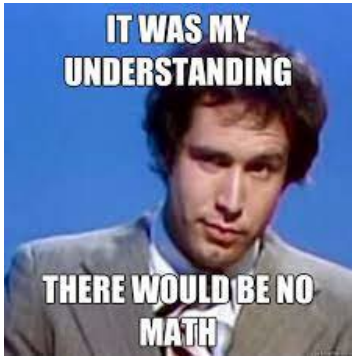
Kelley School of Business

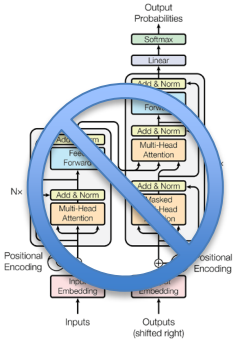
Text Analysis: A Primer

INDIANA UNIVERSITY – Aaron F. McKenny – August 4, 2023

Goals

1. Shared vocabulary
2. Shared understanding of NLP landscape
3. Familiarity with how computers understand and analyze text





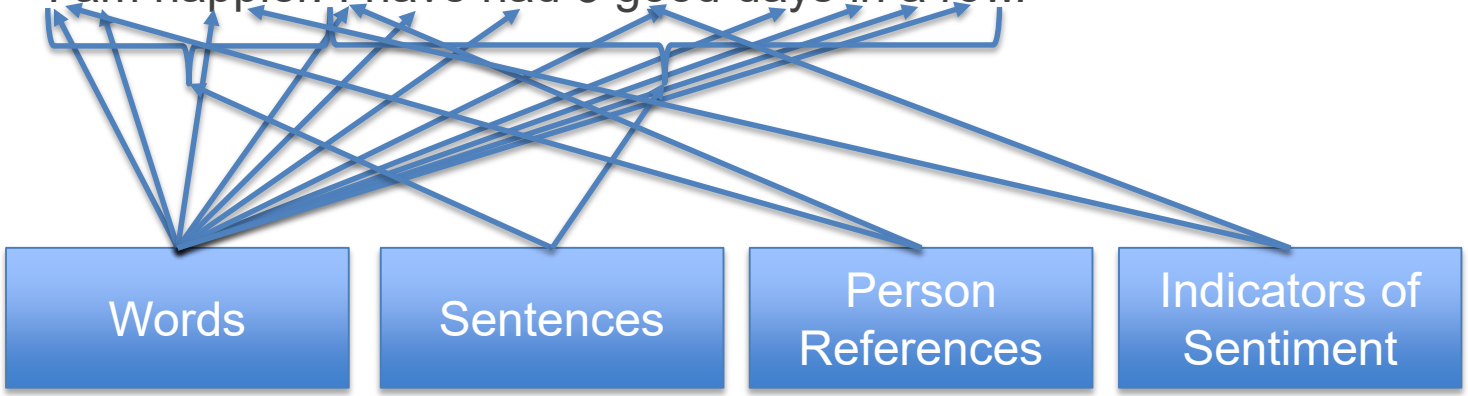
How Computers Interpret Text

Be The Computer…

What is the meaning of this passage?
…or at least tell me how many words are presented here?

What You See vs What the Computer Sees

“I am happier. I have had 3 good days in a row.”



What You See vs What the Computer Sees

"I am happier. I have had 3 good days in a row."

…but is most of the meaning from understanding the individual characters?

Where *is* the Meaning?

