

# **NATURAL LANGUAGE PROCESSING 101**

**Sarah Rodenbeck, Lead Research Data Scientist**

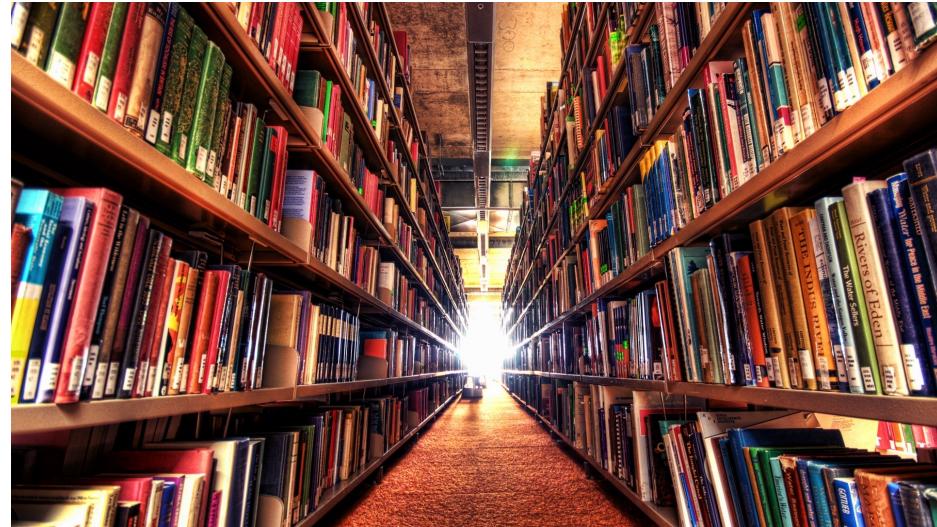
**Rosen Center for Advanced Computing**

# *Agenda*

- NLP Primer
- Mechanics & Evolution of NLP
- Getting started with NLP

# NLP Primer

# *Text data is everywhere!*



# Natural language processing

---

Article   Talk

---

From Wikipedia, the free encyclopedia

*This article is about natural language processing done by computers. For the natural language processing in the brain.*

**Natural language processing (NLP)** is an [interdisciplinary](#) subfield of [linguistics](#), [computer science](#) and [intelligence](#) concerned with the interactions between computers and human language, in particular computers to process and analyze large amounts of [natural language](#) data. The goal is a computer "understanding" the contents of documents, including the [contextual](#) nuances of the language within technology can then accurately extract information and insights contained in the documents as well as organize the documents themselves.

Challenges in natural language processing frequently involve [speech recognition](#), [natural-language understanding](#) and [natural-language generation](#).

---

## History [edit]



With a sufficiently large corpus of  
text data, models can learn the  
patterns of language

# *When have you used NLP?*

Google

I'll be I'll tomo  
ill tomorrow X



Cisco Quarantine  
Spam Quarantine Notification Yes  
Spam Quarantine Notification Hello :

**P** PURDUE  
UNIVERSITY®



Hey Siri

what is natural language processing

All Images Videos News Maps Shopping Settings

All regions Safe search: moderate Any time

**natural language processing**

**noun**

- 1. a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages
- 2. the branch of information science that deals with natural language information

Wiktionary, Creative Commons Attribution/Share-Alike License.

More at Wordnik

<https://www.ibm.com/topics/natural-language-processing>

**What is Natural Language Processing? | IBM**

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

**Natural language processing**

Natural language processing is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. [Wikipedia](#)

# *What can you do with NLP?*

Core Domain	Description	Example
Text Classification	Grouping documents into categories	Spam Filter
Information Extraction	Identifying information from text	Automatic calendar event creation when times mentioned
Information Retrieval	Finding relevant information	Search Engines
Question Answering Systems	Answering questions based on a natural language question	Siri/Alexa
Machine Translation/Summarization	Converting a sequence of text to another with the same meaning	Google Translate
Natural Language Generation	Generate new text based on a prompt	Chat GPT

# *What can't you do with NLP?*

## ChatGPT



### Examples

"Explain quantum computing in simple terms"

"Got any creative ideas for a 10 year old's birthday?"

"How do I make an HTTP request in Javascript?"



### Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



### Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

Logical Reasoning

NLP doesn't truly understand  
language!

# Challenges of NLP - Ambiguity

## Lexical Ambiguity

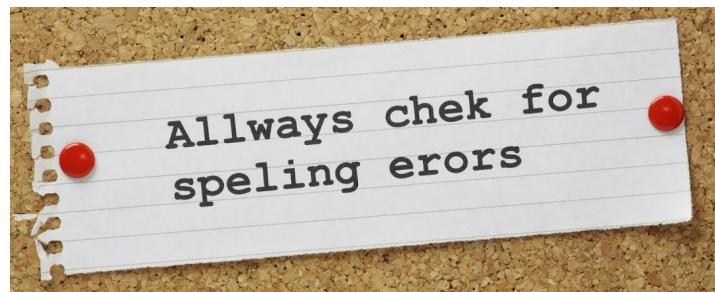
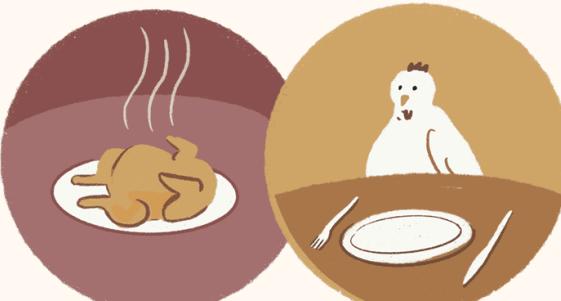
The presence of two or more possible meanings within a single word.



