CSCE 581: Introduction to Trusted AI

Lectures 21 and 22: Supervised ML (Text Processing), Trust Issues

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

1ST AND 3RD APRIL, 2025

Carolinian Creed: "I will practice personal and academic integrity."

Credits: **Copyrights of all material reused acknowledged**

Organization of Lectures 21, 22

- Introduction Section
  - Recap from Week 10 (Lectures 19 and 20)
  - Announcements and News
- Main Section
  - L21: Supervised ML (Text)
  - L22: Classification, Trust Issue
- Concluding Section
  - About next week – Lectures 23, 24
  - Ask me anything

Recap from Week 10 (Lectures 19, 20)

- We looked at
  - L19 AI - Unstructured (Text): Representation, Common NLP Tasks
  - L20: Natural Languages/ Language Models and their Impact on Text/ AI

AI Trust: News

- Browser extensions of LLM-based chatbots leak information. “"In our study, we emphasize that not only do these assistants collect users' highly sensitive data but that this information is also shared with their own servers as well as third-party trackers, which can be further utilized to target highly personalized and sensitive ads to the user.”

- Article - https://www.theregister.com/2025/03/25/generative_ai_browser_extensions_privacy/
- Paper - Big Help or Big Brother? Auditing Tracking, Profiling, and Personalization in Generative AI Assistants, https://arxiv.org/abs/2503.16586, 2025

Table 1: Overview of studied AI browser assistants sorted by their popularity.

Legend: Personal Data : Personally Identifiable Information (**PII**), Personal Communications (**PC**), Financial Information (**FI**).
Web Data : User Activity (**UA**), Web History (**WH**), Website Content (**WC**), Location (**LOC**). No Data : **No Data** collected.

Extension Name	Install Counts	Supported Model(s)	Default Model	Invocation Mode	Response Mode	Data Disclosures	SDK Version
Sider: ChatGPT Sidebar	4M		sider	Automatic	Server-side	PII WC	4.35.0
Monica - Your AI Copilot	2M		gpt-4o-mini	Mixed	Server-side	PII UA PC FI	7.6.0
ChatGPT for Google	2M		gpt-4o-mini	Mixed	Client-side	PII UA PC FI	5.5.1
Merlin Ask AI	1M		gpt-4o	Mixed	Server-side	PII LOC	7.3.2
MaxAI: Chat with Webpage	800K		gpt-4o-mini	Manual	Server-side	PII UA	6.7.1
Perplexity - AI Companion	500K		perplexity	Manual	Server-side	No Data	1.0.21
HARPA AI	400K		harpa-v1-smart	Manual	Server-side	PII UA WH WC	9.6.2
Wiseone - AI Copilot	90K		gpt-4o	Manual	Server-side	PII WC	1.7.2
TinaMind - AI Assistant	50K		gemini-1.5-pro	Manual	Server-side	PII UA PC	2.14.2
Copilot: AI Assistant	30K		gpt-4o-mini	Automatic	Server-side	PII	1.5.73

AI Trust: News

Table 2: Data collection and exfiltration behavior of assistants in public and private online spaces of a user. *Exfiltration legend*: Full Webpage : Page text, title, location, hyperlinks. Server-fetch Webpage : Page title, location, server-fetched file's upload location. Plain Webpage : Page text, title, location. Partial Webpage : Partial content or missing details. *Response legend*: ✔: Response with Relevant Details. ✔✗: Missing some details in Response. ⊘ : Response restricted. ✘: No response generated.

