Data Science for Public Policy

Large Language Models

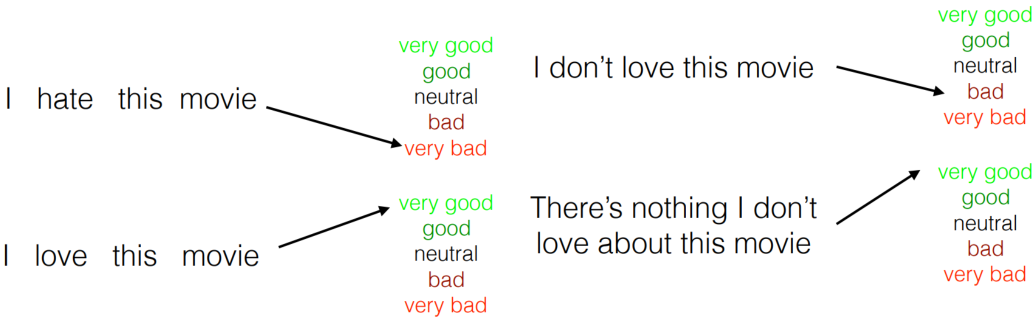
ETHZ Zurich

03/04/2025

Text as Data Recap

- Last time we covered *traditional* NLP methods
  - Bag of words
  - Topic Model
  - Word Embeddings
  - Dependency parsing
- While these methods can solve many research tasks, they still have some limitations

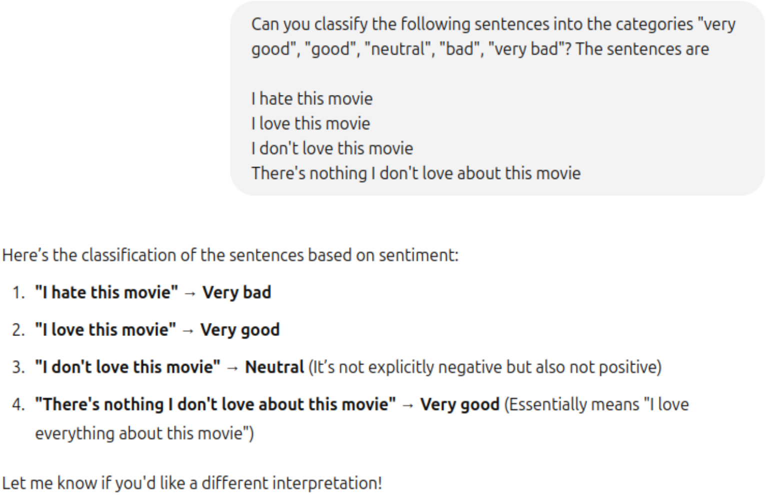
The Classic Sentence Classification Problem



Source: Graham Neubig slides.

- bag-of-words or word embedding models won't capture the importance of "don't love" or "nothing I don't love"
- N-grams have a large feature space (especially with 4-grams) and don't share information across similar words/n-grams.

How about Large Language Models?



Can you classify the following sentences into the categories "very good", "good", "neutral", "bad", "very bad"? The sentences are

I hate this movie
I love this movie
I don't love this movie
There's nothing I don't love about this movie

Here’s the classification of the sentences based on sentiment:

1. **"I hate this movie"** → **Very bad**
2. **"I love this movie"** → **Very good**
3. **"I don't love this movie"** → **Neutral** (It’s not explicitly negative but also not positive)
4. **"There's nothing I don't love about this movie"** → **Very good** (Essentially means "I love everything about this movie")

Let me know if you'd like a different interpretation!

What are the Most Popular LLMs?