- Sentences **Sentence Segmentation**
- Words/Tokens **Tokenization**
  - And some words convey more meaning than others **Stopword/Non-word Character Removal**

Transformed: [[happier], [3, good, days, row]]

Bigger Vocabulary, Greater Complexity

Happy: Happy, happier, happiest, happily, happiness

Be: Be, is, am, are, was, were, been

Same meaning, but different conjugations/inflections/words

Stemming – Happier→Happi, Were→Were

Lemmatization – Happiness→Happy, Were→Be

Transformed: [[happy], [3, good, day, row]]

Challenges With Processing Language

- Are these words different? “Have” “have”
  - The computer thinks so by default *(“H” vs “h”)*
  - Often normalize the casing of words
- Are these ‘words’? “☺”, “:)”, “3”
  - Decide how to treat emoji/emoticons/numerals

Transformed: [[happy], [three, good, day, row]]

A Final Catch

Most NLP applications don't work with words

[[happy], [three, good, day, row]] ← *simpler, but still uninterpretable*

Solution: Use numeric vectors based on the words used ("vector space model")

- Document-Term Matrix →
- One-Hot Vectors
- Word Embeddings

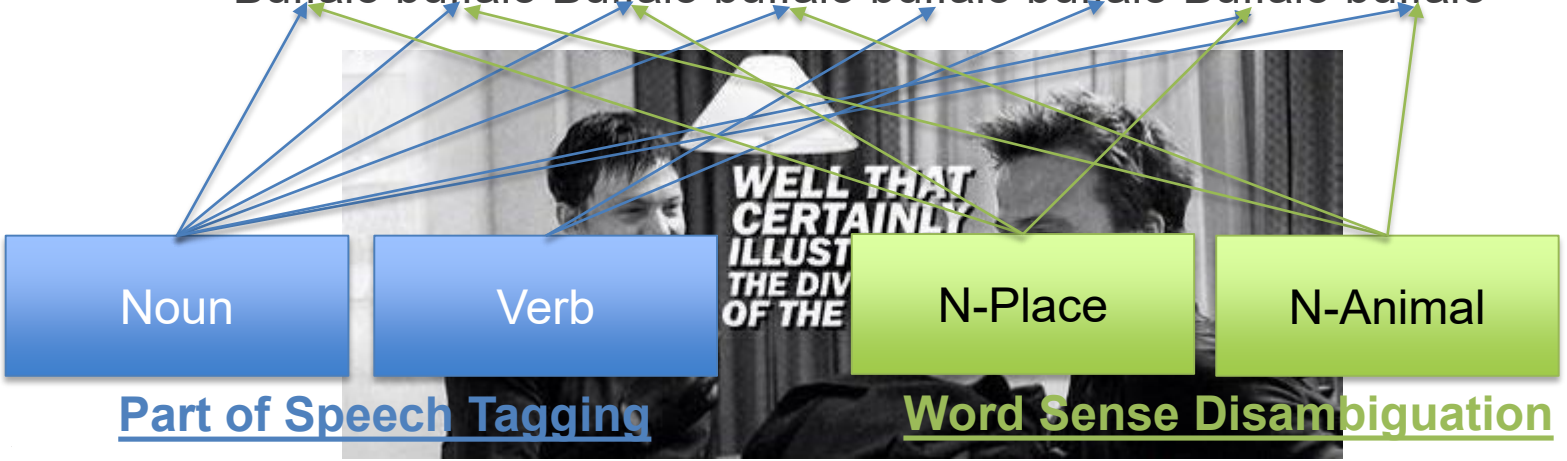
	day	good	happy	row	three
S1.	0	0	1	0	0
S2.	1	1	0	1	1

What Can We Do With This?

