COMP 4332 / RMBI 4310
Big Data Mining (Spring 2019)

Project 1: Sentiment Classification

TA: Xin Liu (xliucr@connect.ust.hk)

Sentiment Analysis

- Generally modeled as **classification** or regression task
  - predict a binary or ordinal label

Sentiment Analysis

- Simplest task:
  - Is the attitude of this text positive or negative?

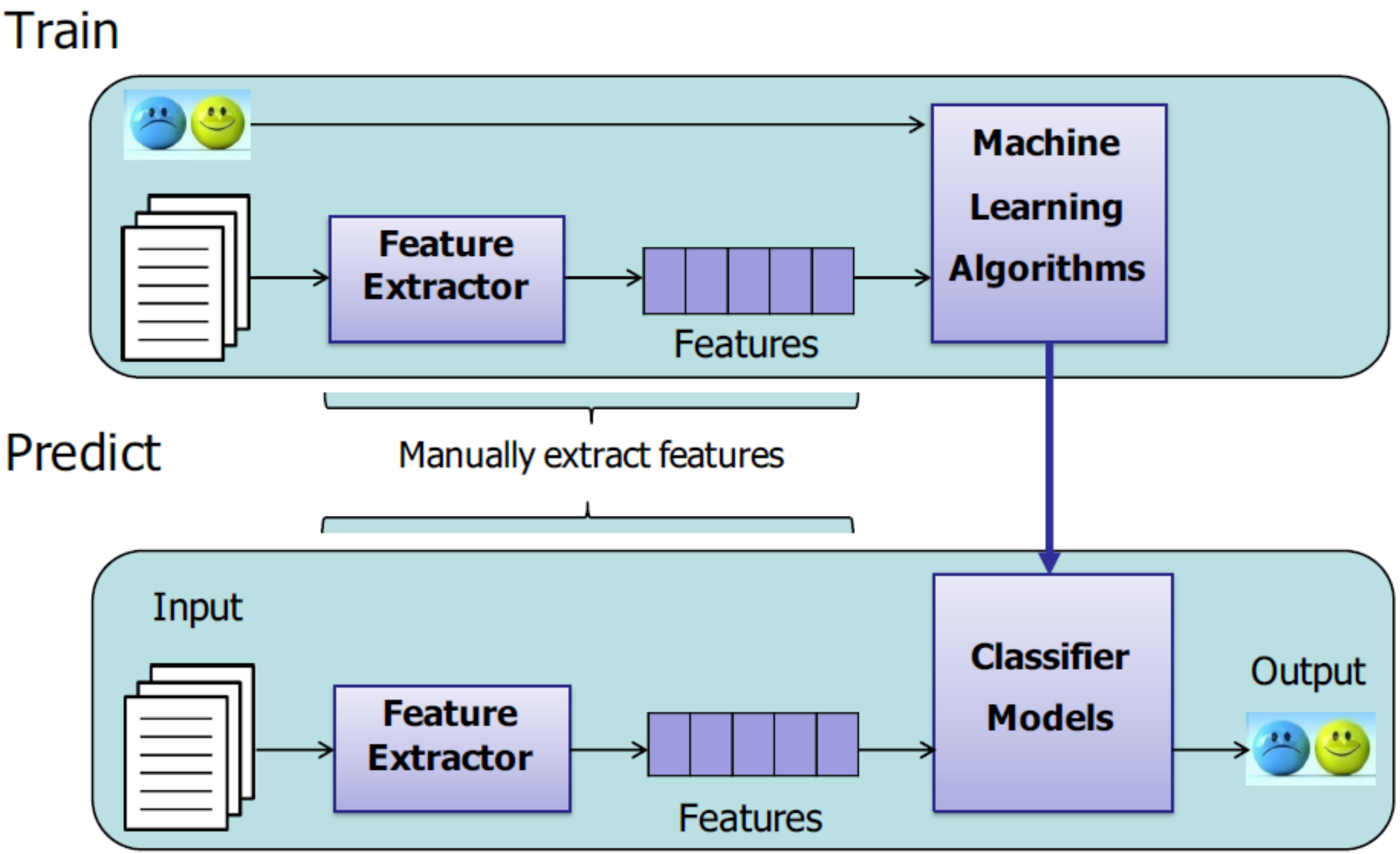
- **More complex:**
  - **Rank the attitude of this text from 1 to 5**
  - (3/5) The room was clean and everything worked fine – even the water pressure
  - (1/5) …the worst hotel I had ever stayed at ...