- Type of generative AI that produces text
- AI systems trained to predict the next word given preceding text
- Typically fine-tuned to follow human instructions and generate responses aligned with human preferences
- Based on deep neural networks with billions of parameters
- Built on transformer models (with attention mechanisms, which endogenously assign varying degrees of importance to different words)

→ Each of these elements has value on its own: let's take a step back

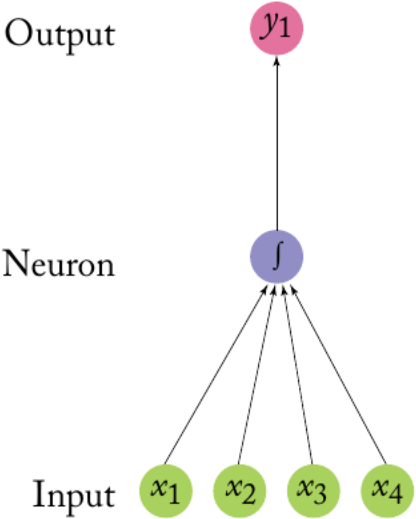
Large Language Models from the Start: Sequence Data

- The real breakthrough from deep learning for NLP:
  - moving from bag-of-words representations to sequence representations.
  - Rather than inputting **counts over words *x***, take as input a **sequence of tokens** $\{w_1, ..., w_t, ... w_n\}$.

→ The position and order of appearance of each token matters!

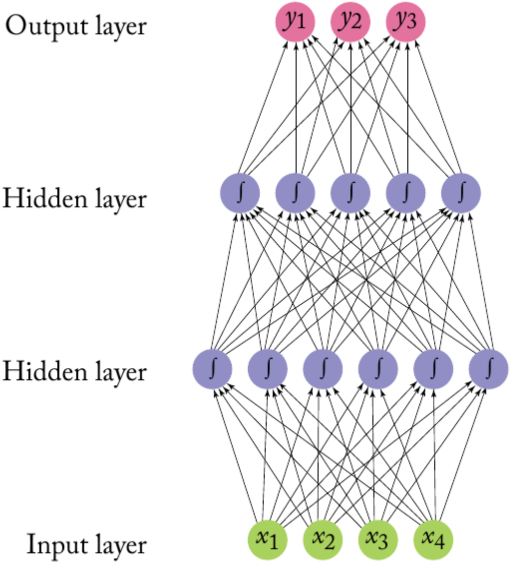
- Main advantages compared to *traditional* methods:
  - Learn different meanings depending on the context (e.g., “sun light” vs. “the rock is light”
  - Directly capture dependencies between words (even if distant)
  - Can be used as direct input in deep-learning models

A "Neuron"



- Transforms the input to be processed efficiently by applying the dot product:
    - multiplies each input by a learned weight (parameter or coefficient)
    - sums these products
- applies a non-linear "activation function" to the sum
    - enables the model to learn complex patterns and relationships
    - (e.g., the $\int$ shape indicates a sigmoid transformation)
- passes the output.

Neural Net Example (Multi-Layer Perceptron)



- A multilayer perceptron (also called a feed-forward network or sequential model) stacks neurons horizontally and vertically.
- alternatively, think of it as a stacked ensemble of logistic regression models.
- this vertical stacking is the “deep” in “deep learning”!

Example: Universal Sentence Encoder (USE)

- Produces embeddings that are sensitive to word order and context
- Neural net architecture with embeddings pre-trained on:
  - Identifying co-occuring sentences
  - Identifying message-response pairs (Henderson et al 2017)
  - Some supervised learning tasks (see Cer et al 2018)

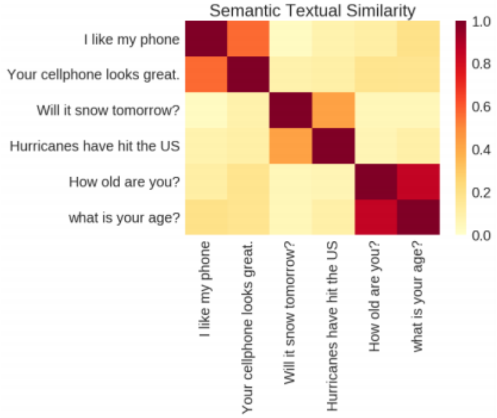


Figure 1: Sentence similarity scores using embeddings from the universal sentence encoder.

