DESCRIPTION-ONLY SUPERVISION: CONTRASTIVE LABEL-EMBEDDING ALIGNMENT FOR ZERO-SHOT TEXT CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot text classification (ZSC) aims to assign labels without task-specific annotation by exploiting the semantics of human-readable labels. In practice, embedding-based ZSC often falls back on training a linear probe, reintroducing annotation costs. We propose *description-only supervision*, a simple, compute-efficient alternative that requires only a handful of natural-language descriptions per label. We lightly finetune a base embedding model with a contrastive objective that pulls each label verbalizer toward its associated descriptions while pushing it away from others, using a multi-positive formulation to capture the many-to-one label-description relation. Across four benchmarks (topic, sentiment, intent, emotion) and ten encoders (22M-600M parameters), as few as five descriptions per label yield consistent gains, improving macro-F1 by **+0.10** on average over zero-shot baselines. Compared to a few-shot SetFit baseline with 8 examples per class, our method attains higher mean performance with substantially lower variance across 20 runs, indicating improved stability in low-data regimes. The approach preserves the dual-encoder advantage (pre-encodable labels/documents), avoids labeled documents entirely, and adds minimal engineering overhead.

1 INTRODUCTION

Text classification remains a central task in Natural Language Processing (NLP), supporting a wide range of applications such as sentiment analysis across domains, topic categorization of diverse document types, and intent detection in dialogue systems (Maas et al., 2011; Zhang et al., 2015b; Coucke et al., 2018; Larson et al., 2019; Sebastiani, 2002; Aggarwal & Zhai, 2012). Formally, the objective is to assign one or more labels from a predefined set to each text sample using only the information contained in the text itself. While progress in supervised learning has led to substantial improvements in classification accuracy, these approaches rely on large-scale, high-quality annotated datasets. Constructing such datasets is often prohibitively expensive and time-consuming, particularly in specialized domains where expert annotation is required (Settles, 2012; Ratner et al., 2017).

Zero-shot text classification (ZSC) has emerged as a compelling alternative, enabling models to assign labels that were not observed during training ((Yin et al., 2019)). ZSC methods exploit the semantic relationships between input texts and candidate labels, typically leveraging pretrained language models that encode these relationships based on extensive pretraining over large corpora ((Brown et al., 2020; Liu et al., 2023)). A widely adopted approach is to prompt large language models (LLMs) with the input text and candidate label verbalizers, allowing the model to rank or score each label. While effective, this strategy incurs considerable computational cost and latency, limiting its practicality for large-scale or real-time applications ((Brown et al., 2020; Schick & Schütze, 2021; Liu et al., 2023)).

Concurrently, text embedding models have seen substantial progress ((Reimers & Gurevych, 2019; Gao et al., 2021; Muennighoff et al., 2023)). These models map textual inputs to dense vector spaces, positioning semantically similar texts close together. This structure enables efficient similarity-based retrieval and, in principle, supports zero-shot classification by embedding both input texts and candidate label representations into a shared space and applying nearest-neighbor matching ((Reimers

& Gurevych, 2019; Gao et al., 2021; Fei et al., 2022)). However, while such zero-shot approaches are theoretically feasible, their performance in practice is often limited, especially on challenging or fine-grained classification tasks. As a result, it is common to further adapt embedding models for classification by training a linear probe or classifier head using labeled data ((Neelakantan et al., 2022; Enevoldsen et al., 2025; Chung et al., 2025)), thereby reintroducing the need for annotated resources and undermining the zero-shot premise.

A parallel strand of research leverages external language and knowledge resources, including dictionary-style definitions, encyclopedic entries such as Wikipedia, and lexical ontologies such as WordNet, to provide semantic structure for zero-shot or "dataless" text classification. Early work introduced lexical resources to enrich text representations and label semantics ((Miller, 1995), see also (Scott & Matwin, 1998)), while Wikipedia-based methods mapped texts and labels into concept spaces using explicit semantic representations ((Gabrilovich & Markovitch, 2007)) and later demonstrated gains in downstream classification ((Wang et al., 2009)). More generally, dataless classification methods formalized how labels and documents can be compared via semantic proxies rather than task-specific annotations ((Chang et al., 2008)), and subsequent approaches operationalized label names and short natural-language descriptions as supervision signals for improved zero-shot performance ((Gao et al., 2023; Chai et al., 2020; Meng et al., 2020)).

Building on these insights, we introduce a description-only contrastive alignment framework specifically designed for embedding models in the zero-shot setting. Our approach requires only a handful of natural-language descriptions per label, each clarifying the types of documents a given label is intended to capture. We employ a contrastive objective that explicitly pulls each label verbalizer toward its associated descriptions, while pushing it away from unrelated descriptions. In this way, the model learns to capture the many-to-one correspondence between labels and their natural-language descriptions, fostering more robust and discriminative representations. Our formulation draws inspiration from foundational work in contrastive learning, such as DrLIM, InfoNCE, SimCLR, and CLIP, but adapts these ideas to the alignment of textual label verbalizers with natural-language descriptions ((Hadsell et al., 2006; van den Oord et al., 2018; Chen et al., 2020a; Radford et al., 2021b)).

2 Related Work

Zero-shot and "dataless" text classification. Early research in dataless classification replaced labeled data with semantic proxies such as label names, seed words, or external knowledge bases (e.g., WordNet, Wikipedia), enabling documents and labels to be compared in a shared semantic space (Miller, 1995; Scott & Matwin, 1998; Gabrilovich & Markovitch, 2007; Chang et al., 2008; Wang et al., 2009). More recent methods frame ZSC as textual entailment between input texts and label verbalizers, often leveraging pretrained language models to provide the entailment signal (Yin et al., 2019). Another line explores natural-language label descriptions (e.g., definitions or short summaries) as supervision, showing improved robustness and transfer across domains (Chai et al., 2020; Meng et al., 2020; Gao et al., 2023). Despite these advances, most approaches rely on cross-encoder architectures, which require jointly encoding each document with every candidate label at inference. This results in inference costs that scale linearly with the number of labels and prevents caching of document embeddings, making such methods impractical for large label sets or real-time deployment.

