Data science project description

50.038 Computational Data Science

Group: 3 members. Register here by the end of week 3: https://bit.ly/2mg9ZpR

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. 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. 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 can also be written as a scientific conference or journal paper.

- **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.
- **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.
- **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.

Tip: you can use Overleaf for easy collaborative writing in LaTeX.

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 25 marks.

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

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/SenticNet/conv-emotion to download the dataset to experiment.

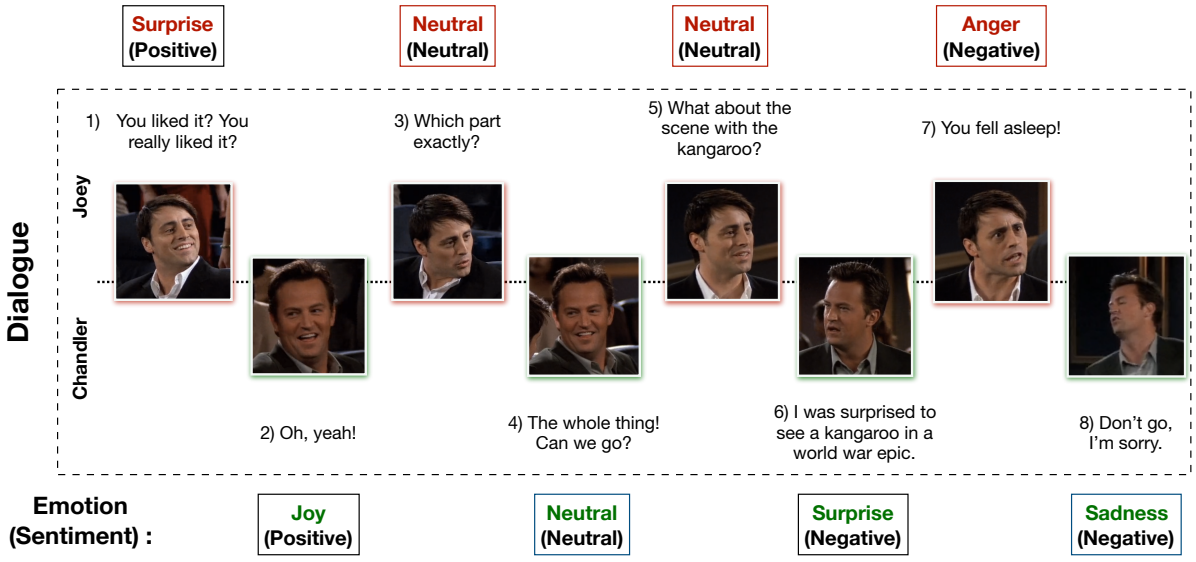

Figure 1: The task of emotion recognition in conversations.

Audio spoofing detection Automatic speaker verification, such as every other biometric system, is vulnerable to spoofing attacks. Using only a few minutes of recorded voice from a genuine client of a speaker verification system, attackers can develop a variety of spoofing attacks that might trick such systems. Detecting these attacks using the audio cues present in the recordings is an important challenge. Dataset can be found here: https://www.idiap.ch/dataset/avspoof.

Speaker recognition Related to the above topic is speaker recognition, which is a competition of the upcoming IEEE ASRU conference in Singapore mid-December: https://www.nist.gov/itl/iad/mig/nist-2019-speaker-recognition-evaluation

Music emotion prediction Emotion prediction (in terms of valence / arousal or key words) from music. In addition to DEAM dataset, please ask TA Raven for access to a custom collected dataset.

Bitcoin 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 (tweets, news, price data, market data), will be interesting to build either indicators or predictive models.

Music transcription - Either in terms of chords or polyphonic music. This is an important, and very difficult problem. At each moment in time, we predict if there is a note onset and which pitch height is being played. Datasets include musicnet, MAPS and the huge Maestro

Personality detection Predict personality type of a person from his/her writing. Download the dataset from here - https://github.com/SenticNet/personality-detection

7 Awards

EPS Computer Systems Awards price – A total of 2,500 SGD will be awarded to the top three projects!!!

Tip: It is recommended that the students write their project as a technical report and upload to the arxiv. Something similar to this is often practiced at Stanford e.g., check out this project - https://github.com/alpv95/MemeProject and the corresponding paper - https://arxiv.org/abs/1806.04510. It was a project assignment of the cs224n class at Stanford CS which got so much attention from the press and deep learning community.