Multilingual Encoders

- The multilingual sentence encoder (**MUSE**) expands the USE model to sixteen languages, in a single embedding model.
    - Trained on a similar array of tasks in all languages, so that it can be used out-of-the-box.
- Facebook's LASER encoder produces vectors for 90 languages with a single model.
    - bidirectional LSTM architecture
    - trained on multilingual machine translation task

Another Step towards Modern LLMs: Transformers

- Since a 2017 paper (Vaswani et al, 2017), deep learning for NLP has been transformed by a new class of models: **transformers.**
- Standard approach:
  - represent documents as counts over words/phrases, shares over topics, or the average of word embeddings.
- Recurrent neural nets can process whole documents:
  - but they have to sweep through the whole document processing it word-by-word in sequence, so they learn too slowly.
- Transformers overcome these limitations. Intuitively, using self-attention:
  - they process all words at once, making them faster and more efficient
  - they can capture long-range dependencies between words, improving performance on complex NLP tasks

Autoencoding Language Models: BERT

- BERT = Bidirectional Encoder Representations from Transformers
- Training task: Masked language modeling
  - 15% of words masked (randomly)
  - if masked: replace with [MASK] 80% of the time, a random token 10% of the time, and left unchanged 10% of the time.
  - model has to predict the original word.
- Corpus:
  - 800M words from English books (modern work, from unpublished authors), by Zhuet al (2015)
  - 2.5B words of text from English Wikipedia articles (without markup)
  - Architecture: The largest BERT model has ≈ 340M parameters to learn (a stack of transformer blocks with a self-attention layer and an MLP.)
- BERT attention observes all tokens in the sequence, reads backwards and forwards (bidirectional)
- It can be fine-tuned and achieve great results on many tasks, e.g., text classification

Application: Climate-Related Corporate Disclosures (Bingler, Kraus, and Leippold 2021)

- Fine-tunes RoBERTa ("Robust BERT") to classify texts related to corporate climate disclosures (using hand-annotated sample).

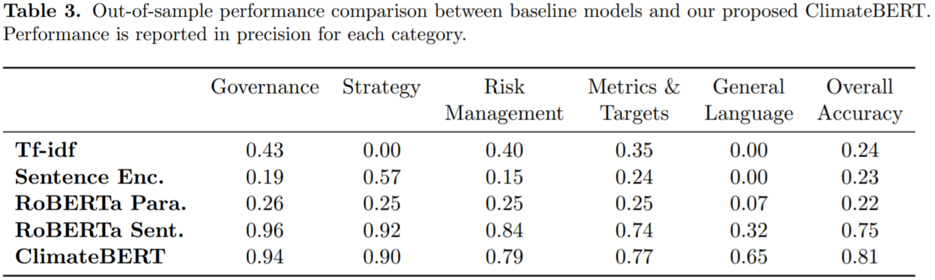


Table 3. Out-of-sample performance comparison between baseline models and our proposed ClimateBERT. Performance is reported in precision for each category.

	Governance	Strategy	Risk Management	Metrics & Targets	General Language	Overall Accuracy
Tf-idf	0.43	0.00	0.40	0.35	0.00	0.24
Sentence Enc.	0.19	0.57	0.15	0.24	0.00	0.23
RoBERTa Para.	0.26	0.25	0.25	0.25	0.07	0.22
RoBERTa Sent.	0.96	0.92	0.84	0.74	0.32	0.75
ClimateBERT	0.94	0.90	0.79	0.77	0.65	0.81

- model applied to large sample, shows that most disclosures are about more subjective / less verifiable aspects of climate disclosures.

Autoregressive Language Models: GPT

- e.g. GPT = “Generative Pre-Trained Transformer”:
- Task: guess the next token having read all the previous ones.
- Unlike BERT, during training, attention heads only view previous tokens, not subsequent tokens.
- Ideal for text generation.