	Category	WebPage	Sider	Monica	CFG	Merlin	MaxAI	Perplexity	Harpa	Wiseone	TinaMind	Copilot
Public Spaces	News Platforms	cnn.com	✔	✔	✔	✔	✔	✘	✔	✔	✔	✘
	Open Forums	reddit.com	✔✗	✔	✔✗	✔	✔✗	✔✗	✔	✔✗	✔✗	✔
	Informative Articles	wikipedia.org	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘
	E-commerce Website	amazon.com	✔	✔	✔	✔	✔✗	✘	✔	✘	✔✗	✘
	Sports Websites	espn.com	✔	✔	✔	✔	✔	✔	✔	✘	✔✗	✘
	Travel Platforms	expedia.com	✔	✔	✔	✔	✔	✘	✔	✘	✔	✔
	User-generated Media	youtube.com	✔✗	✘	✔	✔	✔✗	✘	✔✗	✔	✔	✔
	Kids Website	nickjr.com	✘	✔	✔	✔	✔	✔	✔	✘	✔	✘
	Misinformation Website	infowars.com	✔	✔	✔	✔	✔	✘	✔	✔	✔	✔
	Violence Material	guns.com	✔	✔	✔	✔	✔	✘	✔	✔	✔	✘
Private Spaces	Health Portal	university health portal	✔	✔	✔	✔	✔	✘	✔✗	✔	✔✗	✘
	Email Account	mail.google.com	✔	✔	✔✗	✔	✔✗	✘	✔✗	✔	✔	✘
	Social Media Platform	facebook.com	✔	✔	✔	✔	✔	✘	✔	✔	✔✗	✘
	Adult Content	pornhub.com	✔✗	✔✗	✔✗	✔✗	✔✗	⊘	✔✗	⊘	✔✗	✘
	Online Streaming Service	netflix.com	✔✗	✔✗	✔✗	✔✗	✔✗	✘	✔✗	✘	✔✗	✘
	Government Website	irs.gov	✔✗	✔✗	✔✗	✔✗	✔✗	✘	✔	✔✗	✔✗	✘
	Dating Service	tinder.com	✘	✔	✔✗	✔✗	✔✗	✘	✔	✘	✔✗	✘
	Financial Service	chase.com	✘	✔✗	✔✗	✔✗	✔✗	✘	✔✗	✘	✘	✘
	Educational Platform	canvas.instructure.com	✔	✔	✔	✔	✔	✘	✔	✔✗	✔✗	✔
	Messaging Platform	slack.com	✔	✔	✔	✔	✔	✘	✔	✔✗	✔	✘

"One of the most shocking findings was that GenAI browser assistants were freely able to collect and share data to their own servers on authenticated **health** portals. They were able to answer follow-up questions ranging from patient details to entire medical history. Collection of PHI without appropriate user consent is in clear violation of HIPAA [14]. "

"Moreover, **student's academic records** including assessment scores, exam performances, overall grades – were all collected and shared with browser assistant's servers demonstrating violation of FERPA [34] that aims to protect these attributes for a student."

AI Trust News: LLM-based Summaries

Bloomberg Has a Rocky Start With A.I. Summaries

- https://www.nytimes.com/2025/03/29/business/media/bloomberg-ai-summaries.html?unlocked_article_code=1.7k4.oSbJ.htksCZb-DQ3E&smid=nytcore-android-share
- The newspaper chain Gannett uses similar A.I.-generated summaries on its articles,
- The Washington Post has a tool called "Ask the Post" that generates answers to questions from published Post articles

Questions: why deploy automated summaries ? If so, why LLM-based and not classical, extractive summaries, which are guaranteed from original source? What is the impact on long-term credibility?

Project Status and Timeline

- Office Hours: 3-4pm (M), 10-11am (Th)
- Finish project presentations by Apr 22
- Project presentations
  - Apr 22 (Tu) Project presentation
  - Apr 24 (Th) Project presentation
- Project delivered

