Hate Speech Detection, groupd 2: Keith Fleming, Kiran Git, John Lazenby
2 sources of data for hate speech: Twitter and Fox News
N grams, RNN, BERT
Tested the models and neither models worked well in terms of generalization.
Trained on fox and tested on Twitter and the other way around
Issues with tokenization
Issues with imbalanced data and use different methods in order to do so with the 3 different models
Difficulty of hate speech classification
How did you get these datasets ? The labels were already done from previous research where they did have speech. Crowdsourcing platforms to label the twitter data.
The outline was from Canvas

Image Generation for recipes, group 3
Big GAN architecture, training took about 5 days.
Objective: generate images from text in order to create the images from the recipe.
New about their approach: the type of data fed and how the images will evolve depending on the network input.
Evaluation: Inception Score of GANs + FID: measure the diversity of images in the training set.

Group 4: Amber Li, Emma Liu, Julia Wang, Kate Xu
Chatbot for 3 way conversations
Tried different models: Seq2Seq models with different embedding sizes and w or w/o attention
Corpuus based response by generative models, or can also work on IR-based models.
PersonA-Chat dataset.
University of Chicago template for the poster

Group 5: Task-oriented dialog: structured Seq2Seq Generation
Task oriented dialog dataset. Specific seq2 seq generation task. Very diverse dataset. Used Seq2Seq approaches for prediction. Map from the user input to dialog flow
Use CopyNet
Performed Structural ANalysis doing hidden state clustering and found a lot of structure in these hidden states. Looked at the activation of individual neurons in sentences.

Group 6: Semantically searching large codebases
Objecive: given a natural language query, find the most relevant code from an existing corpus
Code (--> python parser) + docstrings (--> BPE) + subtokenization to decrease the size of the vocab. Why did not they do matching only based on the docstring ? Unclear. Data: Codebase from Github
Lot of questions to question their approach: why not just similarity

Used Co-occurence metrics after having been through the Bag of Words models. This is a matching model, take a docstring. Objective is the representation of knowledge for the docstrings.

Group 13: Stock Return prediction using NLP on Unstructured Financial Text
Look for unconventional data sources in order to explain market behaviour, predict short term day-to-day price movements. Used EDGAR database.