Data science project description

50.038 Computational Data Science

Group: 3 members. Register here by October 5th: https://docs.google.com/spreadsheets/d/15UouWMgDAPdaLI_YJljrl02aZLF1RH0Q8MnmXy6BnIg/edit?usp=sharing. Groups can be across cohorts (as long as you find a time to all be there for the project presentations during cohort classes).

Initial presentation: Week 8

Final presentation: Week 13

Report: Week 12

Submission: Report in PDF form through eDimension

1 Objective

The main objective of this project is to equip and familiarize students with the necessary skills to successfully complete a data science project, including data collection and processing, data exploration and visualization, identifying and formulating problems, developing algorithms and models, designing experimental evaluations and discussing results, scientific writing and working in teams.

2 Project Overview

For this project, students select a data science problem, such as those listed in the examplar section, or a newly proposed one. Based on their problem description, students then find (multiple) datasets, and implement innovative multi-modal solutions. Students will form a team comprising of exactly three members, and are expected to deliver two presentations and submit a final report which compares multiple approaches. Code should be submitted on github or bitbucket and a link should be put in the report. Details about the presentations and report are provided in the following sections.

3 Initial and Final Presentations

Two presentations are to be delivered for this project, and each team will be allocated 10 min (time to be confirmed based on number of groups) for each presentation. Details of the presentations are:

- For the initial check-off (Week 8), the teams should describe the type of dataset selected or collected, the problem they aim to address, data visualisation, and a preliminary naive model implemented based on one dataset.
- For the final presentation (Week 13), the teams should briefly describe their datasets and problem, and elaborate more on the algorithms used, the type of evaluation, results obtained and their implications.

4 Final Report and Required Sections

Teams are expected to submit a report of max. 6,000 words, comprising the following sections. The report should be written in the style of a scientific conference or journal paper using the scientific typesetting system LaTeX (e.g. `overleaf.com`).

- **Dataset and Collection:** Describe the *type of datasets* being used and the *source* where it is obtained from. If applicable, mention any data collection methodology or APIs used. Students are free to select existing datasets, or collect their own datasets.
- **Data Pre-processing:** Describe any pre-processing or data cleaning steps applied on the dataset.
- **Problem and Algorithm/Model:** Motivate and describe the problem that this project aims to address. Some examples of problems are predicting whether a stock will rise or fall over X days, or predicting the volume of stock activities on a specific day. Also, describe the algorithm or model that is used for solving the earlier defined problem, and briefly situate it in its context by citing related work (tip use bibtex in combination with Google Scholar).
- **Evaluation Methodology:** Describe the methodology that is used to evaluate the effectiveness of the proposed algorithm. This section should cover how the dataset is being used in training and evaluation, and the types of evaluation metrics used. Can you compare with existing models/work?
- **Results and Discussion:** Describe the results obtained and discuss the implications of these results or any other main findings observed during this project.

Apart from the sections listed above, teams are also welcome to include any other sections they deem necessary such as a brief literature review.

Since this project will be written as a scientific paper, good projects can be submitted to a venue in co-authorship with their respective supervisors if the model contributes to new knowledge. The professors will be available for supervision during the lab times and via email/Discord. You may also work with dedicated class-external supervisors.

5 Deliverables and Grading

This project is worth a total of 40 marks. The deliverables and grading of this project is further divided into the following components:

- An initial presentation as described in Section 3. This component serves to provide feedback and check that there is progress. It is worth 5 marks.
- A final presentation as described in Section 3. This component is worth 10 marks
- A final report as described in Section 4. This component is worth 20 marks.
- Peer review. 5 free points if you work together well in team. Each team member grades the others.

The initial and final presentations will be conducted during the lectures of the respective weeks. Detailed schedules will be provided nearer to the presentations. The report is to be submitted in PDF format via eDimension.

6 Project exemplar

You are free to (and in fact encouraged!) to propose your own free topic. The below only serve as examples of possible projects.

Multimodal sentiment analysis Here, the task is to detect sentiment of a person speaking in a video. Students are expected to utlize facial expressions, audio and textual features in the classification model. The features are given in this link https://github.com/soujanyaporia/multimodal-sentiment-analysis but it is encouraged that the students should write codes to extract features from the raw videos. Finally, these features should be fused for sentiment prediction. Fusion can be performed in several ways. Concatenation is the simplest method of future fusion.

Multimodal sarcasm detection Similar to multimodal sentiment analysis task. However, in this setting, instead of detecting sentiment, students will detect sarcasm in the videos. Dataset can be downloaded from this link - https://github.com/soujanyaporia/MUStARD.

Emotion recognition in conversation This is a challenging yet popular task where the goal is to classify emotion of each utterance in a conversation (Figure 1). Students can visit this link - https://github.com/declare-lab/conv-emotion to download the dataset to experiment.

