Textual analysis using Python

Session 2

Ties de Kok
University of Washington

UNIVERSITY *of* WASHINGTON

FOSTER SCHOOL OF BUSINESS

Session #2 – what will we cover?

- “Left-over” topics
  - Topic modeling
  - Text similarity
  - Word embeddings

- Non-generative LLMs → BERT models

- Generative LLMs → GPT models

Conceptual introduction

Many major developments in last 5 years

1. **Transfer learning**

➔Kicked off by word embeddings (e.g., word2vec)

2. **Larger & larger models + training datasets**

➔Made possible by GPU advancements

3. **Ability to retain (“understand”) textual context**

➔ Made possible by algorithmic advancements

Key developments

> **Word / sentence embeddings**

> **LLMs → Transformer models**

Application #1: Non-generative → BERT

Application #2: Generative → GPT

Transfer learning

Without transfer learning:

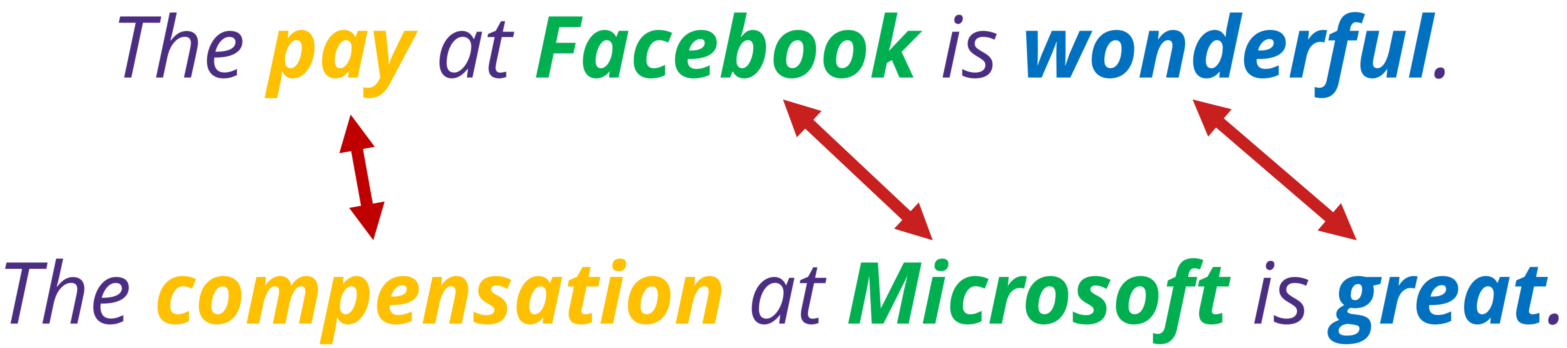
Every problem → start from scratch

With transfer learning:

