LLMs in Action: Scanning, Detecting, & Monitoring Fake & Unsafe Content

Floriana Zefi
Senior Machine Learning Scientist

Booking.com | pyladies

Amsterdam, 20 Nov 2024

Agenda

Workshop Part I , Floriana Zefi | (30 mins)

Introduction, Business Use Case, LLMs

Break , All | (15 mins)

Workshop Part II , Floriana Zefi | (50 mins)

Hands-On Workshop

Quiz and Closing, Floriana Zefi | (10 mins)

Quiz and Prize!

Part I

Introduction, Business Use-Case and LLMS

Floriana Zefi

PhD, Machine Learning Scientist

PhD in High Energy Astrophysics

Msc in Astrophysics

Msc in Physics

Machine Learning Scientist

Data Scientist

Assistant Professor

Business Use Case: Leveraging Technology for Fraud Prevention

100M
monthly active app users

318M+
verified guest reviews and
24/7
customer service
in **45**
languages and dialects

Since 2010, Booking.com has welcomed
4.5B+
guest arrivals

29M
total reported listings worldwide

7.2M
options in homes, apartments and other unique places to stay

140 offices in **70** countries over
5,000 employees in Amsterdam

174,000
destinations around the world

30
different types of places to stay, including homes, apartments, B&Bs, hostels, farm stays, bungalows, even boats, igloos and treehouses

Car hire available in **155+** countries and pre-booked taxis in over **600+** cities across **130+** countries

Fake Reviews

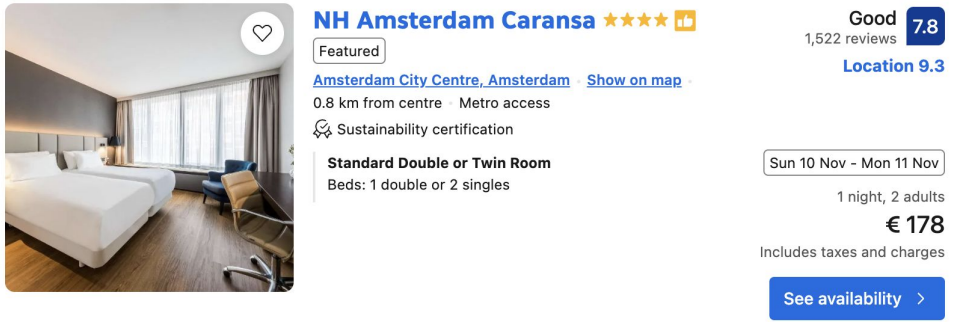
Fake reviews are fabricated customer feedback designed to mislead users about a product or service's quality

Types:

- **Positive Boosting**: Inflated reviews to raise ratings.
- **Negative Bombing**: Harmful reviews to damage a competitor.
- **Incentivized Reviews**: Paid or rewarded reviews with biased feedback.

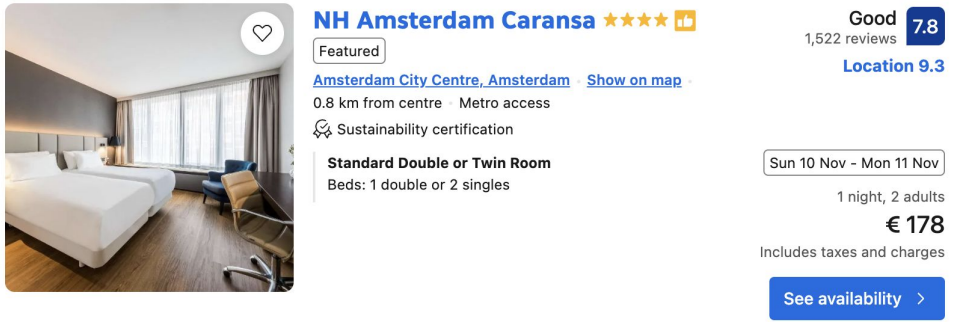
Impact:

- **For Businesses**: Loss of trust, unfair competition, and financial impact.
- **For Consumers**: Poor decisions, safety risks, and reduced confidence in reviews.



Why Detecting Fake Content Matters?

