BTZSC

A Benchmark for Zero-Shot Text Classification

Ilias Aarab
European Central Bank

25 March 2026

Why Zero-Shot Classification?

Text classification is everywhere in practice:

- Sentiment (Kowsari et al. 2019)
- Topic (Sebastiani 2002)
- Intent (Kowsari et al. 2019)
- Emotion (Kowsari et al. 2019)

Yet labeled data is expensive to build, especially in specialized domains (Sebastiani 2002; Kowsari et al. 2019).

Zero-shot classification removes the need for task-specific labels (Yin, Hay, and Roth 2019).

> **Zero-shot classification (ZSC)**
>
> Score the match between an input text and each label description, then pick the best match.

Four Model Families for ZSC

- **NLI Cross-Encoders**: entailment scoring of text-label pairs (Yin, Hay, and Roth 2019; Lewis et al. 2020).
- **Rerankers**: relevance scoring of text-label pairs, adapted from IR (Nogueira, Jiang, and Lin 2020).
- **Embedding Models**: cosine-similarity matching in shared embedding space (Reimers and Gurevych 2019; Wang et al. 2024).
- **Instruction-tuned LLMs**: multiple-choice prompting with next-token scoring (Brown et al. 2020).

Practical Question

Which approach should we use, and when?

The Gap: Fragmented Evaluation

Existing work evaluates ZSC **in isolation**:

- **Single family** — only NLI cross-encoders in early benchmarks (Yin, Hay, and Roth 2019)
- **Narrow scope** — topic-only evaluations (Gretz et al. 2023)
- **Not pure zero-shot** — MTEB uses supervised probes (Muennighoff et al. 2023)
- **Missing families** — small-LLM studies omit embeddings and rerankers (Lepagnol et al. 2024)

This makes it **impossible to compare** across model families under a consistent protocol.

> **Our Goal**
>
> A **unified benchmark** — shared datasets, shared metrics, strict zero-shot protocol — comparing all four families side by side.

BTZSC: Benchmark Design

22 public datasets, 38 models

Task	# Datasets	# Classes
Topic	11	2–56
Sentiment	6	2–3
Intent	3	2–77
Emotion	2	6–32

Design criteria:

- Task & domain diversity
- Class cardinality: 2 to 77
- Document length: 8 to 282 tokens
- Strictly zero-shot protocol
- Label verbalizers from (Laurer et al. 2023)