Few-shot learning. Few-shot methods fine-tune compact encoders on small labeled sets, bridging the gap between zero-shot and fully supervised learning. SetFit exemplifies this paradigm: it trains a sentence encoder contrastively, followed by a lightweight classifier head, achieving strong results with limited supervision and modest compute (Tunstall et al., 2022). Parameter-efficient fine-tuning (e.g., adapters, LoRA) further reduces trainable parameters (Houlsby et al., 2019; Hu et al., 2022), but these methods remain dependent on labeled examples, in contrast to purely description-driven zero-shot approaches.

In-context learning with large models. Large language models (LLMs) can perform zero- or few-shot classification via in-context learning (ICL), where label names and demonstration examples are provided in the prompt (Dong et al., 2024; Luo et al., 2024). While effective out of the box, ICL has limitations: sensitivity to demonstration choice, prompt length constraints, and high inference

costs. Comparisons show that fine-tuned encoders can be more stable and compute-efficient for sustained deployment on targeted tasks (Mosbach et al., 2023).

Embedding models for retrieval and ZSC. Recent embedding models trained with large-scale contrastive or instruction-tuning objectives (e.g., SBERT, SimCSE, E5, GTE, BGE, EmbeddingGemma, Qwen3-Embedding) provide strong transfer across retrieval and classification benchmarks (Reimers & Gurevych, 2019; Gao et al., 2021; Wang et al., 2022; Li et al., 2023; Xiao et al., 2023; Google DeepMind & Google Research, 2025; Zhang et al., 2025). These dual-encoder architectures allow independent encoding of documents and labels, enabling efficient nearest-neighbor classification. However, their naïve zero-shot performance often lags on fine-grained tasks, so the typical remedy is to train a linear probe or lightweight classifier head on top of frozen embeddings, which reintroduces the need for labeled data and departs from the zero-shot premise (Muennighoff et al., 2023; Neelakantan et al., 2022).

Contrastive learning for label-description alignment. Contrastive learning objectives such as InfoNCE (van den Oord et al., 2018), SimCLR (Chen et al., 2020a), and CLIP (Radford et al., 2021b) align paired views while separating negatives. Adapting this principle to text-only settings allows label names or descriptions to be treated as natural-language supervision signals. Our work follows this line: by aligning label verbalizers with small sets of curated descriptions, we encode many-to-one label-description correspondences in the embedding space, without relying on labeled documents.

3 Contrastive Label-Embedding Alignment

The core idea of our approach is exemplified in Figure 1. We start with a base text-embedding model $f_\theta : \text{text} \to \mathbb{R}^d$, pretrained on large-scale corpora with self-supervised objectives. Our method requires only a small set of natural-language descriptions per label, which elucidate the types of documents the label should encompass as a short paragraph. We then lightly finetune the embedding model using a contrastive learning objective that aligns each label verbalizer with its associated descriptions while repelling unrelated ones. Specifically, we combine InfoNCE over descriptions and verbalizers with a multi-positive variant that captures the many-to-one relation between labels and their descriptions.

More formally, let $\mathcal{Y} = \{1, \dots, L\}$ be the label set. For each $y \in \mathcal{Y}$ we assume (i) a short *verbalizer* v_y (e.g., for $y =$ Sports we have the verbalizer $v_y =$ "This news snippet is about sports."), and (ii) a small set of *label descriptions* $\mathcal{D}_y = \{d_y^k\}_{k=1}^{K_y}$. Table 3 in the appendix shows an example for AGNews (Zhang et al., 2015a). The set of all descriptions and its size are

$$\mathcal{D} = \bigcup_{y \in \mathcal{Y}} \mathcal{D}_y, \qquad D = \sum_{y \in \mathcal{Y}} K_y.$$

We use a single encoder f_θ with pooling map $\pi(\cdot)$[1] and ℓ_2 normalization,

$$e(t) = \frac{\pi\big(f_\theta(t)\big)}{\big\|\pi(f_\theta(t))\big\|_2} \in \mathbb{R}^d,$$

so cosine similarity equals the dot product. With temperature $\tau > 0$, the similarity between a description d and verbalizer v is

$$s(d, v) = \frac{e(d)^\top e(v)}{\tau}.$$

Batch structure and anchors. A training batch forms the cross-product between all descriptions $\mathcal{D}$ and all verbalizers $\{v_1, \dots, v_L\}$, yielding the score matrix

$$S \in \mathbb{R}^{D \times L}, \qquad S_{y\ell}^k = s\big(d_y^k,\, v_\ell\big).$$

We view *descriptions as row-anchors*: each row (a single d_y^k) should prefer its own label y over all other labels. Dually, *verbalizers are column-anchors*: each column ℓ should gather probability mass

[1] We use the pooling native to the pretrained model, e.g., CLS-token, mean, or last-token pooling.

from its positives $\{d_\ell^k\}_{k=1}^{K_\ell}$ while discounting descriptions of other labels. This row/column duality is important: rows enforce one-positive discrimination, while columns implement multi-positive aggregation.

Rowwise InfoNCE. Each description d_y^k has a single positive *for that row*, namely v_y, and $L-1$ negatives. The induced distribution over labels is

$$p(\ell \mid d_y^k) = \frac{\exp\{S_{y\ell}^k\}}{\sum_{j=1}^{L} \exp\{S_{yj}^k\}}.$$

The rowwise InfoNCE objective averages the cross-entropy against the correct label y:

$$\mathcal{L}_{\text{rows}} = \frac{1}{D} \sum_{y \in \mathcal{Y}} \sum_{k=1}^{K_y} \Big(\log \sum_{j=1}^{L} e^{S_{yj}^k} - S_{yy}^k \Big). \tag{1}$$

Intuitively, equation 1 pulls each d_y^k toward v_y while pushing it away from $v_{\ell \neq y}$.

