EVALUATION OF OPINION DETECTION SYSTEMS

PERFORMANCE AND EVALUATION

- Performance depends on :
  - the type of task :
    - Ex: target identification, polarity classification
  - the number and type of classes
    - positive vs. negative
    - multi-class
    - fear vs. anger more subtle than fear vs. joy
  - the train/test corpus (diversity of data)

PERFORMANCE AND EVALUATION

- Provide comparison of Human vs. system performance
- See evaluation campaings :
  - Semeval : series of evaluations of computational semantic analysis systems including sentiment analysis tasks

EVALUATION METRICS FOR CLASSIFICATION SYSTEMS

- Accuracy
- Recall
- Precision
- F1-score
- AUC-ROC
- Mean Square Error
- RandIndex

ACCURACY

Predicted class / Actual class	Positif	Négatif
Positif	**TP**	FN
Négatif	FP	**TN**

- Accuracy : percentage of correctly predicted instances

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

=> Not appropriate for imbalanced data sets

EVALUATION METRICS FOR CLASSIFICATION SYSTEMS

- Recall for class c
  - how many documents of class c in the test dataset have been correctly predicted as class c?
    - R = (number of system's correct assignments to class c) / (number of documents labelled c)
    - R = TP/ (TP+FN)
      - A system that tends to infrequently assign class c (high system *silence* for class c) will have a low recall
      - Also called sensitivity of the model or detection rate

EVALUATION METRICS FOR CLASSIFICATION SYSTEMS

- Precision for each class
  - how many documents predicted as class c correspond to correct predictions?
  - P = (number of system's correct assignments to class c) / (number of system's assignments to class c)
  - P = TP/(TP+FP)
    - A system that tends to allocate class c too frequently (system *noise* is high for class c) will have a low precision

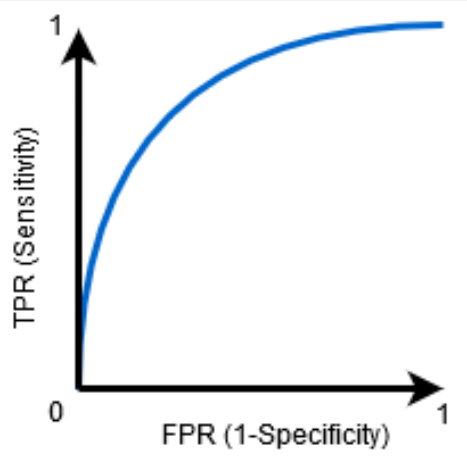
EVALUATION METRICS FOR CLASSIFICATION SYSTEMS

- F-score for each class
  - F-score : harmonic mean between recall and precision = 2×(P ×R) / (P +R)
- Multiclass : average over the classes
  - => macro-F1 : harmonic mean between average precision and average recall obtained on each class : gives equal importance to each observation
  - => micro-F1: harmonic mean between micro-recall and micro-precision
    - Micro-precision and micro-recall use all the correctly predicted instances in a multiclass case : gives each class equal importance.

$$\frac{TP1+TP2}{TP1+TP2+FP1+FP2} \qquad \frac{TP1+TP2}{TP1+TP2+FN1+FN2}$$

AUC – AREA UNDER THE CURVE

- Area Under the ROC (Receiver Operating Characteristic) Curve
  - ROC : $TPR_s = f(FPR_s)$ for different classification threshold s
  - TPR True positive ratio = recall (ratio of documents of class c in the test dataset that have been correctly predicted as class c)
  - FPR False positve ratio = FP/(FP+TN) (ratio of documents not in class c in the test set that have been wrongly predicted as class c)
- A metric for measuring the quality of a model independently from the used classification threshold.
  - The higher the AUC, the better



EVALUATION METRICS FOR REGRESSION

- Mean square error:
  - Can be used to measure the divergence between the predicted sentiment labels and actual sentiment labels (sentiment classif. viewed as a regression task)
  - Can be used as a loss for the training