Digital asset price prediction traditional stock markets are highly sensitive to public opinion. This is even more so the case for cryptocurrencies. An analysis of different modalities (news, price data, market data), will be interesting to build either indicators or predictive models. One approach would be to perform clustering per time periods, so as to determine the market type. It can then be tested if this allows us to build better predictive models. (Advisor/more info: Prof. D. Herremans)

NLP for stock market indicators there are lots of models out there to predict the stock market. The challenge, however, is to extract meaningful information out of the vast amount of data available to detect 'useful' information:

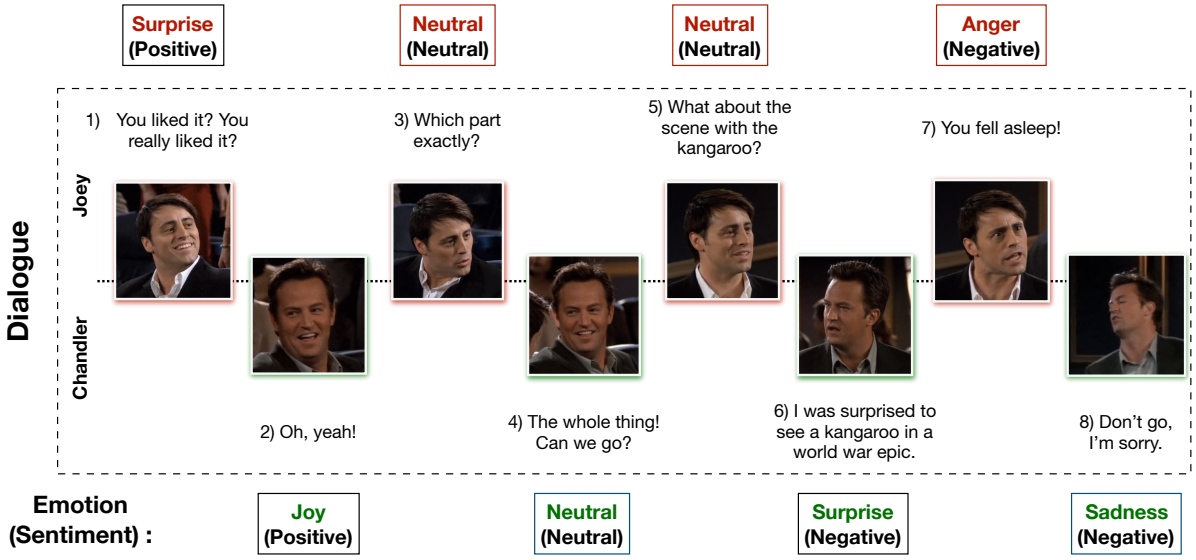

Figure 1: The task of emotion recognition in conversations.

- predict when a market crash will be happening, or predict when we move from a bull to a bear market based on information from Twitter, such as mentions of taper to predict US 10YR treasury daily yields; or Tweets from top crypto influencers to see if it improves a basic price model.

In some markets, a basic trend strategy (turtle) would be good, in others, a reversal strategy works better. Can we predict which market we are in? We can create 'indicators' for this task. (Advisor/more info: Prof. D. Herremans)

Music recommendation for branding About a decade ago there was study that showed that people buy more wine if a drink store played French music. When they played German music, they sold more beer. Music is a powerful influencer for purchase behaviour. In the project, you can collect a dataset of keywords that fit music, and build a predictive/recommendation system. Which music will make this shampoo commercial feel luxurious?. A possible collaboration with the Dutch company Sounders Music is possible. (Advisor/more info: Prof. D. Herremans)

Automatic Speech Recognition (ASR) related topics For beginners: Familiarize yourselves with a simple ASR model which takes Melspectrogram as input and predicts the phonemic sequence using the TIMIT dataset. AMAAI has a pytorch template to kickstart your project. https://github.com/KinWaiCheuk/pytorch_template Once you are familiar with your first ASR, you can start exploring any (one or more) of the following ideas:

- Experiment with different Melspectrogram parameters, such as n_fft, n_mels. Then report which parameters give the best or worst results, and their respective training time.
- Experiment with other spectrogram types such as STFT and CQT. Which spectrogram type is better for speech?

- Modify the model so that it predicts English characters or English words instead of the phonetics. Does the model predict characters or words better than phonetics? What is the reason for it?
- After training a good model for phonetic predictions, can you apply the same model to other tasks in the same dataset such as speaker recognition, or dialect classification? You will need to change the classifier for other tasks and freeze the weight for the feature extractor part.
- Explore different model architectures to make the prediction accuracy as high as possible
- Train a ASR model on other languages. For example, German or Polish in Multilingual LibriSpeech (MLS) (Or any other languages that you like). How does your model performance in various languages? What are the difficulties for training a ASR on other languages and how do you overcome them? You might want to check what datasets are available in the AMAAI lab first https://github.com/KinWaiCheuk/AudioLoader

Contact kinwai_cheuk@mymail.sutd.edu.sg for more info.

Automatic Music Transcription (AMT) related topics For beginners: Familiarize yourselves with a simple AMT model using the MAPS dataset AMAAI has a pytorch template to kickstart your project. https://github.com/KinWaiCheuk/pytorch_template Once you are familiar with your first AMT, you can start exploring any (one or more) of the following ideas:

- Experiment with different Melspectrogram parameters, such as n_fft, n_mels. Then report which parameters gives the best or worst results, and their respective training time.
- Experiment with other spectrogram types such as STFT and CQT. Which spectrogram type is better for AMT?
- Modify the model so that it predicts also the onsets of the musical notes. Does this extra task help with AMT performance?
- After training a good model on piano music (MAPS dataset), does the same model perform well for other musical instruments?
- Explore different model architectures to make the prediction accuracy as high as possible
- Train a AMT model that predicts not only the pitch, but also the musical instrument of that pitch. You need to use other datasets for this direction, such as MusicNet. You might want to check what datasets are available in the AMAAI lab first https://github.com/KinWaiCheuk/AudioLoader

Contact kinwai_cheuk@mymail.sutd.edu.sg for more info.

Deep audio representation learning – multitask Building on some of the resources from above: can we create clever embeddings for audio that are powerful to solve a multitude of tasks. Based on the HEAR challenge in NeurIPS, co-organized by Prof. Herremans. (advisors: prof. DH and Cheuk Kin Wai)
https://neuralaudio.ai/hear2021-holistic-evaluation-of-audio-representations.html