Poster notes

This should be on the poster

# Our Project (Abstract / Idea)

* What was the idea

# Our Data

* What Data did we have
* Problems? And challenges with the data
* Preprocessing and cleaning

# Path 1: unsupervised

Steps

1. Cleaning and tokenizing the data
2. Calculate embeddings using T5-base (~40000 dims per sample)

Then we did

1. t-SNE to reduce dimensionality of embeddings from 40000 to 2 for plotting
2. k-means on the embeddings
   * Find k with elbow method (show plot on poster)
   * Show resulting clustering on poster
   * Wordcloud for one cluster

# Path 2: supervised

We annotated 10% of the data

As classification problem (as baseline and exploration)

* + Tf-idf embeddings
  + Svm
  + (put results on poster)
  + What else did you do here zoe? Just add it 😊

As Seq2Seq mapping

Idea:

* + Fine tune a T5-model with summerisation decoder to summarize to a single or a few words
  + If I (Max) get some results we would put them here but I do not currently (Tuesday noon) think this will happen
  + If we did not do the seq2seq mapping and do not have results for that we can either not mention this approach at all or (my preferred way) mention the approach as a potential idea