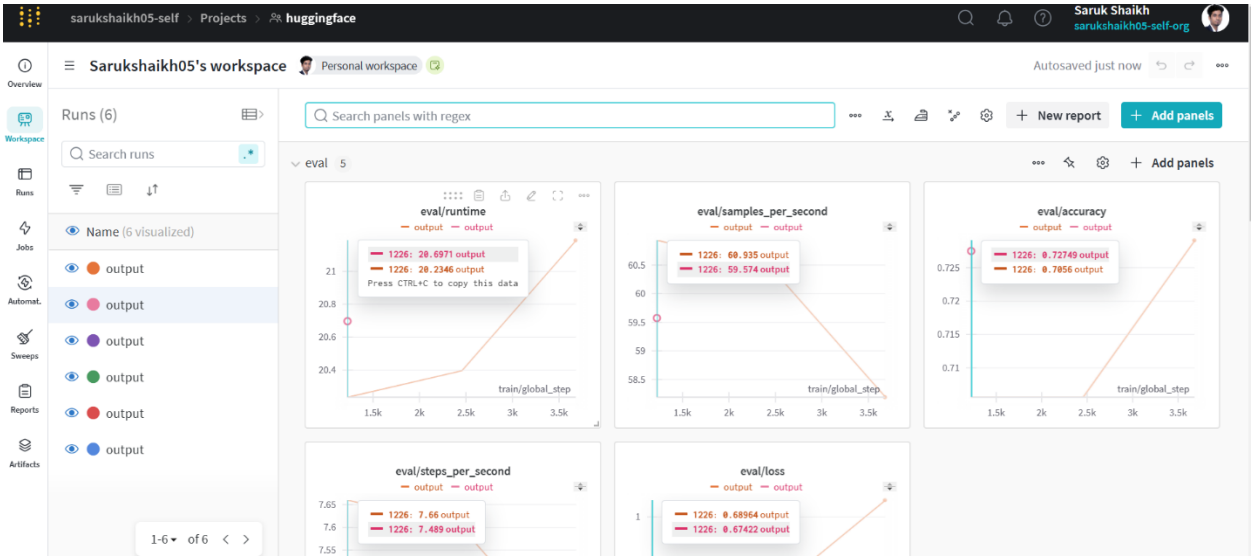
Model Selection

The choice of the model is essential for achieving a good balance between **accuracy** and **efficiency**.

Model Choice: distilbert-base-uncased-finetuned-health_facts

This specific model is a fine-tuned variant of DistilBERT on health-related claims, optimized for fact-checking tasks. Key reasons for this choice:

- **Performance**: DistilBERT retains 97% of BERT's language understanding while being 60% faster.
- **Efficiency**: Reduced model size allows faster inference and lower memory usage, making it suitable for production deployment.
- **Specialized Fine-tuning**: This model has been further trained on health facts, making it more adept at recognizing health claim veracity patterns than general-purpose BERT models.

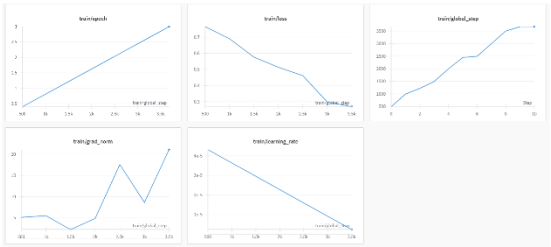


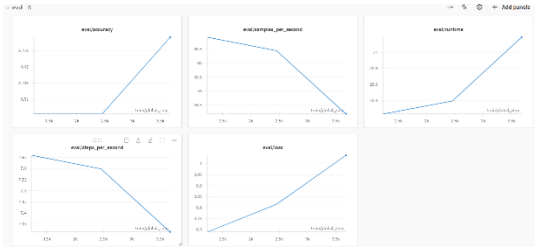
Model Evaluation

To evaluate the fine-tuned model, we focus on both **accuracy** and **explainability**:

1. **Evaluation Metrics**:
    - **Accuracy**: Measures the proportion of correctly classified claims.
    - **F1 Score**: Balances precision and recall, especially important for imbalanced classes (e.g., if "unproven" or "mixture" claims are less frequent).
2. **Cross-Validation**:
    - To ensure robustness, we can use cross-validation on the training set, which helps estimate model performance across different splits.
    - Cross-validation reduces overfitting risk and gives a better picture of expected accuracy on unseen data.
3. **Explainability**:

- The model's attention weights can be inspected to understand which parts of the text contribute most to the prediction, providing insights into model decisions.





