

COMP 4332 / RMBI 4310

Big Data Mining (Spring 2021)

Project 1: Sentiment Analysis

TA: Zihao WANG (zwanggc@cse.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

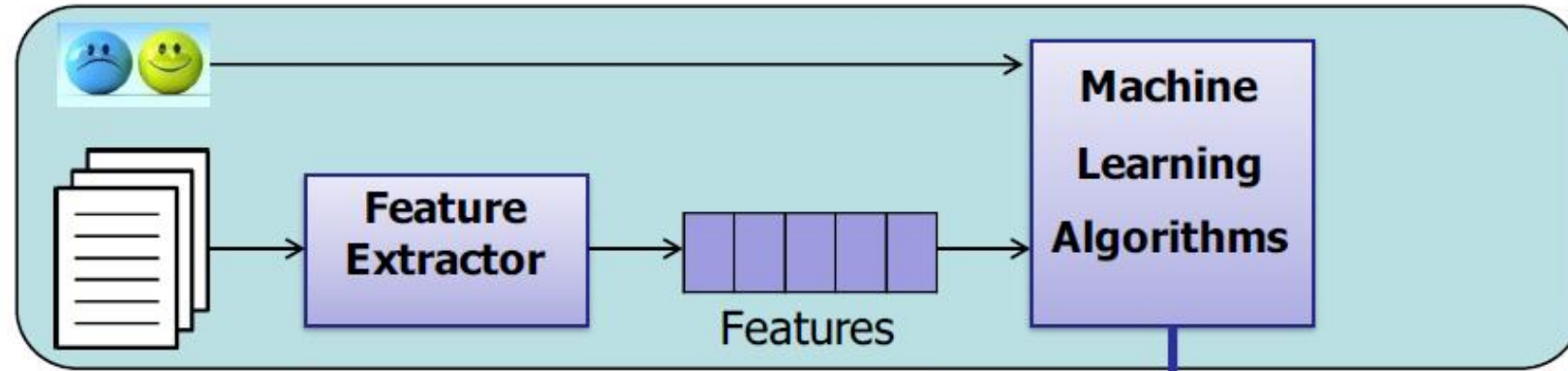
Pipeline

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

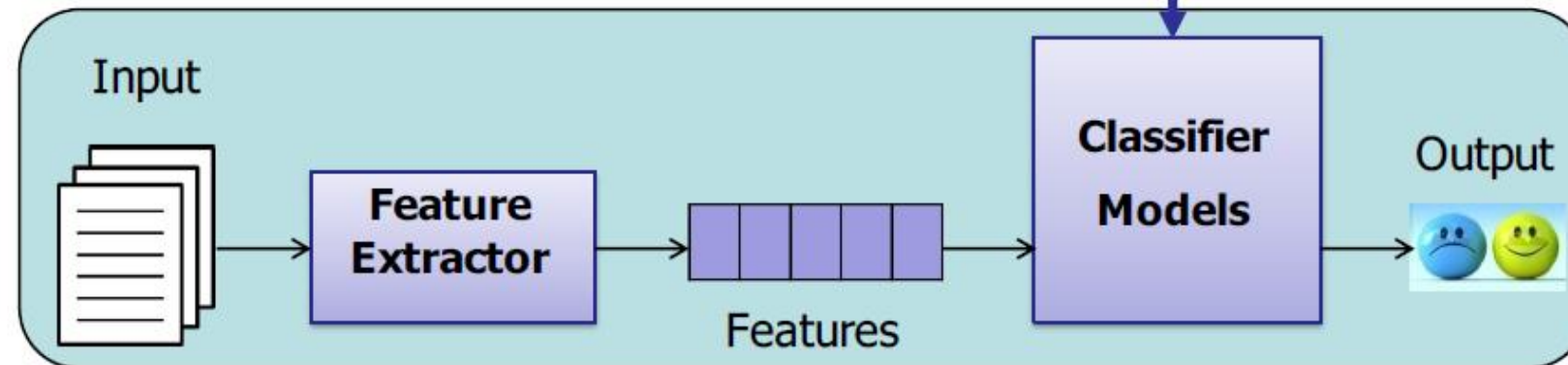
For more information and examples, please refer to `demo.ipynb`