Columnwise multi-positive InfoNCE. Each verbalizer v_ℓ has a *set* of positives $\mathcal{D}_\ell = \{d_\ell^k\}_{k=1}^{K_\ell}$; all d_y^k with $y \neq \ell$ are negatives. Define

$$Z_\ell = \sum_{y \in \mathcal{Y}} \sum_{k=1}^{K_y} \exp\{S_{y\ell}^k\}, \qquad Z_\ell^+ = \sum_{k=1}^{K_\ell} \exp\{S_{\ell\ell}^k\}.$$

The columnwise objective maximizes the aggregated positive mass against the global normalizer:

$$\mathcal{L}_{\text{cols}} = \frac{1}{L} \sum_{\ell=1}^{L} \big(\log Z_\ell - \log Z_\ell^+ \big). \tag{2}$$

This *multi-positive* term optimizes the *set-level* probability of a label's positives: strong descriptions can compensate for weaker or idiosyncratic ones (log-sum-exp behaves as a smooth max), and its gradient

$$\frac{\partial \mathcal{L}_{\text{cols},\ell}}{\partial S_{\ell\ell}^k} = \frac{e^{S_{\ell\ell}^k}}{Z_\ell} - \frac{e^{S_{\ell\ell}^k}}{Z_\ell^+}$$

induces *adaptive within-positive weighting* proportional to $e^{S_{\ell\ell}^k}/Z_\ell^+$, down-weighting outliers while emphasizing representative descriptions. Because the loss is computed per label and depends on the ratio Z_ℓ^+/Z_ℓ, it is also stable to the raw number of descriptions per class. Collectively, these properties pull each verbalizer toward the high-density region ("cloud") of its own descriptions while repelling it from other labels' descriptions.

Final objective. We use a simple symmetric combination

$$\mathcal{L} = \tfrac{1}{2}\, \mathcal{L}_{\text{rows}} + \tfrac{1}{2}\, \mathcal{L}_{\text{cols}}. \tag{3}$$

All off-diagonal pairs act as in-batch negatives, yielding an $O(DL)$ softmax per batch. Unit-norm embeddings constrain optimization to the hypersphere, and the temperature τ controls the sharpness of both row- and column-softmax distributions. We fix $\tau = 0.07$ following common practice in contrastive learning (Gao et al., 2021; Chen et al., 2020b; Radford et al., 2021a).

Inference. Given a document x, we compute $e(x)$ and score labels by similarity to verbalizers:

$$\text{score}(y \mid x) = e(x)^\top e(v_y), \qquad \hat{y} = \arg\max_{y \in \mathcal{Y}} \text{score}(y \mid x).$$

Geometric intuition. The rowwise term contracts each description toward its own verbalizer and expands margins to other labels. The columnwise term simultaneously moves each verbalizer toward the *barycenter* of its descriptions while pushing it away from non-matching descriptions. Together they produce a progressive alignment: initially scattered verbalizers and descriptions coalesce into

tight, label-specific clusters with widened inter-label separation. Figure 1 illustrates this on AGNews (Zhang et al., 2015a) with the canonical `all-MiniLM-L6-v2` model. In the *left* panel, stars (verbalizers) sit off-center relative to the document clouds, and class regions partially overlap. The *middle* panel depicts the learning forces: each description triangle d_y^k is pulled toward v_y and pushed away from other verbalizers; each v_y is pulled toward the barycenter of $\{d_y^k\}_k$ and repelled from other labels' descriptions. After optimization, the *right* panel shows verbalizers relocated near the densest part of their label's description cloud and larger inter-label margins. Although training uses only verbalizers and descriptions, the shared encoder is updated, globally reshaping the feature space: documents with similar semantics align to their label's "attractor direction," reducing within-class dispersion and increasing between-class separation. The 2-D UMAP view renders this as tighter, better-separated clouds in the right panel.

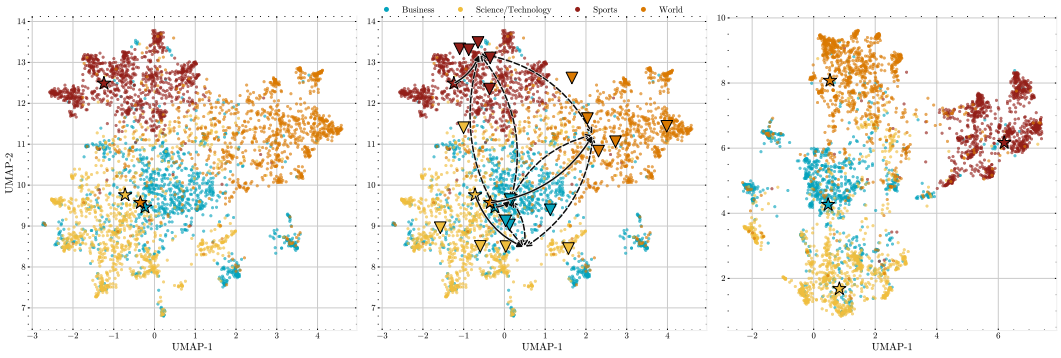


Figure 1: **AGNews (Zhang et al., 2015a).** Left: embeddings before finetuning (stars denote label verbalizers). Middle: schematic of our training forces (triangles denote label descriptions). Right: embeddings after finetuning.

3.1 Hyperparameters

Batching and training length. Because description sets are small, we treat one sweep over all description-verbalizer pairs as an *epoch* and, when memory is constrained, use gradient accumulation so a single optimizer update corresponds to one logical sweep. We cap the maximum iterations liberally and apply early stopping on the training loss with *patience* $= 10$ and tolerance $\Delta = 10^{-5}$: training halts if the loss fails to improve by at least Δ over 10 consecutive checks.

Learning rate and uniformity selection. Performance is sensitive to the learning rate (LR): too-aggressive LRs can trigger *representation collapse* (especially mode collapse (Bardes et al., 2021)) in our small-data regime, while simply reducing the LR avoids hard collapse but can stall progress and undercut alignment. We therefore adopt the view that contrastive learning balances *alignment* and *uniformity* on the hypersphere (Wang & Isola, 2020). In our setup, alignment is enforced by the supervision signal (descriptions $\leftrightarrow$ verbalizers), so the key is to *preserve uniformity* so the embedding space does not degenerate. Therefore, we select the LR using a label-free uniformity criterion computed on an unlabeled pool $\mathcal{X}_u = \{x_i\}$ from the target domain.

