1. Project Overview

Tribe Dynamics is a San Francisco-based startup that measures social media engagement for beauty, fashion, and lifestyle brands. Online content creation led by digital influencers (beauty and fashion bloggers) is one of the key predictors of offline revenue in this industry. Tribes provide software, data, and consulting services to some of the most well-known brands, private equity firms and Fortune 500 companies in the space. Tribe’s proprietary data is also regularly featured in leading industry research through publications such as WWD, CEW, and The Business of Fashion. Tribe is a team of writers, strategists, data scientists, and developers that has been creating the next generation of social influencer marketing tools since 2012.

Tribe Dynamics reports metrics to clients split by product category. Labels such as skincare, haircare, or cosmetics are automatically determined for every social media post using text classification. Our training data set is mostly in English because we originally started by tracking brands in the U.S. As we expand internationally to Europe and Asia, we must be able to build the same classification tools for other languages. We want to investigate how the latest techniques in transfer learning and NLP can improve our text classification accuracy across different languages.

Objectives. Build a human-in-the-loop pipeline for training a foreign language text classifier given Tribe’s English text training data, limited training data in other languages as well as appropriate supplementary data sets collected by the students.

Learning Outcomes. Students will learn the fundamental principles and applications of natural language processing, transfer and active learning. They will also gain experience in designing machine learning frameworks, pipelines and products that efficiently (and meaningfully) incorporate human feedback.

Deliverables. The students will produce:

(1) A pipeline that:
   (a) takes the training data set
   (b) produces an appropriately visualized summary of the output of an intermediate classifier or the training data
(c) generates targeted queries for human input and modifies/trains a classifier based on this input

(2) A scientific report documenting, justifying and analyzing theoretical underpinning of their model, implementation as well as model evaluation. The evaluation of the model must include results showing performance comparison of the model with reasonable benchmark models.