Pipeline

Train



Predict



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

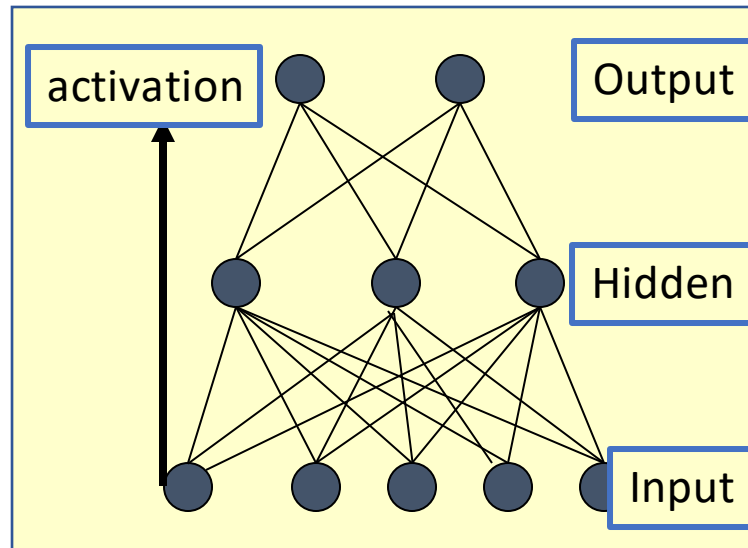
- Hilton Hotel
- Youth Hostel

- data mining

Classification

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

Multi Layer Perceptron



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

CNN

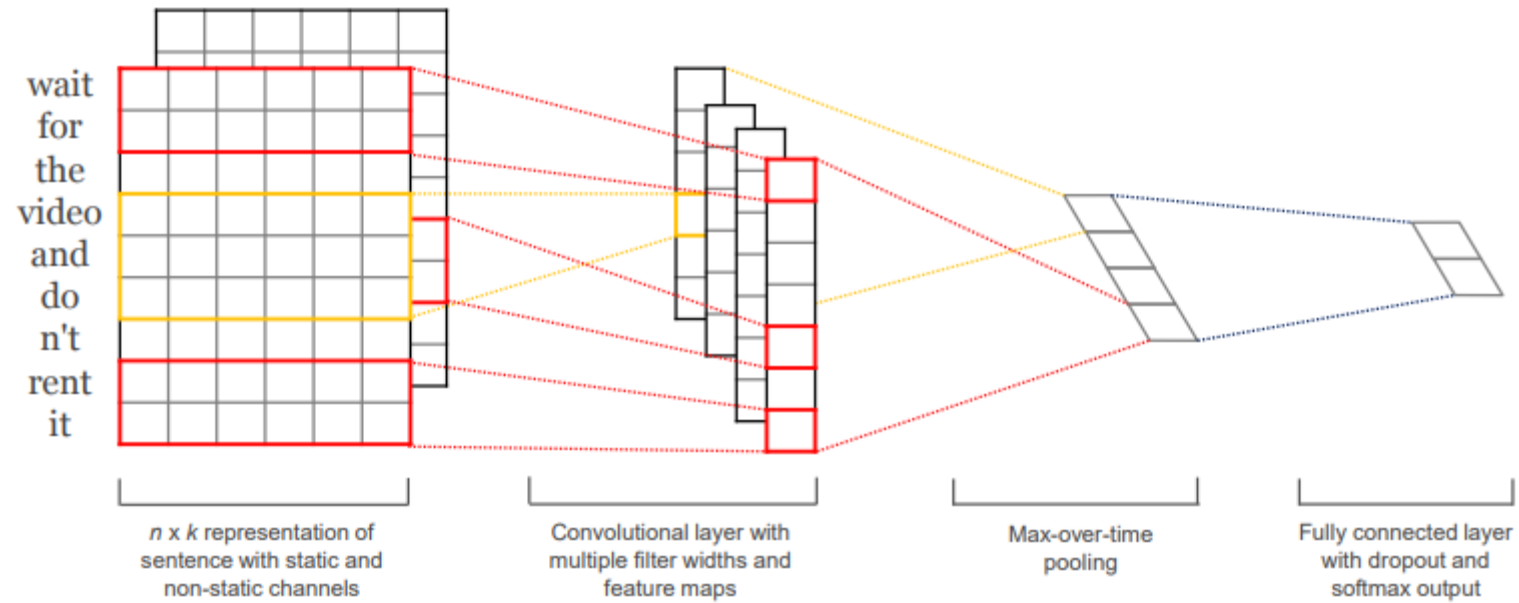
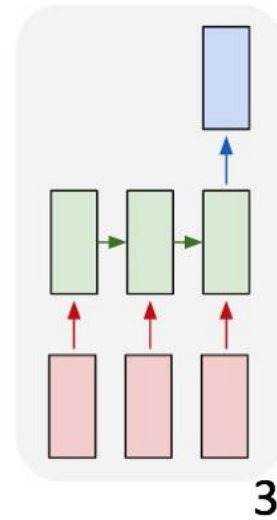