Let $z_i = e(x_i)$ be ℓ_2-normalized embeddings and $t > 0$ a scale parameter. Define

$$\mathcal{L}_{\text{uni}}(t) \;=\; \log \mathbb{E}_{i \neq j}\Big[e^{-t\,\|z_i - z_j\|_2^2} \Big] \;\approx\; \log\left(\frac{1}{M} \sum_{m=1}^{M} e^{-t\,\|z_{i_m} - z_{j_m}\|_2^2} \right), \tag{4}$$

where (i_m, j_m) are random distinct indices from $\mathcal{X}_u$. Lower values correspond to more uniform (i.e., less collapsed) embeddings. To select the learning rate, we run short warmups at candidate values and choose the one that *minimizes* $\mathcal{L}_{\text{uni}}(t)$; following Wang & Isola (2020), we fix $t = 2$. This criterion is label-free, computationally inexpensive, and in practice lower values correlate with stronger downstream performance. Figure 3 in the appendix illustrates this correlation across a range of models and datasets. As a fallback, reusing the base model's pretraining LR provides a safe, though non-optimized, choice.[2] Figure 2(b) illustrates AGNews document embeddings from

[2]This heuristic proved effective across many model-dataset combinations we tested.

all-MiniLM-L6-v2 at the LR chosen by this procedure; embeddings are reduced with PCA to $\mathbb{R}^3$ and projected onto the unit sphere $\mathbb{S}^2$ via ℓ_2-normalization.[3]

4 Experimental Setup

We evaluate our method across a diverse set of text-classification tasks, including topic classification (AGNews (Zhang et al., 2015a)), emotion recognition (EmotionDAIR (Saravia et al., 2018)), sentiment analysis (RottenTomatoes (Pang & Lee, 2005)), and fine-grained intent detection (Banking77 (Casanueva et al., 2020)), for a total of four datasets. Table 2 in the appendix provides the source and further details for each dataset. For each dataset and each class, we write exactly 5 short descriptions that characterize typical documents in that class. We do not tune these descriptions; they are generic summaries intended to capture the label semantics, and we leave optimizing description quality to future work. Table 3 in the appendix shows an example for AGNews (Zhang et al., 2015a). For Banking77 (Casanueva et al., 2020), we evaluate on six card-related intents to probe fine-grained classification.

We experiment with ten pretrained text-embedding models spanning a range of architectures and sizes (roughly 22M-600M parameters). Table 4 in the appendix summarizes the models used.

Training. We use the AdamW optimizer (Loshchilov & Hutter, 2019). We train for at most 1000 iterations with early stopping (patience $= 10$, tolerance $\Delta = 10^{-5}$), evaluating the stopping criterion every 10 steps. We sweep learning rates over $\{1, 3, 5\} \times \{10^{-4}, 10^{-5}, 10^{-6}\}$ and select the LR with the best uniformity score (Eq. 4). The uniformity score is computed on a pool of $50{,}000$ *pairs* (i, j) sampled from the *test subset* of the target-domain documents; these pairs are used when evaluating $\mathcal{L}_{\text{uni}}(t)$. For increased stability given the small training dataset, we apply a linear warmup of the learning rate during the first 50% of training steps.

Evaluation. We report macro F_1 to accommodate both binary and multi-class datasets (Sokolova & Lapalme, 2009).

5 Results and Analysis

We begin with the overall effect of label alignment across all models and datasets. Averaging over ten encoders and four benchmarks (Table 1), the method yields a consistent absolute macro-F1 improvement of $+\mathbf{0.10}$. This result demonstrates that aligning label verbalizers with a compact set of semantically rich descriptions provides a broadly applicable and reliable gain in zero-shot transfer.

The magnitude of improvement, however, varies by dataset. The largest benefits are observed on **AGNews**, where the mean increase reaches $+\mathbf{0.13}$ (range $+0.03$ to $+0.29$). Substantial gains also occur on **Banking77** (mean $+\mathbf{0.11}$, range $+0.02$ to $+0.29$) and **EmotionDAIR** (mean $+\mathbf{0.09}$, range $+0.01$ to $+0.30$). By contrast, **RottenTomatoes** records the smallest average improvement ($+\mathbf{0.06}$), but also the widest spread (from no gain up to $+0.35$), reflecting a strong dependence on the underlying encoder's initial quality.

Turning to dataset-specific winners, different models achieve the top post-finetuning performance. On **AGNews**, *Qwen3-Embedding-0.6B* reaches $\mathbf{0.84}$, while on **Banking77**, *GTE-base-en-v1.5* attains $\mathbf{0.96}$. For **EmotionDAIR**, *Qwen3-Embedding-0.6B* again leads at $\mathbf{0.58}$, and on **RottenTomatoes**, *GTE-large-en-v1.5* achieves $\mathbf{0.87}$. Considering macro-averaged F1 across tasks, *BGE-large-en-v1.5* and *GTE-large-en-v1.5* tie at the top with $\mathbf{0.79}$, closely followed by *GTE-base-en-v1.5*, *E5-large-v2*, and *Qwen3-0.6B* at $\mathbf{0.78}$.

The cost-benefit profile shows that both compact and larger encoders benefit, albeit in different ways. Smaller models often realize the largest relative improvements: for example, *all-MiniLM-L6-v2* improves by $+\mathbf{0.31}$ on average ($0.38 \rightarrow 0.69$), including $+\mathbf{0.36}$ on RottenTomatoes, while *embeddinggemma-300m* gains $+\mathbf{0.13}$ ($0.61 \rightarrow 0.74$). At the same time, the strongest model in the pool, *Qwen3-0.6B*, records an average $+\mathbf{0.09}$ and achieves best-in-class results on two datasets,

[3] We avoid UMAP because its locality-crowding parameters can arbitrarily distort interpoint distances, making it unsuitable for objectively visualizing uniformity.

underscoring that substantial gains are not limited to smaller encoders. In contrast, families such as E5 and GTE start from already high baselines (0.82-0.87 on RottenTomatoes), which naturally constrains the headroom for further improvement and results in more modest deltas.

Examining family-level trends, the improvements are relatively stable. The E5 models consistently show average deltas between $+\mathbf{0.05}$ and $+\mathbf{0.07}$, BGE models between $+\mathbf{0.07}$ and $+\mathbf{0.08}$, and GTE models between $+\mathbf{0.04}$ and $+\mathbf{0.06}$. Despite these modest increments, GTE remains among the strongest performers after finetuning. Meanwhile, Gemma and Qwen models perform above expectation given their parameter counts, with Qwen notably securing the top scores on AGNews and EmotionDAIR.

