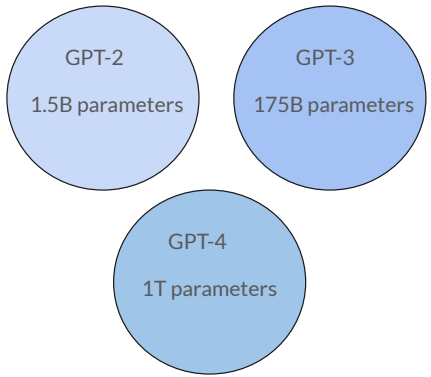
Optimization of Large Language Models for Real-Time Sentiment Analysis : Knowledge Distillation for Sentiment Classification

Idris NECHNECH, Youssef ENNOURI, Younes OUDINA

2025/12/3

Context: The LLM Paradox for Real-Time Applications

An explosion of LLMs parameters



The Critical Need: Instant Analysis

- Real-time applications, like content moderation on live platforms (Twitter, YouTube), must detect and filter thousands of toxic comments per second.

The Ideal Tool: Large Language Models (LLMs)

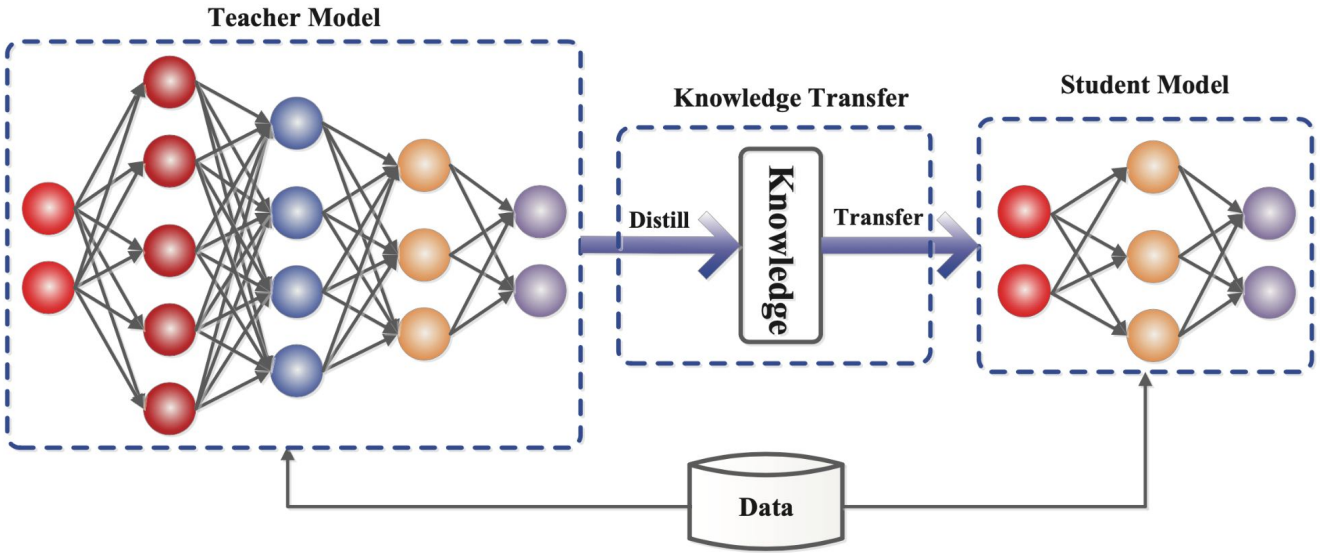
- LLMs are state-of-the-art for understanding the complex nuances of human language (e.g., sarcasm, irony, hate speech).

The Paradox: Performance at a Price

- Models are growing exponentially larger
- Massive VRAM usage, and high energy costs.
- High latency makes the most powerful models impractical for instant responses.