GPT: Main Training Steps

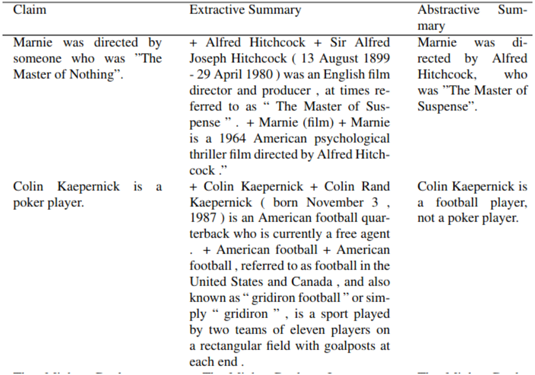
1. Pre-Training: Calculate conditional probability distribution over words given the preceding words, based on its training data
   - Self-Supervised Learning: the model is fed text fragments; parameters are adjusted to predict continuation on terabytes of data
   - Neural nets learns language structure: syntactic structures, relationships between words and concepts they represent, context of sentences
2. Instruction Fine-Tuning: improving the model to follow human instructions
   - According to pretrained model, a likely continuation of "What's your name?" may be "And how old are you?" → This is not the answer we want
   - Supervised learning: Feeding the model millions of examples of how to respond to thousands of different instructions for tasks like summarization, answering questions etc.
3. Reinforcement learning: improving the model by incorporating human feedback
   - Feedback from human raters tells the model how different responses compare
   - Makes model better aligned with human preferences
   - Noisy process (for example, it is part of the reason why LLMs have learned to sound authoritative even when they hallucinate)

GPT = Generative Pre-Trained Transformer

- GPT-1: the first autoregressive transformer model (2018)
  - trained on the Books corpus.
  - train on a language modeling task, as well as a multi-task that adds a supervised learning task.
- GPT-2 (2019):
  - all articles linked from Reddit with at least 3 upvotes (8 million documents, 40 GB of text)
  - dispense with supervised learning task, make some other architectural adjustments
  - make model much bigger
- GPT-3 (2020):
  - use an even bigger corpus (Common Crawl, WebText2, Books1, Books2 and Wikipedia)
  - make model much, much bigger

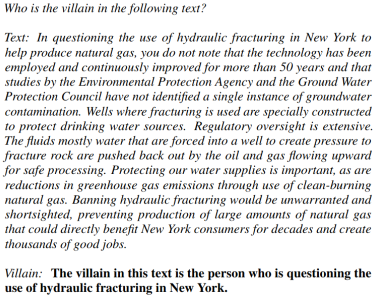
Applications: GPT-3

Stammbach and Ash (2020),
“e-FEVER: Explanations and Summaries for Automated Fact Checking”



Claim	Extractive Summary	Abstractive Summary
Marnie was directed by someone who was "The Master of Nothing".	+ Alfred Hitchcock + Sir Alfred Joseph Hitchcock (13 August 1899 - 29 April 1980) was an English film director and producer , at times referred to as “ The Master of Suspense ” . + Marnie (film) + Marnie is a 1964 American psychological thriller film directed by Alfred Hitchcock .”	Marnie was directed by Alfred Hitchcock, who was "The Master of Suspense".
Colin Kaepernick is a poker player.	+ Colin Kaepernick + Colin Rand Kaepernick (born November 3 , 1987) is an American football quarterback who is currently a free agent . + American football + American football , referred to as football in the United States and Canada , and also known as “ gridiron football ” or simply “ gridiron ” , is a sport played by two teams of eleven players on a rectangular field with goalposts at each end .	Colin Kaepernick is a football player, not a poker player.

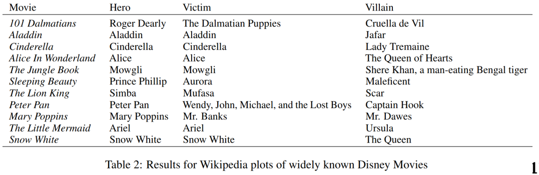
Stammbach, Antoniak, and Ash (2022),
"Heroes, Villains, and Victims, and GPT-3"



Who is the villain in the following text?

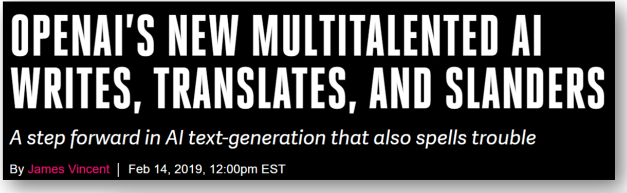
Text: In questioning the use of hydraulic fracturing in New York to help produce natural gas, you do not note that the technology has been employed and continuously improved for more than 50 years and that studies by the Environmental Protection Agency and the Ground Water Protection Council have not identified a single instance of groundwater contamination. Wells where fracturing is used are specially constructed to protect drinking water sources. Regulatory oversight is extensive. The fluids mostly water that are forced into a well to create pressure to fracture rock are pushed back out by the oil and gas flowing upward for safe processing. Protecting our water supplies is important, as are reductions in greenhouse gas emissions through use of clean-burning natural gas. Banning hydraulic fracturing would be unwarranted and shortsighted, preventing production of large amounts of natural gas that could directly benefit New York consumers for decades and create thousands of good jobs.

