COMP 4332 / RMBI 4310
Big Data Mining (Spring 2024)

Project 1: Sentiment Analysis

TA: Zhaowei Wang (zwanggy@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 ...

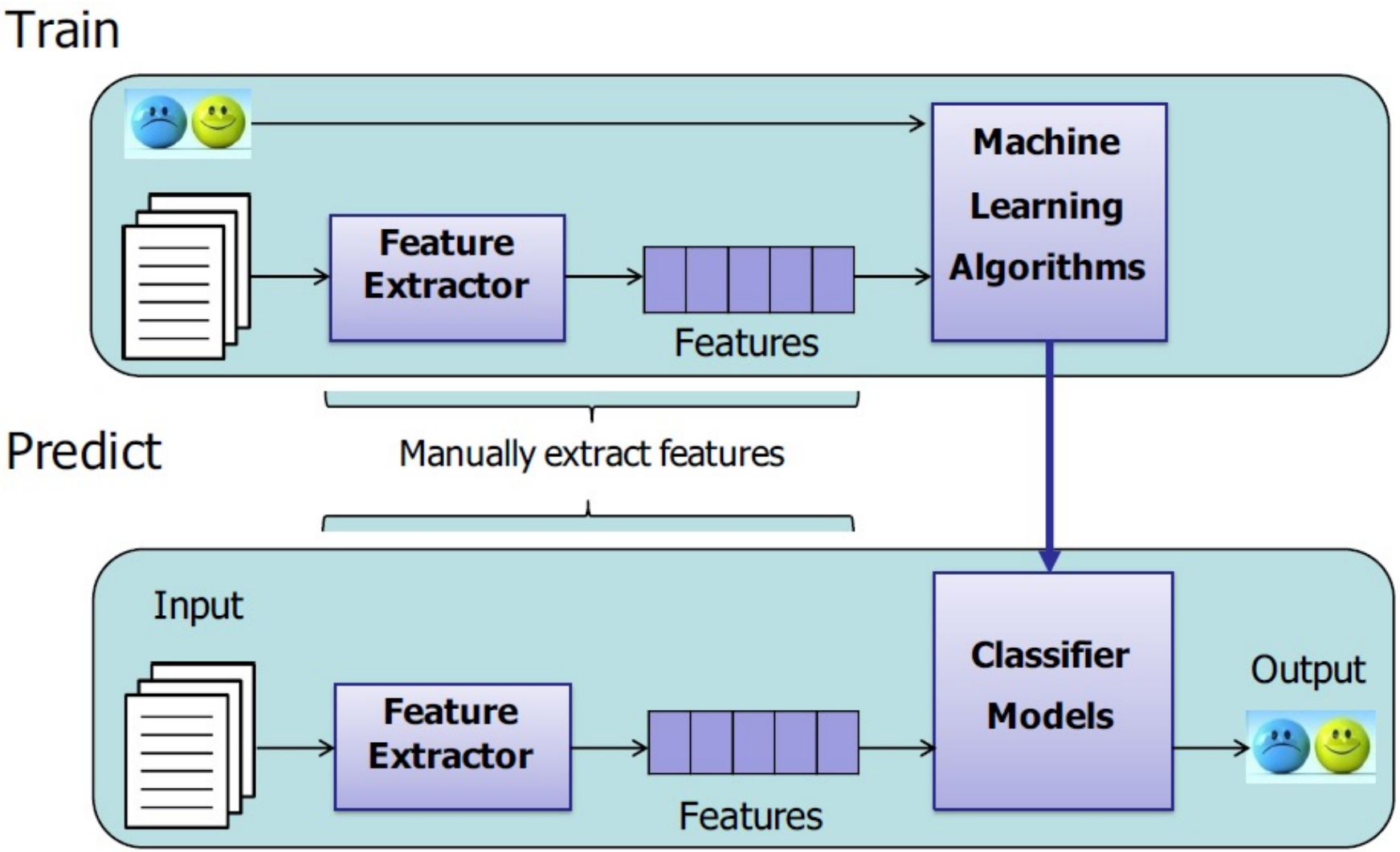
- **Advanced:**
  - Detect the target, source, or complex attitude types

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

Pipeline


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

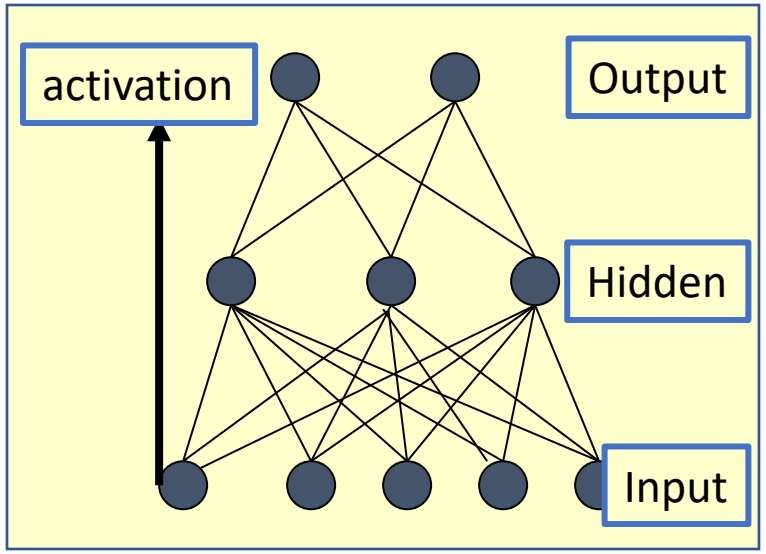
Feature Extraction

- **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

Classification

- Naïve Bayes
- Logistic Regression
- Support Vector Machine
- **Deep Learning: RNN, CNN, BERT, GPT**