- **Platform Integrity**: Builds user trust by ensuring content authenticity
- **Improved User Experience**: Real reviews help customers make informed decisions
- **Compliance**: Many regions require platforms to prevent misleading content
- **Ethical Practices**: Levels the playing field for legitimate businesses



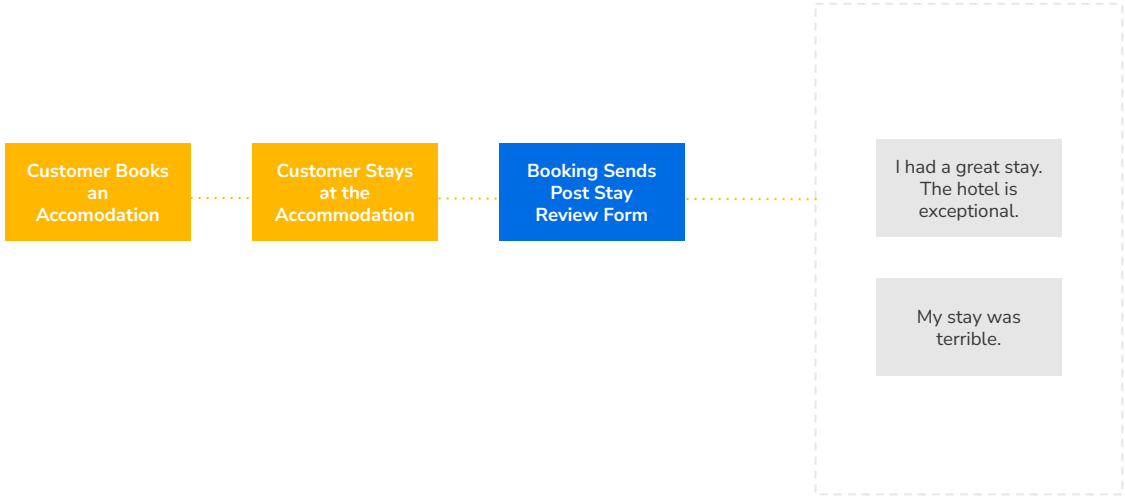
Problem Statement

Customer Books an Accomodation

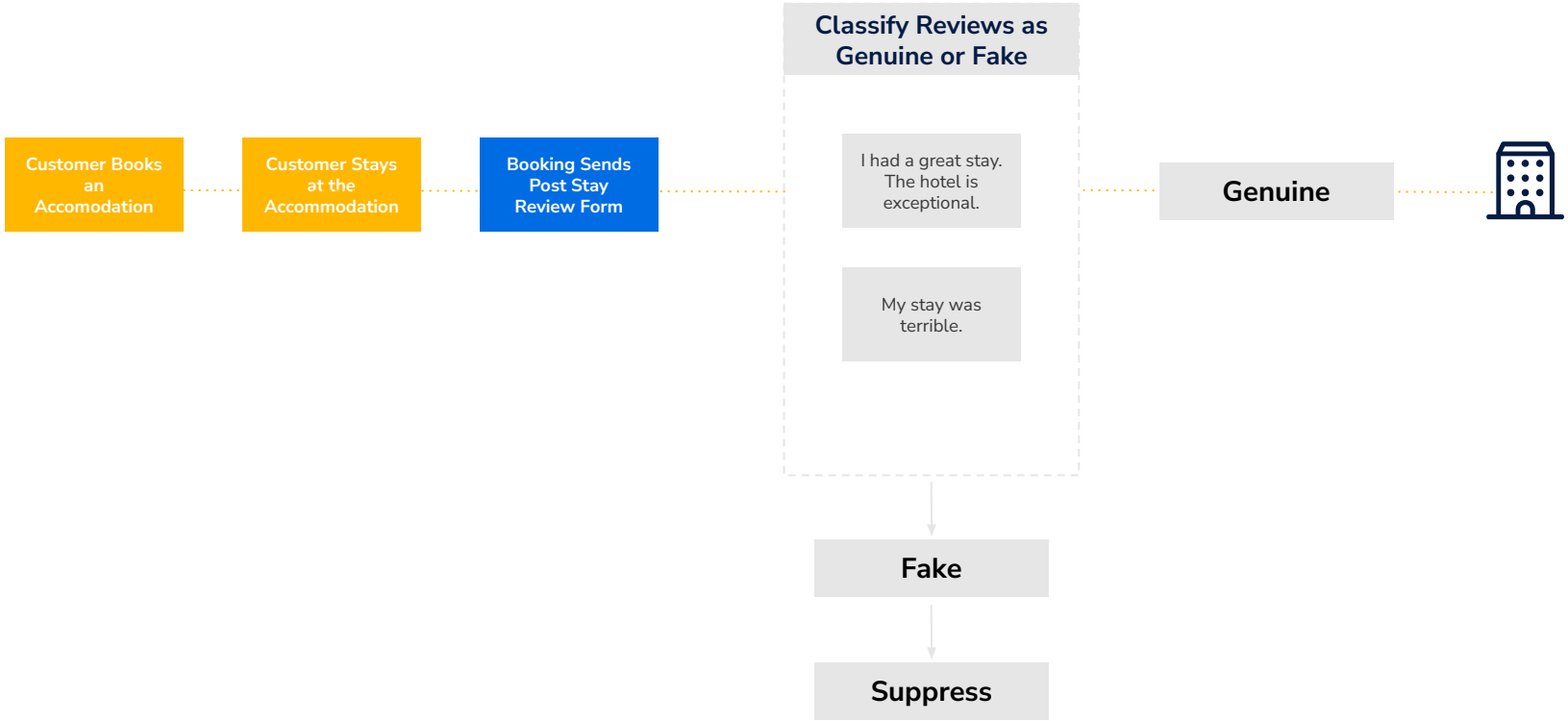
Customer Stays at the Accommodation

Booking Sends Post Stay Review Form

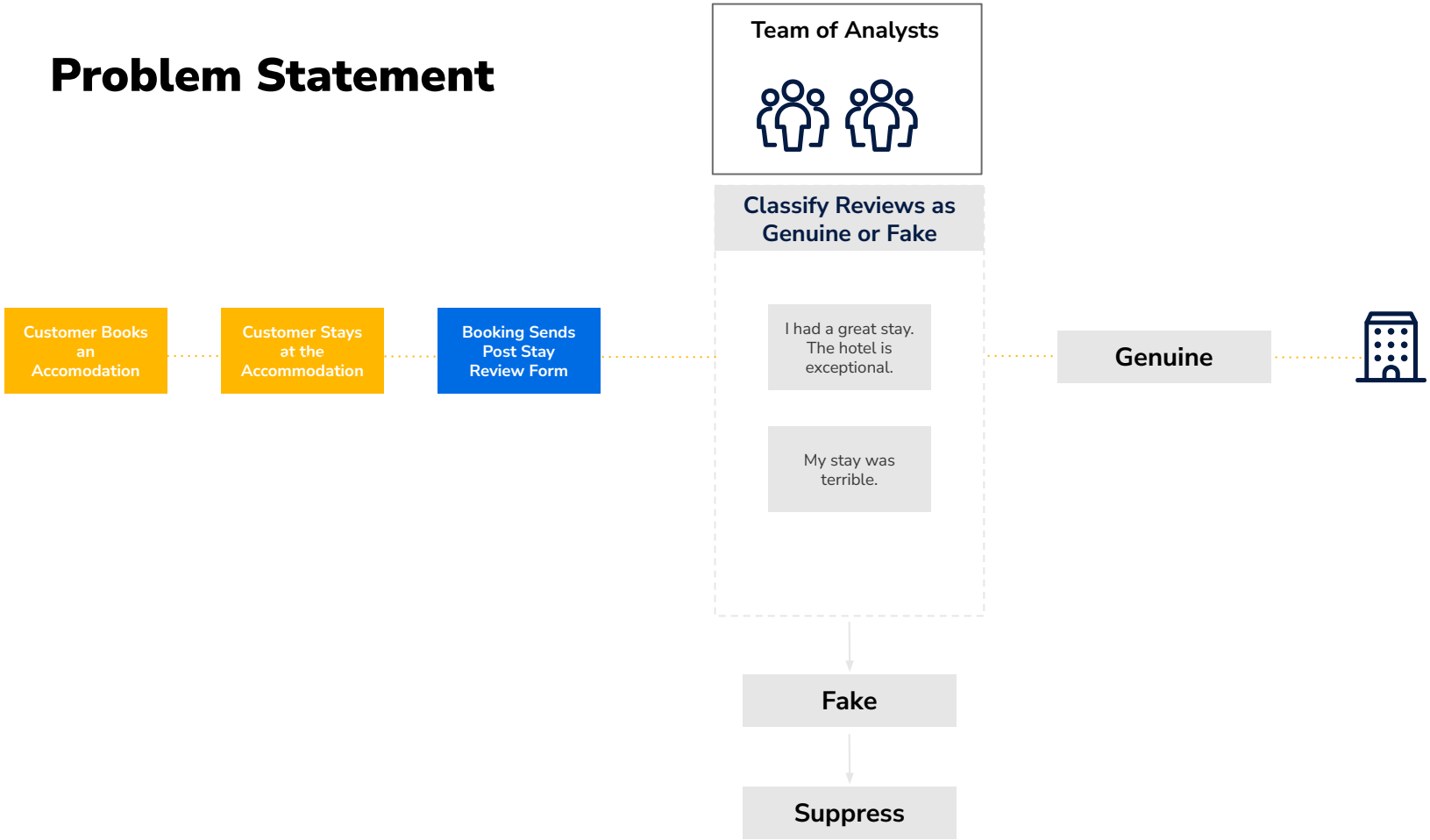
Problem Statement



Problem Statement



Problem Statement