P. J. Maliakel, S. Ilager and I. Brandic, *Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings*, IPDPS, 2025.

A solution: Knowledge Distillation



J. Gou et al., Knowledge Distillation: A Survey, IJCV, 2021.

The Proving Ground: IMDb Dataset

The Dataset: We use the IMDb dataset, a widely-recognized benchmark for sentiment analysis.

- It contains 50,000 movie reviews, evenly split for training and testing.

The Task: Binary Classification

- The goal is simple and clear: classify each review as either **positive** or **negative**.

The Teacher Model: Setting the Gold Standard

Our Expert: `roberta-large`

- 355 million parameters, known for its deep understanding of language.

Specialization for the Task

- We first fine-tuned this model on the IMDb dataset, turning it into a highly specialized expert for movie review analysis.

The Performance Benchmark

- This expert teacher achieved a final accuracy of **95.88%**.
- This score becomes our "gold standard"—the target for our smaller, faster student models to aim for.

A. Albladi et al. — TWSSenti: A Novel Hybrid Framework for Topic-Wise Sentiment Analysis on Social Media Using Transformer Models (2025, arXiv).

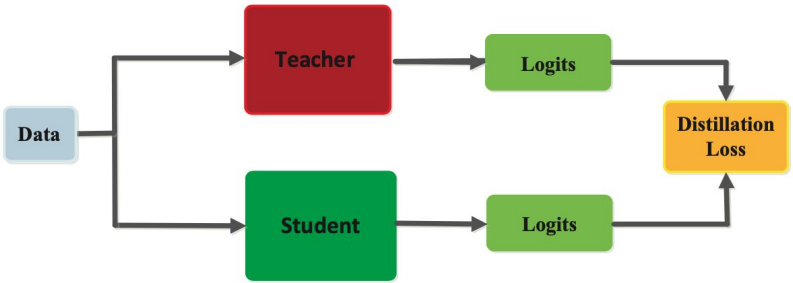
The Students: A Range of Compact Models

- **Our Goal:**
  To find the best trade-off between size, speed, and performance.
- **The Candidates:**
  We selected four well-known, compact models to act as our "students," each with a different size

Student	Number of parameters
DistilRoBERTa	82M
DistilBERT	66M
MiniLM	33M
TinyBERT	14.5M

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (Sanh et al., 2019)

The Distillation Process



$$\mathcal{L}_{overall} = \alpha \frac{1}{T^2} \mathcal{L}_{distill} + (1 - \alpha)\mathcal{L}_{CE}$$

$$\mathcal{L}_{distill} = \frac{1}{m} \sum_{i}^{m} D_{KL}(P(x_i) \,||\, Q(x_i))$$

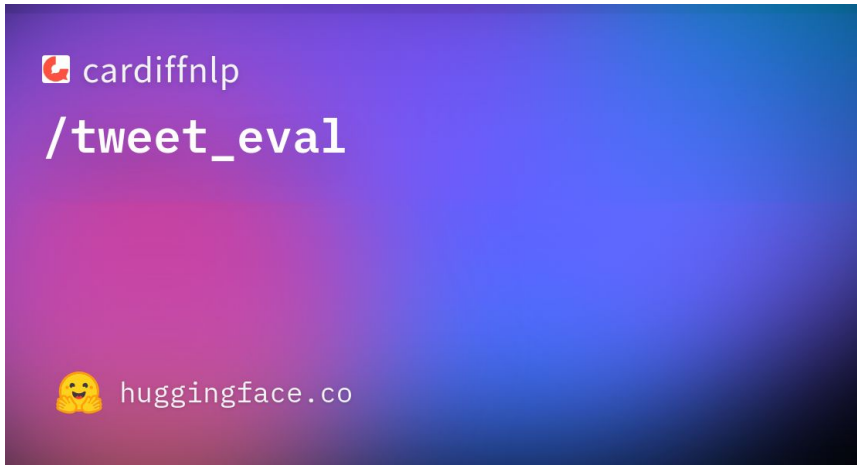
J. Gou et al., Knowledge Distillation: A Survey, IJCV, 2021.

