The aim of this Master‘s Thesis will be to design a machine learning model that will be able to classify input data in the form of merchant URL into the main category and corresponding subcategories.

One of the challenges is to have a complete categorization of the whole available market, where each merchant can be classified into a main category based on what they offer, and then into more specific sub-categories, e.g., Eco, Zero Waste or Bike Sharing.

Although the available database has categorised merchants mainly from the CEE region, we want to create a global coverage, thus the demand for accurate automated categorization is increasing. Thanks to the collected data, it is possible to use modern machine learning approaches to process it.

The work assumes an existing dataset of a small number of manually categorised merchants. The student should build a merchant dataset based on the web information. On this partially labelled dataset he will test selection strategies for active learning, based on misclassification error and based on a Bayesian approach

Today, data is an essential part of our lives. From data we can relive the past, improve the present and predict the future. In our line of business, we strive to enrich payment transactions and use the client's payment history to precisely target her/his potential future needs in a convenient and categorizable form.

The aim of this work is to test active learning methods on the merchant categorization based on their web pages.

Although our available database now numbers over 100,000 manually categorized merchants mainly from the CEE region, we want to create a global coverage, thus the demand for accurate automated categorization is increasing. Thanks to such a large volume of data, it is possible to use modern machine learning approaches to process it.

As for the data and the NDA, can we publish the results?

Would be i.e. 500 categorised merchants be a sufficient sample to consider them as a result? We will be ok publishing some sample as a result but not all of them.

Your (Tapix) preference is to keep the majority of the data private? Yes.