  Apr 29 (Tu) Project in Github

19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview

Introduction Section

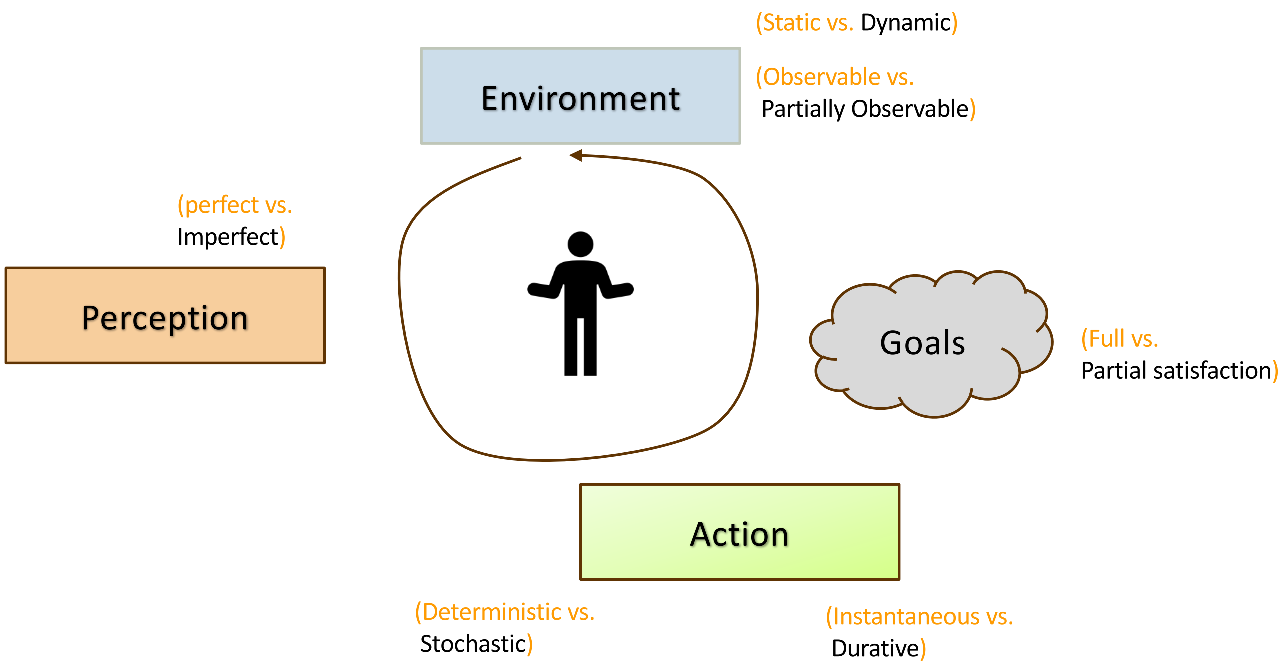
Announcement: Change to
Student Assessment

Tests	Undergrad	Grad
Course Project – report, in-class presentation	600	600
Quiz – 2 quizzes	200	200
Final Exam	200	100
Additional Final Exam – Paper summary, in-class presentation		100
Total	1000 points	1000 points

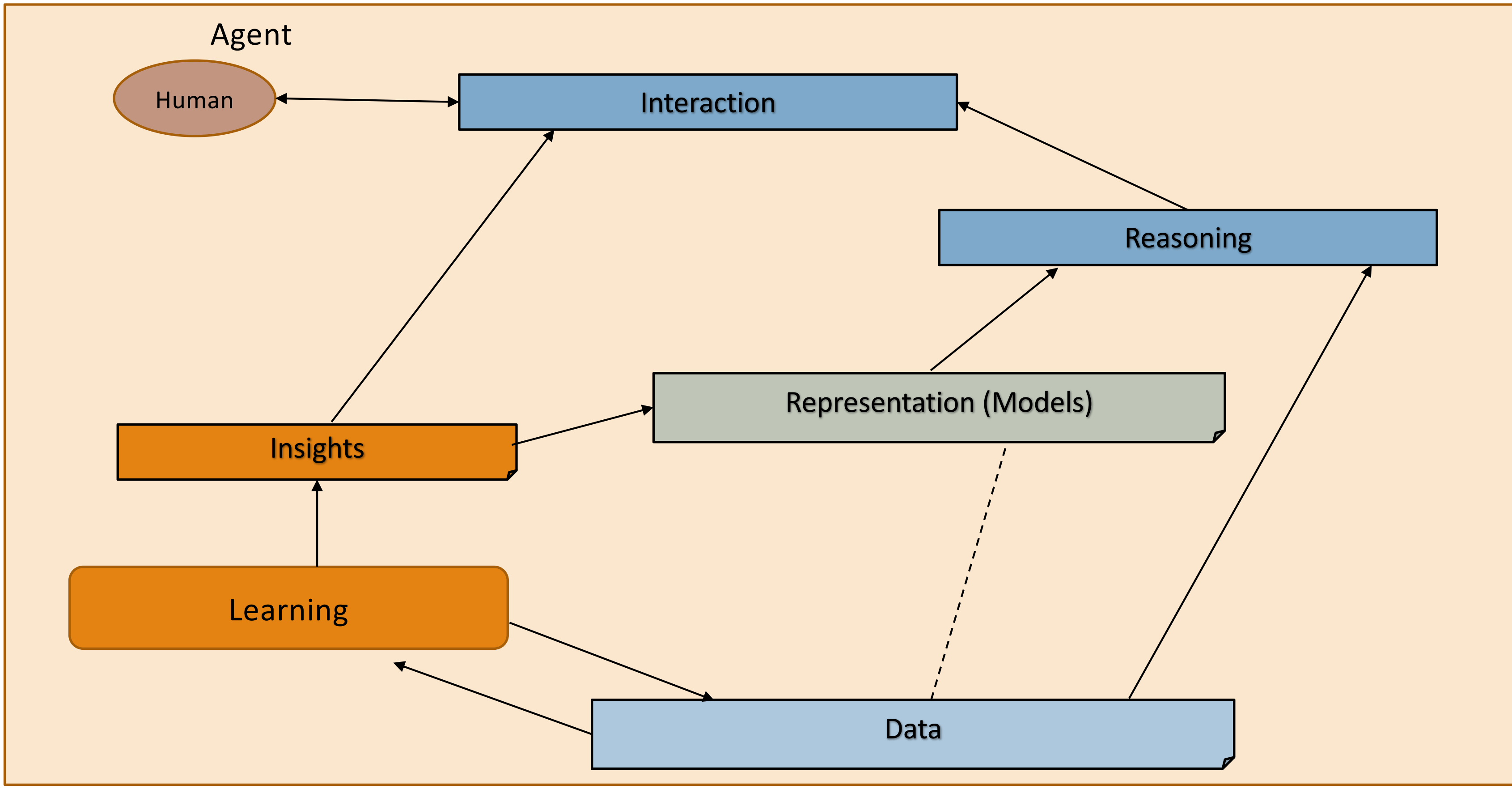
A = [920-1000]
B+ = [870-919]
B = [820-869]
C+ = [770-819]
C = [720-769]
D+ = [670-719]
D = [600-669]
F = [0-599]