Step 1: extract natural language relations from a large body of text

Step 2: use step 1 as a starting point for specific task

Word embeddings



*The **pay** at **Facebook** is **wonderful**.*

*The **compensation** at **Microsoft** is **great**.*

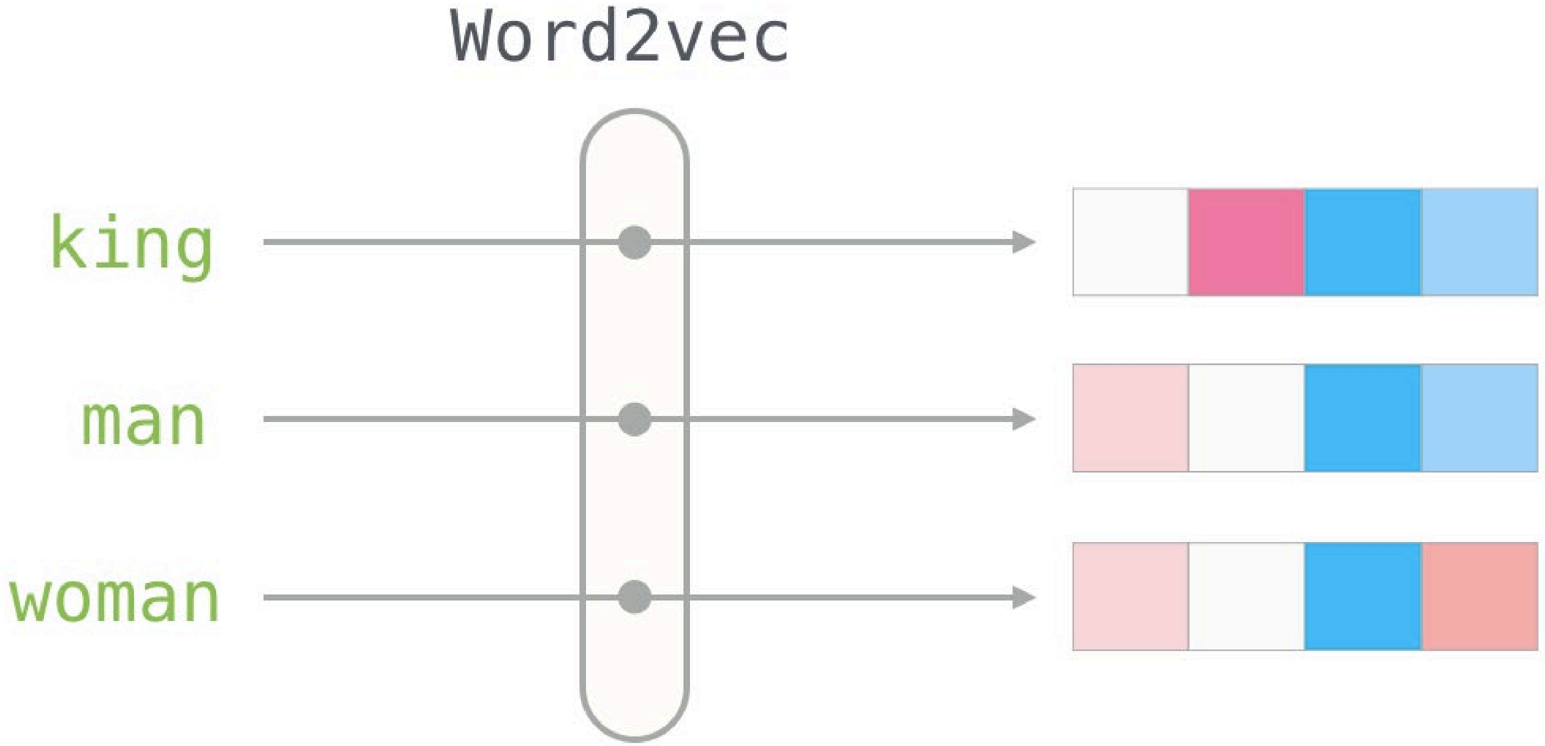
Bag of words perspective:

→ All words are independent

Word embeddings:

→ Treat similar words as similar

Word embeddings

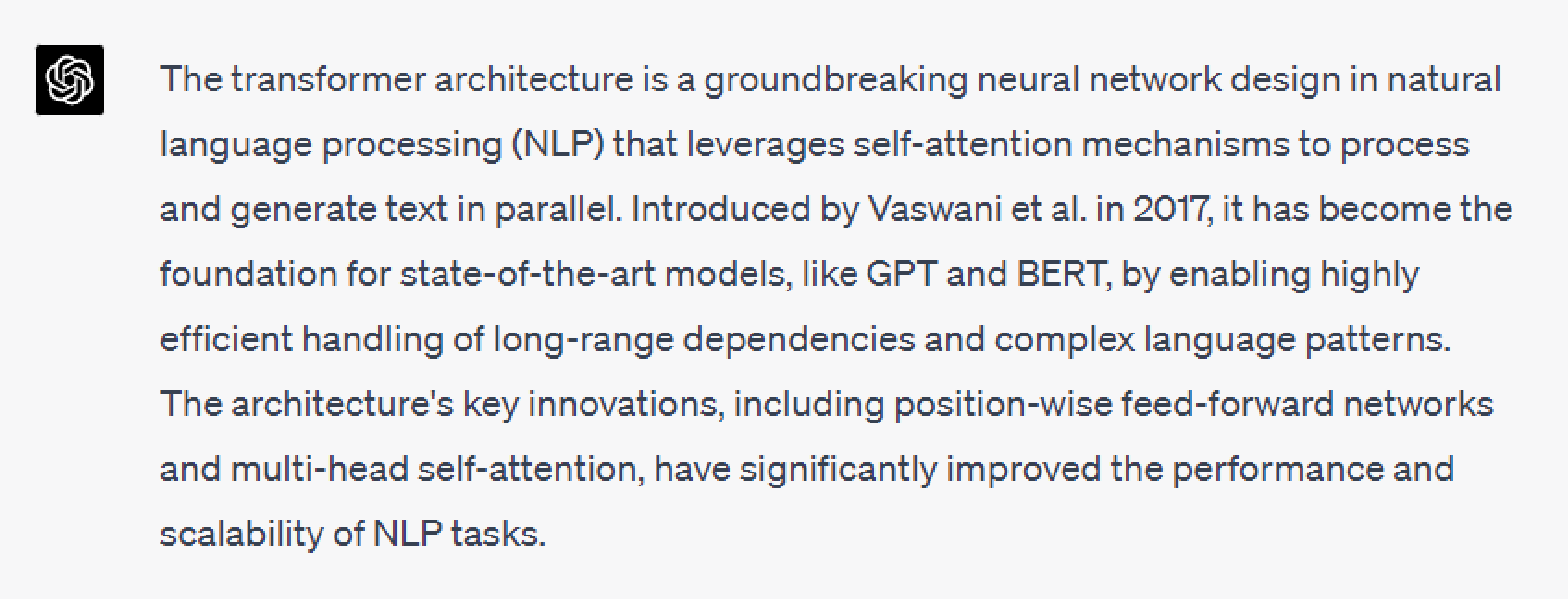


Source: https://jalammar.github.io/illustrated-word2vec/

Transformers

Transformers

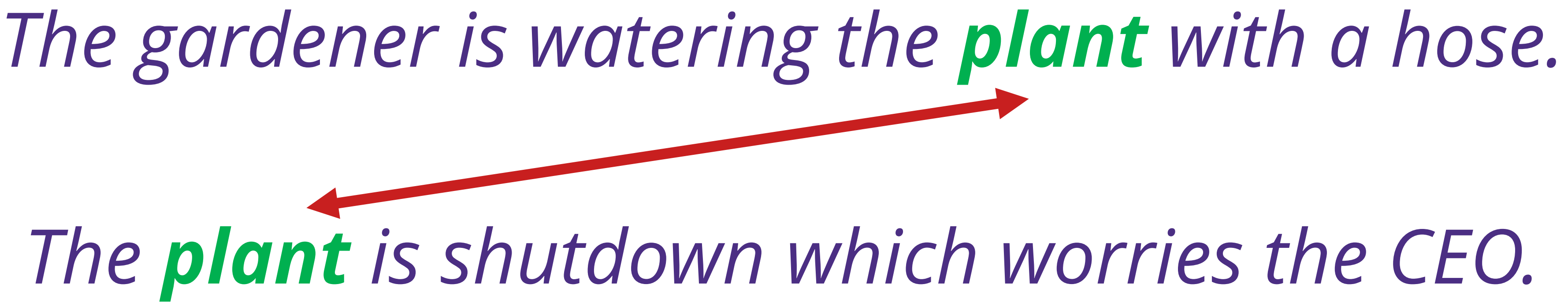
What are transformers???



The transformer architecture is a groundbreaking neural network design in natural language processing (NLP) that leverages self-attention mechanisms to process and generate text in parallel. Introduced by Vaswani et al. in 2017, it has become the foundation for state-of-the-art models, like GPT and BERT, by enabling highly efficient handling of long-range dependencies and complex language patterns. The architecture's key innovations, including position-wise feed-forward networks and multi-head self-attention, have significantly improved the performance and scalability of NLP tasks.

My simpler take on transformers:

The transformer architecture makes it possible to efficiently model and represent complex relationships in text.

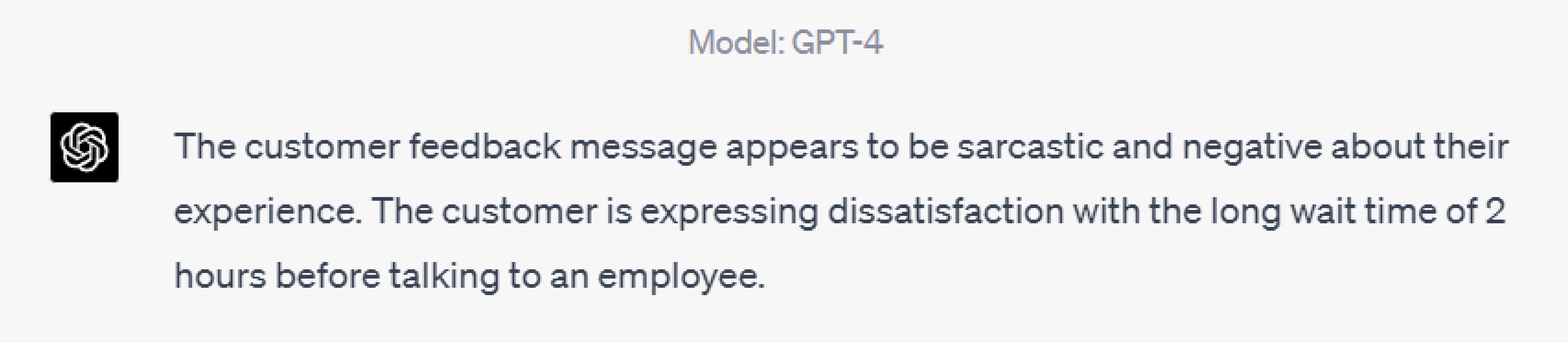


*The gardener is watering the **plant** with a hose.*

*The **plant** is shutdown which worries the CEO.*

Context matters:

Calling the customer service hotline was a wonderful experience, I love ***waiting 2 hours before*** *talking to an employee, really wonderful.*



Model: GPT-4

The customer feedback message appears to be sarcastic and negative about their experience. The customer is expressing dissatisfaction with the long wait time of 2 hours before talking to an employee.

Application #1: non-generative

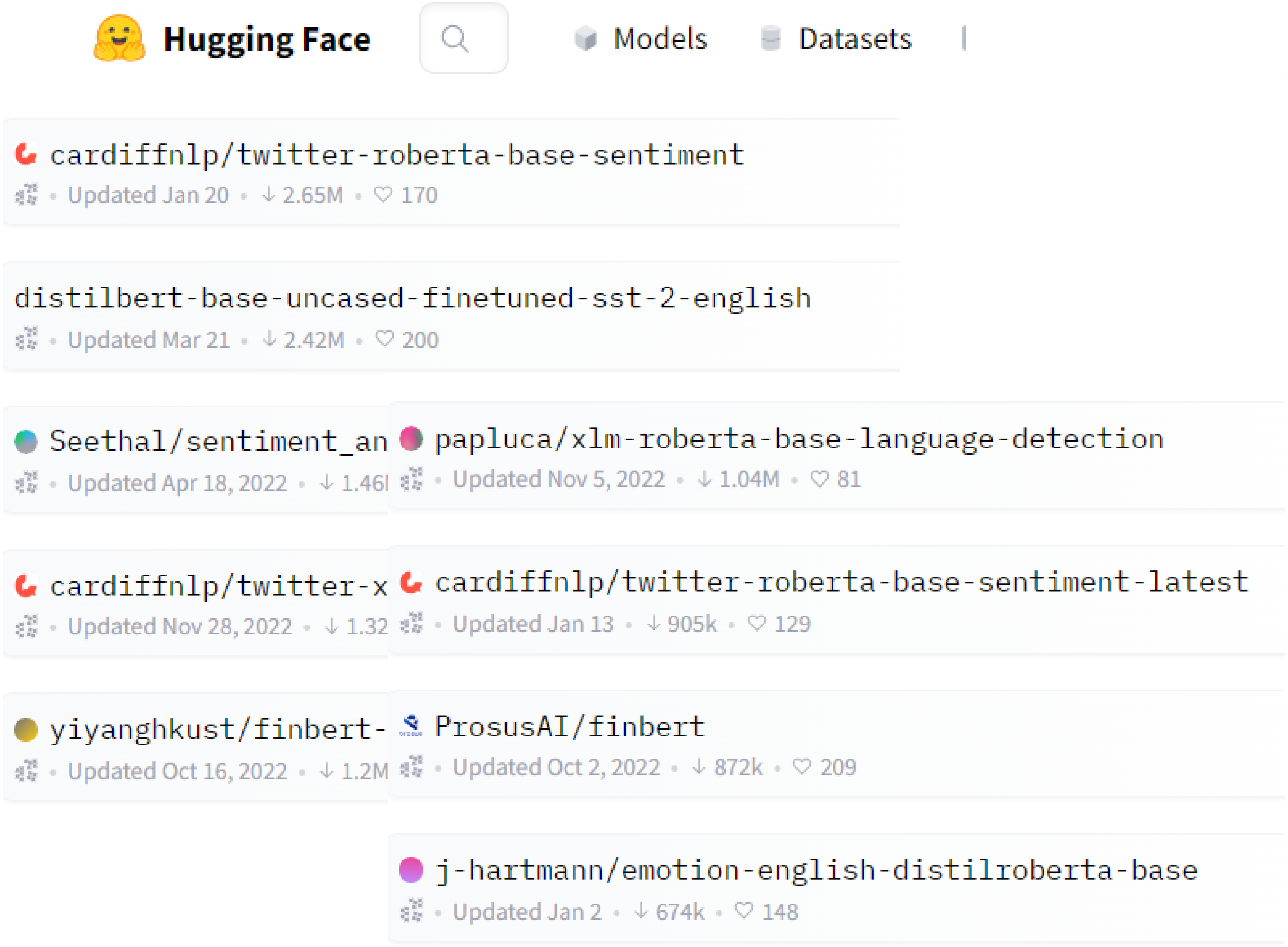
Application #1:

➔ transformers make ML pipelines more powerful

➔ This is the “BERT wave”

BERT models provide a better way to represent text in a machine learning pipeline

BERT model explosion



Hugging Face

Models

Datasets

cardiffnlp/twitter-roberta-base-sentiment
Updated Jan 20 • ↓ 2.65M • ♡ 170

distilbert-base-uncased-finetuned-sst-2-english
Updated Mar 21 • ↓ 2.42M • ♡ 200

Seethal/sentiment_an
Updated Apr 18, 2022 • ↓ 1.46l

papluca/xlm-roberta-base-language-detection
Updated Nov 5, 2022 • ↓ 1.04M • ♡ 81

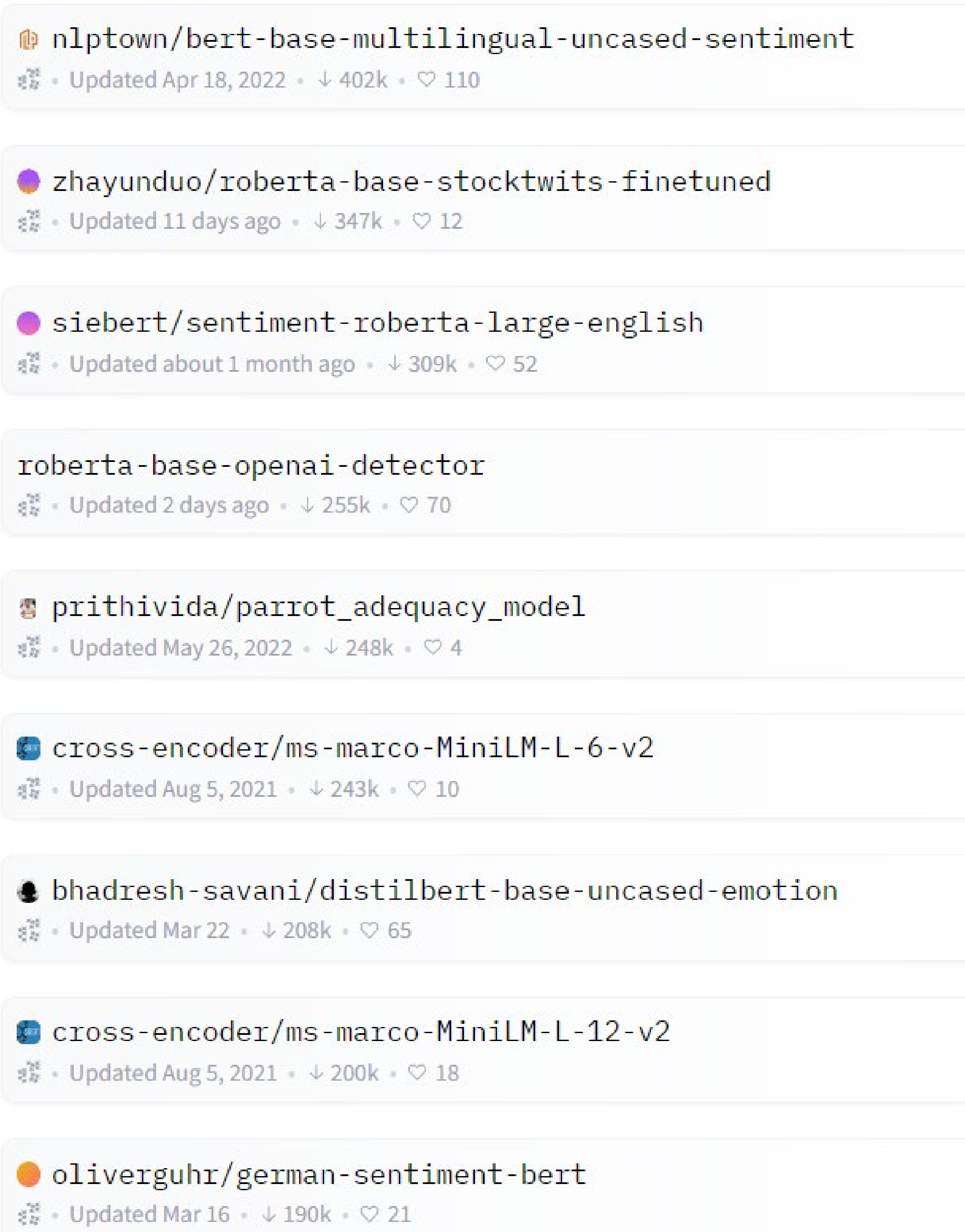
cardiffnlp/twitter-x
Updated Nov 28, 2022 • ↓ 1.32