Challenges of Detecting Fake Reviews Manually

Time-Consuming Process

Manually reviewing is a labor-intensive and time-consuming task

Subjectivity

Human reviewers might have different interpretations of what constitutes a fake review

Volume Overload

With a large volume of reviews being submitted daily, moderators can easily become overwhelmed

Limited Scalability

As the number of reviews grows, scaling manual review processes becomes increasingly impractical

Approaches for Automatically Detecting Fake Reviews

How to automatically detect genuine and fake reviews?

Identify Suspicious Patterns - Machine Learning Algorithms

Behavioural Analysis and Features - Behavioural / Outliers

Analyze Text for Authenticity - NLP

Approaches for Automatically Detecting Fake Reviews

How to automatically detect genuine and fake reviews?

Identify Suspicious Patterns - Machine Learning Algorithms

Behavioural Analysis and Features - Behavioural / Outliers

Analyze Text for Authenticity - NLP

Leveraging LLMs for Fake Reviews Detection

NLP and Evolution of LLMs

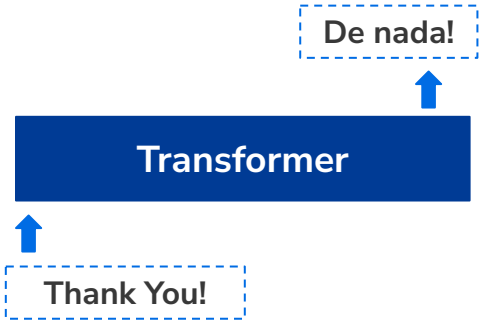
NLP is the science of teaching computers to understand, interpret, and respond to human language naturally

The Rise of Large Language Models (LLMs)