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

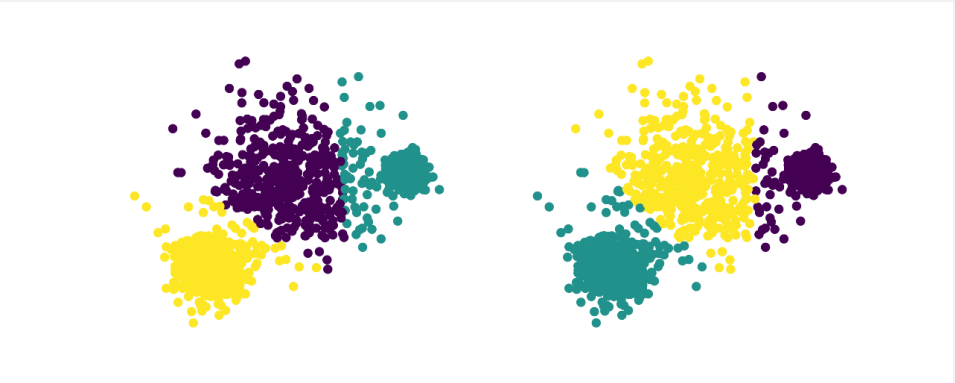
n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

EVALUATING CLUSTERING

- RandIndex: a sort of percentage of agreement between the clustering and a ground truth partition.
- the proportion of pairs of elements that are jointly grouped or jointly separated.



EXAMPLE OF PERFORMANCE FOR ABSA

CAN: Constrained Attention Network (RNN)

IMN: Interactive Multi-Task Learning Network (CNN)

COMPARISON STUDY FOR THE SEMEVAL 2014 DATA SET ON THE RESTAURANT DOMAIN

Method	Accuracy	Macro-F1	Method	Accuracy	Macro-F1
CNN	77.95%		AF-LSTM [48]	75.40%	
PF-CNN [33]	79.20%		PRET+MULT [49]	79.11%	69.73%
PG-CNN [33]	78.93%		Inter-aspect dependencies [50]	79.00%	
GCAE [34]	77.28%		LSTM+SynATT+TarRep [51]	80.63%	71.32%
TNet-LF [36]	80.79%	70.84%	Soft label strategy [52]	80.98%	71.52%
TNet-AS [36]	80.69%	71.27%	CAN [53]	**83.33%**	**73.23%**
IMN [37]	**83.89%**	**75.66%**	ATLS [54]	82.86%	
TD-LSTM [40]	75.60%		AdaRNN [55]	60.42%	
AE-LSTM [43]	76.60%	66.45%	PhraseRNN [56]	66.20%	62.21%
AT-LSTM [43]	76.20%		MemNet [58]	80.95%	
ATAE-LSTM [43]	77.20%	65.41%	Tensor DyMemNN [59]		58.61%
BILSTM-ATT-G [44]	79.73%	69.25%	Holo DyMemNN [59]		58.82%
IAN [45]	78.60%		RAM [61]	80.23%	70.80%
MGAN [46]	81.25%	71.94%	Cabse [64]	80.89%	
AOA-LSTM [47]	81.20%		CMA-MemNet	81.26%	68.64%
			FCMN [65]	82.03%	

Liu, Haoyue, et al. "Aspect-Based Sentiment Analysis: A Survey of Deep Learning Methods." *IEEE Transactions on Computational Social Systems* 7.6 (2020): 1358-1375.

PERFORMANCE COMPARISON

- RNN models seem to perform better than CNN
- But execution and training time is higher than CNN for LSTM and its variants.
- See SOA: *Sentiment analysis using deep learning architectures: a review*

TO GO FURTHER

- The Curse of Performance Instability in Analysis Datasets: Consequences, Source, and Suggestions Xiang Zhou, Yixin Nie, Hao Tan, Mohit Bansal
- Reliable characterizations of NLP systems as a social responsibility, talk by Christopher Potts at ACL 2021 http://web.stanford.edu/~cgpotts/talks/potts-acl2021-slides-handout.pdf
  - Important question: "is there anything about the composition of the dataset that might impact future uses"