- Advanced:
  - Detect the target, source, or complex attitude types

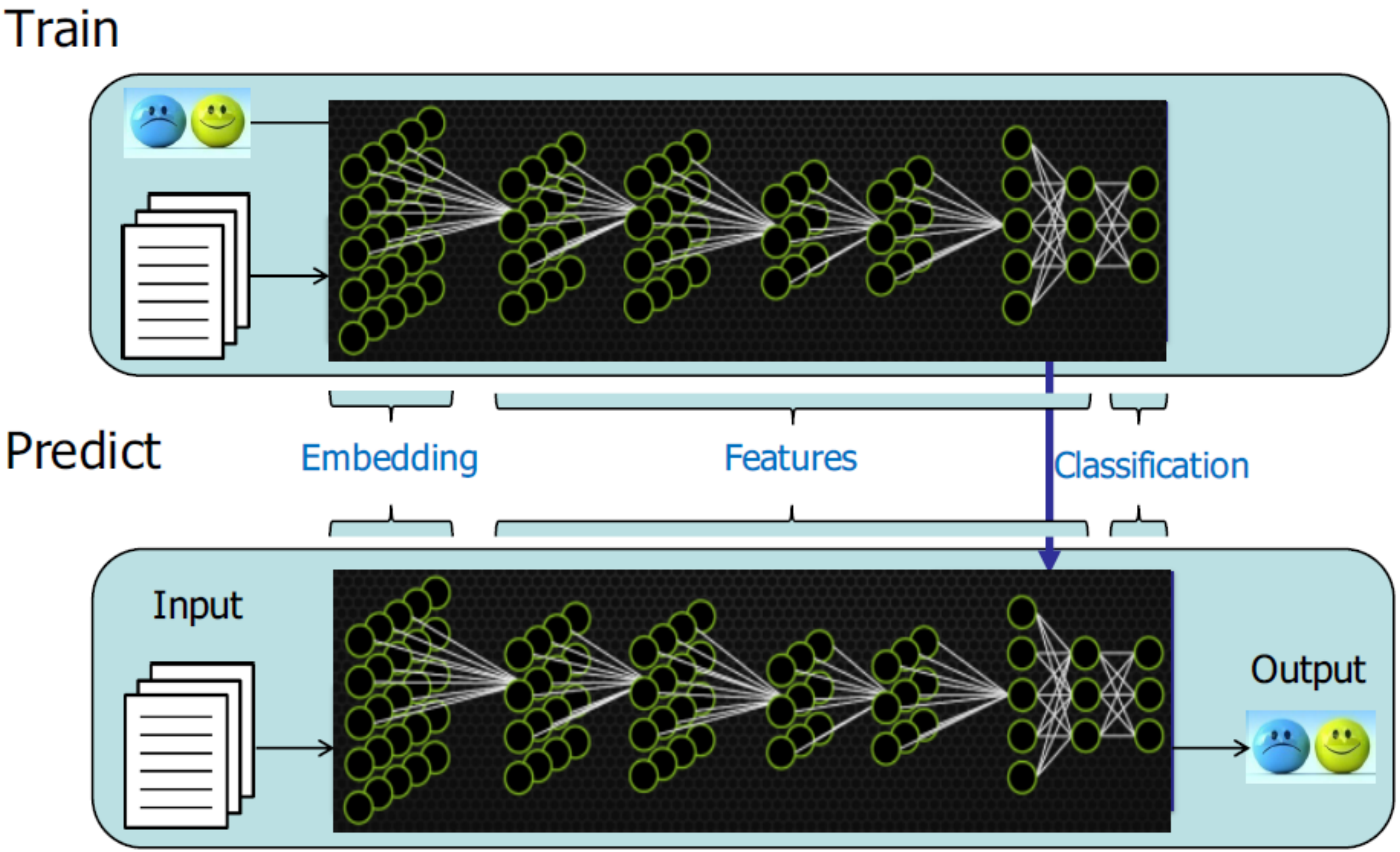
Basic Pipeline

- Tokenization: Split document into words (tokens)
- Feature Extraction: Find useful features
- Classification: Classification via different classifiers

General Pipeline



End2end Pipeline



Basic Pipeline

- Tokenization: Split document into words (tokens)
- Feature Extraction: Find useful features
- Classification: Classification via different classifiers

