COMP 4332 / RMBI 4310
Big Data Mining (Spring 2023)

Project 1: Sentiment Analysis

TA: Van Quyet DO (vqdo@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

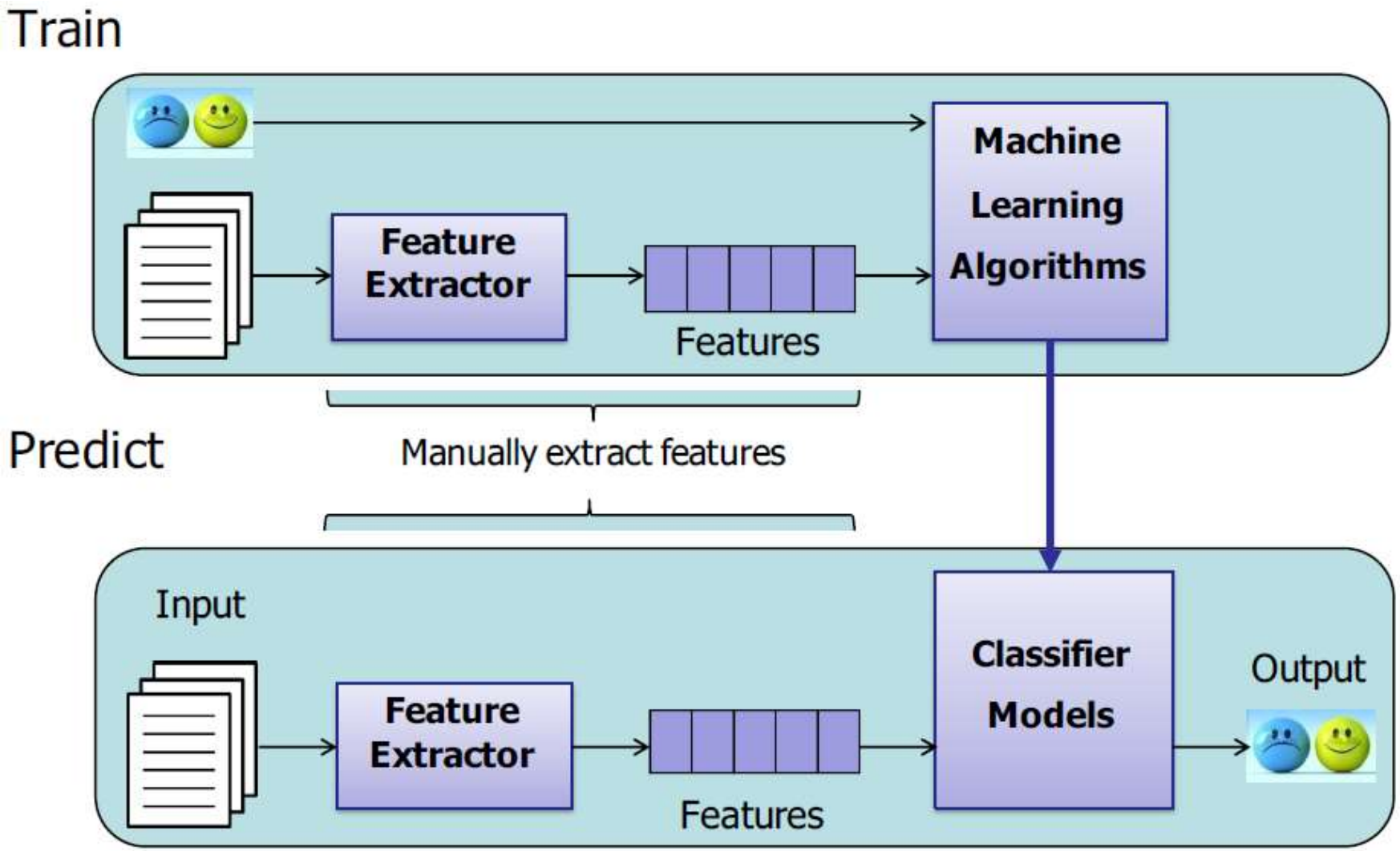
Pipeline

- **Data Loader:** Load data from disks
- **Feature Extraction:** Find useful features
- **Learning:** Classification via different classifiers