Final Results: The Performance vs. Size & Speed & GPU Usage

Model	Parameters	Compression Ratio	Accuracy	vs. Teacher	Inference Time (ms)	GPU Usage (W)
RoBERTa-Large	355 M	x 0.00	95.88%	-0.00%	12.10	161
DistilRoBERTa	82 M	x 4.33	92.80%	-3.08%	2.60	92
DistilBERT	66 M	x 5.31	91.64%	-4.24%	2.35	93
MiniLM	33 M	x 10.65	91.98%	-3.90%	4.54	84
TinyBERT	15 M	x 24.76	88.24%	-7.64%	1.80	82

Results after hyperparameter tuning, Computation done on a single Nvidia GeForce RTX 3090

Onto more complex tasks



The Dataset: We use the Tweet Eval dataset

- **Multi-class Sentiment Classification** (59,9k samples)

Onto more complex tasks

Model	Parameters	Compression Ratio	Accuracy	Vs Teacher	Inference Time (ms)	GPU Usage (W)
RoBERTa-Large	355 M	x 0.00	75.55 %	0.00 %	9.05	208
DistilRoBERTa	82 M	x 4.33	74.40 %	-1.15 %	2.63	110
DistilBERT	66 M	x 5.31	75.65 %	+0.10 %	2.40	101
MiniLM	33 M	x 10.65	75.60 %	+0.05 %	4.63	88
TinyBERT	15 M	x 24.76	73.35 %	-2.20 %	1.81	83

Results after hyperparameter tuning, Computation done on a single Nvidia GeForce RTX 3090

Smaller but better ?

$$\mathcal{L}_{overall} = \alpha \frac{1}{T^2} \mathcal{L}_{distill} + (1 - \alpha)\mathcal{L}_{CE}$$

$$\mathcal{L}_{distill} = \frac{1}{m} \sum_{i}^{m} D_{KL}(P(x_i) \,||\, Q(x_i))$$

J. Gou et al., Knowledge Distillation: A Survey, IJCV, 2021.

Model	Accuracy (%)
DistilRoBERTa	0.7335
DistilBERT	0.7395

Accuracy after performing full-fine tuning without KD

Smaller but better ? x2

Sample : “Ayyye I just purchased my Ed Sheeran tickets for tmrw but I may not even go…”

Hard Label: 0

Soft Labels : [0.59585476 0.38153255 0.02261272]

Here, the teacher is doubting between negative (0) and neutral (1)

Sample : “Typical formula action film: a good cop gets entangled in a mess of crooked cops and Japanese gangst...”

Hard Label: 0

Soft Labels : [0.9946672 0.00533289]

Here, the teacher is sure of himself in choosing negative (0)

Smaller but better ? x2

$$\text{Confidence} = \frac{1}{N}\sum_{i=1}^{N}\hat{p}_i, \quad \text{where } \hat{p}_i \text{ is the confidence for sample i}$$

We use the Maximum Softmax Probability.
For instance:

$$\sigma\begin{bmatrix}0.59585476\\0.38153255\\0.02261272\end{bmatrix} = \begin{bmatrix}0.42179\\0.34043\\0.23777\end{bmatrix} \longleftarrow \hat{p}_i$$

Dataset	Confidence (%)
IMDb	98.68
TweetEval	81.38

D. Hendrycks et al., A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, ICLR, 2017.
L. Yang et al., Rethinking the Knowledge Distillation From the Perspective of Model Calibration.
A. K Menon et al., Statistical Perspective on Distillation, PMLR 2021.