"I saw her duck."

## Syntactic Ambiguity

The presence of two or more possible meanings within a single sentence or sequence of words.



I **ran** to the store because we **ran** out of bread.

Can I **run** something past you?

That house is really **run** down.

The animal didn't cross the street because **it** was too tired.

-----versus-----

The animal didn't cross the street because **it** was too wide.

I love taking tests 😊

# *Challenges of NLP - Language*

## Teen Slanguage

### **WIG (NOUN/INTJ.)**

Expression of surprise/shock/amazement  
"I just got free tickets to the show!"  
"OMG, wig!"

### **T/TEA (NOUN)**

Gossip  
"Girl, spill the T!"

### **YEET (INTJ.)**

Expression of excitement  
"We're going out tonight, yeet!"

### **SNATCHED (ADJ.)**

Looking good, attractive  
"I think she likes you, she said you're snatched."

### **SALTY (ADJ.)**

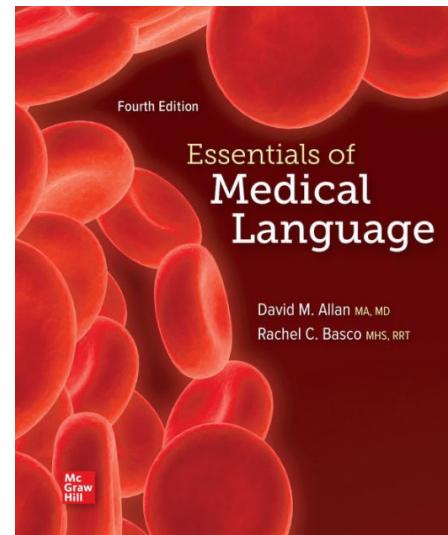
Angry/bitter/upset  
"They're just salty because they lost."

### **CUFFED (ADJ.)**

Dating/In a relationship  
"It's cuffing season."

### **BOP (ADJ.)**

Good/cool  
"I love that dress, it's a total bop!"



WILL KNIGHT PARESH DAVE

BUSINESS MAR 29, 2023 12:01 PM

## In Sudden Alarm, Tech Doyens Call for a Pause on ChatGPT

Tech luminaries, renowned scientists, and Elon Musk warn of an “out-of-control race” to develop and deploy ever-more-powerful AI systems.

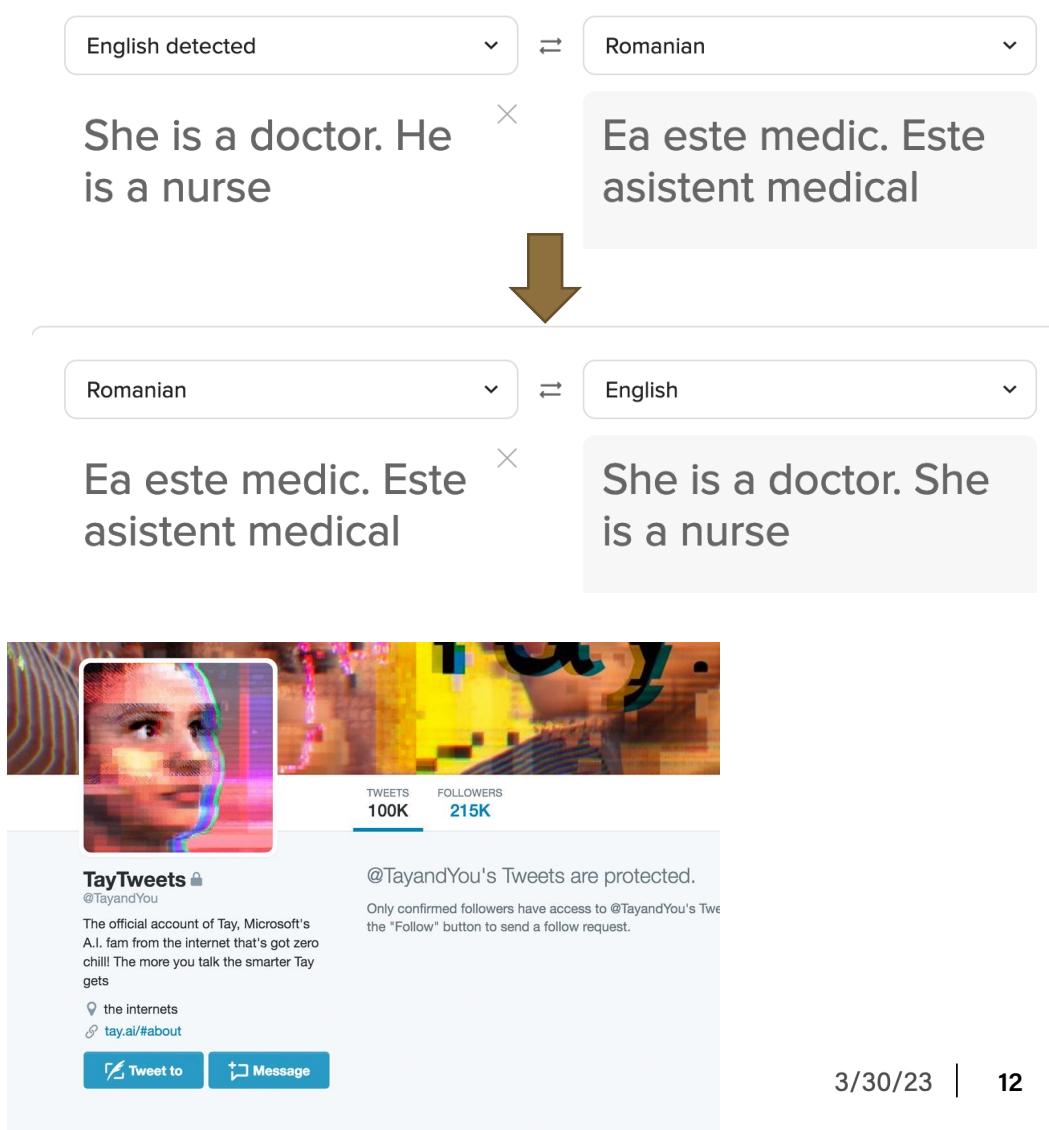


TIME

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

A thumbnail image from TIME magazine shows several people working at computer monitors in what appears to be a call center or data processing environment.