Villain: **The villain in this text is the person who is questioning the use of hydraulic fracturing in New York.**



Movie	Hero	Victim	Villain
101 Dalmatians	Roger Dearly	The Dalmatian Puppies	Cruella de Vil
Aladdin	Aladdin	Aladdin	Jafar
Cinderella	Cinderella	Cinderella	Lady Tremaine
Alice In Wonderland	Alice	Alice	The Queen of Hearts
The Jungle Book	Mowgli	Mowgli	Shere Khan, a man-eating Bengal tiger
Sleeping Beauty	Prince Phillip	Aurora	Maleficent
The Lion King	Simba	Mufasa	Scar
Peter Pan	Peter Pan	Wendy, John, Michael, and the Lost Boys	Captain Hook
Mary Poppins	Mary Poppins	Mr. Banks	Mr. Dawes
The Little Mermaid	Ariel	Ariel	Ursula
Snow White	Snow White	Snow White	The Queen

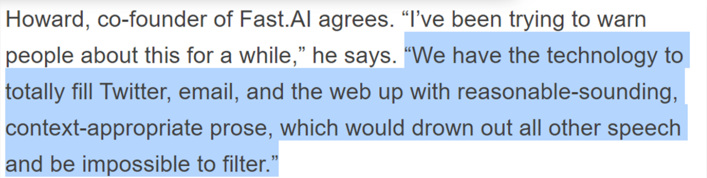
Table 2: Results for Wikipedia plots of widely known Disney Movies



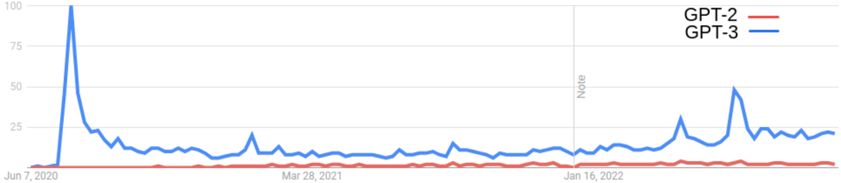
OPENAI'S NEW MULTITALENTED AI WRITES, TRANSLATES, AND SLANDERS

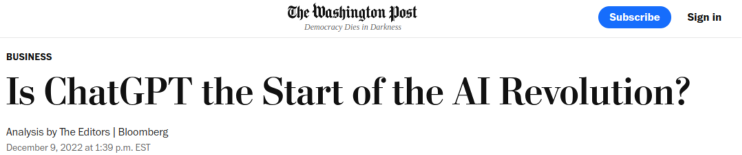
A step forward in AI text-generation that also spells trouble

By James Vincent | Feb 14, 2019, 12:00pm EST



Howard, co-founder of Fast.AI agrees. "I've been trying to warn people about this for a while," he says. "We have the technology to totally fill Twitter, email, and the web up with reasonable-sounding, context-appropriate prose, which would drown out all other speech and be impossible to filter."





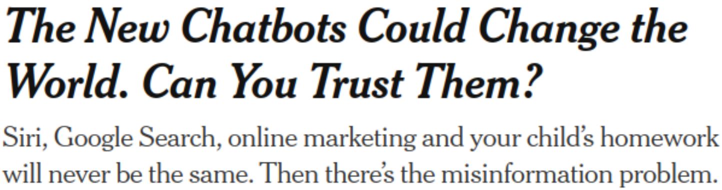
The Washington Post
Democracy Dies in Darkness

Subscribe Sign in

BUSINESS

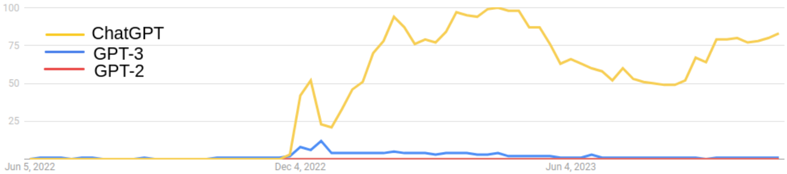
Is ChatGPT the Start of the AI Revolution?