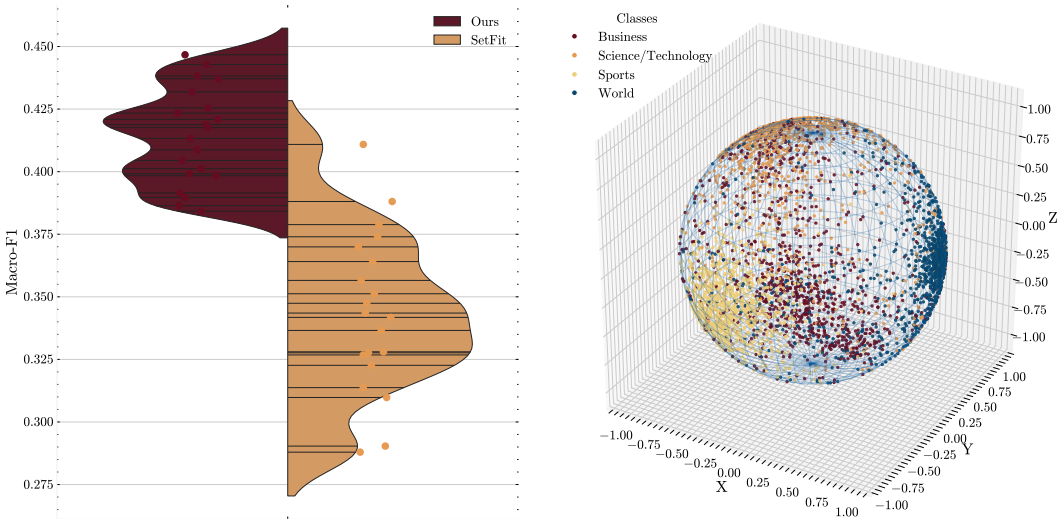
The distribution of improvements also provides insight into dataset difficulty. **EmotionDAIR** emerges as the most challenging benchmark: even the best finetuned model reaches only $\mathbf{0.58}$ macro-F1. This suggests that emotion recognition may require not only richer descriptions but also multiple complementary ones per label, so that different linguistic manifestations of the same emotion are adequately represented. In contrast, **AGNews** and **Banking77** benefit most strongly from description alignment, consistent with the fact that topical and intent-based semantics are well captured by concise definitions. On **RottenTomatoes**, the degree of improvement inversely correlates with the encoder's baseline quality: weaker models gain considerably, while stronger ones improve only marginally.

Few-shot comparison. To contextualize our zero-shot description-only alignment, we compare against SetFit (Tunstall et al., 2022), a widely used few-shot method for embedding models. SetFit combines a contrastive pretraining stage with a lightweight classifier head. Following the original setup, we train SetFit on EmotionDAIR with 8 samples per class and repeat the procedure 20 times with different random draws of the training set. Our approach, by contrast, uses 5 descriptions per class and generates 20 variations of the descriptions. Figure 2(a) shows that our method achieves a higher average macro-F1 and, more importantly, exhibits substantially smaller variance. While SetFit can occasionally match or exceed our performance, it displays a long tail of poor outcomes, reflecting its sensitivity to the specific few-shot samples selected.

In summary, description-only finetuning yields consistent performance gains across a diverse set of encoders and tasks. The largest improvements occur on topic and intent classification datasets, while emotion recognition remains comparatively difficult. The method is particularly attractive in low-compute settings, where smaller models realize disproportionate benefits, yet even state-of-the-art encoders record large positive gains.

Model	AGNews	Banking77	EmotionDAIR	RottenTomatoes	Avg
MiniLM					
all-MiniLM-L6-v2	0.47	0.59	0.13	0.34	0.38
trained	0.76 (+0.29)	0.88 (+0.28)	0.43 (+0.30)	0.69 (+0.36)	0.69 (+0.31)
E5					
e5-base-v2	0.76	0.80	0.45	0.84	0.71
trained	0.80 (+0.04)	0.94 (+0.14)	0.47 (+0.02)	0.83 (+0.00)	0.76 (+0.05)
e5-large-v2	0.77	0.79	0.43	0.85	0.71
trained	0.80 (+0.03)	0.94 (+0.15)	0.53 (+0.10)	0.85 (+0.00)	0.78 (+0.07)
BGE					
bge-base-en-v1.5	0.63	0.86	0.42	0.82	0.68
trained	0.82 (+0.18)	0.95 (+0.09)	0.46 (+0.04)	0.82 (+0.00)	0.76 (+0.08)
bge-large-en-v1.5	0.75	0.84	0.45	0.82	0.72
trained	0.81 (+0.06)	0.94 (+0.10)	0.55 (+0.10)	0.85 (+0.03)	0.79 (+0.07)
GTE					
gte-base-en-v1.5	0.73	0.87	0.44	0.84	0.72
trained	0.83 (+0.10)	**0.96** (+0.09)	0.50 (+0.06)	0.85 (+0.01)	0.78 (+0.06)
gte-modernbert-base	0.75	0.88	0.46	0.82	0.73
trained	0.81 (+0.06)	0.95 (+0.07)	0.47 (+0.01)	0.85 (+0.03)	0.77 (+0.04)
gte-large-en-v1.5	0.72	0.93	0.40	0.87	0.73
trained	0.83 (+0.11)	0.95 (+0.03)	0.51 (+0.10)	**0.87** (+0.01)	0.79 (+0.06)
Qwen					
Qwen3-Embedding-0.6B	0.64	0.89	0.48	0.76	0.69
trained	**0.84** (+0.20)	0.91 (+0.02)	**0.58** (+0.09)	0.81 (+0.04)	0.78 (+0.09)
Gemma					
embeddinggemma-300m	0.53	0.81	0.50	0.59	0.61
trained	0.72 (+0.18)	0.93 (+0.12)	0.57 (+0.07)	0.73 (+0.14)	0.74 (+0.13)

Table 1: Main results by model family. Each model has a base row (zero-shot) and a trained row, with F1 scores and improvements reported in percentage points. Best trained F1 per dataset is **bold**. For each model, the largest improvement is underlined. Averages are macro-averages across datasets; trained averages include the mean improvement in parentheses.