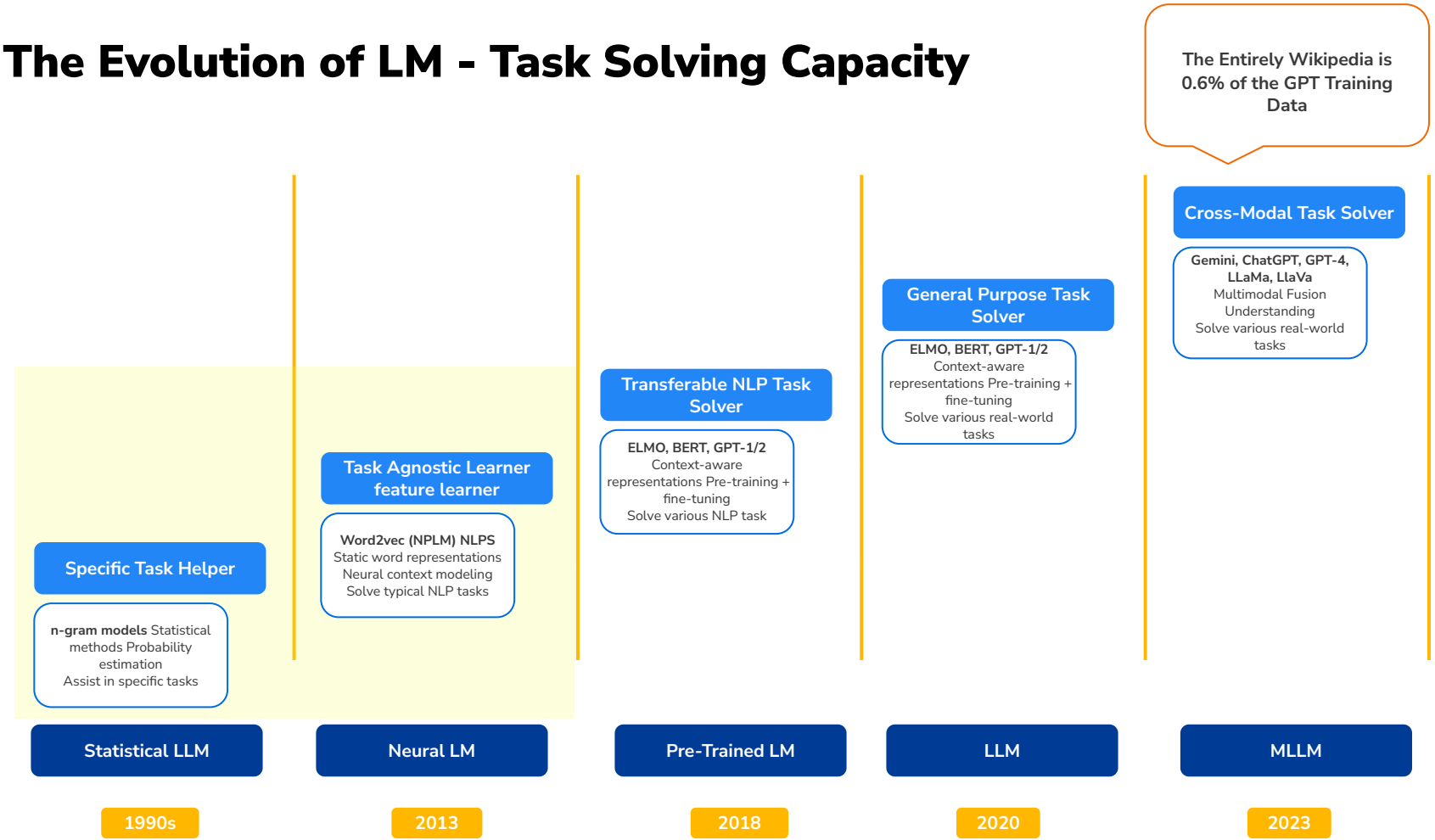
LLMs are advanced NLP models trained on massive datasets, capable of a wide array of language tasks, from question-answering to content generation

Game-Changing Innovation: "Attention Is All You Need" (2017)

- **Key Paper** by Vaswani et al: **Introduced the Transformer Architecture**
- Core Concept is the Attention Mechanism - Enables models to focus on relevant words in a sentence, handling complex dependencies in language and improving performance