Change: 4 quizzes to 2; no best of 3

Intelligent Agent Model



Relationship Between Main AI Topics (Covered in Course)



High Level Semester Plan (Adapted, Approximate)

CSCE 581 –

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification, **Large Language Models**
- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

Increased focus on LLMs and projects now

AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability

Main Segment

Recap – ML, Classification / Supervised ML, Metrics

Machine Learning – Insights from Data

- Descriptive analysis
  - Describe a past phenomenon
  - **Methods**: classification (feedback from label), clustering, dimensionality reduction, anomaly detection, neural methods, reinforcement learning (feedback from hint/ reward)

- Predictive analysis
  - Predict about a new situation
  - **Methods**: time-series, neural networks

- Prescriptive analysis
  - What an agent should do
  - **Methods**: simulation, reinforcement learning, reasoning

- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

Concepts

- **Input data**: data available
  - **Training data**: used for training a learning algorithm and get a model
    - [Optional] **Validation data**: used to tune parameters
  - **Test data**: used to test a learning model

- **Classification problem**
  - Separating data into classes (also called labels, categorical types)
  - One of the attributes is the class label we are trying to learn
  - Class label is the **supervision**
- **Clustering problem**
  - We are trying to learn grouping of data
  - There is no attribute indicating membership in the groups (hence, **unsupervised**)
- **Prediction problem**
  - Learning value of a continuous variable

Reference: https://machinelearningmastery.com/difference-test-validation-datasets/
https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf

Many Method Types and Classifiers

- Individual methods
  - Decision Tree (J48), R1, One-R
  - Naïve Bayes
  - …
- Ensemble
  - Bagging: Aggregate classifiers (“bootstrap aggregation” => bagging)
    - Random Forest
    - Samples are chosen with replacement (bootstrapping), and combined (aggregated) by taking their average
  - Gradient Boosting: aggregate to turn weak learners into strong learners
    - Boosters (aggregators) turn weak learners into strong learners by focusing on where the individual weak models (decision trees, linear regressors) went wrong
    - Gradient Boosting

Source:
- Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber
- https://towardsdatascience.com/getting-started-with-xgboost-in-scikit-learn-f69f5f470a97

Metric Types

- **Effectiveness**: what the **user** of a system sees, primarily cares about
- **Efficiency**: what the **executor** in a system sees, primarily cares about

***Effectiveness* Metrics**

***Efficiency* Metrics**

Metrics: Accuracy, Precision, Recall

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Accuracy =
(TP+TN)/
(TP+FP+FN+TN)

Precision =
(TP)/
(TP+FP)

Recall =
(TP)/
(TP+FN)