Understand Contextualized Meaning

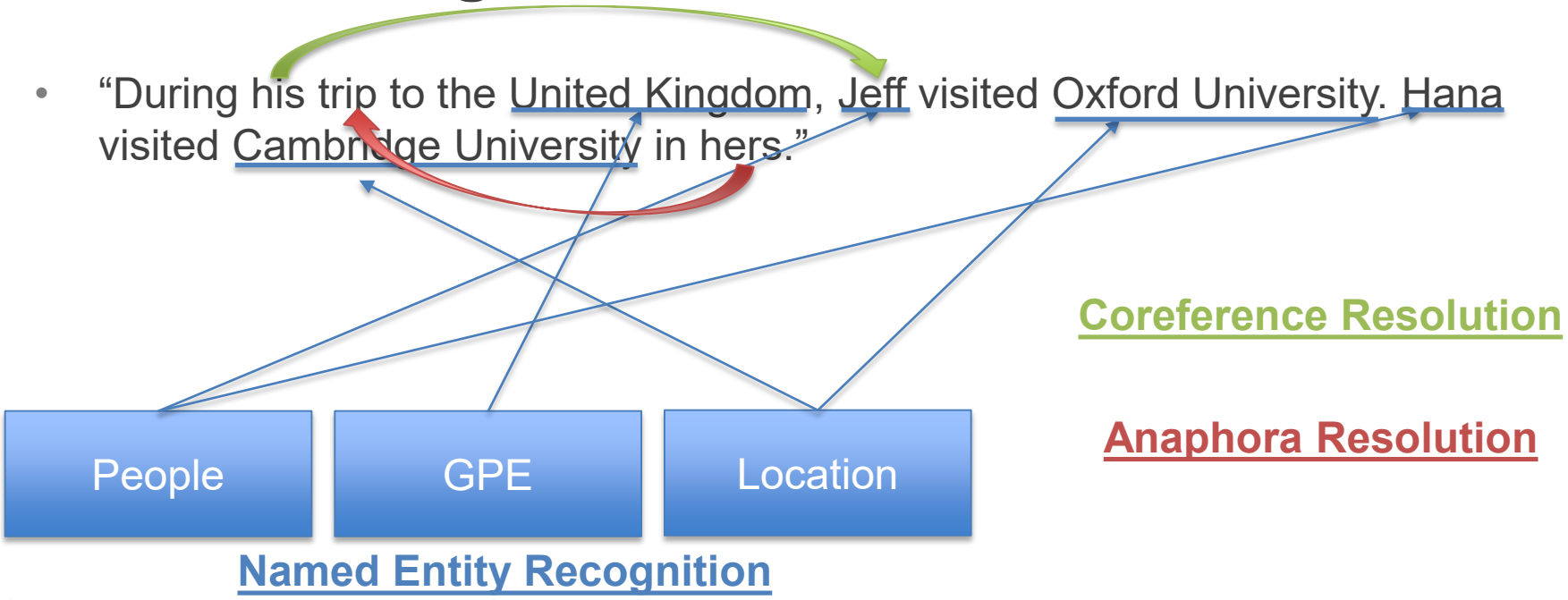
- Older approaches assume words have a single meaning/use

"Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo"



Who Is Doing What With/to Whom?

- "During his trip to the United Kingdom, Jeff visited Oxford University. Hana visited Cambridge University in hers."



Investigate What Is Being Discussed

Text Summarization

Topic 1: CEO, executive, manage
Topic 2: supply, source, procure, chain
Topic 3: manufacture, create, develop

Topic Modeling

Label/Score Texts Based on Their Contents

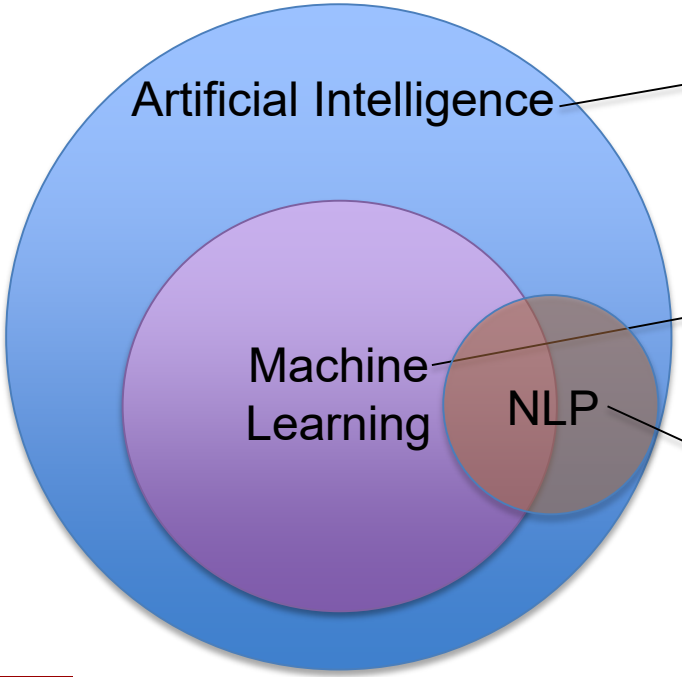
Text Classification **Regression Analysis**