The Evolution of LM - Task Solving Capacity



Transformers Architecture

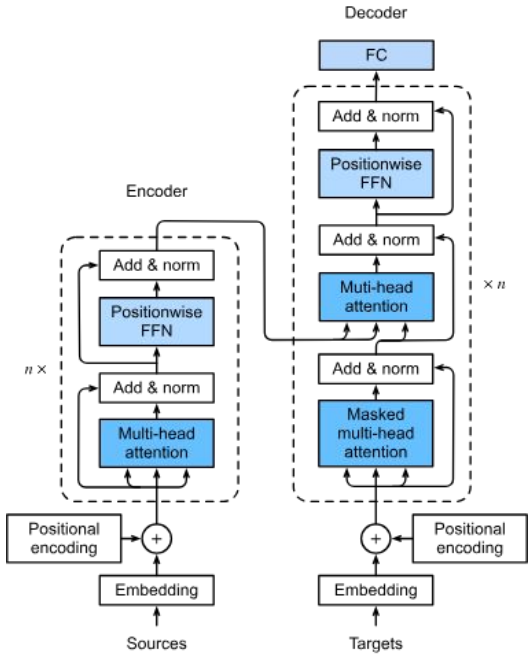


Image Credits

Transformers Architecture

Understands input by analyzing words before and after each word

Predicts masked words (e.g.,“The cat [MASK] on the mat.”) to capture context

Good at analyzing text (BERT) but not generating text

Encoder Only Models

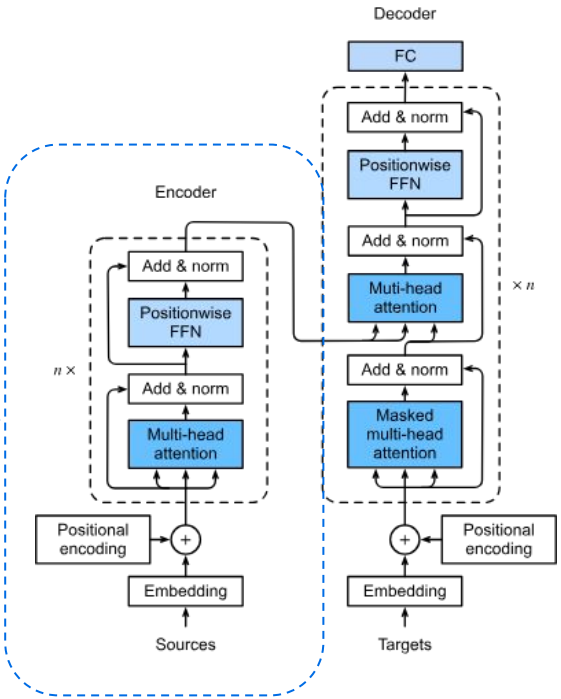


Image Credits

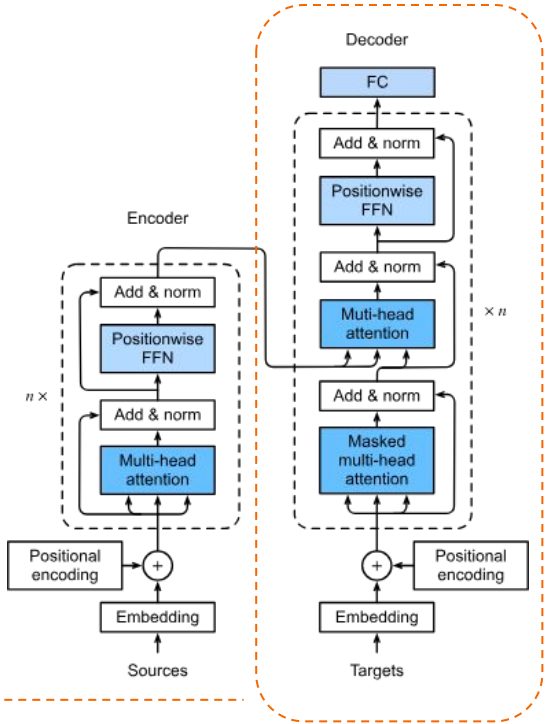
Transformers Architecture

Understands input by analyzing words before and after each word.

Predicts masked words (e.g.,"The cat [MASK] on the mat.") to capture context.

Good at analyzing text (BERT) but not generating text

Encoder Only Models



Generates output by predicting the next word in a sequence based on previous words.

Excels at creating text (e.g., The sun is shining, and the"→ "The sun is shining, and people are happy."

Good at generating text but not analyzing input (e.g., GPT).

Decoder Only Models

Image Credits

Transformers Architecture

Understands input by analyzing words before and after each word.

Predicts masked words (e.g.,"The cat [MASK] on the mat.") to capture context.