Multi Layer Perceptron



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

CNN

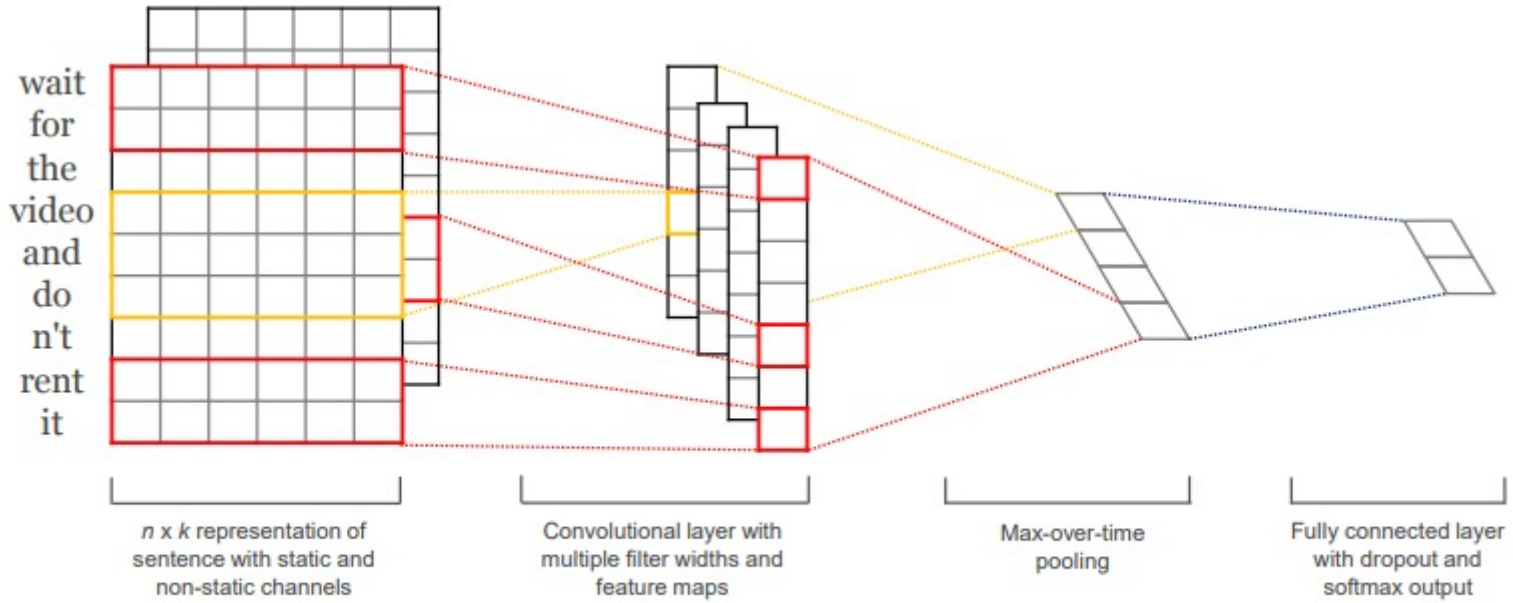
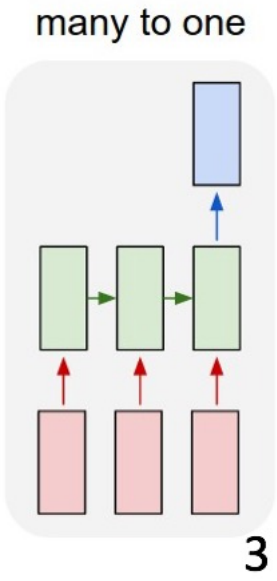


Figure 1: Model architecture with two channels for an example sentence.

RNN



Dataset

- Training data: 18000 reviews
- Validation data: 2000 reviews
- Test data: 4000 reviews
- Label (integer, to be predicted): 1-5 (the larger, the more positive)
- Text: We need to predict whether a piece of text is positive
- Format: CSV format (each column separated by commas)

```
id,text,label
AWYITJ9IUMYKH_524,"Two Wolfgang Petersen directed films together in one package is all you could want, w

""Air Force One"" in particular is excellent.",5
ANDZLSFNII2EW_12768,"For fans of the series and the movies
this film is a must. It continues The
Wrath of Khan but not at the same level
of interest. Anyway is a good movie",4
```

Evaluation

- Accuracy 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

- [March 16, 2024] Project starts
- [March 23, 2024] TA will release the validation performance of the easy and hard baselines
- [April 6, 2024, 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 6, 2024
- 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 **Accuracy**) on the test set.

Grading Rule

Grade	Classifier (80%)	Report (20%)
50%	Example code in tutorials or in Project 1 without any modification	Submission
75%	A method that can outperform the easy baseline	Algorithm you used
95%	A method that can outperform the hard baseline	Detailed explanation and analysis, such as explorative data analysis, hyperparameters and ablation studies
100%	A method that can outperform the hard baseline with at least one excellent idea	Excellent idea, detailed explanation and solid analysis

Thank You and Good Luck