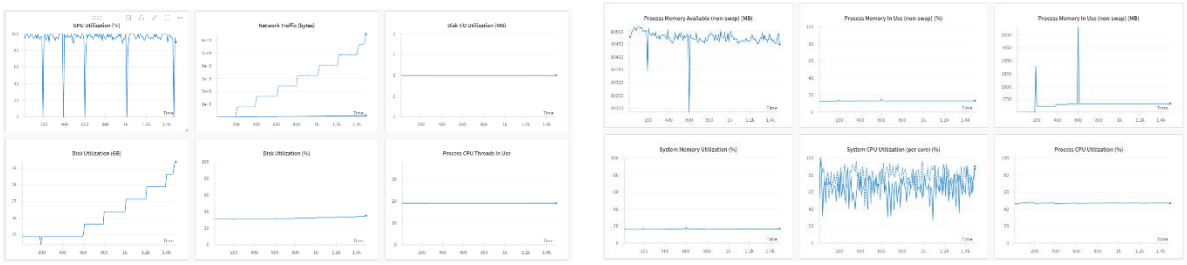
Model Optimization

Further optimization of the model can be performed using the following strategies:

1. **Hyperparameter Tuning**:
   - Conduct a grid search or use Bayesian optimization (e.g., optuna) for hyperparameter tuning.
   - Key hyperparameters include learning rate, batch size, and the number of training epochs.
2. **Data Augmentation**:
   - Use synthetic data generation techniques to expand the dataset, particularly for minority classes (e.g., "unproven").
   - Techniques such as paraphrasing or back-translation can introduce slight variations in claims, improving the model's generalizability.
3. **Regularization Techniques**:
   - Use dropout or weight decay to prevent overfitting. These techniques can help the model generalize better, especially when working with small datasets.
4. **Ensemble Models**:
   - Combine predictions from multiple models (e.g., a voting classifier with BERT and RoBERTa variants) to improve overall accuracy.
   - This approach can capture more diverse perspectives, though at the cost of increased computational overhead.

Monitoring and Updates

Monitoring is essential to ensure that the model performs well over time and to identify when retraining is needed due to data drift or new patterns in incoming data.



Monitoring with Weights & Biases (WandB)

1. **Real-time Monitoring**:
    - Track metrics like accuracy, F1 score, and loss during training and evaluation.
    - Monitor resource usage, such as GPU memory and processing time, which helps optimize infrastructure.
2. **Alerting**:
    - Set up alerts for significant drops in accuracy or increases in loss. This helps promptly detect and address issues before they impact production performance.
3. **Model Drift Detection**:
    - Monitor the distribution of input data features over time. If substantial changes are detected (e.g., more "unproven" claims), it may indicate a need to update the model.

Update Strategy

- **Scheduled Retraining**: Retrain the model periodically (e.g., every 3-6 months) to incorporate new data, especially if there are updates in public health information or new claim types.

- **Incremental Learning**: Fine-tune the existing model on small batches of new data without full retraining to reduce computational load.

Conclusion

This project demonstrates a robust pipeline for automated fact-checking of public health claims. By leveraging the PUBHEALTH dataset and fine-tuning a specialized version of DistilBERT, we have created a scalable solution that can assess the veracity of health-related claims. Here are the key takeaways:

1. **Efficient Model Selection**: Choosing distilbert-base-uncased-finetuned-health_facts strikes an optimal balance between performance and efficiency, enabling the system to perform reliable predictions with reduced computational costs.
2. **Comprehensive Pipeline**: The project encompasses end-to-end processes, from data ingestion to API deployment, making it straightforward to integrate into larger systems or production environments.
3. **Model Evaluation and Optimization**: By using cross-validation, accuracy, and F1 score as core evaluation metrics, the model's effectiveness is rigorously assessed. Optimization techniques like hyperparameter tuning and data augmentation further enhance model robustness and accuracy.
4. **Effective Monitoring and Maintenance**: Utilizing tools like Weights & Biases (WandB) for tracking metrics and resource utilization allows for proactive monitoring, ensuring that the model remains performant over time. With scheduled retraining and monitoring, this solution can adapt to new data and changing patterns, ensuring that it continues to deliver accurate results.
5. **Scalability and Flexibility**: The Docker-based deployment and optional Kubernetes configuration make the system highly scalable, capable of handling increased load in a cloud-based or distributed environment.