**Capstone Project II: Proposal**

**Problem statement**

Many enterprises need to improve their products and services, either if they are sold online or in a more traditional way, need to rely on interaction with and feedback from their customers. Hence, the customer’s “voice” about the service or product might be of extensive value in order to re-design the way a company is marketing and servicing its products or it might provide important information for the development for new product series. Therefore, extracting and summarizing information of product reviews is key for many companies to remain competitive. However, many NLP applications focus on classifying reviews to be either positive or negative or do provide another kind of sentiment quantification. In turn, these measures are important to summarize the impression and feelings of customers over a wide range of products.

**Approach for possible solutions**

To classify reviews into sentiment patterns a Deep Learning approaches could get used, however, classifications might not provide the enough information for further improvements of products. Thus, besides classifying reviews, this project needs to develop a method to summarize product reviews. The approach to solve above targets is twofold. First, Deep Learning is applied to the solve the classification task, i.e. to predict the ratings a reviewer gives a product. In a second step, extractive summarization techniques get developed leveraging the encodings provided by the network. Thus, the model of this project should be able to 1) classify a review and 2) to provide a framework to generate summaries of certain reviews which could get selected by a user. Basically, step 1 is necessary because many companies might not have a clear functionality for customers to rate products on scale, moreover, in other applications – where the company relies, for example, on datasets like email messaging form clients, there are not any possibilities to give categorical ratings without the need of a human. Thus, step 1 provides an important output which then helps to segment reviews without the need of labelling the data beforehand in a company. In step 2 the Deep Learning embeddings for words could get used to generate summaries.

**Key Steps:**

**Data crawling:** generating random selection of Amazon reviews about apps and electronic entertainment products

**EDA:** analysing features of the dataset and taking care of interdependencies. Looking at text structure and perform cleaning steps.

**Work on model development:** possibilities to obtain embeddings via pretrained models; setting up DL classifier; ways to extract summaries. Taking care of imbalances and polarity of sentiment.

**Define test/train sets**

Taking care of imbalances and polarity of sentiment.

**Evaluation**