(a) EmotionDAIR results

(b) AGNews embeddings

Figure 2: (a) Performance comparison of our approach with SetFit (Tunstall et al., 2022) on the EmotionDAIR dataset (Saravia et al., 2018), across 20 runs with different sampled training sets. (b) Visualization of AGNews document embeddings after finetuning `all-MiniLM-L6-v2`, projected onto the hypersphere using PCA; colors indicate class membership.

6 Conclusion and Future Work

We introduced *description-only supervision* for zero-shot text classification with embedding models, a contrastive label-embedding alignment method that relies only on short, human-written descriptions per label. By aligning label verbalizers with their descriptions via a multi-positive contrastive objective, the approach yields consistent, architecture-agnostic improvements over naïve zero-shot use of embeddings, averaging **+0.10** macro-F1 across ten encoders and four datasets. Compared with a few-shot SetFit pipeline using 8 labeled examples per class, our method attains higher average performance with markedly smaller variance across repeated runs, despite using no labeled documents, making it well suited for settings with tight annotation and compute budgets.

Looking ahead, we see value in examining how performance scales with the number and diversity of descriptions per class, providing clearer guidance on how much description-level supervision is needed in practice. Another promising direction is to investigate the role of hyperspherical uniformity more deeply, both in its correlation with downstream performance and in ways it can be integrated directly into the training objective. Extending the approach to multilingual settings offers a natural testbed for evaluating generality and robustness across languages. Finally, tightening the theoretical connection between uniformity, alignment dynamics, and generalization may yield sharper insights into why description-only supervision is effective and how it can be further improved.

Reproducibility Statement

We will release the full codebase under an MIT license, including preprocessing scripts, training and evaluation routines, the uniformity-based learning-rate selection code, and all logging utilities required to regenerate figures and tables. The label-description sets and the exact sampling protocol used for the uniformity metric will be made publicly available. Fine-tuned checkpoints for all reported models will be released in Hugging Face format, and we will document the exact pretrained encoder revisions used.

All hyperparameters (optimizer, learning-rate grids, batch sizes, gradient accumulation, early-stopping criteria) are specified in the main text at the point of use; the appendix provides additional in-depth results. Experiments were run on NVIDIA A100 80 GB GPUs, with inference carried out in `bfloat16`. We provide pinned package versions and configuration files to recreate the software environment.

We do not fix random seeds during training. Instead, we verified empirically that the results and conclusions are robust to stochasticity in initialization and sampling. We rely only on publicly available datasets and pretrained encoders, which are properly cited. To our knowledge, there are no legal or technical restrictions that would prevent exact reproduction of our results.

References

Charu C. Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining Text Data*, pp. 163–222. Springer, 2012. doi: 10.1007/978-1-4614-3223-4_6.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. URL https://arxiv.org/abs/2105.04906.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

Diyi Chai, Hongliang Fei, and Ping Li. Description based text classification with reinforcement learning. In *Proceedings of ICML*, volume 119 of *Proceedings of Machine Learning Research*,

pp. 1385–1395, 2020. URL https://proceedings.mlr.press/v119/chai20a/chai20a.pdf.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pp. 1597–1607, 2020a. URL https://proceedings.mlr.press/v119/chen20j/chen20j.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020b.

Isaac Chung, Imene Kerboua, Marton Kardos, Roman Solomatin, and Kenneth Enevoldsen. Maintaining mteb: Towards long term usability and reproducibility of embedding benchmarks. *arXiv preprint arXiv:2506.21182*, 2025.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. In *arXiv preprint arXiv:1805.10190*, 2018. URL https://arxiv.org/abs/1805.10190.

Xinwei Dong, Shujian Huang, Jiajun Chen, et al. A survey on in-context learning. *EMNLP 2024*, 2024. URL https://aclanthology.org/2024.emnlp-main.64/.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

Yu Fei, Zhao Meng, Ping Nie, Roger Wattenhofer, and Mrinmaya Sachan. Beyond prompting: Making pre-trained language models better zero-shot learners by clustering representations. In *Proceedings of EMNLP*, pp. 8560–8579, 2022. URL https://aclanthology.org/2022.emnlp-main.587.pdf.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. The benefits of label-description training for zero-shot text classification. In *Proceedings of EMNLP*, 2023. URL https://aclanthology.org/2023.emnlp-main.853/.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, 2021. URL https://arxiv.org/abs/2104.08821.

Google DeepMind and Google Research. Introducing embeddinggemma: Best-in-class open multilingual text embeddings under 500m parameters. Google Developers Blog, 2025. URL https://developers.googleblog.com/en/introducing-embeddinggemma/. Accessed Sep 2025.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of CVPR*, pp. 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of ICML*, 2019. URL https://arxiv.org/abs/1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv:2106.09685, 2022. URL https://arxiv.org/abs/2106.09685.

Steven Larson, Anish Mahendran, Andrew Lee, and et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP*, pp. 1311–1316, 2019. URL https://aclanthology.org/D19-1131.pdf.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Building Efficient Universal Classifiers with Natural Language Inference, December 2023. URL http://arxiv.org/abs/2312.17543. arXiv:2312.17543 [cs].

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. arXiv:2308.03281, 2023. URL https://arxiv.org/abs/2308.03281.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in nlp. *ACM Computing Surveys*, 2023. URL https://arxiv.org/abs/2107.13586.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*, 2024. URL https://arxiv.org/abs/2401.11624.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of ACL-HLT*, pp. 142–150, 2011. URL https://aclanthology.org/P11-1015/.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, and Jiawei Han. Text classification using label names only: A language model self-training approach. In *Proceedings of EMNLP*, 2020. URL https://aclanthology.org/2020.emnlp-main.724.pdf.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Marius Mosbach, Nicolas Meier, Michael A. Hedderich, and Dietrich Klakow. Few-shot fine-tuning vs. in-context learning: A fair comparison on challenging datasets. In *Findings of ACL*, 2023. URL https://aclanthology.org/2023.findings-acl.779/.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Reza Rezagholizadeh, Giuseppe Attanasio, Noémie Leprêtre, Daniel J. Beutel, Milad Moradi, Yacine Jernite, Douwe Kiela, et al. Mteb: Massive text embedding benchmark. In *Proceedings of EACL*, 2023. URL https://aclanthology.org/2023.eacl-main.148/.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. arXiv:2201.10005, 2022. URL https://arxiv.org/abs/2201.10005.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*, 2005.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, 2021b. URL https://arxiv.org/abs/2103.00020.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *PVLDB*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. URL https://aclanthology.org/D19-1410.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of EMNLP 2018*, pp. 3687–3697, Brussels, Belgium, 2018. doi: 10.18653/v1/D18-1404.