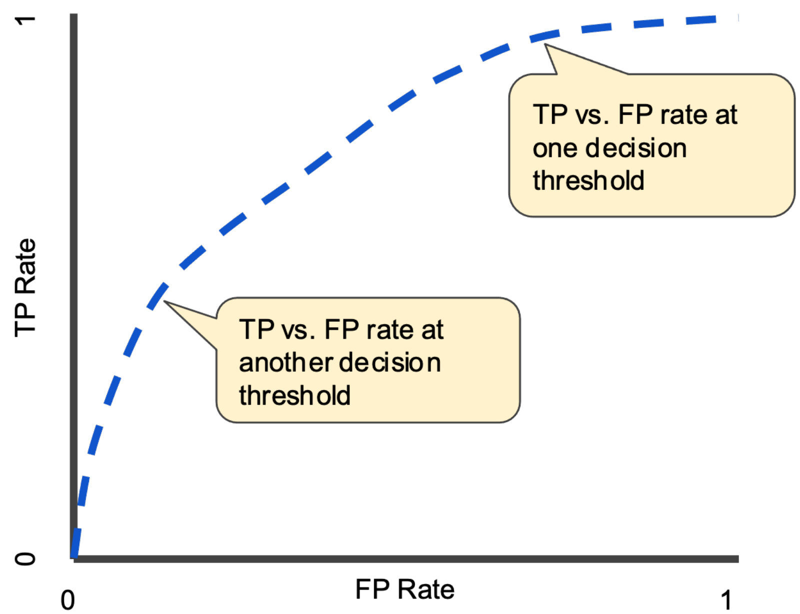
F1 Score: *Harmonic Mean*

1/F1 = 1/Precision + 1/Recall

F1 = 2*(Recall * Precision) /
(Recall + Precision)

ROC – Receiver Operating Characteristic curve

An ROC curve plots TPR vs. FPR at different classification thresholds



True Positive Rate = Recall =
(TP)/
(TP+FN)

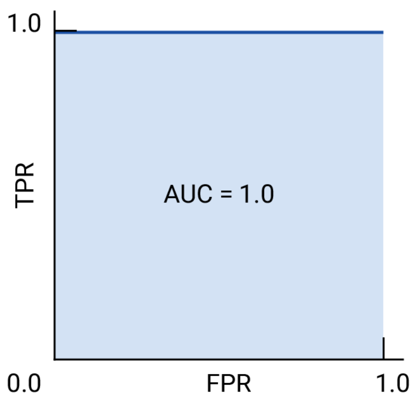
False Positive Rate =
(FP)/
(FP+TN)

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

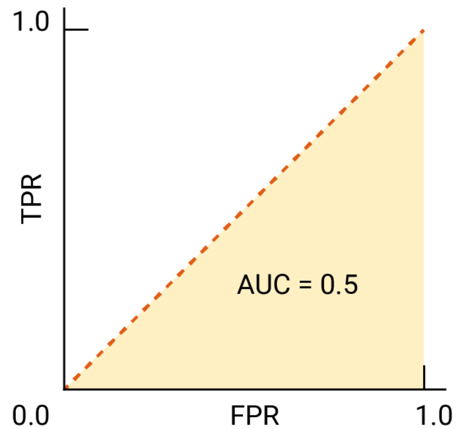
Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

AUC/ ROC Examples

ROC and AUC of a perfect system



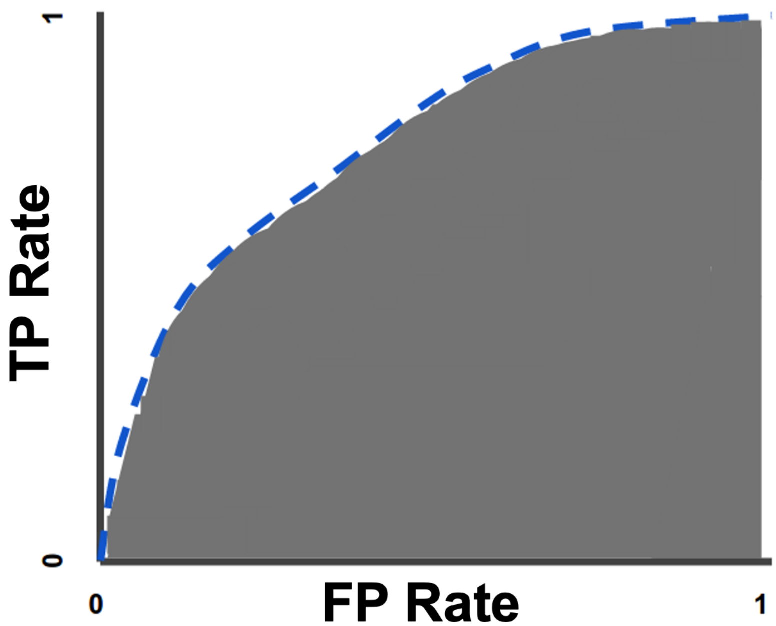
ROC and AUC of completely random guesses



