TRANSLATION TASK DRAFT

--- MAKE IT CONNECT TO CONTEXT ---

Our dataset primarily consists of reviews from Danish establishments, a significant portion of which were originally written in Danish. Consequently, our dataset contains content in both English and Danish languages. While this bilingual aspect might not pose a problem in some scenarios, it is widely acknowledged that managing multilingual data can be challenging for various downstream tasks. Therefore, we opted to standardize the entire dataset by translating all Danish reviews into English.

To accomplish the translation task, we chose to utilize an open-source translation model from the well-known NLP collaboration platform, 'Hugging Face.' This decision enabled us to seamlessly integrate the translation step into our data collection pipeline, leading to a more efficient and automated collection process.

Upon visual inspection, the model appears to perform well; however, to accurately assess its performance, an additional round of manual annotations is necessary, preferably by native Danish speakers. There exists exhaustive literature supporting different metrics to assess performance in translation models, such as BLEU (Bilingual Evaluation Understudy), NIST (from the National Institute of Standards and Technology), METEOR (Metric for Evaluation of Translation with Explicit Ordering), LEPOR (Length-Penalty, Precision, n-gram Position Difference Penalty and Recall), METEOR (Metric for Evaluation of Translation with Explicit ORdering), TER (Translation Edit Rate, and Perplexity.

We believe assessing the translation model for this particular application is crucial, we leave this task outside the scope of processing the dataset for now. --- WORK A BIT HERE ---

In addition to basic accuracy, several other challenges may arise, including:

* Reviews in languages other than Danish or English
* Reviews containing slang, mixed languages, and/or emojis.
* Misspellings, and so on.