English detected ↗ Romanian

She is a doctor. He is a nurse

Ea este medic. Este asistent medical

Romanian ↗ English

Ea este medic. Este asistent medical

She is a doctor. She is a nurse

A screenshot of a neural network interface demonstrating machine translation. It shows two rows of text pairs. The top row translates "She is a doctor. He is a nurse" from English to Romanian ("Ea este medic. Este asistent medical"). The bottom row translates "Ea este medic. Este asistent medical" from Romanian back to English ("She is a doctor. She is a nurse"). A large orange arrow points downwards between the two rows, indicating the direction of translation. The interface includes dropdown menus for language selection and a search bar.

**TayTweets** • @TayandYou

The official account of Tay, Microsoft's A.I. fan from the internet that's got zero chill! The more you talk the smarter Tay gets

the internets  
tay.ai/#about

TWEETS 100K FOLLOWERS 215K

@TayandYou's Tweets are protected.  
Only confirmed followers have access to @TayandYou's Tweets. Click the "Follow" button to send a follow request.

[Tweet to](#) [Message](#)

3/30/23 | 12

## Fairness



- Have our algorithms been tested on diverse data?
- Are our algorithms equally performant on all groups?

## Accountability



- How are we holding ourselves accountable if AI makes a mistake?
- What recourse is available and how do we ensure the issue doesn't happen again?

## Transparency



Ethics



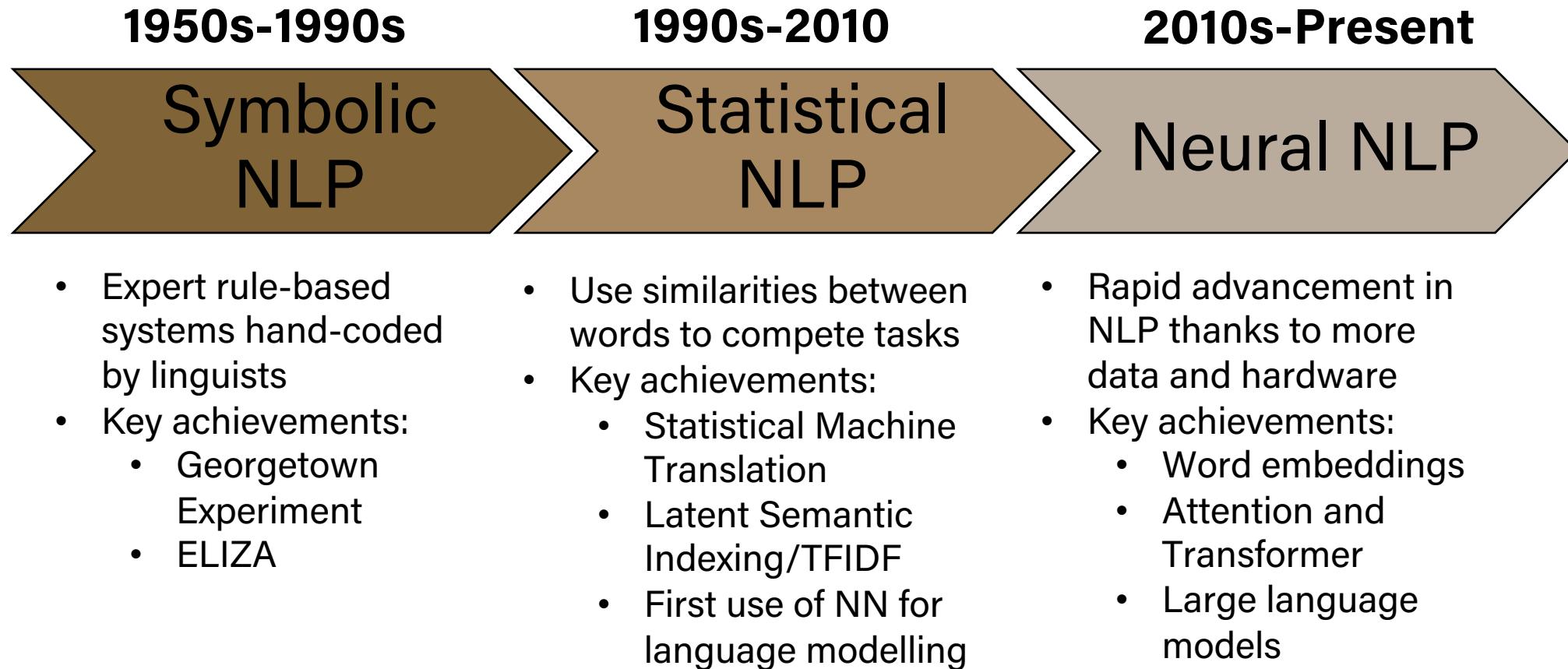
- Are we transparent about how we're using AI?
- Do we allow outside researchers or watchdogs to examine our use of AI?
- Are the applications we're using AI for ethical?