cardiffnlp/twitter-roberta-base-sentiment-latest
Updated Jan 13 • ↓ 905k • ♡ 129

yiyanghkust/finbert-
Updated Oct 16, 2022 • ↓ 1.2M

ProsusAI/finbert
Updated Oct 2, 2022 • ↓ 872k • ♡ 209

j-hartmann/emotion-english-distilroberta-base
Updated Jan 2 • ↓ 674k • ♡ 148



nlptown/bert-base-multilingual-uncased-sentiment
Updated Apr 18, 2022 • ↓ 402k • ♡ 110

zhayunduo/roberta-base-stocktwits-finetuned
Updated 11 days ago • ↓ 347k • ♡ 12

siebert/sentiment-roberta-large-english
Updated about 1 month ago • ↓ 309k • ♡ 52

roberta-base-openai-detector
Updated 2 days ago • ↓ 255k • ♡ 70

prithivida/parrot_adequacy_model
Updated May 26, 2022 • ↓ 248k • ♡ 4

cross-encoder/ms-marco-MiniLM-L-6-v2
Updated Aug 5, 2021 • ↓ 243k • ♡ 10

bhadresh-savani/distilbert-base-uncased-emotion
Updated Mar 22 • ↓ 208k • ♡ 65

cross-encoder/ms-marco-MiniLM-L-12-v2
Updated Aug 5, 2021 • ↓ 200k • ♡ 18

oliverguhr/german-sentiment-bert
Updated Mar 16 • ↓ 190k • ♡ 21

https://huggingface.co/models

Application #2: generative

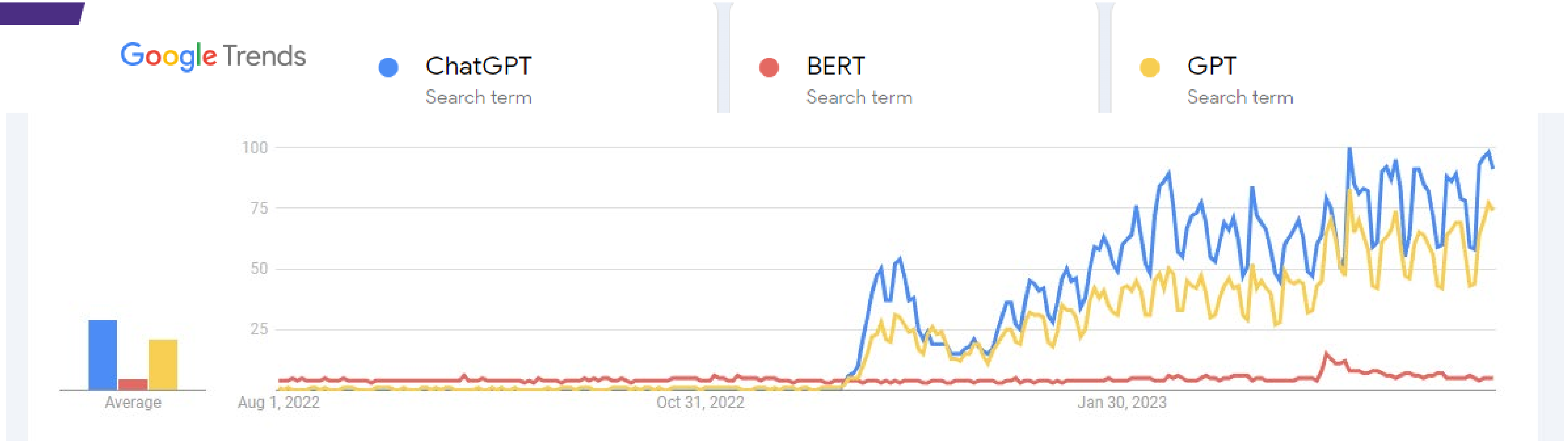
Application #2:

➔ transformers can yield models that generate text

➔ This is the "GPT wave"

Large GPT models provide a way of interaction that is closer to human interaction.