Figure 1: Model architecture with two channels for an example sentence.

RNN

many to one



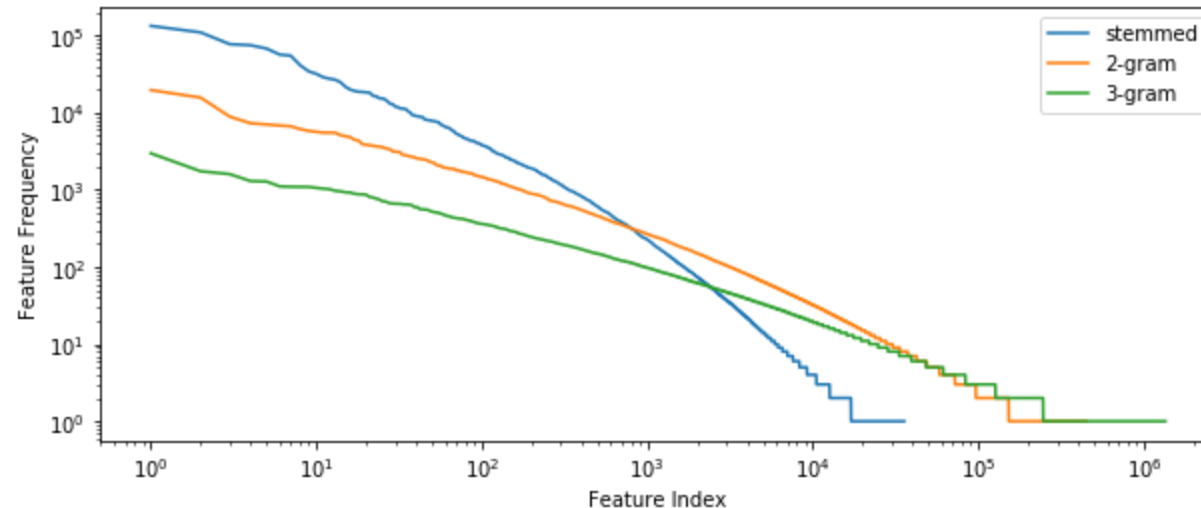
Dataset

- training data: 20000 reviews
- validation data: 2000 reviews
- test data: 2000 reviews
- stars: 1.0-5.0
- given features: `business_id`, `cool`, `date`, `funny`, `review_id`, `text`, `useful`, `user_id`