For more information and examples, please refer to instruction.ipynb

If you want to quickly get familiar with the whole pipeline, please refer to general_pipeline.ipynb

Pipeline


Train
Machine Learning Algorithms
Feature Extractor
Features
Manually extract features
Predict
Input
Classifier Models
Output
Feature Extractor
Features

Feature Extraction

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

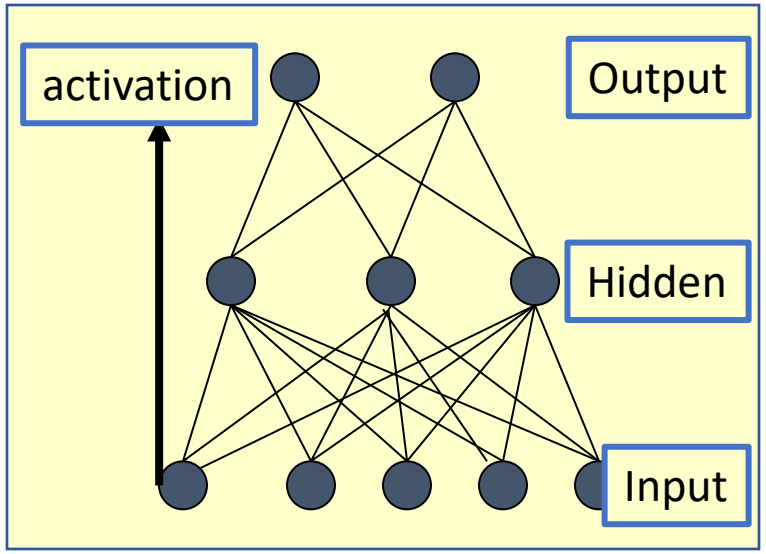
Feature Extraction

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

Classification

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

Multi Layer Perceptron



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

CNN

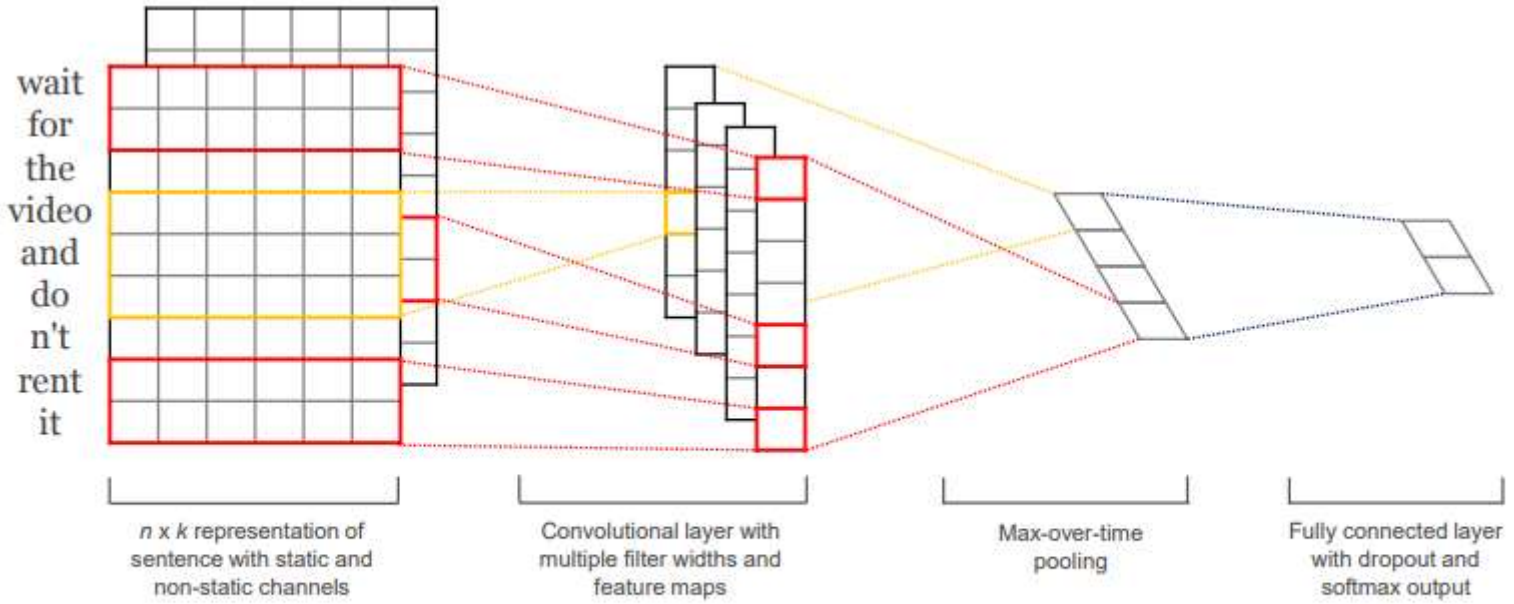
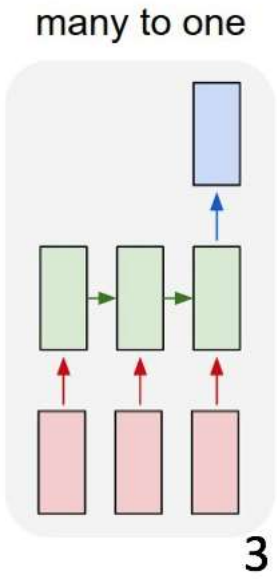