GPT model explosion



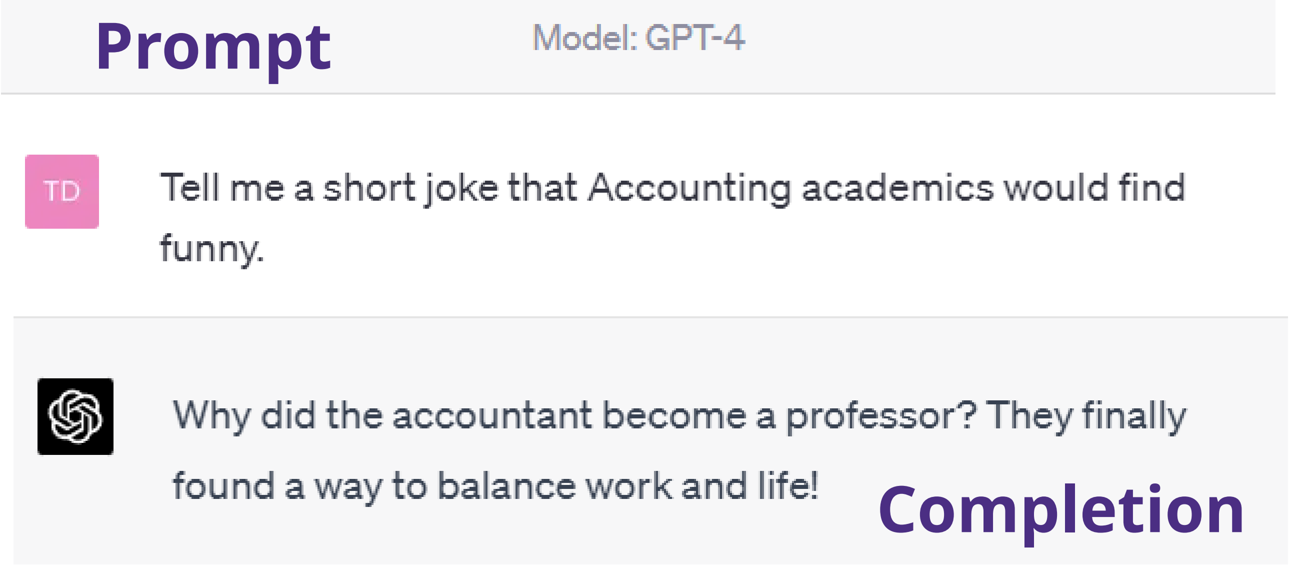
The Jobs Most Exposed to ChatGPT

New study finds that AI tools could more quickly handle at least half of the tasks that auditors, interpreters and writers do now

