Capstone 3 proposal: H&M product category classification

(1) What is the business problem?
The business problem is classifying new products into an existing taxonomy of product categories for online retail stores such as H&M.

(2) Who are the intended stakeholders, and why is this problem relevant to them?
The stakeholders are the company's sales department, as having products correctly classified in an intuitive, easy to navigate taxonomy helps customers find what they are looking for and boost sales. Having correctly classified products also helps with search accuracy.
The company's product data management is also a key stakeholder in this issue as they are the ones responsible for fixing incorrectly classified products and inconsistencies which can be a time consuming manual process that increases operational costs.

(3) Where is the dataset from?
The dataset is from the H&M kaggle competition to provide product recommendations based on previous purchase history of individual customers. Part of the competition's data is a dataset of H&M product metadata and product images. The product categories are provided for each product.

The dataset has 105k rows, where each row is a product. There are 24 columns in the dataset. I like this dataset because it would use both NLP and computer vision. The constraints of this dataset is that there are several product categories that could be used for classification, some more relevant to certain business use cases than others, so determining which one to use is open ended (and might evolve into using a combination of product categories)

(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem?
To build supervised classification models to predict product categories based on product titles, product descriptions and product images.

How do you anticipate that you will evaluate the performance of each of the data science approaches that you envision?
Performance of the models will be evaluated on the metrics (Precision, Recall, Accuracy and F1 score). However other metrics will be considered as needed.

(5) How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?
The company could use the models to automate labeling of new products.
The sales department could use the classification results to evaluate if the current product taxonomy makes sense, or if there is room for improvement. E.g. if slippers are frequently misclassified by the machine learning model as being footwear rather than the correct product category of accessories.
The company's product data management team could use the models results to validate if the product categories of existing inventory are labeled correctly.