Online Resource 1 - Supplementary Information

Article title: Forecasting Fine-Tuning Break-Even Label Budgets in Text Classification from Frozen Embedding Geometry
Journal: Knowledge and Information Systems
Author: Emin Talip Demirkiran
Affiliation: Department of Computer Engineering, Eskisehir Technical University, Iki Eylul Campus, 26555 Eskisehir, Türkiye
Corresponding author e-mail: etd@eskisehir.edu.tr
Contents: Table S1 reports the task corpus and source attribution; Table S2 reports the executed model, prompt, sampling, and training configuration; Table S3 reports post hoc regression diagnostics and influence sensitivities.

Table S1. Task corpus and source attribution

Tasks were fixed before separability or break-even outcomes were observed. The initial set was expanded for pre-outcome power adequacy. Acquisition purposively sought public single-label English tasks spanning class count, size, input length, domain, and task type; the resulting corpus is a heterogeneous benchmark sample rather than a probability sample of all text-classification tasks.

task_id	task	K	N	mean_tokens	repository	attribution
task_ade_corpus_v2	ade_corpus_v2 (medical adverse-drug-event task)	2	3000	18.24	https://huggingface.co/datasets/ade-benchmark-corpus/ade_corpus_v2	Gurulingappa et al. 2012, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects
task_amazon_polarity	Amazon Polarity (binary product-review sentiment)	2	3000	73.87	https://huggingface.co/datasets/fancyzhx/amazon_polarity	Zhang, Zhao, LeCun 2015
task_app_reviews	app_reviews (software/app-store review task)	5	3000	11.5	https://huggingface.co/datasets/sealuzh/app_reviews	Sealuzh App Reviews dataset (software-engineering research; app-store review star-rating classification)
task_banking77	Banking77 (77-class banking-domain intent classification)	77	3000	11.89	https://huggingface.co/datasets/mteb/banking77	Casanueva et al. 2020, Efficient Intent Detection with Dual Sentence Encoders
task_clickbait	clickbait (news headline task)	2	3000	9.02	https://huggingface.co/datasets/marksverdhei/clickbait_title_classification	Chakraborty et al. 2016 clickbait-detection lineage, via marksverdhei/clickbait_title_classification mirror
task_climate_fever	climate_fever (climate-claim verification task)	4	1535	20.19	https://huggingface.co/datasets/tdiggelm/climate_fever	Diggelmann et al. 2020, CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims
task_clinc_oos_small	CLINC150 (small config; 150-intent + out-of-scope virtual-assistant classification)	151	3000	8.36	https://huggingface.co/datasets/clinc/clinc_oos	Larson et al. 2019, An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction
task_emotion	dair-ai/emotion (6-class tweet emotion classification)	6	3000	19.39	https://huggingface.co/datasets/dair-ai/emotion	Saravia et al. 2018, CARER: Contextualized Affect Representations for Emotion Recognition
task_enron_spam	enron_spam (email spam task)	2	2994	289.08	https://huggingface.co/datasets/SetFit/enron_spam	Metsis, Androutsopoulos, Paliouras 2006, Spam Filtering with Naive Bayes -- Which Naive Bayes? (Enron-Spam corpus)
task_fake_news	fake_news (news-article veracity task)	2	3000	403.01	https://huggingface.co/datasets/GonzaloA/fake_news	Binary fake-news classification (GonzaloA/fake_news mirror)
task_financial_phrasebank	Financial PhraseBank (3-class financial-news sentiment)	3	3000	22.93	https://huggingface.co/datasets/warwickai/financial_phrasebank_mirror	Malo et al. 2014, Good debt or bad debt: Detecting semantic orientations in economic texts
task_go_emotions	go_emotions (Reddit comment emotion (filtered to single-label rows only, disclosed) task)	28	3000	12.59	https://huggingface.co/datasets/google-research-datasets/go_emotions	Demszky et al. 2020, GoEmotions: A Dataset of Fine-Grained Emotions (Google Research)
task_hate_speech18	Hate Speech 18 / Stormfront forum-post hate-speech classification	4	3000	17.9	https://huggingface.co/datasets/SetFit/hate_speech18	de Gibert, Perez, Garcia-Pablos, Cuadros 2018, Hate Speech Dataset from a White Supremacy Forum
task_hate_speech_offensive	hate_speech_offensive (general Twitter hate/offensive task)	3	3000	14.12	https://huggingface.co/datasets/tdavidson/hate_speech_offensive	Davidson, Warmsley, Macy, Weber 2017, Automated Hate Speech Detection and the Problem of Offensive Language