<code>business_id</code>	<code>cool</code>	<code>date</code>	<code>funny</code>	<code>review_id</code>	<code>stars</code>	<code>text</code>	<code>useful</code>	<code>user_id</code>
c0J1uIVIHCiefUyWG2wDfw	0	2016-04-19 01:53:32	0	knNl1wnZo3PdLHKr7Bd1JA	5.0	Great spot for Spark Plug shots (espresso, vod...	0	4mSdZyA7hut2s5t5WHR1mA
7m10a1VYV98UUuo_6i0EZg	0	2017-01-15 23:14:09	0	tX4vCH0zH79mqGONyhYziA	5.0	One of the most delicious burgers I've had in ...	0	j3t_Qv2SF1dRsYRVtnpZ0Q
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	DronQM0A01-KIrX3UzaJFA
UIU7tug_Y-qVv_aLt7NN4g	0	2015-05-10 00:51:44	0	skjbbRmy4FiUUJS10msU-A	5.0	Absolutely delicious. They don't skimp on your...	0	V9H524ayC1oMfBT7b3BlhQ

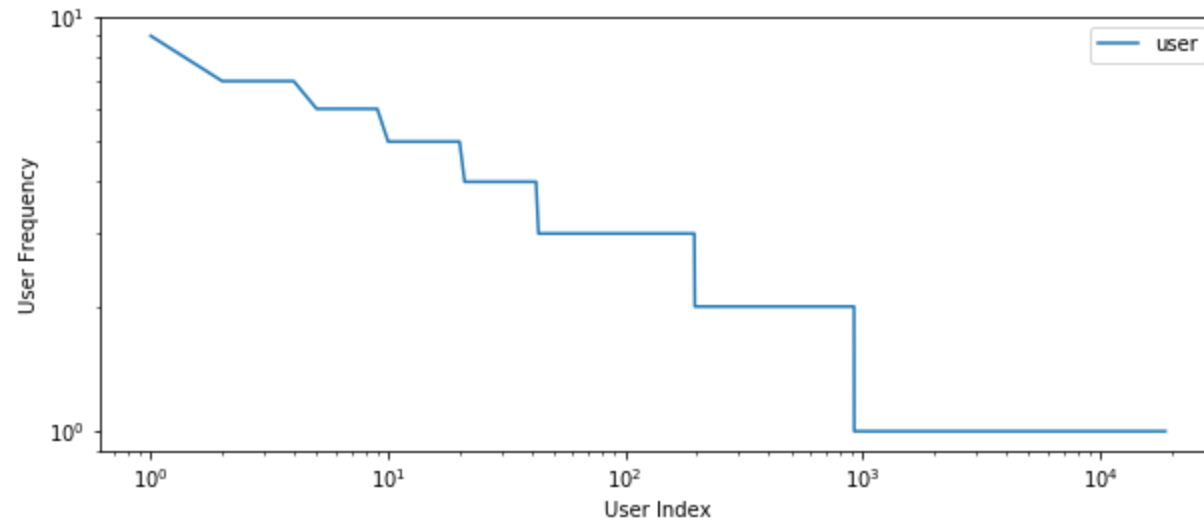
Text Distribution

- stemmed feature size: 35531
- 2-gram feature size: 446252
- 3-gram feature size: 1345361