Figure 1: Model architecture with two channels for an example sentence.

RNN



Dataset

- training data: 18000 reviews
- validation data: 2000 reviews
- test data: 4000 reviews
- stars (integer, to be predicted): 1-5
- given features: `business_id, cool, date, funny, review_id, text, useful, user_id`

business_id	cool	date	funny	review_id	stars	text	useful	user_id
JYgoAQHdJWKPArQDvBEBng	0	2016-03-30 21:40:34	0	JlUcJily24pw5jStCLtavg	5	Best Sunday buffet in the two cities of Charlo...	1	SxLNrLxHm0aEw-kLrbPLew
AASa5G_OHCxGQ0tbjT_2tw	0	2018-08-30 02:50:46	0	xJBrURol6Tm7PCmytXUMyg	4	My friends and I decided to check out this pla...	0	aW22TlXwhkUUqBYFG7fbTA
Z2xuK4BbrD0Qr9dAs7oTVw	0	2017-02-20	0	99wD_l4D6Sw7Kesaq9GPhg	5	This is definitely New York Chinese food! The ...	4	SerdK2DW_2R7z1b9WU97fg
YAMXCiebYV49_B8lDAaLxA	0	2017-04-06 21:56:41	0	AmNFz9svFx9QCSZsUs8JTA	3	Beaucoup moins de choix que son voisin d'en fa...	0	pf4nr7_PlMrHjbmQYbEFcQ
ynvp3qvt3xc321dLKFxpgA	0	2012-03-09 19:30:47	0	nwnlKZN2MWhyL3aKUqY7ig	2	Location is nice, but it is the typical blah H...	2	Mf5TQEqn59k_TapTpfjYdA

Evaluation

- Macro F1 on **test data**
  - You would not get the test labels, but you can use the provided validation set to estimate your model's performance

Important dates

Four weeks in total

- [March 16, 2023] Project starts
- [March 23, 2023] TA will release the validation performance of a weak baseline
- [March 30, 2023] TA will release the validation performance of a strong baseline
- [April 13, 2023, 23:59] Submission deadline

Submission

- Predictions file pred.csv on **test data** (before submitting your test predictions, please make sure you can successfully evaluate your validation predictions on the validation data with the help of evaluate.py)
- Report (1~2 pages)
- Code (Frameworks and even programming languages are not restricted.)
- DDL: April 13, 2023
- Submission: Each **team leader** is required to submit the groupNo.zip file that contains pred.csv, the report, and your team's code on Canvas.
- We will check your report with your code and the model performance (in terms of macro F1) on the test set.

Grading Rule

Grade	Classifier (80%)	Report (20%)
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
80%	a competitive baseline that about half students can surpass	detailed explanation
90%	a very competitive baseline without any special mechanism	detailed explanation and analysis, such as explorative data analysis, hyperparameters and ablation studies
100%	a very competitive baseline with at least one mechanism	excellent ideas, detailed explanation and solid analysis

Thank You and Good Luck