The AUC is 0.5, representing a 50% probability of correctly ranking a random positive and negative example

Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

AUC – Area Under the ROC Curve



- Aggregate measure of performance across all possible classification thresholds.
- Interpretation: probability that the model ranks a random positive example more highly than a random negative example

Not helpful when the **cost** of false negatives vs. false positives are asymmetric

Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Exercise and References

- Google: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
  - Take quiz
- Blogs: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Image credit: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Which ML Classification Method to Choose?

- Blog: "Crawl – Walk – Run"
  - [The Crawl-Walk-Run Approach for AI-based Real World Problem Solving], Feb 2025, Biplav Srivastava

- Reading material:
  - "Which ML to Use" with title: Data-driven advice for applying machine learning to bioinformatics problems https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5890912/
  - "10 tips with title": Ten quick tips for machine learning in computational biology https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3

Discussion: 10 Tips Paper

- Access: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3
- Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining* **10,** 35 (2017). https://doi.org/10.1186/s13040-017-0155-3

The Tips

- Tip 1: Check and arrange your input dataset properly
- Tip 2: Split your input dataset into three independent subsets (training set, validation set, test set), and use the test set only once you complete training and optimization phases
- Tip 3: Frame your biological problem into the right algorithm category
- Tip 4: Which algorithm should you choose to start? The simplest one!
- Tip 5: Take care of the imbalanced data problem
- Tip 6: Optimize each hyper-parameter
- Tip 7: Minimize overfitting
- Tip 8: Evaluate your algorithm performance with the Matthews correlation coefficient (MCC) or the Precision-Recall curve
- Tip 9: Program your software with open source code and platforms
- Tip 10: Ask for feedback and help to computer science experts, or to collaborative Q&A online communities

Examples of Classification with Text

- Sentiment analysis (*assign sentiment classes*)
- Annotation (*assigning entity type to to text*)
- Application specific
  - Fake news
  - Spam email
  - …

When to Use Learning/ When Not

- When
  - There are too many patterns to enumerate
  - When data is plentifully available
  - [**Cost**: annotation to know the class labels, re-learning to avoid model shift]
- When Not
  - There are few, distinct patterns that can be easily enumerated
  - When there may be a high cost of accessing data
  - When there is a need to exert control on the AI development process; e.g., sensitive domain
  - [**Cost**: of knowing the patterns completely]

So, What Changes With Text?

So, What Changes With Text?

Handling of (textual) data before and after applying ML methods!

Adapted from Image Credit: Trustworthy Machine Learning, Kush Varshney

Common Textual Data Processing Steps for ML

- Input: strings / documents/ corpus
- Processing steps (task dependent / optional - *)
  - Parsing
  - Word pre-processing
    - Tokenization – getting tokens for processing
    - Normalization* - making into canonical form
    - Case folding* – handling cases
    - Lemmatization* – handling variants (shallow)
    - Stemming* – handling variants (deep)
  - Semantic parsing – representations for reasoning with meaning *
  - Embedding – creating vector representation*

ML – Supervised (Fake News)

- By Example:
  - https://github.com/biplav-s/course-nl/blob/master/l9-ml-review/Classification%20-%20Fake%20news.ipynb