task_id	task	K	N	mean_tokens	repository	attribution
task_lex_glue_ledgar	lex_glue_ledgar (legal contract provision task)	100	3000	112.54	https://huggingface.co/datasets/coastalcph/lex_glue	Chalkidis et al. 2022, LexGLUE (LEDGAR subtask, orig. Tuggener et al. 2020)
task_lex_glue_scotus	lex_glue_scotus (US Supreme Court opinion task)	13	3000	5975.65	https://huggingface.co/datasets/coastalcph/lex_glue	Chalkidis et al. 2022, LexGLUE (SCOTUS subtask)
task_medical_abstracts	medical_abstracts (medical literature abstract task)	5	3000	181.99	https://huggingface.co/datasets/TimSchopf/medical_abstracts	Schopf, Braun, Matthes 2023, Medical Abstracts TC Corpus
task_multi_nli	multi_nli (cross-genre natural language inference task)	3	3000	30.9	https://huggingface.co/datasets/nyu-mll/multi_nli	Williams, Nangia, Bowman 2018, A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference (MultiNLI)
task_poem_sentiment	poem_sentiment (poetry sentiment task)	4	892	7.1	https://huggingface.co/datasets/google-research-datasets/poem_sentiment	Sheng & Uthus 2020, Investigating Societal Biases in a Poetry Composition System (PoetryFoundation poem sentiment)
task_scicite	scicite (scientific-citation intent task)	3	3000	33.63	https://huggingface.co/datasets/allenai/scicite	Cohan, Ammar, van Zuylen, Cady 2019, Structural Scaffolds for Citation Intent Classification in Scientific Publications (SciCite)
task_sentiment140	sentiment140 (auto-labeled Twitter sentiment task)	2	3000	12.96	https://huggingface.co/datasets/stanfordnlp/sentiment140	Go, Bhayani, Huang 2009, Twitter Sentiment Classification using Distant Supervision
task_setfit_amazon_counterfactual	setfit_amazon_counterfactual (counterfactual-statement detection (Amazon reviews) task)	2	3000	19.74	https://huggingface.co/datasets/SetFit/amazon_counterfactual_en	O'Neill et al. 2021, I Wish I Would Have Loved This One, But I Didn't (Amazon Counterfactual Detection), via SetFit mirror
task_setfit_cr	setfit_cr (customer product review task)	2	3000	19.27	https://huggingface.co/datasets/SetFit/CR	Hu & Liu 2004, Mining and Summarizing Customer Reviews, via SetFit mirror
task_setfit_insincere_questions	setfit_insincere_questions (Quora question insincerity task)	2	3000	12.94	https://huggingface.co/datasets/SetFit/insincere-questions	Kaggle Quora Insincere Questions Classification challenge, via SetFit mirror
task_setfit_qnli	setfit_qnli (question-answering entailment (SQuAD-derived) task)	2	3000	37.64	https://huggingface.co/datasets/SetFit/qnli	Rajpurkar et al. 2016 SQuAD; Wang et al. 2018 GLUE QNLI task, via SetFit mirror
task_setfit_rte	setfit_rte (textual entailment (RTE) task)	2	2490	53.36	https://huggingface.co/datasets/SetFit/rte	Dagan, Glickman, Magnini 2005 / Bar-Haim et al. 2006 RTE challenges; Wang et al. 2018 GLUE RTE task, via SetFit mirror
task_setfit_wnli	setfit_wnli (coreference-based entailment (Winograd) task)	2	635	28.8	https://huggingface.co/datasets/SetFit/wnli	Levesque, Davis, Morgenstern 2012, The Winograd Schema Challenge; Wang et al. 2018 GLUE WNLI task, via SetFit mirror
task_sms_spam	sms_spam (SMS spam task)	2	3000	15.63	https://huggingface.co/datasets/ucirvine/sms_spam	Almeida, Hidalgo, Yamakami 2011, Contributions to the Study of SMS Spam Filtering
task_snli	snli (image-caption-derived natural language inference task)	3	3000	21.2	https://huggingface.co/datasets/stanfordnlp/snli	Bowman, Angeli, Potts, Manning 2015, A Large Annotated Corpus for Learning Natural Language Inference (SNLI)
task_subj	subj (movie-review subjectivity task)	2	3000	24.37	https://huggingface.co/datasets/SetFit/subj	Pang & Lee 2004, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts (Subjectivity dataset), via SetFit mirror
task_tweet_topic_single	tweet_topic_single (Twitter topic task)	6	3000	28.0	https://huggingface.co/datasets/cardiffnlp/tweet_topic_single	Antypas et al. 2022, TweetTopic: Twitter Topic Classification
task_twitter_financial_sentiment	twitter_financial_sentiment (Twitter financial-news sentiment task)	3	3000	12.22	https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment	Twitter Financial News Sentiment dataset (zeroshot/Refinitiv release)
task_yahoo_answers_topics	Yahoo Answers Topics (10-class Q&A topic classification)	10	3000	92.49	https://huggingface.co/datasets/pietrolesci/yahoo_answers_topics	Zhang, Zhao, LeCun 2015
task_yelp_polarity	yelp_polarity (business-review binary sentiment task)	2	3000	132.97	https://huggingface.co/datasets/fancyzhx/yelp_polarity	Zhang, Zhao, LeCun 2015, Character-level Convolutional Networks for Text Classification
task_yelp_review_full	Yelp Review Full (5-class star-rating sentiment)	5	3000	132.92	https://huggingface.co/datasets/Yelp/yelp_review_full	Zhang, Zhao, LeCun 2015, Character-level Convolutional Networks for Text Classification

Table S2. Executed model, prompt, sampling, and training configuration

This table reports the executed configuration rather than only the planned protocol. The FLAN-T5 prompt-length deviation is shown explicitly.