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few-shot text classification and natural language inference. In *Proceedings of EACL*, 2021. URL https://arxiv.org/abs/2001.07676.

Sam Scott and Stan Matwin. Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems (Workshop at COLING-ACL)*, pp. 45–52, 1998.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002. doi: 10.1145/505282.505283.

Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012. doi: 10.2200/S00429ED1V01Y201207AIM018.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Benoit Delangue, Leandro von Werra, Abhishek Thakur, Philipp Schmid, Sylvain Gugger, Omar Sanseviero, and Nils Reimers. Efficient few-shot learning without prompts. In *NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing*, 2022. URL https://arxiv.org/abs/2209.11055.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. arXiv:2212.03533, 2022. URL https://arxiv.org/abs/2212.03533.

Pu Wang, Jian Hu, Hua-Jun Zeng, and Zheng Chen. Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19(3):265–281, 2009. doi: 10.1007/s10115-008-0152-4.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Shitao Xiao, Fan Cui, Yuxiang Zhang, et al. Packed resources for general chinese embeddings (bge) and flagembedding. arXiv:2309.07597, 2023. URL https://arxiv.org/abs/2309.07597.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of EMNLP-IJCNLP*, pp. 3914–3923, 2019. URL https://aclanthology.org/D19-1404.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015a.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of NeurIPS*, 2015b. URL https://arxiv.org/abs/1509.01626.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv:2506.05176, 2025. URL https://arxiv.org/abs/2506.05176.

A Datasets Overview

Table 2 summarizes the datasets used in our experiments, including their domains, number of classes, sources, and license terms. All datasets are publicly available via Hugging Face Datasets[4]. For label verbalizers, we follow the setup of Laurer et al. (2023).

Task	Domain	Dataset	# Classes	Source	License
Emotion	Social media	EmotionDAIR	6	Saravia et al. (2018)	Research/education only
Intent	Banking	Banking77	6[5]	Casanueva et al. (2020)	CC BY 4.0
Sentiment	Movies	RottenTomatoes	2	Pang & Lee (2005)	CC0 1.0
Topic	News	AGNews	4	Zhang et al. (2015a)	Non-commercial

Table 2: Datasets used in the evaluation, covering emotion recognition, intent detection, sentiment analysis, and topic classification.

B Label Descriptions

[4] https://huggingface.co/datasets

[5] For *Banking77*, we restrict evaluation to the six card-related intent classes for fine-grained classification.

Category (verbalizer)	Descriptions
World News *Verbalizer:* "This example news text is about world news."	• Coverage of international affairs and geopolitics: governments, elections, diplomacy, conflicts, treaties, and sanctions. Stories focus on cross-border events and their global implications rather than domestic business or sports outcomes. • News about countries interacting on the world stage: summits, UN resolutions, regional alliances, and humanitarian crises. Emphasis is on state actors, policy decisions, and shifts in international relations. • Reporting on wars, ceasefires, peace talks, and military deployments across regions. Articles highlight causes, stakeholders, civilian impact, and reactions from other nations or international bodies. • Global society and policy issues such as migration, human rights, climate diplomacy, and development aid. Pieces track how multiple countries respond and coordinate. • International incidents and disasters (natural or man-made) where response, accountability, and cross-national coordination are central. Focus remains on worldwide context rather than local business ramifications.
Sports *Verbalizer:* "This example news text is about sports."	• Results, previews, and analysis of professional or amateur competitions: matches, tournaments, standings, and championships. Content centers on performance, tactics, and outcomes on the field. • Athlete-focused updates including injuries, transfers, contracts, and retirements. Stories emphasize team dynamics and competitive impact. • Coverage of leagues and events: scheduling, rule changes, drafts, and officiating controversies. The angle is sporting governance and competitive fairness. • Game recaps and statistical breakdowns highlighting key plays, records, and milestones. The narrative ties individual performances to team results. • Profiles and human-interest features about coaches, players, and training methods. Emphasis is on preparation, strategy, and competitive psychology.
Business *Verbalizer:* "This example news text is about business news."	• Corporate news: earnings, revenue guidance, layoffs, executive changes, and strategic shifts. Articles assess company performance and shareholder impact. • Markets and finance coverage: stocks, bonds, commodities, currencies, and macro sentiment. Focus is on price moves, drivers, and investor reactions. • Mergers, acquisitions, IPOs, and venture funding. Pieces explain valuations, synergies, and regulatory hurdles. • Industry developments such as competition, supply chains, pricing, and business models across sectors. Reporting connects firm-level actions to market structure. • Policy and regulation affecting commerce: antitrust cases, trade policy, taxes, and compliance. The lens is how rules shape corporate behavior and profitability.
Science & Technology *Verbalizer:* "This example news text is about science and technology."	• Scientific research findings across fields like biology, physics, medicine, and climate science. Articles emphasize methods, evidence, and potential applications or limitations. • Technology product and platform news: hardware, software, mobile, cloud, and consumer gadgets. Coverage focuses on features, performance, and user impact. • AI, data science, and computing breakthroughs including models, chips, algorithms, and benchmarks. Stories discuss capabilities, risks, and real-world use cases. • Space and astronomy updates: launches, missions, telescopes, and planetary discoveries. Coverage highlights scientific goals and engineering challenges. • Cybersecurity and privacy incidents: vulnerabilities, breaches, hacks, and defenses. Reporting centers on technical cause, affected users, and mitigations.