# Mechanics & Evolution

# *Quick Review of Key ML Terms*

- **Unsupervised Learning:** neural network used patterns in unlabeled data, e.g., clustering
- **Supervised Learning:** Labelled data used to help the model “learn” how to do a particular task, e.g., classification
- **Transfer learning:** Reusing general information learned from a previous task for a new task; speeds up training and reduces data requirements
  - **Pre-training:** General learning
  - **Fine-tuning:** Tweaking the pre-trained model for a downstream task

# *Evolution of NLP*



# *Representational Learning: Text as Numbers*

## A 4-dimensional embedding

<b>cat</b> =>	1.2	-0.1	4.3	3.2
<b>mat</b> =>	0.4	2.5	-0.9	0.5
<b>on</b> =>	2.1	0.3	0.1	0.4
...				...

Embedding ideally captures:

- Meaning of words
- Similarities/differences between words
- Contextual meaning of words

# *Representational Learning: Text as Numbers*

## **“You shall know a word by the company it keeps”**

- A word's meaning can be understood based on the words it frequently appears close to
- Use the many contexts of a word to build up its representation

*...government debt problems turning into banking crises as happened in 2009...*

*...saying that Europe needs unified banking regulation to replace the hodgepodge...*

*...India has just given its banking system a shot in the arm...*

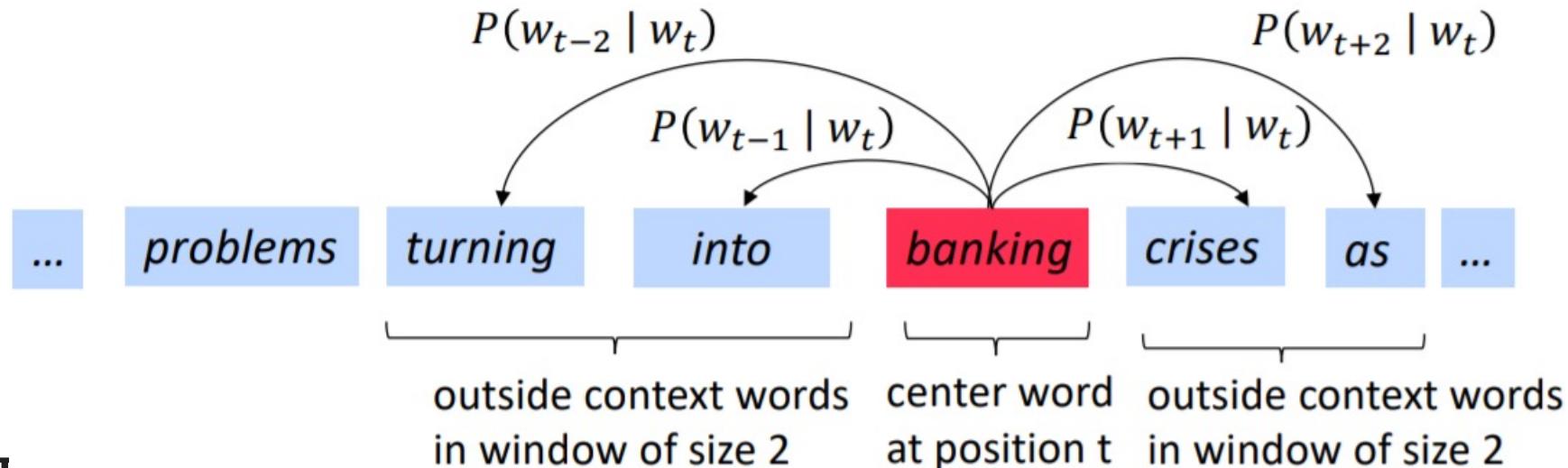


These context words will represent **banking**

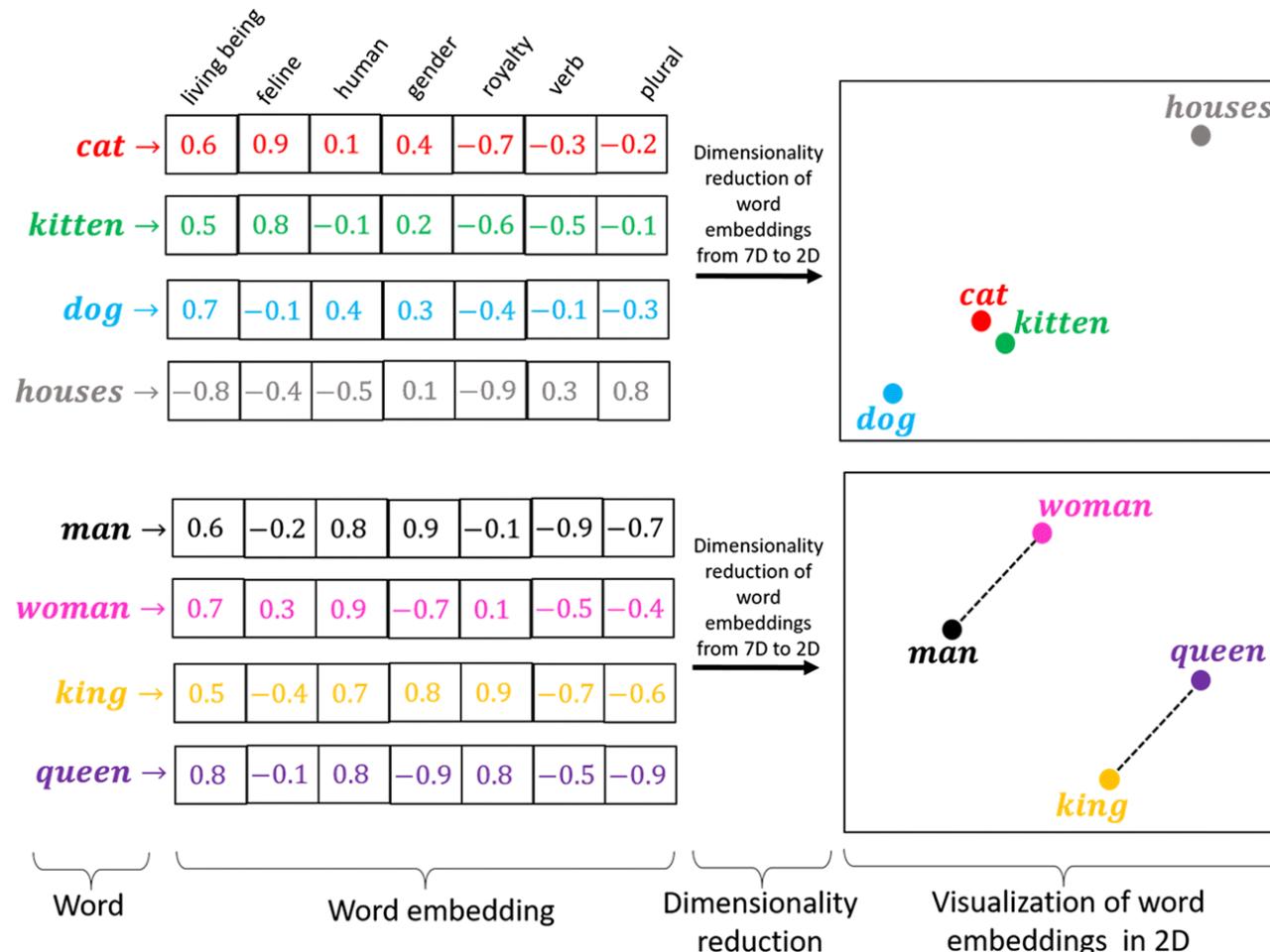
# Representational Learning: Text as Numbers