- Fake news dataset

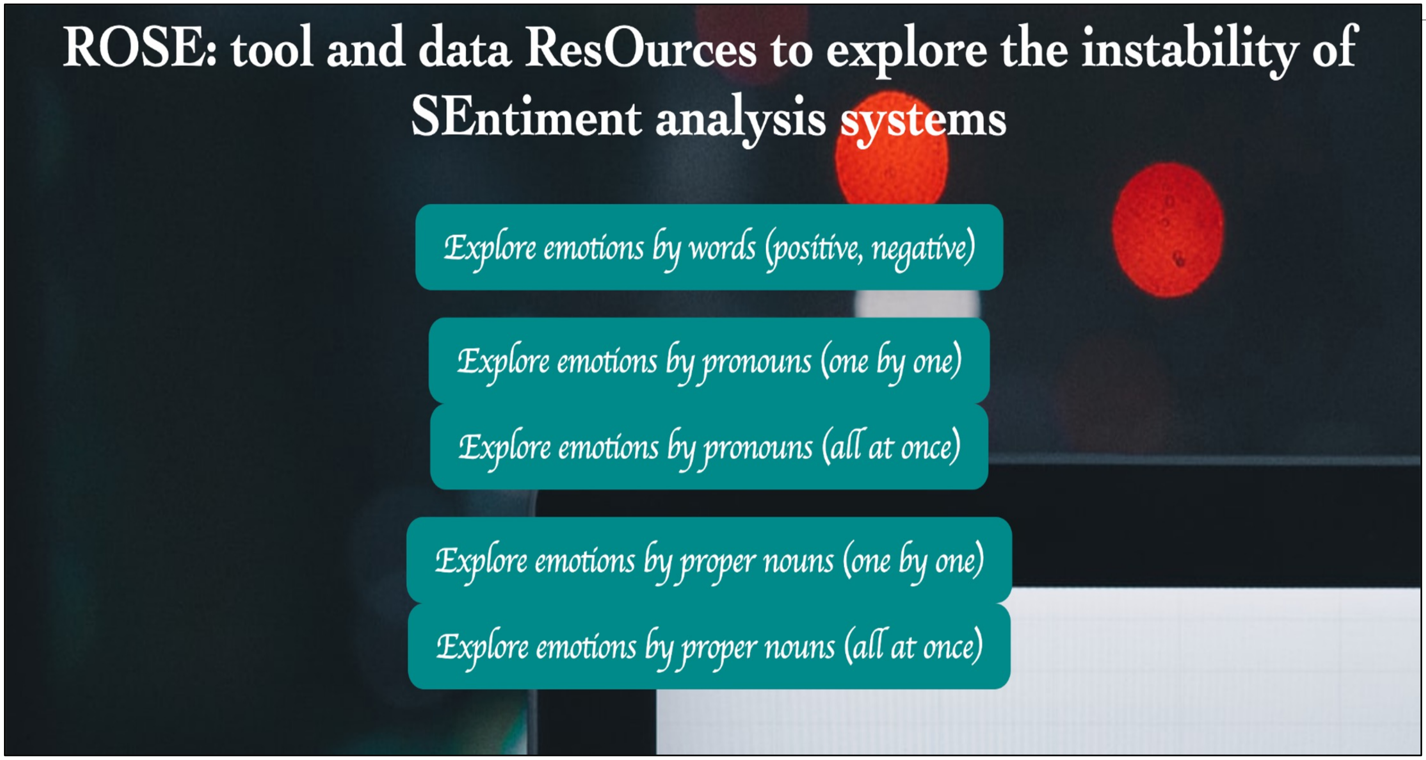
ML – Supervised (Movie Sentiment)

- By Example:
  - Data: IMDB – 50K movie reviews
    http://ai.stanford.edu/~amaas/data/sentiment/
    https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/

- Sample code
  - Classical ML methods:
    - https://www.kaggle.com/code/eminaanapaydn/imdb-sentiment-analysis (MultinomialNB)
    - https://www.kaggle.com/code/youssefemad004/sentimentanalysisml (MultinomialNB, RandomForest)
  - LLM/ BERT classifier:
    - https://www.kaggle.com/code/tcc3281/bert-sentiment