Open Science

All publicly released: Datasets · Code · Leaderboard

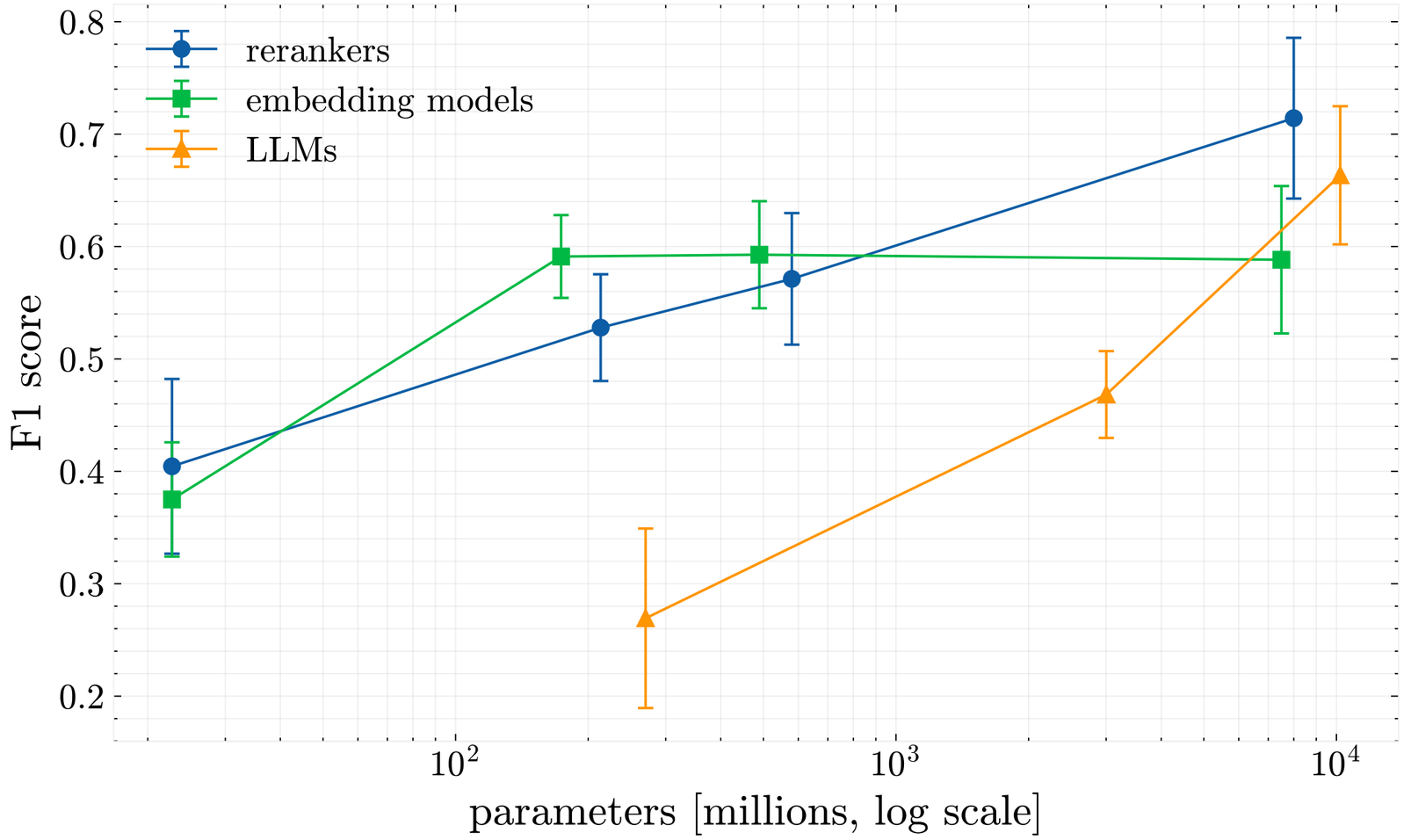
Some Insights

Family	Best Model	Avg F_1	Avg Acc
Reranker	Qwen3-Reranker-8B	**0.72**	**0.76**
LLM	Mistral-Nemo-12B	0.67	0.71
Embedding	GTE-large-en-v1.5	0.62	0.64
NLI Cross-Enc.	DeBERTa-v3-L-NLI	0.60	0.62

Key insight: Model *family* matters more than *size*, but scaling effects are family-dependent.

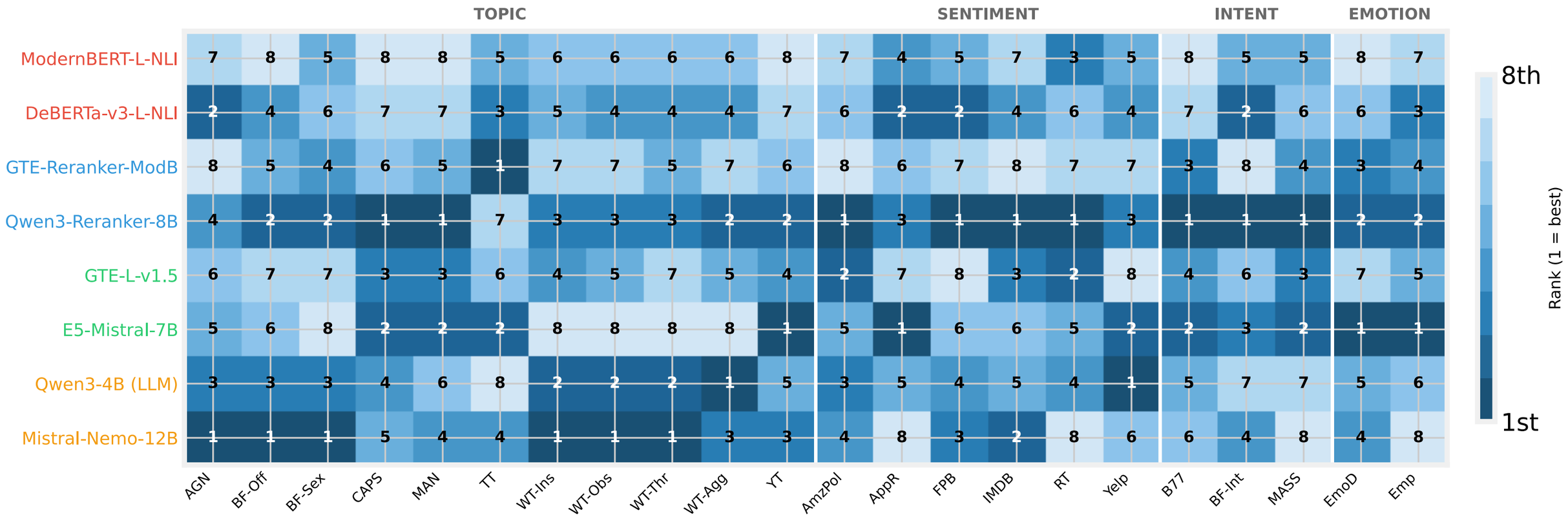
- **Architecture dominates at matched scale**: Qwen3-Reranker-8B (F_1=0.72) vs Qwen3-Embedding-8B (F_1=0.59) — same backbone, +13 pts
- **NLI cross-encoders remain competitive**: DeBERTa-v3-L-NLI (F_1=0.60, ~300M) matches Qwen3-Reranker-0.6B (F_1=0.61) despite its modern backbone and large-scale training mix
- **Embeddings are efficient but task-sensitive**: GTE-large (F_1=0.62) rivals similar-size cross-encoders and rerankers, but doesn’t gain with scaling

Scaling Behavior



- **Rerankers:** monotonic gains with scale
- **LLMs:** steepest scaling — poor at small scale, rapid jump between 3B and 8B
- **Embeddings:** saturate around $F_1 \approx 0.60–0.62$

Rankings Shift Across Datasets



No single model wins everywhere — the best family **changes with the dataset**.