How to increase the accuracy ?

$$\mathcal{L} = \alpha \frac{1}{T^2} \mathcal{L}_{KL}(T^* || P^S) + (1 - \alpha)\mathcal{L}_{CE}$$

$$T^\star(x_i) = \begin{cases} P^T(x_i) & \text{if } \underset{c}{\text{argmax}}(P^T(x_i)) = y_i \\ 1_{y_i} & \text{if } \underset{c}{\text{argmax}}(P^T(x_i)) \neq y_i \end{cases}$$

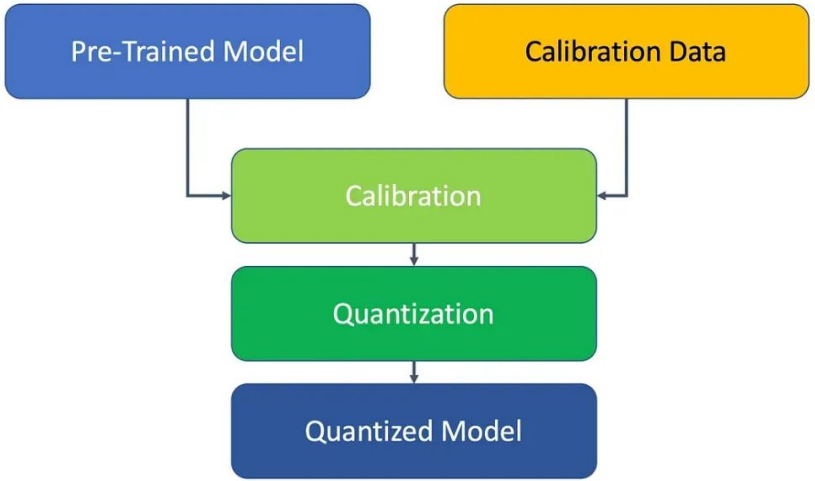
Table 5: Student accuracy as a function of the teacher's loss function

Student	Teacher (CE Loss)	Teacher (MLE Loss)	(MLE - CE)
DistilRoBERTa	0.7440	0.7480	+0.0040
DistilBERT	0.7565	0.7550	-0.0015
MiniLM	0.7560	0.7550	-0.0010
TinyBERT	0.7335	0.7390	+0.0055

M. Martinez-Garcia et al., Teacher privileged distillation: How to deal with imperfect teachers?, Knowledge-Based Systems, 2025
Sh. Mohajer Hamidi et al., How to Train the Teacher Model for Effective Knowledge Distillation, CoRR 2024.

Thank you !

Appendix 1: Another idea for Inference: Quantization



Post-Training Quantization (PTQ)

Diving deeper into Quantization Realm : Introduction to PTQ and QAT - Pratima Rathore

Appendix 1: Another idea for Inference: Quantization

Table 1: Performance Comparison: **DistilRoBERTa**

Metric	Base Model	Quantized Model	Change (%)
Average Inference Time (s)	0.00260	0.00281	+8.1%
Peak GPU Memory (MB)	1076.06	1066.06	-0.9%
Average GPU Power (W)	92.46	183.90	+98.9%

Table 2: Performance Comparison: **DistilBERT**

Metric	Base Model	Quantized Model	Change (%)
Average Inference Time (s)	0.00235	0.00257	+9.4%
Peak GPU Memory (MB)	928.06	910.06	-1.9%
Average GPU Power (W)	92.88	168.24	+81.1%

Table 3: Performance Comparison: **MiniLM**

Metric	Base Model	Quantized Model	Change (%)
Average Inference Time (s)	0.00454	0.00343	-24.5%
Peak GPU Memory (MB)	1108.06	910.06	-17.9%
Average GPU Power (W)	84.28	158.50	+88.1%

Table 4: Performance Comparison: **TinyBERT**

Metric	Base Model	Quantized Model	Change (%)
Average Inference Time (s)	0.00180	0.00134	-25.5%
Peak GPU Memory (MB)	1108.06	942.06	-15.0%
Average GPU Power (W)	81.97	146.48	+78.7%

Appendix 2: Github repository

https://github.com/Idrisdesu/knowledge-distillation-for-sentiment-analysis