## How are embeddings actually created?

- Unsupervised training on large corpus of text
  - Randomly initialized vectors for each word in corpus
  - Train to maximize similarity (dot product) of target and context word vectors (Word2Vec)
  - Add global statistics about corpus (co-occurrence probabilities) to improve embeddings (GloVe)



# Representational Learning: Text as Numbers



## Word2Vec/GloVe Embeddings Capture:

- ✓ Meaning of words
- ✓ Similarities/differences between words
- ✗ Contextual meaning of words

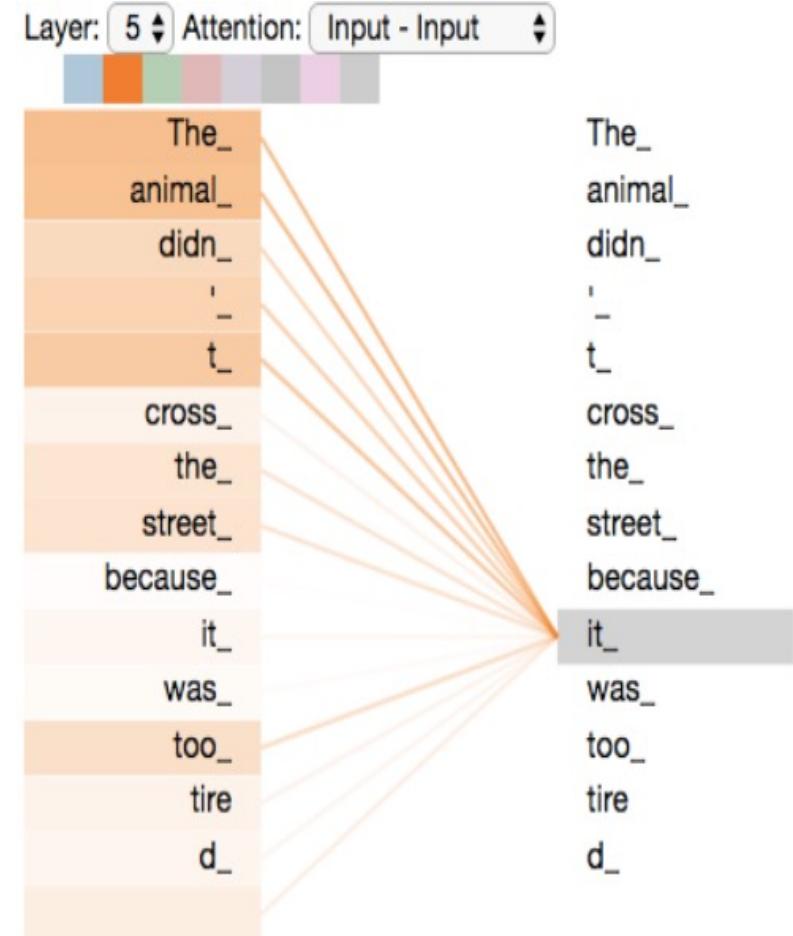
# *Representational Learning: Text as Numbers*

## How are conditional embeddings actually created?

- Unsupervised pre-training on large corpus of text
- Run pre-processed text through the pre-trained model to dynamically generate embeddings for each word → “fine-tuning” the embeddings
- ELMo/BERT/other conditional embeddings satisfy all of our requirements!

“After stealing money from the **bank vault**,  
the **bank robber** was seen fishing on the  
Mississippi **river bank**.”

