**Analytics for Unstructured Data**

**F2025 Assignment 3 (Group work)**

**Due: 12th October 2025 by 11:59 p.m.**

**Is a Picture Worth a Thousand Words?**

**You have two alternatives (Choose One)**

**Alternative 1**

**You are an analytics consultant to a brand (Choose one, e.g., Zara, H&M, National Geographic, etc.). Your objective is to help the brand increase engagement on its Instagram page.**

**Task A:** Scrape Instagram.py to fetch ~500 posts from the brand’s Instagram page. Fetch (i) image URLs, (ii) post caption (the text description of a post), and (iii) # likes. Fetching comments is difficult and you can easily get blocked by Insta. Using a dynamic VPN like ExpressVPN is highly recommended.

**Task B**: Using the image URLs, obtain **image labels** (text) from Google Vision (cloud service) or other services such as Azure. You can also use an LLM through its API. You will need an account, though.

**Task C:** Create a column called **binary** (lowercase only) where value =1 (stands for high engagement) or 0 (stands for low engagement) based on whether the number of likes is above or below the median value.

**Task D:** Run a logistic regression with **binary** as the dependent variable, and the image\_labels as independent variables. You can use a BoW model for text. What is the accuracy (show the confusion matrix) of this prediction model? The idea is to be able to predict the engagement level for an image.

Accuracy = 1 - # prediction errors / total # cases

What accuracy do you get by using the post\_caption words as the independent variables instead of image\_labels? Finally, what accuracy do you get by combining (concatenating) the image\_labels and post\_caption and using them together as independent variables? What can you conclude from your analysis?

**Task E:** Perform topic modeling (LDA) on the original image\_labels. Choose an appropriate number of topics. You may want to start with 4-5 topics, but adjust the number up or down depending on the word distributions you get. **Decide on suitable names for each topic.**

Now sort the data from high to low number of likes (don’t use the binary column, use the actual number of likes), and consider the highest and the lowest quartiles of likes. What are the main differences in the **average** topic weights of images across the two quartiles (e.g., greater weight of some topics in the highest versus lowest quartiles)? Show the main results in a table.

**Task F:** What advice would you give to the brand if it wants to increase engagement on its Instagram page based on your findings?

**Alternative 2**

**You are an analytics consultant to a non-profit organization (e.g., a pet rescue society, a children’s hospital, etc.). Your objective is to help the organization raise more funds in its gofundme.com campaigns.**

**Task A:** Scrape ~1000 fundraising campaigns in a particular category (e.g., Golden Retriever rescue society, an organization raising funds for wildfire victims, etc.). Fetch (i) image URLs, (ii) text description of a post, (iii) $ raised, (iv) how long the campaign has been running.

**Task B**: Using the image URLs, obtain **image labels** (text) from Google Vision (cloud service) or other services such as Azure. You can also use an LLM through its API. You will need an account, though.

**Task C:** Create a column called **binary** (lowercase only) where value =1 (stands for high $$) or 0 (stands for low $) based on the median value.

**Task D:** Run a logistic regression with **binary** as the dependent variable, and the image\_labels as independent variables. You can use a BoW model for text. Figure out how you can use the duration of the campaign in the model. What is the accuracy (show the confusion matrix) of this prediction model? The idea is to be able to predict the engagement level for an image.

Accuracy = 1 - # prediction errors / total # cases

What accuracy do you get by using the description words as the independent variables instead of image\_labels? Finally, what accuracy do you get by combining (concatenating) the image\_labels and word descriptions and using them together as independent variables? What can you conclude from your analysis?

**Task E:** Perform topic modeling (LDA) on the original image\_labels or word descriptions depending on which category gave higher accuracy in D. If the combination of image\_labels and description words gave the highest accuracy, use them together. Choose an appropriate number of topics. You may want to start with 4-5 topics, but adjust the number up or down depending on the word distributions you get. **Decide on suitable names for each topic.**

Now sort the data from high to low $ raised (don’t use the binary column, use the actual $), and consider the highest and the lowest quartiles of $. What are the main differences in the **average** topic weights of images across the two quartiles (e.g., greater weight of some topics in the highest versus lowest quartiles)? Show the main results in a table.

**Task F:** What advice would you give to the organization if it wants to increase the $ raised based on your findings?