Table 3: AGNews (Zhang et al., 2015a) class verbalizer and class descriptions (5 per class).

C Models Overview

Model	Yr	Arch.	Backbone	FT / train data	# P	Pool	Dim
all-MiniLM-L6-v2	2021	enc.	MiniLM	1B paired sentences	22.7M	mean	384
e5-base-v2	2023	enc.	E5 (BERT)	270M synthetic contrastive	110M	mean	768
e5-large-v2	2023	enc.	E5 (BERT)	same as above	335M	mean	1024
bge-base-en-v1.5	2023	enc.	BGE (RoB.)	1.5B pair data, contrastive	137M	CLS	768
bge-large-en-v1.5	2023	enc.	BGE (RoB.)	same as above	434M	CLS	1024
gte-base-en-v1.5	2024	enc.+	GTE	MLM + contrastive pre-train	137M	CLS	768
gte-large-en-v1.5	2024	enc.+	GTE	same as above	434M	CLS	1024
gte-modernbert-base	2024	enc.	ModernBERT	same as above	149M	CLS	768
embeddinggemma-300m	2025	enc.	Gemma 3 (enc.)	Multiling. corpus (320B tok), contrastive	308M	mean	768*
Qwen3-Embedding-0.6B	2025	dec.	Qwen3	synthetic multiling. contrastive	0.6B	last	1024

Table 4: Architectural and training overview of the 10 embedding models used. Columns list publication year (Yr), encoder/decoder architecture (Arch.), backbone, principal fine-tuning (FT) or pre-training data, parameter count (#P), pooling strategy (Pool), and embedding dimensionality (Dim). *For embeddinggemma-300m, dimensionality 768 corresponds to Matryoshka Representation Learning (MRL) with nested sizes 512/256/128, a training scheme enabling shorter embeddings.

D Relationship Between Uniformity and Macro-F1 Across Learning Rates

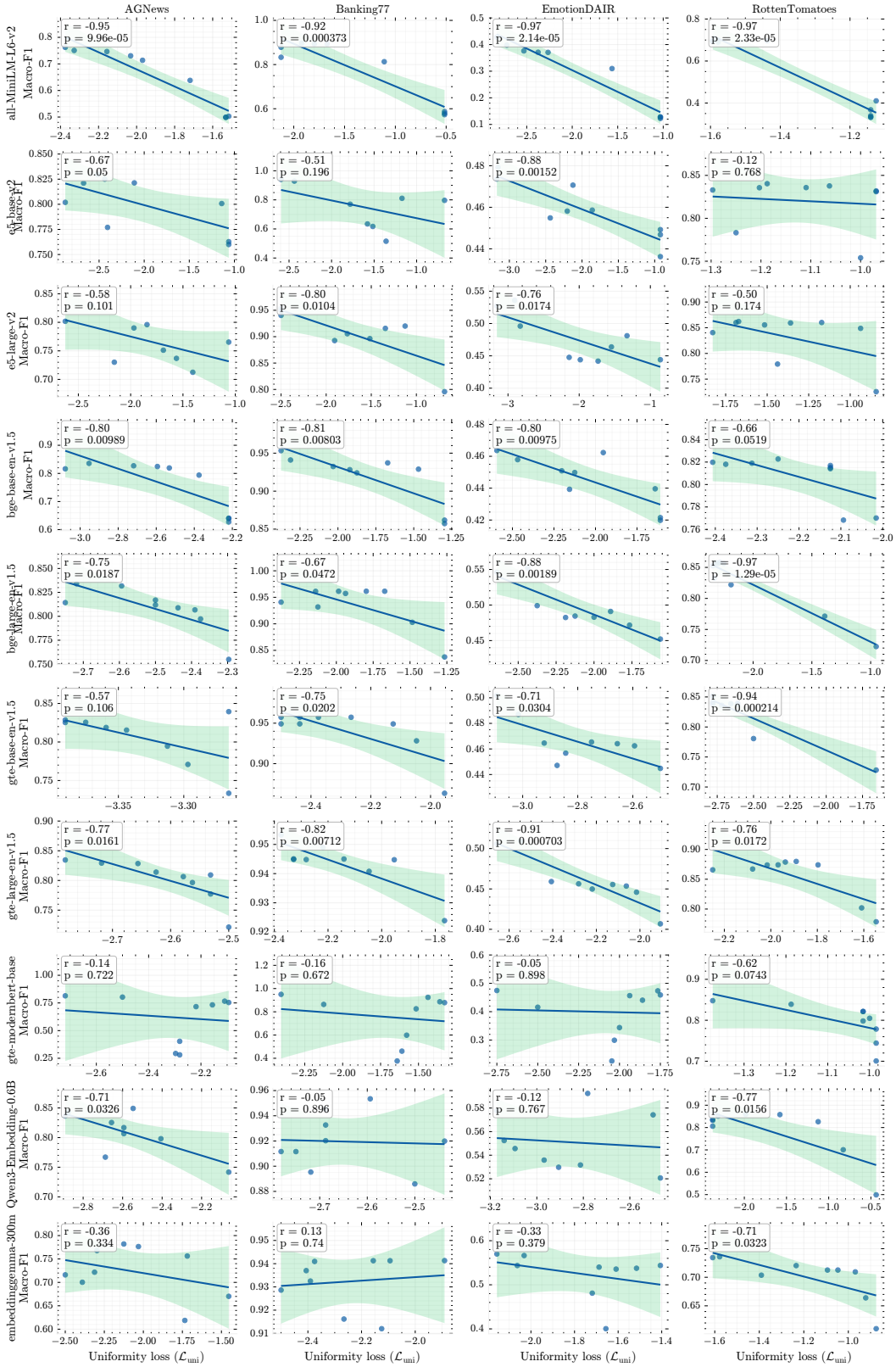


Figure 3: Scatter plots of uniformity loss $\mathcal{L}_{\text{uni}}$ against Macro-F1 performance across datasets. Rows correspond to embedding models, while columns correspond to datasets. Each subplot shows individual runs with a different learning rate (dots), an ordinary least squares regression line with 95% confidence interval (shaded), and the Pearson correlation coefficient between $\mathcal{L}_{\text{uni}}$ and Macro-F1.