Each use of “bank” has a different embedding



## Attributes of text data:

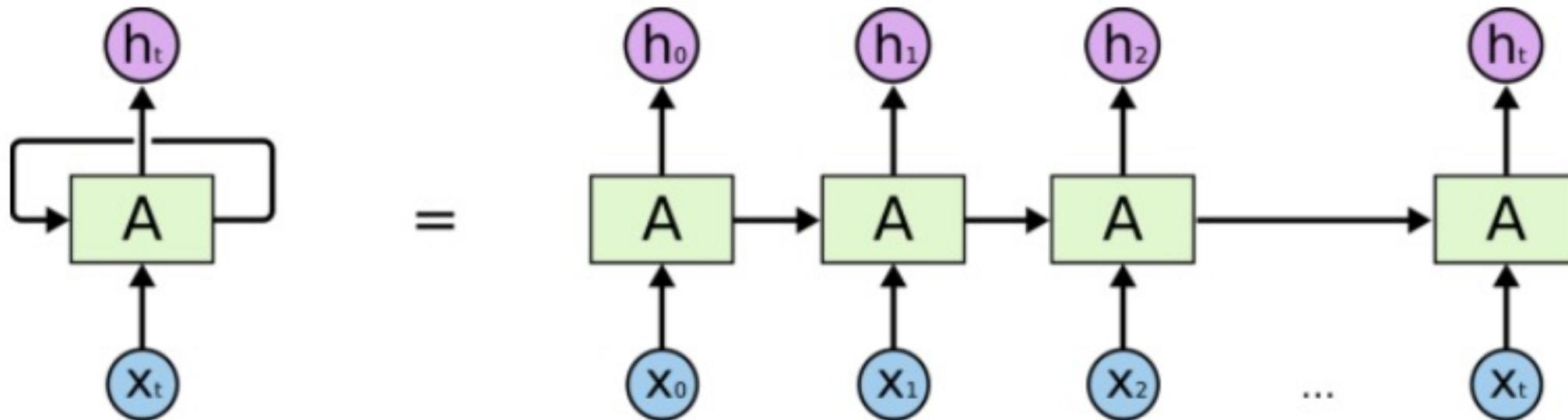
- Word order encodes meaning
- The most relevant information for understanding a word may be near or far away
- Words have differential importance

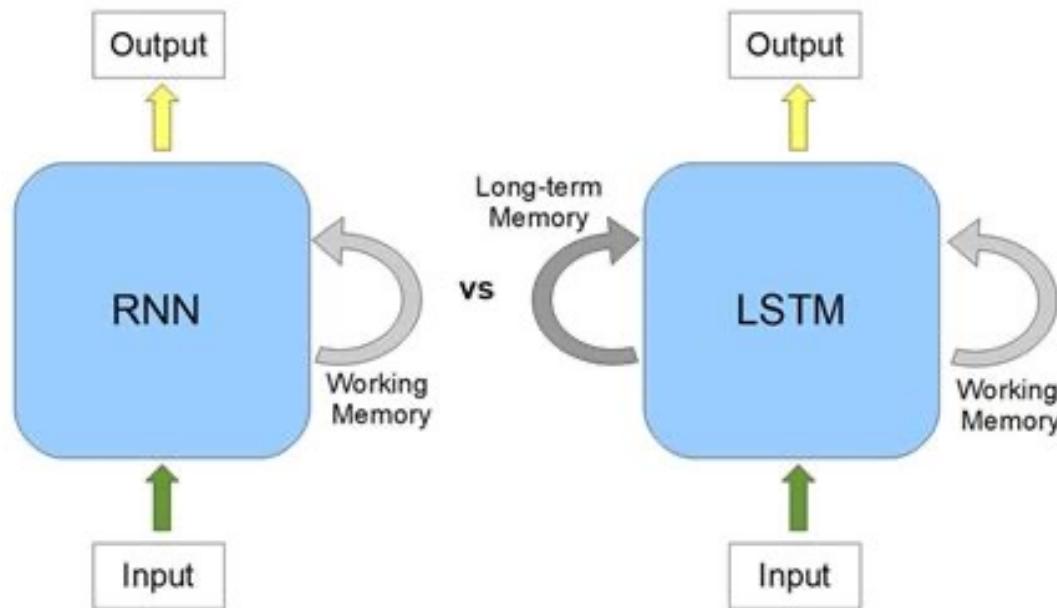
The animal didn't cross the street because **it** was too **tired**.

-----versus-----

The animal didn't cross the street because **it** was too **wide**.

## Recurrent Neural Networks





## Long Short-Term Memory

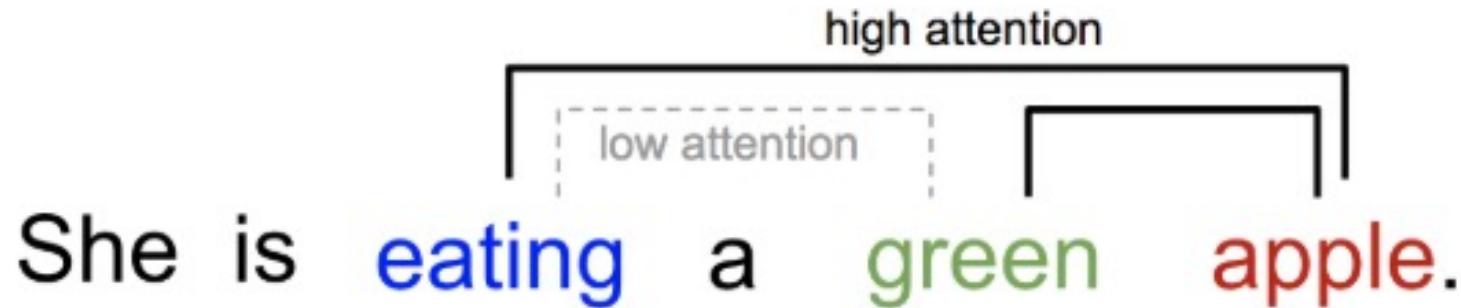
### 1<sup>st</sup> attempt: RNNs and LSTMs