Good at analyzing text (BERT) but not generating text

Encoder Only Models

Processes input to understand its meaning and generates output sequentially.

Transforms input into meaningful output (e.g., "The cat is on the mat" → "Le chat est sur le tapis").

Good at both analyzing and generating text (e.g., T5, BART).

Encoder - Decoder Models

Generates output by predicting the next word in a sequence based on previous words.

Excels at creating text (e.g., "The sun is shining, and the"→ "The sun is shining, and people are happy.")

Good at generating text but not analyzing input (e.g., GPT).

Decoder Only Models

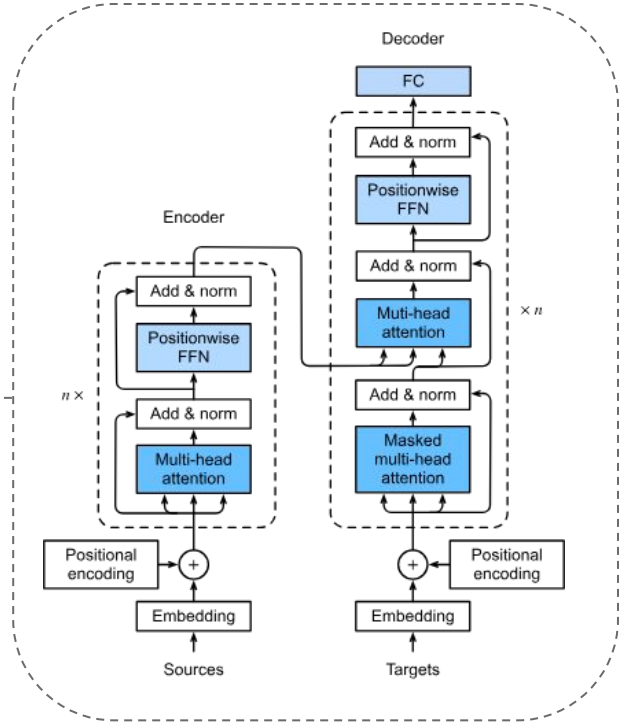


Image Credits

The Power of Pre-Trained Models

- **Pre-Trained Models** are models trained on vast text datasets to understand language structure and meaning.
- These models are then adapted, or fine-tuned, for specific tasks

Examples of Influential Pre-Trained Models

- **BERT (2018)**: Bi-directional encoding captures context from both directions, excelling at understanding nuances in language
- **GPT Series**: Known for generating coherent, human-like text, widely used in conversational AI
- **Additional Models**: T5, RoBERTa, and DistilBERT, each optimizing aspects like speed, efficiency, and task-specific accuracy

Why Pre-Trained Models?

- **Efficiency**: Cuts down training time for new tasks
- **Versatility**: Quickly adaptable to diverse NLP tasks
- **Performance**: Achieves state-of-the-art accuracy across benchmarks

Where to Find Pre-Trained Models?

PyTorch Hub: https://pytorch.org/
TensorFlow Hub: https://www.tensorflow.org/hub
Hugging Face Model Hub: https://huggingface.co/models
OpenAI Models: https://openai.com/