Tokenization

- NLTK
  - tokens = nltk.word_tokenize(text)
- spaCy
  - nlp = spacy.load('en_core_web_sm')
  - tokens = [t.text for t in nlp(text)]
- CoreNLP
  - https://stanfordnlp.github.io/CoreNLP/other-languages.html#python

Feature Extraction

- word frequency or word occurrence
    - This room is clean.
    - `[0,0,1,1,0,1,0,0,1,0,1]`
- random projection
    - In Tutorial 1
- word embedding
    - cbow, skip-gram, GloVe, fasttext
- `contextualized word representation`
    - ELMo, BERT, GPT, GPT-2

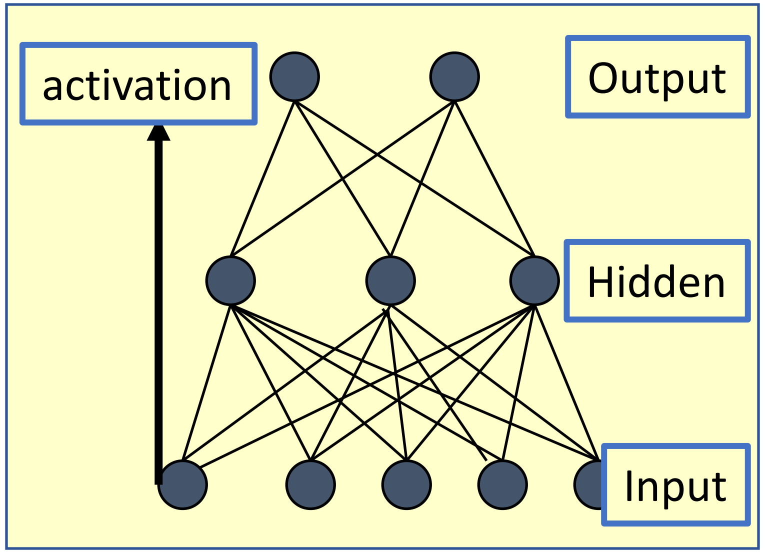
Feature Extraction

- user information
  - nationality
  - age
- date
  - weekday or weekend
  - holiday?
- hotel rating
  - Hilton Hotel
  - Youth Hostel
- data mining