#### Key Features:

- ✓ Word order encodes meaning
- The most relevant information for understanding a word may be near or far away
- Words have differential importance

## Attention

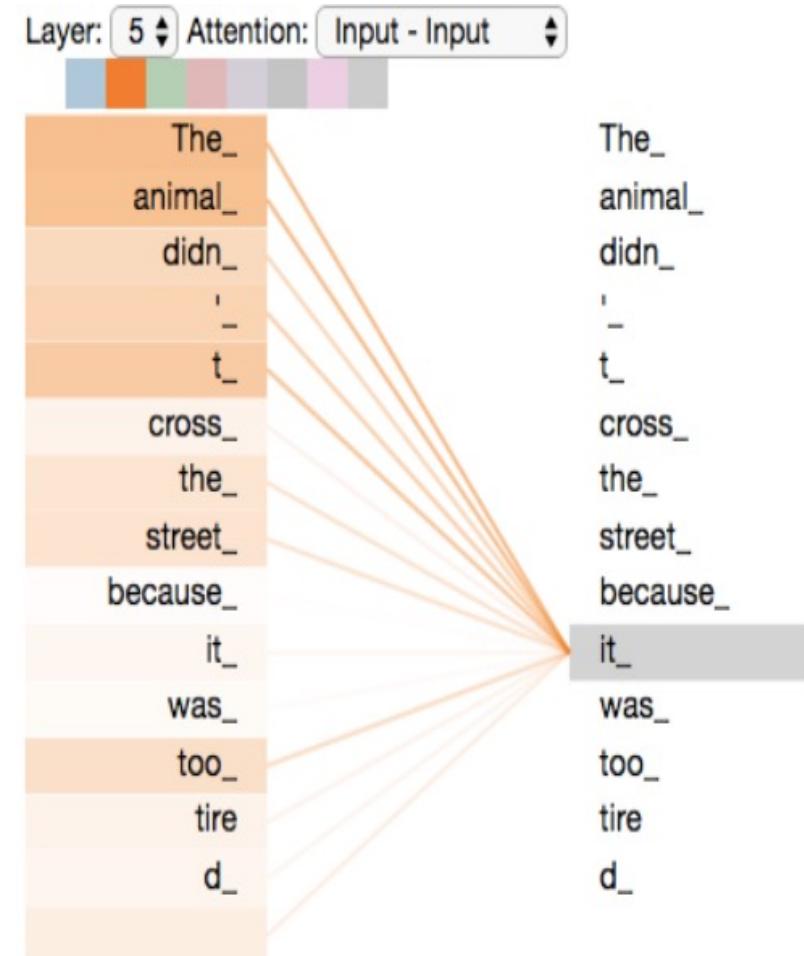
To predict a word, use only the most relevant parts of the input text



## Attention

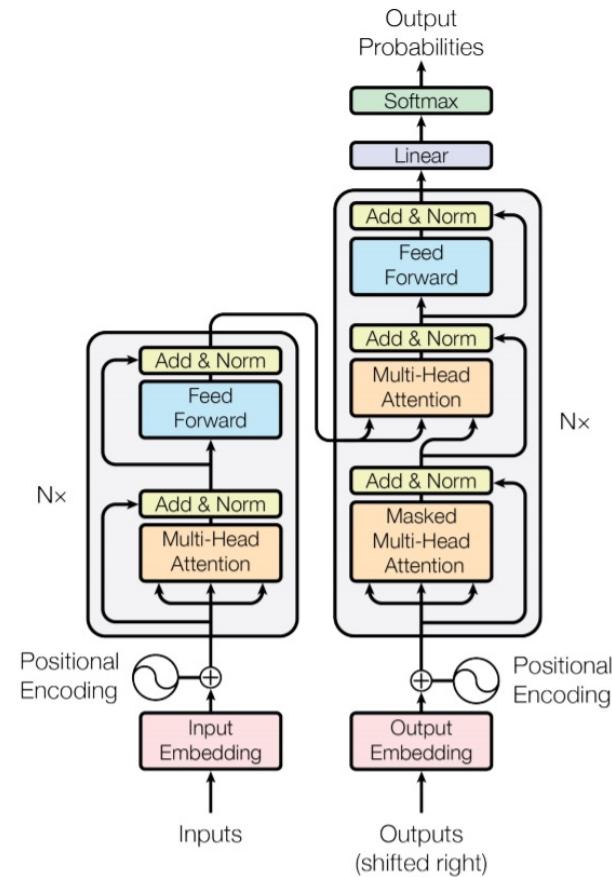
**Self-attention:** relating different positions of a single sequence to itself to compute attention

- Processes each word in the input one at a time (**query**) by looking at all other words in the input sequence (**keys**) for clues that can help the model learn a better encoding for the query (**values**)