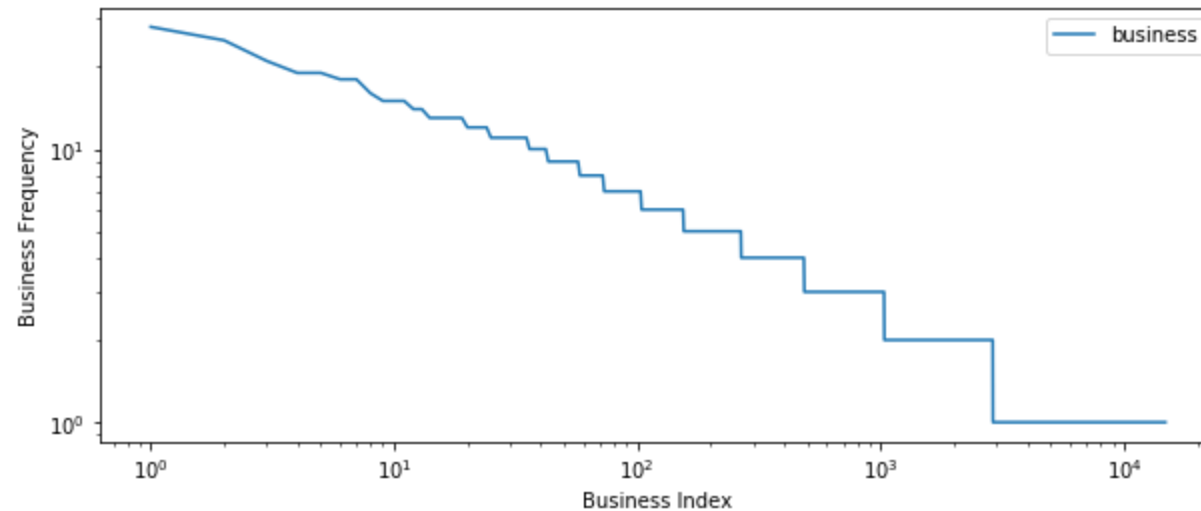
User Distribution

- user size: 35531



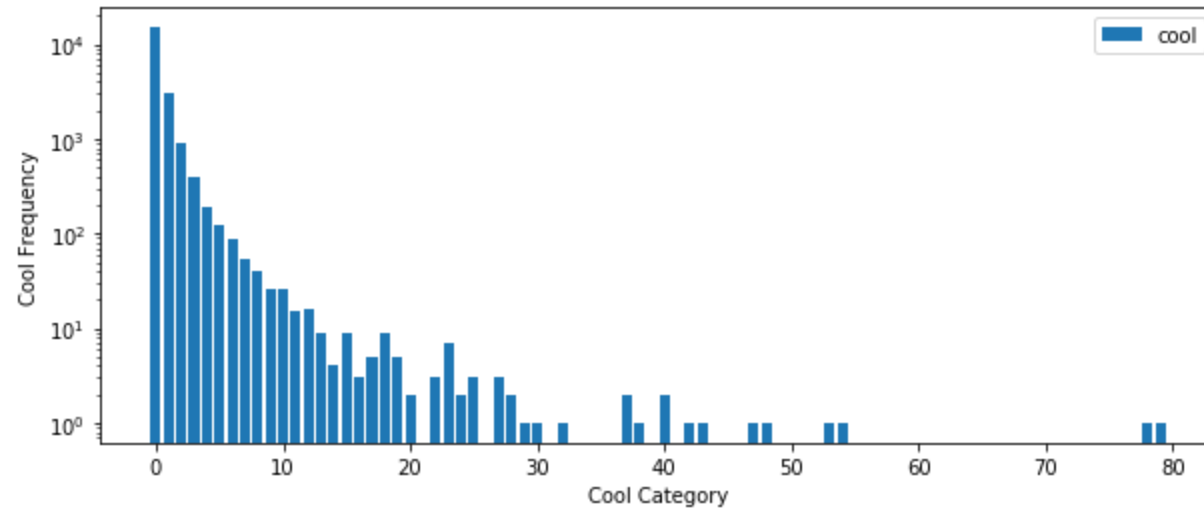
Business Distribution

- user size: 14763



Cool Distribution

- user size: 41



More Distributions

- You can explore by yourselves.

Evaluation

- accuracy on **test data**

Schedule

Three weeks

- [March 16, 2021] Project Start
- [March 23, 2021] Release the validation accuracy of weak baseline
- [March 30, 2021] Release the validation accuracy of strong baseline
- [April 06, 2021, 23: 59] Project due

Submission

- Predictions on **test data** (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 06, 2021
- 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 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