Classification

- Naïve Bayes
- Logistic Regression
- Support Vector Machine
- **Deep Learning**

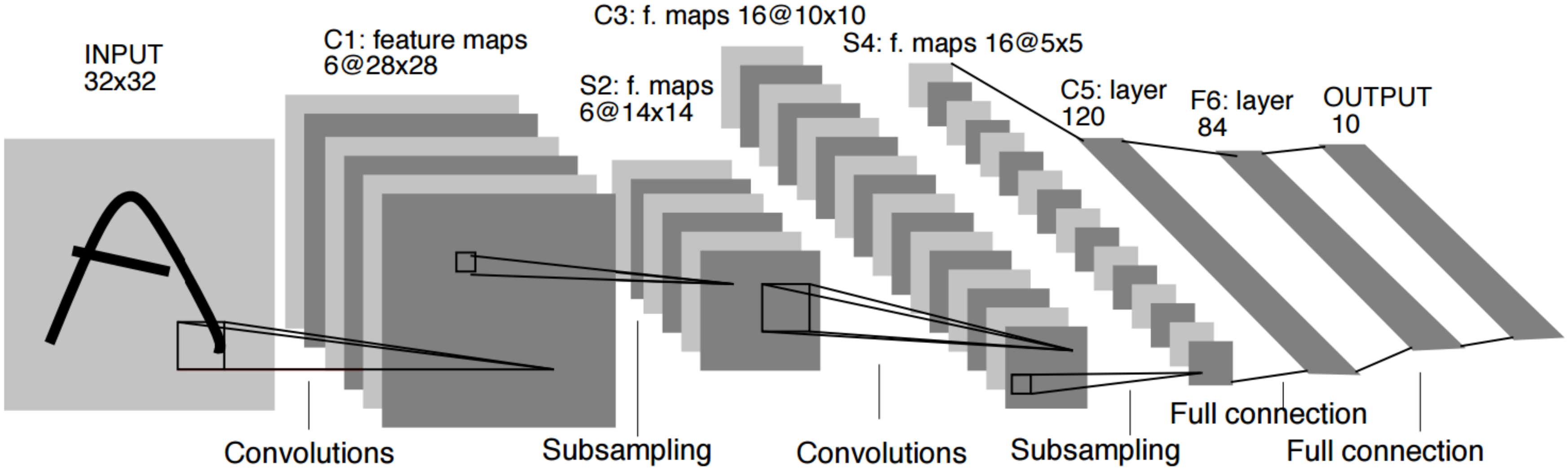
Multi Layer Perceptron



In Tutorial 1

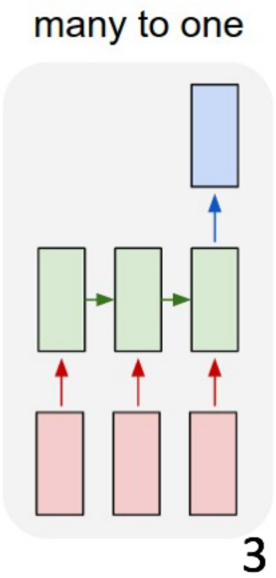
Demo: http://playground.tensorflow.org/

CNN



In Tutorial 2

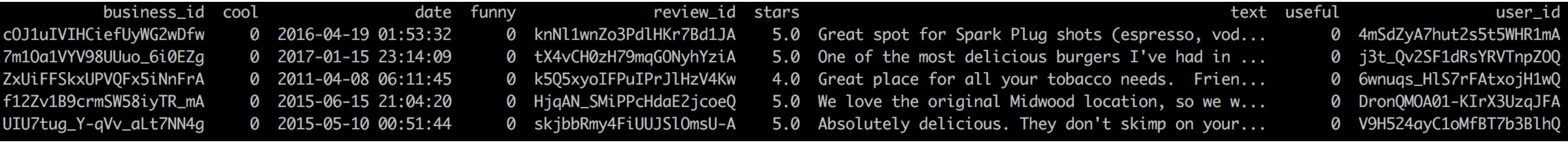
RNN



In Tutorial 3

Dataset

- training data: 100000 reviews
- validation data: 10000 reviews
- test data: 10000 reviews
- rating: 1-5
- given features: `business_id, cool, date, funny, review_id, stars, text, useful, user_id`



```
           business_id  cool                 date  funny               review_id  stars                                               text  useful                 user_id
cOJ1uIVIHCiefUyWG2wDfw     0  2016-04-19 01:53:32      0  knNl1wnZo3PdlHKr7Bd1JA    5.0  Great spot for Spark Plug shots (espresso, vod...       0  4mSdZyA7hut2s5t5WHR1mA
7m1Oa1VYV98UUuo_6i0EZg     0  2017-01-15 23:14:09      0  tX4vCH0zH79mqGONyhYziA    5.0  One of the most delicious burgers I've had in ...       0  j3t_Qv2SF1dRsYRVTnpZOQ
ZxUiFFSkxUPVQFx5iNnFrA     0  2011-04-08 06:11:45      0  k5Q5xyoIFPuIPrJlHzV4Kw    4.0  Great place for all your tobacco needs.  Frien...       0  6wnuqs_HlS7rFAtxojH1wQ
f12Zv1B9crmSW58iyTR_mA     0  2015-06-15 21:04:20      0  HjqAN_SMiPPcHdaE2jcoeQ    5.0  We love the original Midwood location, so we w...       0  DronQMOA01-KIrX3UzqJFA
UIU7tug_Y-qVv_aLt7NN4g     0  2015-05-10 00:51:44      0  skjbbRmy4FiUUJSlOmsU-A    5.0  Absolutely delicious. They don't skimp on your...       0  V9H524ayC1oMfBT7b3BlhQ
```

Evaluation

- accuracy on **test data**

Submission

- predictions on **test data**
- report
- code
- DDL: March 17, 2019
- Submission: Each team Leader is required to submit the groupNO.zip file that contains pre.csv, the report, and your team's code on canvas.
- we will check your report with your code and the accuracy.

Grading Rule

Grade	Classifier (80%)	Report (20%)	Remark
50%	example code in tutorials or in Project 1 without any modification	submission	
60%	an easy baseline that most students can outperform	algorithm you used	release the **accuracy on validation data** on March 3
80%	a competitive baseline that about half students can surpass	detailed explanation	release the **accuracy on validation data** on March 3
90%	a very competitive baseline without any special mechanism	detailed explanation and analysis	release the **accuracy on validation data** on March 10
100%	a very competitive baseline with at least one mechanism	excellent ideas	release the **accuracy on validation data** on March 10

Thank You