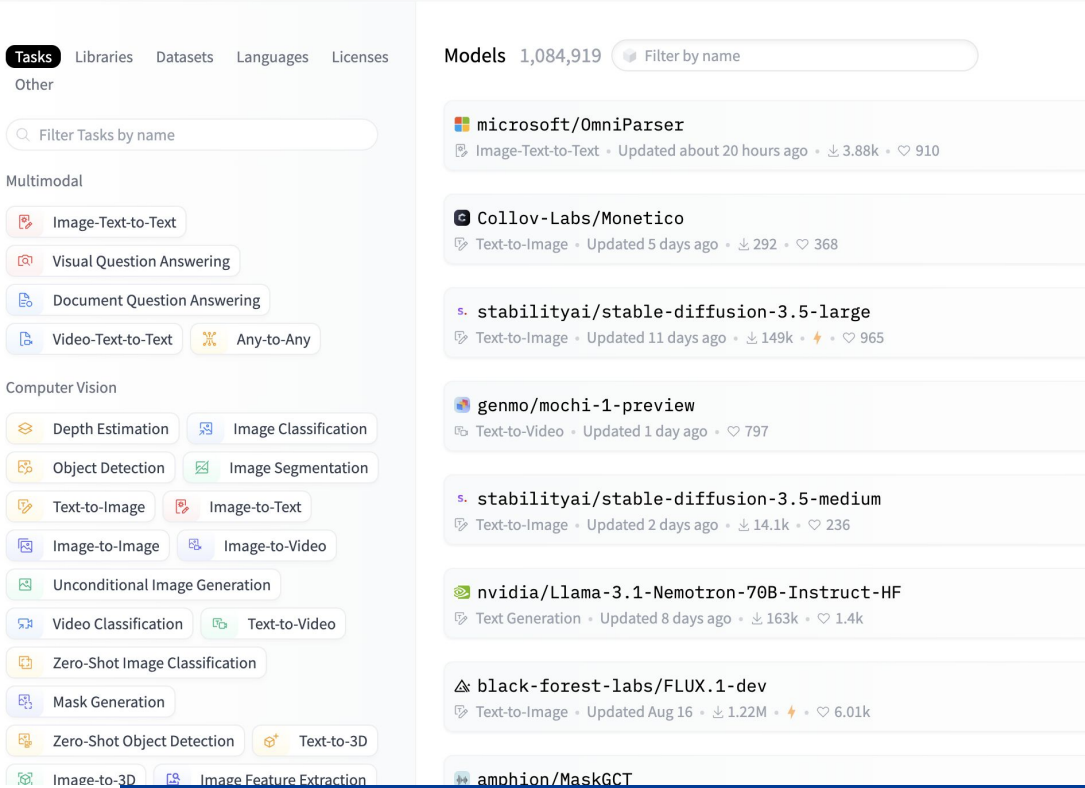
Where to Find Pre-Trained Models?

PyTorch Hub: https://pytorch.org/
TensorFlow Hub: https://www.tensorflow.org/hub
Hugging Face Model Hub: https://huggingface.co/models
OpenAI Models: https://openai.com/

Open Source Models with Hugging Face 🤗

Selecting the Right Model

- Hugging Face Hub is an open platform that hosts models, datasets and ML demos
- To find a model for your project: let's go to the Hugging Face page
- Models suitable for many tasks are available in the "Models Page"



Key Tools for Scaling and Deployment

How to make the process of scaling up and deploying pre-trained models become more efficient and manageable?

Hugging Face

Facilitates experiment tracking, model management, and collaboration

Amazon SageMaker

Provides end-to-end solutions for building, training, and deploying machine learning models at scale

Weights & Biases

Offers pre-trained models and tools for efficient deployment and fine-tuning

Part II

Hands-On Workshop

Instructions

- The second part is a Hand-On Workshop
- Everyone can follow and run the code and exercises by using the notebook in the repository
- Launch the notebook directly: https://colab.research.google.com/github/pyladiesams/llms-scan-reviews-nov2024/blob/master/workshop/LLMs-to-Scan-and-Detect-Fake-Reviews.ipynb

Open the Notebook in Google Colab:

- Click the link above to open the notebook in Google Colab.

Enable GPU Runtime:

- Go to **Runtime** > **Change runtime type** in the Colab menu
- Set **Hardware accelerator** to **GPU**
- Click **Save**

Summary and Conclusions

- Reviews offer deep insights into properties, revealing customer sentiment and highlighting popular discussion topics
- Ensuring reviews are real, not fabricated, is critical for maintaining trust and the integrity of insights

Challenges in Detecting Fake Reviews with LLMs

Lack of Labelled Data

- Difficult to obtain large, accurately labeled datasets
- Genuine reviews vastly outnumber fake ones, causing class imbalance

Sophistication of Fake Reviews

- Fake reviews are increasingly convincing and tailored, blending in seamlessly with genuine ones
- As detection tools improve, so do the methods of those crafting fake reviews

Metadata Manipulation

- Use of VPNs, temporary emails, and fake profiles to obscure identities and bypass detection mechanisms

Thank you

Booking.com