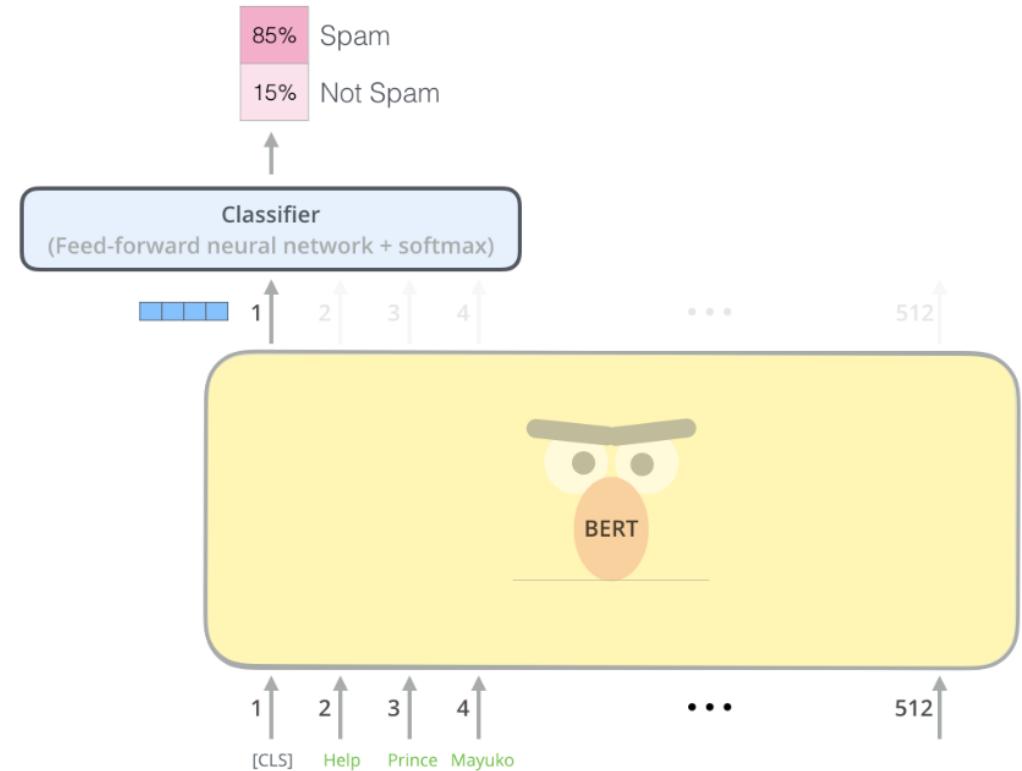
## 2<sup>nd</sup> Attempt: Transformers

- **"Attention is all you Need"**: Seminal NLP paper that presented SOTA results by only using attention mechanisms without recurrence
- Basis of BERT and GPT SOTA models
- Many times faster and parallelizable
- Addresses the issue of differential importance



## BERT Model

- **Bidirectional Encoder Representations from Transformers**
  - Uses both left and right context for training (bi-directional)
  - Language representation model (pre-trained) that can be fine-tuned for a variety of NLP tasks
  - Based on transformer architecture



## GPT

- **Generative Pre-trained Transformer**
  - Uni-directional
  - Draws from corpus of information to generate best results for query
  - Based on transformer architecture



## BERT

vs

## GPT

### Pros

- Suitable for a wide range of NLP tasks
- Can be adapted to a specific domain/task and can learn new information through fine-tuning
- Open-source model

### Cons

- Requires more effort to develop a model

### Pros

- Suitable for a wide range of tasks
- Lower barrier to entry because no fine-tuning required
- Trained on massive corpus of information

### Cons

- Cannot be fine-tuned or learn anything new

# *Review: Key Terms*

- **Embedding:** way to numerically represent the meaning of a word, sentence, paragraph, etc.
- **Language Model:** probabilistic model of words and phrases in a language
- **Transformers:** Architecture based on attention mechanisms
- **Representation Learning:** Based on pattern discovery
- **Generative AI:** Utilizes knowledge to generate data/information

# Getting Started



## Hugging Face

- User-friendly resource to help you get started with NLP
- Transformers python package
- Models/datasets for variety of different tasks



**The AI community building the future.**

Build, train and deploy state of the art models powered by the reference open source in machine learning.

# Getting Started

GLUE Benchmark  
includes many tasks  
to assess general  
language  
understanding

- Linguistic Acceptability
- Paraphrasing
- Semantic Similarity
- Question-Answering
- Sentiment
- And more!

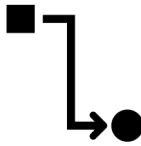
The screenshot shows the GLUE Leaderboard interface. At the top, there are links for GLUE and SuperGLUE. Below the header is a search bar and a "Leaderboard" section. The main content is a table with 9 rows, each representing a model entry. The columns include Rank, Name, Model, URL (with a download icon), and scores for CoLA, SST-2, MRPC, STS-B, QQP, MNLI-m, MNLI-mm, QNLI, RTE, WNLI, and AX. The table highlights the top-performing models from Microsoft and JDExplore.

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	71.5	92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
8	HFL IFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
9	PING-AN Omni-Sinicic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2

## NLP in 3 easy steps



Load pre-trained  
model and data



Tokenize and  
pre-process data



Fine-tune model  
and save checkpoint

HuggingFace Tutorial:

[https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/text\\_classification.ipynb](https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/text_classification.ipynb)

# *Practical Tips*

- Newer NLP approaches generally don't require much manual preprocessing (e.g. older methods like stop word removal and stemming/lemmatization are not usually needed).
- There are a lot of specialized pre-trained models out there (e.g., BERT pre-trained models available for twitter data, medical data, etc.) that are an easy win in boosting performance –research and experiment whenever possible

## General ML:

- As with any ML – garbage in, garbage out! Take the time to ensure sufficient data quality
- The “best” new model may not be the best for you – keep in mind the benefits of using a more established model with more support and try these first
- GPUs not strictly necessary if you are only fine-tuning or doing inference, but will definitely speed up tasks

# *Resources*

- Stanford CS224N NLP with Deep Learning Course:  
<https://youtu.be/rmVRLeJRkl4>
- Variety of excellent explainers on key concepts/architectures:  
<https://jalammar.github.io/>

# ***THANK YOU***

Contact: [rcac-help@purdue.edu](mailto:rcac-help@purdue.edu)