Analysis by The Editors | Bloomberg
December 9, 2022 at 1:39 p.m. EST



The New Chatbots Could Change the World. Can You Trust Them?

Siri, Google Search, online marketing and your child's homework will never be the same. Then there's the misinformation problem.



GPT = Generative Pre-Trained Transformer

- GPT-1: the first autoregressive transformer model (2018)
- GPT-2 (2019)
- GPT-3 (2020)
- GPT-3.5 (2022):
  - Subclass of GPT-3 trained on data up to June 2021
  - Incorporates the base model on which ChatGPT is fine-tuned + is optimized for chat
- GPT-4 (2023):
  - Multimodal model: can take also images as inputs
  - Trained in two stages:
    1. token prediction (like other GPT models)
    2. **reinforcement learning with human feedback**
  - Much, much, much bigger model

A Lot of Potential Applications!

- Fixing OCR errors in digitized text
- Extract structured information from text articles
- Classify articles, e.g., distinguish between believing or skeptical of climate change
- Interpreting old proverbs
- Conducting interview
- Summarize information
- Rank documents on non-trivial metrics through pairwise comparisons
- More on this next week

Limitations

- Hallucinations: LLMs can easily make things up, which limits how much we can leverage their knowledge base
- Weaker in analytic concepts due to their *nature* as LLM (Dreyfuss and Raux, 2025)
- Limits to reproducibility
- Bias (from training data and human feedback)
- Privacy concerns (from training data and human feedback)

The Environmental Impact of LLMs

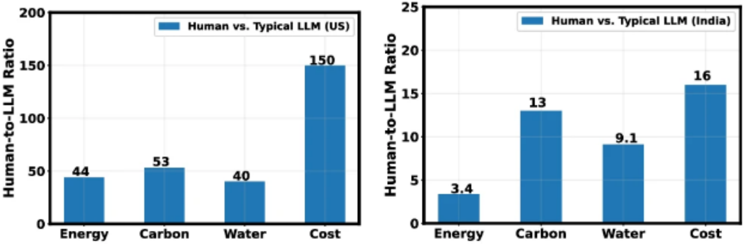
- LLMs require a significant amount of energy at all stages
  - ChatGPT is estimated to consume the energy of 33,000 households
  - Google carbon emissions have increased by 50% to build and sustain new data centers (source)

The Environmental Impact of LLMs

- LLMs require a significant amount of energy at all stages
  - ChatGPT is estimated to consume the energy of 33,000 households
  - Google carbon emissions have increased by 50% to build and sustain new data centers (source)
  - While training GPT-4, OpenAI consumed around 6% of West Des Moins (Iowa) district in a month

The Environmental Impact of LLMs

- LLMs require a significant amount of energy at all stages
  - ChatGPT is estimated to consume the energy of 33,000 households
  - Google carbon emissions have increased by 50% to build and sustain new data centers (source)
  - While training GPT-4, OpenAI consumed around 6% of West Des Moins (Iowa) district in a month
- Recent studies have highlighted how LLMs are a more efficient alternative to human labor (Ren et al., 2025)



(a) Typcial LLM (U.S.)

(b) Typical LLM (India)

The Environmental Impact of LLMs

- LLMs require a significant amount of energy at all stages
  - ChatGPT is estimated to consume the energy of 33,000 households
  - Google carbon emissions have increased by 50% to build and sustain new data centers (source)
  - While training GPT-4, OpenAI consumed around 6% of West Des Moins (Iowa) district in a month
- Recent studies have highlighted how LLMs are a more efficient alternative to human labor (Ren et al., 2025). However:
  - LLMs size is growing ⇒ energy demand will increase
  - In the short term, LLMs are unlikely to substitute humans fully

So What Do We Do?

- Given the fast evolution of these technologies it's hard to make recommendations
- **Users:** Responsible use of GenAI
  - Revert to Google query for simple web searches (more reliable and consumes 10 times less than ChatGPT per query)
  - LLMs are not a one-size-fits-all solution: tailor your NLP method to the research question
- **Companies**: invest in research for *greener* AI systems