Component	Executed configuration	Notes
Task split	Nominal 70% pool / 30% evaluation; seed 42; exact duplicate texts grouped so identical content cannot cross the split	Realized pool fraction may deviate slightly from 0.70 when duplicate groups are present.
all-mpnet-base-v2	sentence-transformers/all-mpnet-base-v2; commit e8c3b32edf; native SentenceTransformers mean pooling; max sequence length 256; inference batch 32	Frozen; no task-specific tuning.
sup-simcse-roberta-base	princeton-nlp/sup-simcse-roberta-base; commit 4bf73c6b5d; final-layer CLS representation; max sequence length 256; inference batch 32	Frozen; no task-specific tuning.
gte-base	thenlper/gte-base; commit c078288308; native mean pooling; max sequence length 256; inference batch 32	Frozen; no task-specific tuning.
Specialized classifier	bert-base-uncased; commit 86b5e0934494bd15c9632b12f734a8a67f723594; fresh K-class head; hidden dropout 0.3; max sequence length 256	Three independent fits per grid point: seeds 42, 43, 44.
Optimizer / schedule	torch.optim.AdamW; learning rate 1e-5; default weight decay 0.01; 10% linear warm-up; linear decay; 10 fixed epochs; no early stopping	No validation-based checkpoint/model selection.
Training batch	Logical batch = max(4, min(32, realized n_train)); when logical batch > 8, gradients accumulated over micro-batches of 8 before one optimizer step	Evaluation batch size 32.
Training-grid sampling	1, 2, 4, 8 examples per class; within each class take min(g, available); separate without-replacement sample for each grid point and seed	Grid subsets were not constrained to be nested across budgets.
Zero-shot comparator	google/flan-t5-base; commit 7bcac572ce; one deterministic run per task	No demonstrations, prompt optimization, retrieval, or task tuning.
Exact zero-shot prompt	Classify the following text into one of these categories: {label_list}. Text: {text} Category:	{label_list} is the comma-separated semantic label-name list recovered for the task.
FLAN-T5 tokenization / decoding	Prompt max length 512 total tokens; batch 16; max_new_tokens=8; do_sample=False; num_beams=1	Executed implementation differs from the planned uniform 256-token cap; disclosed in the manuscript.
Output parsing	Trim + lowercase; case-insensitive exact label-name match first, then substring match; unmatched generation mapped to an impossible class and counted incorrect	No invalid output is silently dropped.
Outcome metric	Macro-F1 over the complete task label set; zero_division=0	Same held-out evaluation set used for supervised and zero-shot systems.
Software environment	Python 3.11.9; torch 2.13.0+cpu; transformers 5.14.1; sentence-transformers 5.6.1; scikit-learn 1.9.0; statsmodels 0.15.0; scipy 1.17.1	CPU-only execution.

Table S3. Post hoc regression diagnostics and influence sensitivities

Panel A uses the prespecified point estimates but recomputes uncertainty with HC3 heteroskedasticity-robust standard errors. These diagnostics were added after outcome inspection and do not change the confirmatory decision.

Encoder	β1	Conventional SE	HC3 SE	HC3 CI low	HC3 CI high	HC3 one-sided p	Breusch-Pagan p	Max leverage	Max Cook D	Max influence task
all-mpnet-base-v2	-41.187	25.164	98.674	-242.706	160.333	0.340	5.33e-06	0.993	16.045	task_lex_glue_scotus
sup-simcse-roberta-base	-19.529	21.976	49.629	-120.885	81.828	0.348	9.49e-07	0.994	13.172	task_lex_glue_scotus
gte-base	-17.323	25.558	64.822	-149.707	115.062	0.396	9.9e-07	0.994	13.054	task_lex_glue_scotus

Panel B. Diagnostic refit after excluding K >= 50 tasks

The excluded tasks are Banking77 (K=77), LEDGAR (K=100), and CLINC150 (K=151). The K >= 50 threshold was not prespecified and this refit is descriptive only.

Encoder	n	β1	SE	one-sided p	partial r
all-mpnet-base-v2	32	-1.720	3.923	0.332	-0.084
sup-simcse-roberta-base	32	0.374	3.831	0.538	0.019
gte-base	32	-1.096	3.835	0.389	-0.055