Cover letter: describe how you could solve a similar problem in a production project for

sentiment analysis task (you may add extra steps to the pipeline if needed)

Referring to my infrastructure in the brief description, I can solve a similar problem in a production project for sentiment analysis task as follows;

1. **Get the data**

Recorded phone conversations (voice clips) need to be transcribed to text, while social media text data can directly be stored as files in the Artifact Storage. This storage can also be in the cloud.

2. **Preprocess the data**

With text format, we still need to preprocess the data e.g. handling missing data and duplicates, removing tags and special characters, and selecting column names and labels to be used as features for training. The main data preprocess step for sentiment analysis task is to vectorize the training data (using NLP technique) so that our model can use the vectors to train on.

3. **Input train data into a model and train the model**

Once we have our data cleaned, we can input the vectors into our model. We could use Logistics Regression algorithm to classify the data into negative, neutral, or positive sentiment. Naïve Bayes, Logistic Regression and Support Vector Machines (SVM) are widely used for large-scale sentiment analysis because they scale well.

We might add 3.1 a step to set up the infrastructure to have the training task, which requires computing power, to be performed in the cloud. Kubeflow supports connection with most of the cloud services. Once the model is trained, the cloud service can save the model in cloud storage. An option for cloud service could be Azure, which has just announced that the fastest public cloud supercomputer is now generally available. It is their [Azure NVIDIA A100 GPU Clusters.](https://azure.microsoft.com/en-in/blog/azure-announces-general-availability-of-scaleup-scaleout-nvidia-a100-gpu-instances-claims-title-of-fastest-public-cloud-super/)

4. **Evaluation of the initial model and optimization of the model**

We can select which performance measurement metrics we want to use (e.g. F-Value, ROC curve). Depending on the metrics and the performance result, we can choose the technique to optimize the model accordingly. We can use many optimization algorithms available in Keras by modifying the loss function.

5. **Use the model to predict is another container and cluster on its own**

When user requests a prediction, they are using the model that has been trained earlier and saved. There will be no training on the spot unless there are layers with different behaviour during training versus inference. The inference pipeline starts with inputting the test data, taking the readily-available model, giving out the prediction, and sending the prediction to the user once a POST request is received (e.g. submit button clicked).

6. **Additional tasks**

Apart from sentiment analysis classification task where prediction can be used to see if customers feedbacks are improving over time, NLP techniques can also be used for tasks such as detecting the most talk-about topic, extracting key phrases to see what customers are interested in or would like to have as a new product, finding similarities between words to include relevant comments, or categorizing the documents. These tasks can be processed and deployed as different endpoints for users.