ChatGPT passes exams from law and business schools

By Samantha Murphy Kelly, CNN Business

Updated 1:35 PM EST, Thu January 26, 2023



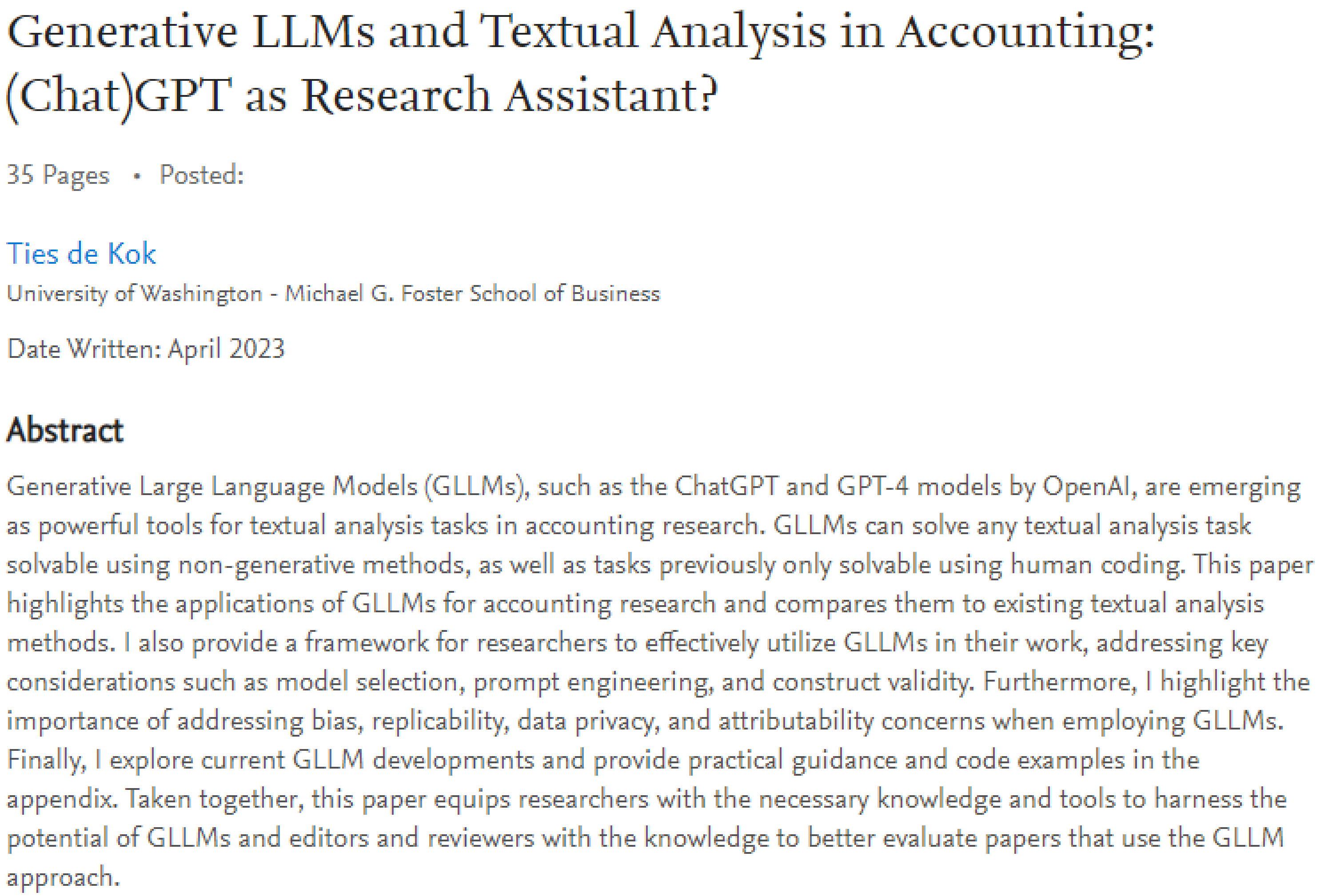
Prompt Model: GPT-4

TD Tell me a short joke that Accounting academics would find funny.

Why did the accountant become a professor? They finally found a way to balance work and life!

Completion

My paper



Generative LLMs and Textual Analysis in Accounting: (Chat)GPT as Research Assistant?

35 Pages • Posted:

Ties de Kok
University of Washington - Michael G. Foster School of Business

Date Written: April 2023

Abstract

Generative Large Language Models (GLLMs), such as the ChatGPT and GPT-4 models by OpenAI, are emerging as powerful tools for textual analysis tasks in accounting research. GLLMs can solve any textual analysis task solvable using non-generative methods, as well as tasks previously only solvable using human coding. This paper highlights the applications of GLLMs for accounting research and compares them to existing textual analysis methods. I also provide a framework for researchers to effectively utilize GLLMs in their work, addressing key considerations such as model selection, prompt engineering, and construct validity. Furthermore, I highlight the importance of addressing bias, replicability, data privacy, and attributability concerns when employing GLLMs. Finally, I explore current GLLM developments and provide practical guidance and code examples in the appendix. Taken together, this paper equips researchers with the necessary knowledge and tools to harness the potential of GLLMs and editors and reviewers with the knowledge to better evaluate papers that use the GLLM approach.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4429658

Why are GLLMs so useful?

Key benefit #1:

Task are communicated to GLLMs using just text.

Key benefit #2:

GLLMs can often handle tasks with little to no training.

Key benefit #3:

GLLMs are powerful and show state-of-the-art performance

Taken together: GLLMs can substitute for manual coding or complex machine learning pipelines ➔ **saves time**

Simple example:

Scenario: we have 25,000 employee reviews and we need to identify the ones that talk about *X. (e.g., work life balance, corporate culture, company reputation)*

Option #1: manually classify all reviews

Option #2: develop a word list or ML approach

Option #3: use a GLLM like ChatGPT

Option #3: using a GLLM like ChatGPT

Zero shot: does this review mention X?

Few shot: does this review mention X? Here are a few examples of yes and no.

Fine-tuning: manually code ~500 reviews and fine-tune the model on this small training dataset

Notebooks

- > Let’s walk through some code!

Thank you!

UNIVERSITY *of* WASHINGTON

W FOSTER SCHOOL OF BUSINESS