"Goldman Sachs analysts are pessimistic regarding the promise of this new object recognition technology to obtain a significant market foothold."

Sentiment	Positive	Neutral	Negative ✓
Intent	Investment Guidance ✓	Financial Disclosure	Other
Future Orientation	5.7/7.0		

How We Accomplish These

AI, Machine Learning, & NLP



Get machine to do things that historically required human intelligence

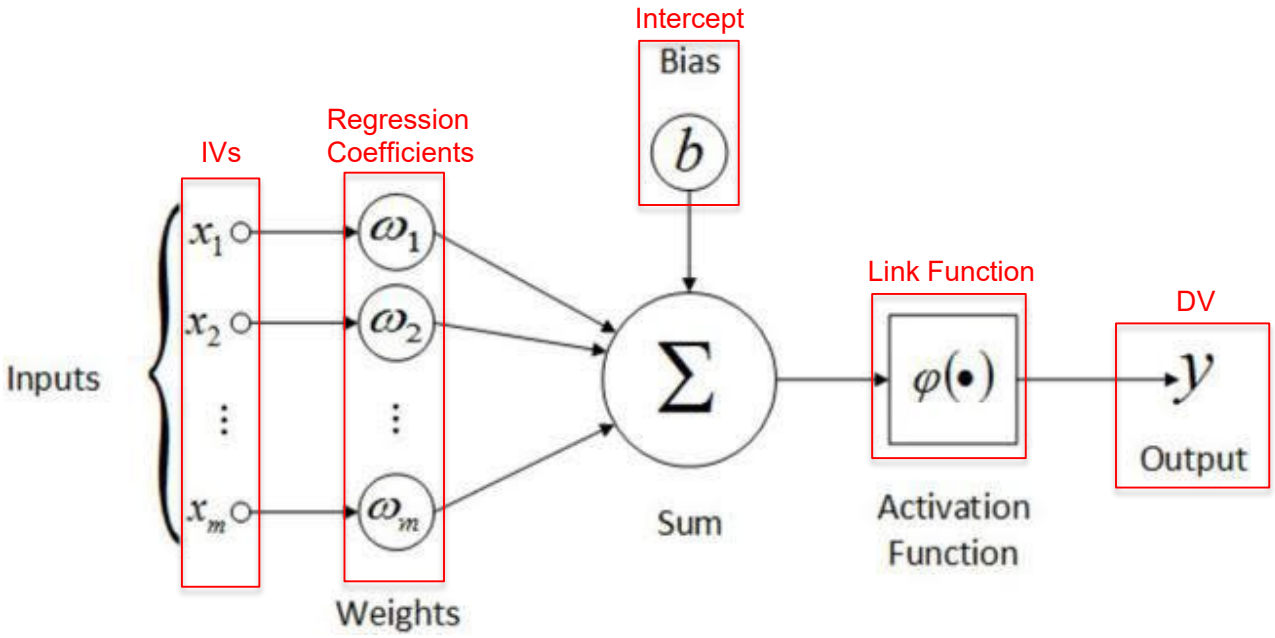
Machine learns how automatically (*more or less*)

