Project Deployment Report

Deployment

- **Deployment pattern:** Containerized deployment dynamically on server
- **Deployment strategy:** Automatic single deployment
- **Justification:** Portable, can be deployed to any environment, and easy to scale, Automatic deployment ensures that the model is always up-to-date and available.
- **Trade-offs:** Containerized deployment can be more complex to set up than traditional deployment methods. It is efficient than virtual server. Single deployment is easy to implement and cost efficient but may cause some risks and
- **Challenges:** The main challenge with Containerized deployment is that it is complex to set up and have to maintain on server.
- **Model trade-offs:** The best model (Bert classifier) is too resource consuming in both training and prediction time; That's why we had to use an inferior model (Binomial Bayesian classifier)

Automation

- **Containerization:** The model is containerized using Docker.
- **CI/CD:** We tried to add CI/CD pipeline using GitHub Actions to automate the deployment process. However, we faced some challenges such as authorization
- **ML workflow:** The ML workflow is automated using Prefect (scheduling and orchestration).
- **Tools and technologies:** The following tools and technologies are used to automate the deployment process:
  - Docker
  - MLFlow
  - Prefect

Monitoring

- **Metrics:** The following metrics are monitored:
  - F1 macro score (main)
  - balanced accuracy
  - accuracy
- **Production metrics:** using grafana and prometheus
- **Model validation:** Model deployment fails if F1 score is under a threshold

Challenges and Trade-offs

- **Challenges:** The main challenges with deployment were:
  - Making trade-offs between cost, performance, and reliability.
  - Strange errors during building and deployment!

- **Trade-offs:** The following trade-offs were made:
    - Containerized deployment is portable, scalable and reproducible but it Can be more complex to set up than serverless deployment, and may not be suitable for models that need to be deployed to devices with limited resources
    - We had to use an inferior model in production