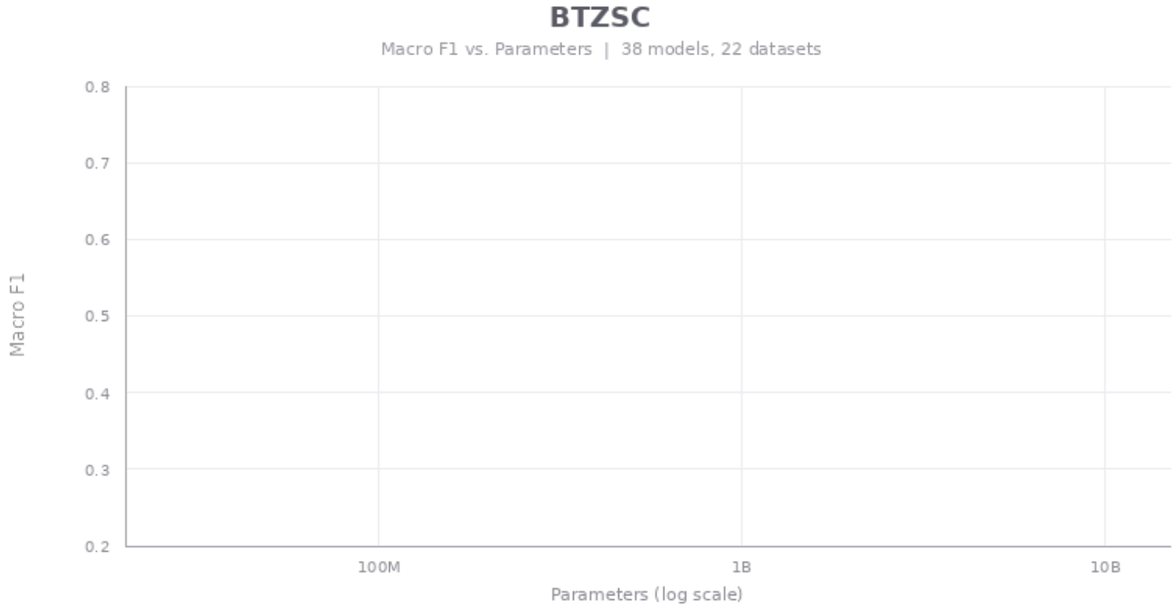
- LLMs excel on **topic** classification
- Rerankers dominate **intent** and **emotion**
- **Sentiment** is comparatively easy for all strong models

Takeaways

1. **Model family matters more than size** — architecture choice drives performance
2. **Scaling behavior varies by family** — rerankers gain steadily, LLMs need 3B+, embeddings saturate, NLI plateaus
3. **No single model dominates** — use **BTZSC** to find the right family for your task

Resources:

- Code: github.com/IliasAarab/btzsc
- Data: huggingface.co/datasets/btzsc
- Leaderboard: huggingface.co/spaces/btzsc



References

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and et al. 2020. “Language Models Are Few-Shot Learners.” *arXiv Preprint arXiv:2005.14165*.

Gretz, Shai, Alon Halfon, Ilya Shnayderman, Orith Toledo-Ronen, Artem Spector, Lena Dankin, Yannis Katsis, et al. 2023. “Zero-Shot Topical Text Classification with LLMs – an Experimental Study.” In *Findings of EMNLP*, 9647–76.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. 2019. “Text Classification Algorithms: A Survey.” *Information* 10 (4): 150. https://doi.org/10.3390/info10040150.

Laurer, Moritz, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. “Building Efficient Universal Classifiers with Natural Language Inference.” arXiv. https://doi.org/10.48550/arXiv.2312.17543.

Lepagnol, Pierre, Thomas Gerald, Sahar Ghannay, Christophe Servan, and Sophie Rosset. 2024. “Small Language Models Are Good Too: An Empirical Study of Zero-Shot Classification.” In *Proceedings of LREC–COLING*.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. “BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension.” In *Proceedings of ACL*.

Muennighoff, Niklas, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. “MTEB: Massive Text Embedding Benchmark.” In *Proc. EACL*, 2014–37.

Nogueira, Rodrigo, Zhiying Jiang, and Jimmy Lin. 2020. “Document Ranking with a Pretrained Sequence-to-Sequence Model.” *arXiv Preprint arXiv:2003.06713*.

Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” In *Proc. EMNLP*.

Sebastiani, Fabrizio. 2002. “Machine Learning in Automated Text Categorization.” *ACM Computing Surveys* 34 (1): 1–47.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. “Multilingual E5 Text Embeddings: A Technical Report.” *arXiv Preprint arXiv:2402.05672*.

Yin, Wenpeng, Jamaal Hay, and Dan Roth. 2019. “Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach.” In *Proc. EMNLP–IJCNLP*, 3914–23.