Get the machine to understand/ manipulate/ generate human language

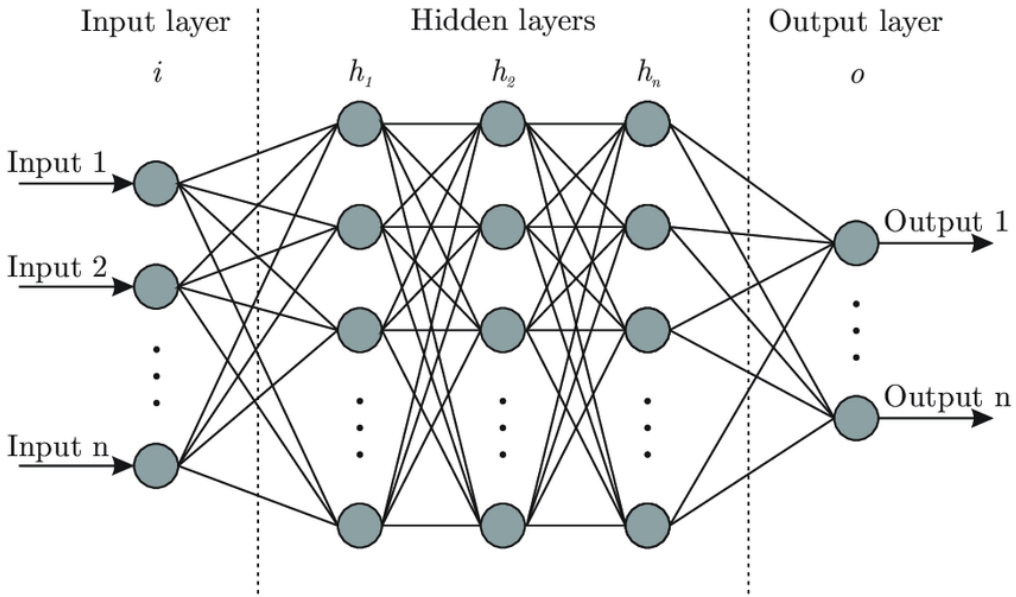
NLP Applications of Machine Learning

1. Unsupervised – Text features as inputs
   - E.g., Clustering, Topic Modeling
2. Supervised – Text features as inputs, researcher-provided data as outputs
   - E.g., Regression Analysis, Text Classification, Named Entity Recognition
3. Others:
   - Reinforcement, Self-Supervised

Neural Networks – Squint… Look Familiar?



Neural Networks



Training NNs From Scratch

1. Many weights and biases to train
   - Sample size requirements
2. Models have to learn
   - Language itself (e.g., English, Spanish, Chinese)
   - Understand/manipulate/generate that language in some desired way

Foundation (Pre-Trained) Language Models

GPT-4

BERT

Meta

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

Transformer-Based Language Models

- A type of neural network architecture
- Use context to interpret meaning of text
  - Uses “attention” to look at an entire sequence (e.g., sentence) of text at once
  - Older approaches (e.g., RNN, LSTM) looked at text sequentially
- Most popular foundation models use transformers
  - GP**T**, BER**T**

So… What About ChatGPT?

Next Word Prediction (using transformers)

1. Encoding step:
   - Preprocess your input text
   - “Attention” and a pretrained NN transform embeddings into contextualized vectors
2. Decoding step:
   - Take output from encoding step (and previously generated tokens) as input
   - Attention and a pretrained NN identify the highest probability next word
   - Repeat decoding step until the next word is an indicator to stop

*(over)simplified

Final Thoughts

“Text analysis moves pretty fast.

If you don’t stop and look around once in a while you could miss it”

-Ferris Bueller (probably)

Upcoming ORM Feature Topic

- Aim: Catalyze innovation in text analytics in organizational research
  - Where do text analyses presently fall short?
  - How can existing practices be improved?
  - How do new techniques open new doors?
  - How might text analysis aid in and support interpretive, qualitative research?
- Format:
  - Full paper submissions – open call
  - Initial submission deadline: August 31, 2024