Trust Issue – Stability of Output

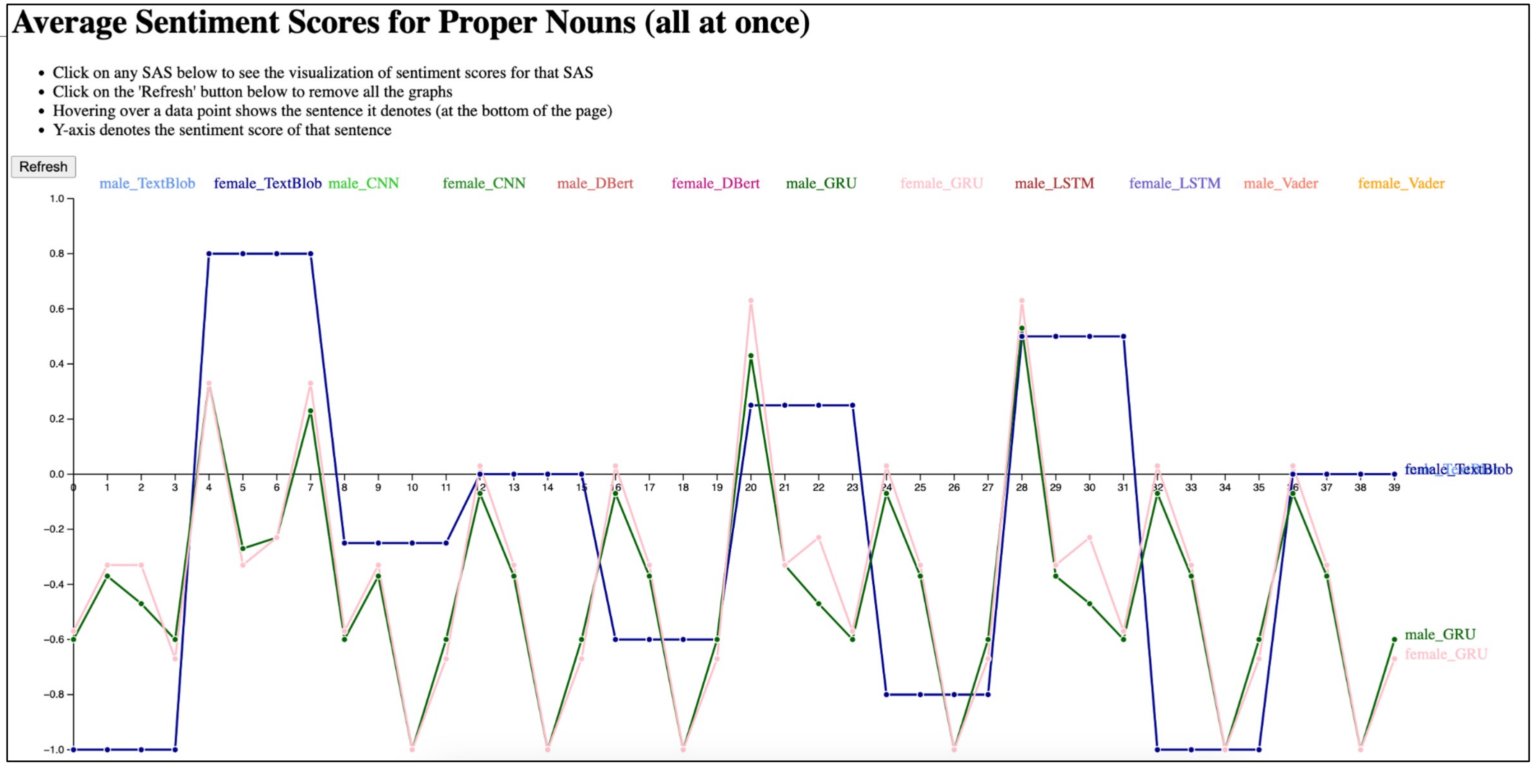
Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems



Scan the code to try our ROSE tool!

References:
1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems



References:

1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Instability of AI is Well Recorded

- [Text] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]
- [Image] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020
- [Audio] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

Project Discussion

Course Project

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
     1. Solution => ML task (e.g. classification), recommendation, text summarization, …
     2. Use a foundation model (e.g., LLM-based) solution as the baseline
  4. (Data) Explore the data for a solution to work
  5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
  - Due to Users
    - Diverse demographics
    - Diverse abilities
    - Multiple human languages
  - Or other impacts
- What one does to mitigate trust issue

Rubric for Evaluation of Course Project

Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

Concluding Section

Week 11 (L21 and 22): Concluding Comments

- We looked at
  - L21: Supervised ML (Text)
  - L22: Classification, Trust Issue

About Next Week – Lectures 23, 24

Lectures 23, 24

- Trust issue – Mitigation – AI Explanation (XAI)
- Trust Issues – Mitigation – AI Rating/ Certification

19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview