Push and Pull: Training Sentence Encoders with Contrastive Losses for Distance-Based Multi-Label Text Classification

Jens Van Nooten, Andriy Kosar
University of Antwerp, Textgain
jens.vannooten@uantwerpen.be, andrew@textgain.com

Abstract

Despite the potential of Distance-Based Classification (DBC), a method that assigns labels to text by measuring semantic similarity between the text and the label representations, it has received very little attention for Multi-Label Text Classification (MLTC). Previous studies have focused on determining optimal thresholds, reaching promising results with contextual sentence encoders. We demonstrate that the performance of these models can be further improved by training them with contrastive losses, i.e., by bringing text representations closer to the corresponding true label representations in an embedding space. Using three supervised contrastive losses and three sentence encoders (Stella, GIST-Large, and BGE), we evaluated our approach on five English datasets (SemEval, BioTech, Reuters, AAPD, and LitCovid) and one Dutch dataset (EventDNA). The results show consistent substantial improvements over base sentence encoders, thereby narrowing the gap between DBC methods and fine-tuned or zero-shot approaches.

Keywords: multi-label classification, semantic similarity, sentence encoders

1. Introduction

Multi-label text classification (MLTC) is a classification problem that consists of predicting one or multiple correct labels from label set L for a single text, where $|L| > 2$. Despite its wide applicability to multiple domains, such as medical (Veeranki et al., 2024), law (Song et al., 2022), news (Lin et al., 2018) and commerce (Deniz et al., 2022), MLTC remains a challenging task due to large, sparse label sets and label ambiguity. To counter these challenges, a wide variety of solutions have been proposed, including a limited number that focus on computationally efficient approaches. One such solution is distance-based classification (DBC), as explored in Veeranna et al. (2016), which is based on the intuition that texts are semantically more similar to correct labels than false labels. This paradigm within classification uses semantic similarity between text and label representations to determine the relevance of a label to a text, making it computationally efficient and adaptable to evolving label sets, in contrast with gradient-based, probabilistic methods, such as fine-tuning pre-trained neural networks.

Most research on DBC focuses on multi-class classification, which usually involves assigning the most similar label in terms of cosine similarity to a text (Chang et al., 2008; Kosar et al., 2023). However, applying DBC to MLTC is difficult due to the multiple true labels for a single text, requiring a threshold value to be determined, as visualized in Figure 1 (Mylonas et al., 2020a; Sarkar et al., 2023).

To address this difficulty, Nooten et al. (2025) proposed calculating label-specific similarity threshold values between a text and labels to perform classification. The authors also highlighted the adaptability of DBC to expanding label sets, the low computational load and requiring little annotated data. While the proposed calibration method provides improvements, it also has limitations in cases where label embeddings do not adequately capture its semantic meaning. To build on this prior work and overcome these limitations, we propose to fine-tune sentence encoders with contrastive losses, enabling the models to explicitly learn relevant text-label similarities. We observed substantial improvements compared to baseline DBC methods on all investigated datasets, in addition to matched or better performance than more expensive classification methods (generative LLMs and *BERT* models).

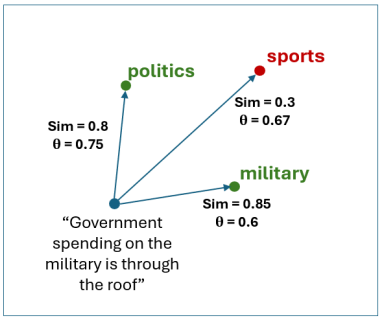

Figure 1: Abstraction of distance-based MLTC. Texts and labels are embedded in a joint label space, after which the cosine similarity is calculated to determine label relevance. If the similarity between a text and a label exceeds a predefined threshold θ, the label is assigned to the text.

Our contributions are as follows:

1. We present a novel DBC-based approach that integrates contrastive learning to maximize the

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7359–7379
11–16 May 2026.

information gained from limited annotated data, achieving major improvements compared to baseline DBC methods and competitive results compared to larger LLMs.
2. We evaluate several contrastive learning losses to determine which performs best for MLTC and demonstrate the overall effectiveness of contrastive learning for DBC.
3. We demonstrate the flexibility and cross-lingual adaptability of our approach through experiments on English and Dutch datasets with varying levels of label granularity.
4. We release the implementation of our approach as an open-source pipeline to support further research and applications for MLTC[1].

2. Related Research

2.1. Multi-label Classification

Multi-label (Text) Classification (ML(T)C) poses some unique challenges compared to traditional single-label classification problems due to the possibility of assigning multiple true labels to a single instance. The most prominent challenges are the explosive number of possible label combinations as the label set size increases (Zhang and Zhou, 2014), modeling label dependencies (Gibaja and Ventura, 2014), label imbalance (Tarekegn et al., 2021) and the high computational cost of probabilistic methods (Li et al., 2024a). Previous studies have addressed these challenges mainly by training models with contrastive losses (Audibert et al., 2025), integrating label correlations (Huang et al., 2024) and augmenting the training data (Song et al., 2023; Van Nooten and Daelemans, 2023).

More recently, the development of sophisticated and well-performing Large Language Models, has inspired researchers to apply them for MLTC. This moved the focus to few-shot or zero-shot multi-label classification (Peskine et al., 2023; Niraula et al., 2024; Ma et al., 2025). However, these papers also highlight that the supervised fine-tuning of BERT models yields superior results.

2.2. Distance-Based Classification

Distance-based Classification (DBC) involves embedding texts and labels in a joint space, after which the similarity between vectors is measured to determine label relevance. Chang et al. (2008) was the first to adopt this classification approach, measuring the cosine similarity between bag-of-words (BOW) representations of texts and Wikipedia class concepts to perform classification. Subsequent studies expanded upon this work by applying it to hierarchical or cross-lingual text classification (Song and Roth, 2014; Song et al., 2016). Recent research has adopted similar approaches, utilizing alternative representations such as encoding texts with neural word embeddings or sentence embeddings (Kosar et al., 2022; Schopf et al., 2023; Kosar et al., 2023).

While most studies incorporate DBC in single-label classification, this approach is unfeasible in a multi-label setting, where each item can have more than one true label. Several studies have addressed this additional challenge by introducing various thresholding methods: if the (cosine) similarity exceeds a pre-defined threshold value, the label is assigned to the text (Sappadla et al., 2016; Veeranna et al., 2016; Mylonas et al., 2020a; Mustafa et al., 2021; Sarkar et al., 2022; Mukherjee and Jha, 2024). However, Nooten et al. (2025) highlighted that previously established uniform thresholding methods, i.e. using the same threshold value for all labels in a dataset, produce suboptimal results. To overcome this limitation, the authors introduced label-specific thresholds, i.e. thresholds optimized for each label separately, which substantially improved the performance of DBC.

2.3. Contrastive Learning

Contrastive learning has been pivotal in many advances in computer vision and NLP tasks (Hu et al., 2024). In NLP, it has been applied to a wide variety of tasks such as text classification, data augmentation, and machine translation. For a comprehensive overview, see (Zhang et al., 2022a).

Moreover, contrastive learning has been widely adopted for classification and clustering problems to train more robust and semantically informed embeddings, including MLTC (Lin et al., 2023; U et al., 2023; Van Nooten and Daelemans, 2025; Zhang et al., 2022b; Zhang and Wu, 2024). For a complete overview, consult Audibert et al. (2025).

In the context of DBC, contrastive learning has been applied but not studied systematically. Kosar et al. (2023) compared the contrastive and alternative loss functions for tailoring sentence embeddings for DBC for topical text classification, demonstrating that the online contrastive loss outperforms cosine similarity and contrastive losses, though it is less effective than the multi-negative ranking loss. In the context of MLTC within DBC, while contrastive losses have been employed, a systematic study of their impact is still lacking.

[1] https://github.com/clips/push_and_pull

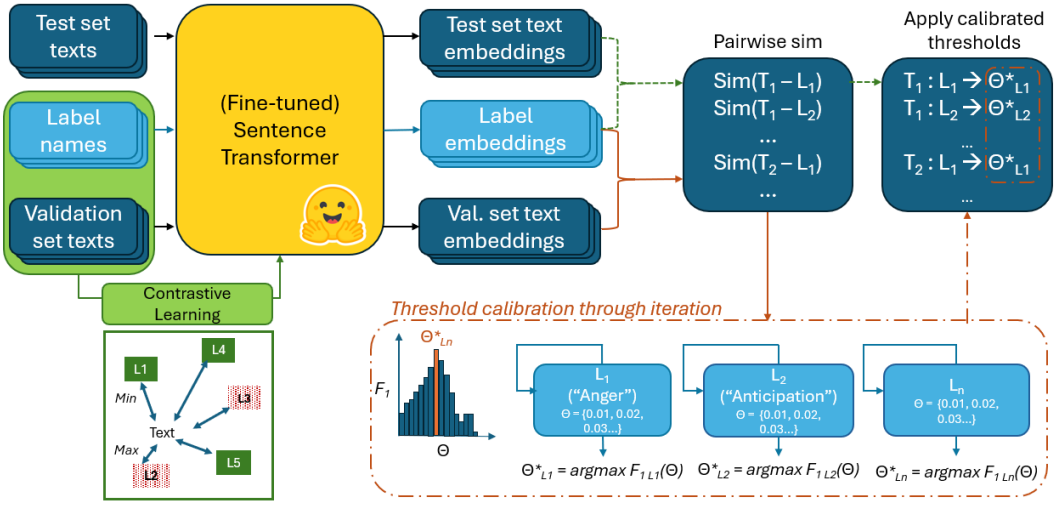

Figure 2: Overview of the contrastive learning, thresholding approach and inference stage. Texts and label representations from the annotated sets are embedded and used for contrastive learning, after which the cosine similarity is used to determine label relevance. Thresholds are optimized based on performance on the same validation set.

3. Methodology

3.1. Datasets

We conduct experiments on six datasets in total, of which five are English and one is Dutch. The latter serves to demonstrate the flexibility and cross-lingual adaptability of our approach. The English datasets include the SemEval 2018 (English subset) dataset (Mohammad et al., 2018), the BioTech news dataset[2], a modified version ("Apté Mod") of the Reuters-21578 dataset (Apt'e et al., 1994)[3], the arXiv Academic Papers Dataset (AAPD) (Yang et al., 2018), and the LitCovid dataset (Chen et al., 2020). All datasets, including the Dutch EventDNA dataset (Colruyt et al., 2022), are summarized in Table 1. It should be noted that we conduct experiments on two versions of the hierarchical EventDNA dataset, namely one version that includes only the top-level labels and another that includes only the second level labels. We created stratified splits for the AAPD, LitCOVID and EventDNA datasets following the method described in Sechidis et al. (2011).

Since the labels in the AAPD dataset are abbreviations, we mapped them to their corresponding descriptive names as provided on the arXiv website. Additionally, we translated the labels in the EventDNA dataset from English to Dutch and reviewed them to ensure their correctness.

[2] https://blog.knowledgator.com/finally-a-decent-multi-label-classification-benchmark-is-created-a-prominent-zero-shot-dataset-4d90c9e1c718

[3] https://huggingface.co/datasets/ucirvine/reuters21578

Dataset	Text Type	Task	N Train	N Val	N Test	N lbls	N lbls/ Text	Mn Tkns	Mdn Tkns
SemEval	tweets	emotion	6,837	886	3,259	11	2.38	27	28
BioTech	news	topics	2,344	414	381	31	1.84	655	572
Reuters	news	topics	7,493	1,323	3,023	118	1.01	180	121
AAPD	abstracts	topics	53,840	1,000	1,000	52	2.41	128	99
LitCOVID	abstracts	topics	21,204	3,756	6,239	7	1.37	303	292
EventDNA (top)	news	topics	1,238	268	265	17	1.54	80	77
EventDNA (btm)	news	topics	1,226	249	296	494	4.46	78	76

Table 1: Statistics for each dataset used for the experiments.

Following Nooten et al. (2025), we adjusted the label names of the BioTech, Reuters and LitCovid datasets, as some labels were ambiguous or unclear, which led to degraded performance on those labels (cf. Table 9, Appendix A).

3.2. Sentence Encoders

We evaluate three state-of-the-art sentence encoders on the English datasets, and one Dutch sentence encoder on EventDNA (cf. Table 2). These models were selected for their relatively small size (less than one billion parameters), which allows them to be deployed on a single *NVIDIA GeForce RTX 2080 Ti* GPU with 11GB of VRAM.

3.3. Contrastive Losses

We evaluate three contrastive losses by fine-tuning the previously mentioned four sentence encoders (cf. Table 2). For each contrastive loss, we assume limited access to the entire training dataset to simulate a realistic low-resource setting. Therefore, we use the annotated validation splits from each dataset as our designated labeled training sets. The supervised contrastive losses that we

Model	Lang.	Size (M)	Embedding Dim.	Max Tokens
(1) Stella	EN	435	8192	512
(2) GIST-Large	EN	335	1024	512
(3) BGE	EN	109	768	512
(4) BERTje-nli	NL	110	768	512

Table 2: Overview of Embedding Models[5]

explore are the following[4]:

Label Margin Loss (L_{LM}) Given are text embedding T, label embedding set L where L^+ is the subset of true labels for T and L^- is the false set of labels. Both T and L are obtained by embedding text t and label set l using a sentence encoder M:

$$T = M(t)$$
$$L = \{M(l_i) \mid l_i \in \mathcal{L}\}$$

This loss minimizes the cosine distance (cos) between T and L^+, in addition to maximizing the distance between T and k most similar labels from L^-. A margin θ is forced between L^+ and L^- in the embedding space. Thus, L_{LM} for a single text embedding T can be defined as follows:

$$\mathcal{L}_{\mathcal{LM}} = ReLu\left(\cos(T, L^-) - \cos(T, L^+) + \theta\right) \quad (1)$$

The optimal value for k is determined by grid-searching multiple values and choosing the value that yields the highest classification performance on the validation set.

Text Margin Loss (L_{TM}) This loss function is the same as the label margin loss, but the margin θ is forced between a text embedding T and the k most similar labels from L^-. Thus, L_{TM} for a single text embedding T can be defined as follows:

$$\mathcal{L}_{\mathcal{TM}} = \left(\cos(T, L^+) - ReLu(\theta - \cos(T, L^-))\right) \quad (2)$$

Pairwise Loss (L_P) Inspired by the work of Reimers and Gurevych (2019), we fine-tune sentence encoders using a Pairwise Contrastive Loss (L_P). For this loss, we construct all possible pairs between texts in a minibatch and all labels from a dataset. If a text-label pair is true ($\gamma = 1$), the squared cosine distance is minimized. In the case of a negative text-label pair ($\gamma = 0$), the cosine distance is subtracted from a margin parameter θ. The final loss is the sum of both terms and can therefore be defined as follows:

$$\begin{aligned} L_P = & \gamma \cdot cos(T, L^-)^2 \\ & + \\ & (1-\gamma) \cdot ReLu(\theta - cos(T, L^+))^2 \end{aligned} \quad (3)$$

Since a single processing step constructs a large number of text-label pairs when the label sets are large (Reuters, EventDNA), we use a batch size of four to maintain efficiency.

3.4. Threshold Calibration

Label-specific Thresholds Following Nooten et al. (2025), we use small annotated partitions of the datasets (the validation sets) to calibrate label-specific thresholds (cf. Figure 2). First, we create text embeddings T from texts t using the embedding model M. Using the same embedding model, we obtain label name embeddings L from label set l. Then, we calculate the pairwise cosine similarity between each embedding in T and L, thereby obtaining the similarity between each possible embedding in T and L. In order to perform DBC with multiple true labels, a threshold θ is calculated to determine label relevance. To calculate the optimal threshold θ_{max} for texts and labels, we approach MLTC as a sequence of n binary classification problems, where $n = |L|$. For each text-label embedding pair in a binary classification task, we iterate over all thresholds within the range of 0.0 and 1.0 with increments of $.01$. For each threshold, we apply it for classification on the annotated subset and calculate the F1-score of the positive class. The threshold that yields the highest score is then selected and applied for inference on the test set.

Normalized 0.5 Additionally, as a baseline, we experiment with uniform thresholds. Similarly to Nooten et al. (2025), we first normalize the pairwise cosine similarity scores between T and L using min-max normalization. Then, we apply a uniform threshold of 0.5 across all labels to perform classification. This approach adopts the common 0.5 similarity threshold in a way that takes model- and dataset-specific similarity scales into account (Li et al., 2024b; Abdi et al., 2025; Nooten et al., 2025).

3.5. Baselines

We compare our method to several baselines, conducting zero-shot and few-shot experiments with generative LLMs and fine-tuned BERT-like models.

[4]The optimal hyperparameters for each model and dataset can be found in Table 10.

[5]Model details:
(1) https://huggingface.co/NovaSearch/stella_en_400M_v5
(2) Solatorio (2024), https://huggingface.co/gist-large-embedding-v0
(3) Xiao et al. (2024), https://huggingface.co/BAAI/bge-base-en-v1.5
(4) Kosar et al. (2023), https://huggingface.co/textgain/tags-allnli-GroNLP-bert-base-dutch-cased

Model Name	Size (B)
Qwen3 (Yang et al., 2025)	1.7
	14
Gemma3 (Team, 2025)	12
Geitje (Vanroy, 2024b)	7
Fietje-Instruct (Vanroy, 2024a)	2
RoBERTa-Large (Liu et al., 2019)	.355
RobBERT-Large (Delobelle et al., 2020)	.355

Table 3: Overview of LLMs used for the baseline experiments. Underlined models are used for the experiments on the Dutch dataset.

Generative LLMs The models that we used for the experiments are summarized in Table 3. We opted for models that could be deployed on a single *NVIDIA GeForce RTX 2080 Ti* GPU using the Ollama software package[6]. We approached the problem as a Binary Relevance (BR) classification (Zhang et al., 2018)[7] and implemented structured outputs to ensure consistent model outputs that could be processed more easily. We provided the LLMs with the same adjusted label names to ensure comparability with our DBC approach.

The prompts for both the zero-shot and few-shot settings can be found in Appendix B. To select in-data examples, we used a greedy example selection method, since including one example per available label led to incoherent model outputs, as observed in An et al. (2025). For all models, we included two in-context examples that were assigned the highest number of distinct labels in order to maximize label coverage. We also experimented with including a higher number of in-context examples, but this consistently caused more hallucinations, especially with the Reuters, AAPD and BioTech datasets. Since including random examples instead of performing a greedy search yielded negligible performance differences, we opted for a greedy search method to represent as many labels as possible in the in-context examples.

Fine-tuned LLMs Due to the widespread success of fine-tuning in improving classification performance, we also fine-tune *RoBERTa-Large* (Liu et al., 2019) on all English datasets and *RobBERT-Large* (Delobelle et al., 2020) on the Dutch datasets by updating all encoder layers and the classification head. We optimize the number of epochs and the learning rate. The optimal hyperparameters for each model are listed in Table 10, Appendix C. We fine-tune these models on two versions of the training datasets, namely the entire training dataset and a stratified sample of the training data that is equal in size to the validation set, following the stratification method described in Sechidis et al. (2011). This offers a more fair comparison to our proposed DBC method. We report the average results and standard deviation across 5 experiments with different random seeds.

We employ a learning rate scheduler with linear decay, the AdamW optimizer and minimize the Binary Cross Entropy (BCE) Loss during fine-tuning.

[6] https://ollama.com/

[7] We refrained from using label ranking prompts, since this requires an additional step to determine a relevance cut-off for the labels. We also experimented with generating lists of label predictions. However, the models generated unrelated label names for larger sets of labels.

3.6. Evaluation

The effectiveness of our models is evaluated on held-out test sets using both classification and ranking metrics. For the classification metrics we opt for macro-averaged and micro-averaged F1-scores, in addition to Exact Match Ratio (EMR). The latter metric measures the proportion of instances for which the predicted label set exactly matches the true label set. In terms of ranking metrics, we chose NDCG@k (where $k = \{3, 5\}$) to measure the ranking quality and Precision@1 to measure whether the predicted label most similar to a text is a true label.

4. Results and Discussion

In the following sections, we discuss the results of the experiments. Additionally, we compare the computation times across all evaluated methods and analyze cosine similarity distributions. Moreover, we conduct learning curve experiments to gain insight into the number of samples required for our method to be effective.

4.1. Effectiveness of Contrastive Learning

Comparison between Losses The classification results for all contrastive losses can be found in Table 4 and the ranking results can be found in Table 12, Appendix E. In general, we observe that training sentence encoders with our contrastive losses consistently increases the performance on all datasets. However, we observe a difference in effectiveness among the different contrastive losses. Across all datasets and models, the pairwise contrastive loss (L_P) outperforms the other baseline contrastive losses in terms of classification metrics. In terms of ranking metrics, L_P also achieves the highest results compared to other contrastive losses, with a few exceptions on SemEval, BioTech and AAPD. Moreover, L_P was the most effective on the LitCovid dataset, leading to an increase of $\sim$ 50 EMR points, $\sim$ 30 macro-F1 points and $\sim$ 30 micro-F1 points across all sentence encoders. It is also noteworthy that L_P causes

Model		SemEval			BioTech			Reuters			AAPD			LitCovid			Avg. Results			
		EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	Clf. Rank	Avg EMR	Avg MaF1	Avg MiF1
GIST	Base	7.43	45.15	53.37	2.36	22.56	32.48	9.92	31.59	38.02	12.00	41.06	49.97	15.18	51.34	56.31	11.83	9.38 (4.84)	38.34 (11.37)	46.03 (10.28)
	L_{LM}	21.29	55.17	66.74	1.05	27.80	33.11	7.28	35.21	46.61	20.90	42.37	53.63	67.35	74.32	83.92	7.00	23.57 (25.99)	46.97 (18.32)	56.80 (19.43)
	L_{TM}	20.28	54.70	65.68	4.99	26.87	33.14	48.92	35.29	59.69	19.40	42.12	52.46	67.00	71.58	83.09	6.50	32.12 (25.18)	46.11 (17.50)	58.81 (18.28)
	L_P	23.32	57.57	68.74	9.45	37.83	46.43	78.20	44.94	83.47	29.30	46.96	64.54	74.27	79.38	87.18	2.20	42.91 (31.29)	53.34 (16.18)	70.07 (16.31)
BGE	Base	3.07	42.10	48.62	5.51	19.72	30.75	16.67	32.97	43.43	9.40	34.77	40.61	20.50	48.08	56.53	13.27	11.03 (7.38)	35.53 (10.70)	43.99 (9.56)
	L_{LM}	10.89	48.61	60.09	2.36	25.02	31.50	56.07	34.78	60.30	14.80	36.57	44.18	59.45	67.50	78.80	9.57	28.71 (26.92)	42.50 (16.30)	54.97 (17.96)
	L_{TM}	10.52	49.83	60.64	0.00	24.81	30.65	27.56	34.27	58.69	18.70	38.47	48.21	63.57	72.43	81.17	9.57	24.07 (24.31)	43.96 (18.27)	55.87 (18.48)
	L_P	24.39	55.87	68.73	4.99	31.17	43.54	80.12	44.89	85.21	17.90	40.09	54.33	74.80	80.44	87.78	3.70	40.44 (34.56)	50.49 (18.97)	67.92 (19.19)
Stella	Base	3.34	41.20	48.96	1.31	26.10	34.45	33.21	37.14	55.54	8.70	41.65	46.60	17.10	58.30	61.68	11.10	12.73 (12.97)	40.88 (11.58)	49.45 (10.25)
	L_{LM}	15.40	55.16	65.19	6.82	31.15	40.56	57.00	38.71	67.66	20.10	44.14	55.33	70.62	76.29	85.05	4.67	33.99 (28.05)	49.09 (17.54)	62.76 (16.39)
	L_{TM}	15.68	53.42	64.05	0.00	32.28	38.00	7.21	38.90	56.62	23.40	44.81	58.13	71.61	71.25	85.27	6.17	23.58 (28.25)	48.13 (15.08)	60.41 (16.98)
	L_P	23.35	57.21	68.38	22.31	41.18	53.43	80.75	47.60	85.49	34.90	53.37	69.58	76.01	82.18	88.01	1.47	47.46 (28.70)	56.31 (15.68)	72.98 (14.12)

Table 4: Classification results (Exact Match Ratio, macro-averaged F1 and micro-averaged F1) across all datasets from all models trained with contrastive losses. Light green and dark green indicate the best performance within each model and best overall performance respectively.

Method	Model	SemEval			BioTech			Reuters			AAPD			LitCovid			Average Results		
		EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	Avg EMR	Avg MaF1	Avg MiF1
DBC (normalized 0.5)	Stella (L_P)	13.75	57.07	66.12	19.42	42.04	56.76	69.4	38	71.75	23.1	44.36	59.2	73.7	81.23	87.56	39.87 (29.15)	52.54 (17.55)	68.28 (12.28)
DBC (optimized)	Stella (L_P)	15.99	**57.2**	66.83	22.31	41.18	53.43	80.75	47.6	85.49	34.9	53.37	69.58	76.01	82.18	88.01	45.99 (30.39)	**56.54** (29.15)	72.67 (14.26)
ZS-LLM	Qwen3:14b	16.26	49.65	60.41	13.91	45.88	45.52	57.26	**64.45**	72.59	4.8	42.98	42.52	33.03	61.7	64.48	25.05 (20.69)	52.93 (9.61)	57.10 (12.77)
FS-LLM	Qwen3:14b	15.71	48.97	59.46	14.44	46.84	47.2	59.78	58.86	73.13	9.8	44.17	46.76	39.25	63.47	66.31	27.79 (21.23)	52.46 (8.29)	58.57 (11.63)
Fine-tuned LLM (sample)	RoBERTa	26.61	50.36	68.52	27.21	27.7	56.67	78.13	14.06	82.78	30.04	32.06	61.22	75.1	78.83	87.12	47.42 (26.71)	40.60 (25.0)	71.26 (13.28)
Fine-tuned LLM	RoBERTa	**29.51**	55.19	**71.29**	**52.76**	**52.64**	**73.11**	**85.03**	32.68	**88.75**	**40.94**	**55.45**	**72.54**	**77.04**	**82.27**	**88.38**	**57.06** (23.55)	55.65 (17.65)	**78.81** (8.93)

Table 5: Best classification results (EMR, macro-averaged F1 and micro-averaged F1) on all datasets from the different methods. The best results per dataset across all approaches are marked in **bold**.

the most substantial increase in EMR across all datasets. This indicates that the models trained with this loss are better at predicting exactly matching label sets.

The superior results of L_P can be attributed to constructing text-label pairs with all labels from a dataset (not limited to the labels present in a minibatch), thereby maximizing the information extracted from a minibatch during training. In contrast, the other loss functions only take true positives and a fixed k number of hard negatives into account during fine-tuning, thereby potentially disregarding crucial information to be obtained from other negative labels.

Regarding the individual models, we observe that *Stella* yields the best average results with the lowest standard deviation between datasets. The effectiveness of this model can be attributed to the higher number of parameters and higher embedding dimensions compared to the other sentence encoders. It should be noted that *GIST-Large* outperforms *Stella* on the SemEval dataset.

To confirm that the observed improvements are statistically significant, we conducted a formal test at the level of individual label decisions using a logistic generalized linear mixed model (GLMM). The model was fit for each combination of datasets and models, and accounted for some labels being inherently harder to predict. Because all four methods are applied to the exact same pairs (instance, label), the comparison is fully paired by design. The effect size is reported as an odds ratio (OR): an OR of 2.0 means the method has twice the odds of a correct prediction compared to the reference. P-values are Holm-Bonferroni corrected for multiple comparisons within each combination. A result is considered significant if the corrected p-value falls below 0.05. Full results are provided in Appendix H, Tables 15 and 16.

The analysis supports previously reported F1 scores. Out of 90 pairwise comparisons, 87 are statistically significant. All methods significantly improve over the base threshold, with L_P achieving the largest effect sizes. L_P also significantly outperforms both L_{LM} and L_{TM} in all 15 dataset/model combinations. L_{LM} and L_{TM} are statistically indistinguishable in three cases (BioTech/BGE, LitCovid/Stella, and Reuters/Stella). The overall ranking is consistent: $L_P > L_{LM} \approx L_{TM} >$ Base.

Comparison with Baselines The best results for each classification approach (DBC, zero-shot/few-shot with LLMs[8] and fine-tuned *BERT*-models) are summarized in Table 5. On average, our DBC approach with CL and optimized label-specific thresholds outperforms all other methods in terms of macro-averaged F1 scores. Additionally, our method achieves the second highest micro-F1 scores and third highest EMR scores, only being surpassed by fine-tuned *RoBERTa-Large* models trained on all available training data. However, our method achieves higher macro- and micro-averaged F1-scores than *RoBERTa-Large* when the latter model is fine-tuned on a smaller sample. Since this sample is equal in size to the validation data that was used for calibrating thresholds and training the contrastive models, these results high-

[8]All results from the experiments with calibrated uniform thresholds can found in Table 11, Appendix D. The results for generative LLMs can be found in Table 13, Appendix F.

light the effectiveness and resourcefulness of our approach.

It should be noted that DBC shows the highest standard deviation across datasets, which indicates that the effectiveness of DBC fluctuates more between datasets than the other approaches. For example, DBC shows high results on LitCovid, but lower results on BioTech compared to other approaches.

Concerning individual datasets, it can be observed that DBC falls behind on the other approaches on the BioTech dataset. This can be attributed to the decision boundaries being optimized for *RoBERTa* during fine-tuning (i.e., *RoBERTa* is fine-tuned for the task itself, not for semantics), which entails that the model might effectively learn label dependencies during training. The differences in semantics between the labels might still be too subtle, even after fine-tuning the sentence encoders with contrastive losses. This, in combination with the skewed label distribution of the BioTech dataset, may have resulted in suboptimal text representations.

4.2. Computation Times

The main advantage of our approach is the low computational cost. As mentioned in Section 3.4, we conducted all experiments on a single *NVIDIA GeForce RTX 2080 Ti* GPU. Table 6 presents the computation times (in hours) for each of the classification methods explored. The results are the sum of the duration of fine-tuning a sentence encoder (*Stella* trained with L_P in this case), calibrating the optimal thresholds and obtaining predictions for the test sets. These results are compared to fine-tuning *RoBERTa-Large*, zero-shot and few-shot experiments with various generative LLMs. The table demonstrates that our method is substantially faster than other approaches: On average, DBC is 16 times faster than *RoBERTa* (and 3 times faster when fine-tuned on a sample), 3 times faster than the smallest generative LLM, and 20 times faster than the largest generative LLM. The most notable time difference is observed for the Reuters dataset, where our method only took .57 hours to run, whereas fine-tuned models and generative LLMs took between 2.7 and 49 hours. Processing this dataset was the most time-consuming for generative LLMs, which is likely due to the large label set and structured output formatting. The faster computation times of DBC combined with the competitive classification results show great potential for other use cases.

Method	Model	SemEval	BioTech	Reuters	AAPD	LitCovid	Total
DBC	Stella (L_P)	**0.18**	0.27	**0.57**	**0.39**	1.62	**3.03**
ZS-LLM	Gemma3:12b	2.58	0.89	26.98	4.70	4.99	40.14
	Qwen3:1.7b	0.70	**0.21**	6.07	0.99	**1.62**	9.58
	Qwen3:14b	2.01	0.71	21.21	3.88	4.02	31.83
FS-LLM	Gemma3:12b	3.70	0.93	37.00	5.14	6.74	53.50
	Qwen3:1.7b	0.73	0.24	12.63	1.03	1.33	15.96
	Qwen3:14b	2.18	0.75	49.94	3.72	4.16	60.75
Fine-tuned (1 seed)	RoBERTa	6.44	4.27	13.47	13.78	11.06	49.02
Fine-tuned (1 seed; sample)	RoBERTa	1.02	0.97	2.74	2.21	2.22	9.16

Table 6: Computation time (in hours) per model. The fastest approach per dataset is marked in bold.

model	SemEval	BioTech	Reuters	AAPD	LitCovid
GIST-Large	0.11	0.08	0.26	0.23	0.09
GIST-Large (L_{LM})	0.35	0.24	0.51	0.36	0.48
GIST-Large (L_{TM})	0.35	0.22	0.55	0.36	0.48
GIST-Large (L_P)	0.34	0.29	0.57	0.37	0.41
BGE	0.07	0.06	0.27	0.18	0.08
BGE (L_{LM})	0.20	0.19	0.50	0.26	0.43
BGE (L_{TM})	0.20	0.19	0.51	0.30	0.42
BGE (L_P)	0.33	0.27	0.51	0.29	0.55
Stella	0.08	0.09	0.35	0.22	0.05
Stella (L_{LM})	0.37	0.23	0.63	0.42	0.50
Stella (L_{TM})	0.34	0.19	0.57	0.39	0.47
Stella (L_P)	0.33	0.31	0.54	0.39	0.05

Table 7: Gap score from all models and datasets. Light green and dark green indicate the highest score within a single model and overall highest score respectively.

4.3. Cosine Similarity Distribution Analysis

In order to demonstrate the effectiveness of contrastive learning, we plot two cosine similarity distributions for each dataset: one showing the similarities between text embeddings and true labels and another showing the similarities between text embeddings and false labels. Figure 3illustrates how texts become more similar to their true labels and less similar to their false labels.

To further analyze these distributions, we calculate a "gap"-score that represents the average difference between the min-max normalized similarity scores (cf. Section 3) for texts and their true labels on the one hand, and texts and corresponding incorrect labels on the other hand. We hypothesize that contrastive learning increases the difference (gap) between these two similarity distributions. As shown in Table 7, contrastive learning increases this difference, although no consistent pattern can be observed across models. Moreover, there does not appear to be a direct correlation between observed scores and classification results, since *Stella (L_P)* does not produce the highest gap scores, but does produce the highest classification results in general. Nonetheless, these analyses provide an intuitive explanation for the general effectiveness of contrastive learning within DBC.

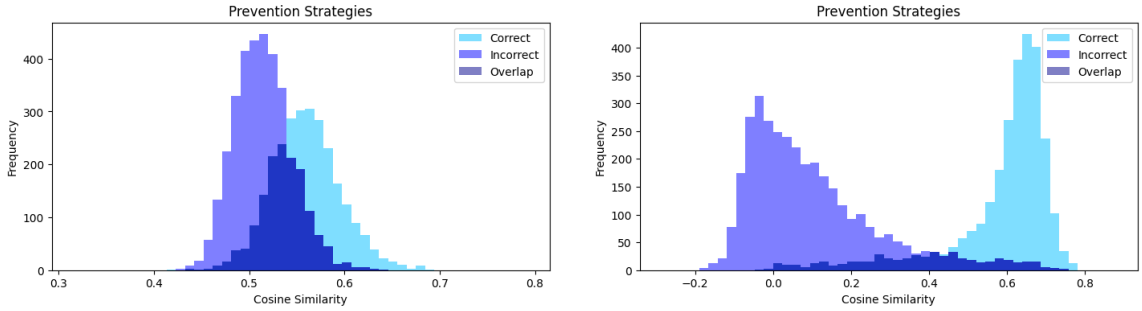

Figure 3: Visualization of the cosine similarity distributions for the *Prevention Strategies* label from the LitCovid dataset.

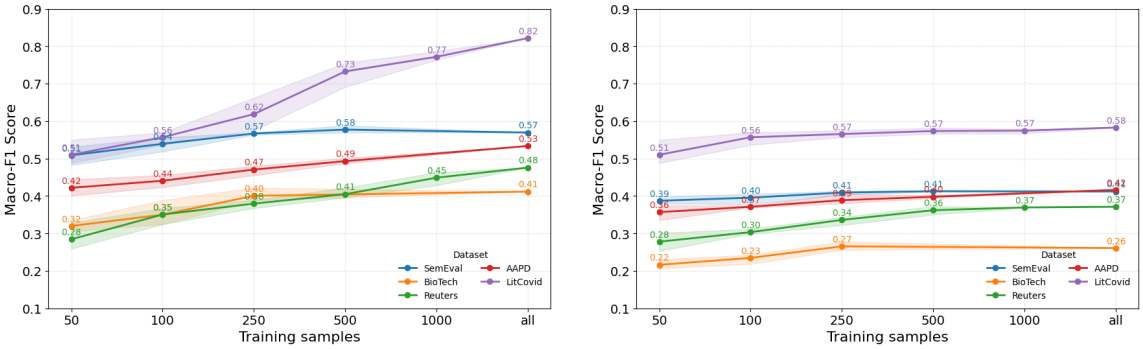

Figure 4: Results of the learning curve experiments with *Stella* trained with L_P. The x-axis shows the number of training samples. The left figure includes results for both contrastive learning and threshold calibration, while the right figure shows results for threshold calibration only. The figures display the average, minimum and maximum values across three random draws for each sample size.

Model	Method	EventDNA (level 1)						EventDNA (level 2)						Average Results		
		EMR	MaF1	MiF1	NDCG@3	NDCG@5	P@1	EMR	MaF1	MiF1	NDCG@3	NDCG@5	P@1	Ranking Rank	Clf. Rank	Avg. Rank
BERTje-nli	Base	9.81	45.54	52.49	83.81	86.35	80.75	0.00	17.23	18.34	34.35	32.58	40.54	4.33	4.17	4.25
	L_{LM}	7.55	50.96	58.22	82.32	85.77	84.15	0.00	20.05	22.99	40.24	39.29	46.62	3.50	3.08	3.29
	L_{TM}	7.55	51.85	57.17	81.90	85.87	84.53	0.00	19.66	21.58	39.54	39.28	48.31	3.67	3.42	3.54
	L_P	43.77	51.80	72.26	88.21	90.07	87.17	0.00	**21.48**	**38.31**	**52.76**	**50.33**	**61.15**	**1.00**	**1.50**	**1.25**
RobBERT-Large (trained on sample)	fine-tuned	46.82 (1.28)	20.17 (2.21)	69.82 (1.08)	/	/	/	N/A	N/A	N/A	/	/	/	/	/	/
RobBERT-Large	fine-tuned	53.58 (1.19)	51.19 (3.24)	78.23 (0.51)	/	/	/	N/A	N/A	N/A	/	/	/	/	/	/
Fietje-Instruct	zero-shot	10.57	20.19	47.4	/	/	/	N/A	N/A	N/A	/	/	/	/	/	/
Fietje-Instruct (few-shot)	few-shot	9.05	13.03	51.69	/	/	/	N/A	N/A	N/A	/	/	/	/	/	/
Geitje	zero-shot	21.87	43.98	53.62	/	/	/	N/A	N/A	N/A	/	/	/	/	/	/
Geitje (few-shot)	few-shot	37.74	38.82	63.24	/	/	/	N/A	N/A	N/A	/	/	/	/	/	/

Table 8: Classification and ranking results for all approaches on the Dutch EventDNA dataset. The results with fine-tuned LLMs include standard deviations across five random seeds. Light green and dark green indicate the best performance within a single model and best overall performance respectively.

4.4. Learning Curve Experiments

In order to show the resource-efficiency of our approach, we conducted learning curve experiments with various sample sizes. These samples are used to train sentence encoders and calibrate label-specific thresholds. For each sample size, we take three different random samples and ensure that the same sample is used for both contrastive learning and threshold calibration. We use the same samples to calibrate thresholds with the baseline sentence encoder.

For these experiments, we opted for *Stella*, since this model yielded the best classification results on average across all datasets. Figure 4 shows the macro-averaged F1-scores obtained on the test set for each sample size. As can be observed, L_P generally shows steeper learning curves compared to the baseline, thereby effectively making use of the limited annotated examples[9].

[9]Experiments with L_{LM} and L_{TM} where all negative labels per text are considered, did not yield better results than an optimized value for k.

4.5. Results on EventDNA

The classification results and ranking results with Dutch models on the EventDNA dataset can be found in Table 8. The patterns observed in the English datasets are also consistent with those found in the Dutch EventDNA dataset, in that DBC shows promising results compared to the baselines. Additionally, we also observe that L_P produces the best results out of the three contrastive losses across both levels of the dataset.

Compared to other baselines, *RobBERT-Large* yields the highest EMR and micro-F1 scores on the first-level labels of EventDNA. However, the models failed at learning the large number of labels at the second level, which indicates that adaptations to the training procedure are required. Similarly, all generative LLMs failed to produce any predictions, which can be attributed to the small model size, the large label set combined with the required structured output, and the prompt design. However, these results highlight the flexibility and effectiveness of our method on datasets with large label sets.

5. Conclusion

In this study, we fine-tuned three state-of-the-art sentence encoders using three supervised contrastive losses to enhance the performance of distance-based multi-label text classification. Although all models showed substantial performance improvements, we observed that the pairwise contrastive loss (L_P) yielded the highest gain on all datasets, thus narrowing the gap between distance-based methods and supervised fine-tuning. We highlight that our proposed method is computationally more efficient than other approaches (generative LLMs and supervised fine-tuning of *BERT* models) and requires little annotated data to be effective, especially compared to baseline DBC methods. Our findings can also be applied to other similarity-based tasks, such as zero-shot (multi-label) classification, ranking or retrieval-augmented generation.

6. Limitations and Future Work

This study is subject to several limitations. First, the comparison between our proposed DBC method and fine-tuned models should be interpreted with caution, as the fine-tuned *BERT* models were evaluated in a default fine-tuning setup without applying training optimizations such as mixed-precision training. However, we still hypothesize that our method would be substantially faster. Additionally, we hypothesize that the efficiency of our method and *BERT* models could be further improved by integrating adapter modules (Poth et al., 2023).

We also recognize that there are more advanced prompt engineering techniques for performing MLTC with generative LLMs, such as chain-of-thought prompting, including label explanations, and improved sample selection methods using retrieval. Future work should explore these methods to provide stronger comparisons.

Moreover, we recognize that our method disregards label correlations during the contrastive fine-tuning, threshold calibration, and inference stages. We hypothesize that incorporating label correlations in these stages would further improve the performance.

7. Acknowledgements

This research was funded by the Flemish government under FWO IRI project CLARIAH-VL and the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

8. Bibliographical References

Silvana Abdi, Mahrokh Hassani, Rosalien Kinds, Timo Strijbis, and Roman Terpstra. 2025. HalluRAG-RUG at SemEval-2025 task 3: Using retrieval-augmented generation for hallucination detection in model outputs. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 846–851, Vienna, Austria. Association for Computational Linguistics.

Bang An, Shiyue Zhang, and Mark Dredze. 2025. RAG LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5444–5474, Albuquerque, New Mexico. Association for Computational Linguistics.

Chidanand Apt'e, Fred Damerau, and Sholom M. Weiss. 1994. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*. To appear.

Alexandre Audibert, Aurélien Gauffre, and Massih-Reza Amini. 2025. Multi-label contrastive learning : A comprehensive study.

Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification.

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. 2022. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification.

Xunxin Cai, Meng Xiao, Zhiyuan Ning, and Yuanchun Zhou. 2023. Resolving the imbalance issue in hierarchical disciplinary topic inference via llm-based data augmentation. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1424–1429.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2020. LitCovid: an open database of COVID-19 literature. *Nucleic Acids Research*, 49(D1):D1534–D1540.

Camiel Colruyt, Orphée De Clercq, Thierry Desot, and Veronique Hoste. 2022. Eventdna: a dataset for dutch news event extraction as a basis for news diversification. *Language Resources and Evaluation*, 57:189 – 221.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model.

Jiawen Deng and Fuji Ren. 2023. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 14(1):475–486.

Emre Deniz, Hasan Erbay, and Mustafa Coşar. 2022. Multi-label classification of e-commerce customer reviews via machine learning. *Axioms*, 11(9).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingwu Fan and Changsheng Qiu. 2023. Hierarchical multi-label text classification method based on multi-level decoupling. In *2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, pages 453–457.

Eva Gibaja and Sebastián Ventura. 2014. Multi-label learning: a review of the state of the art and ongoing research. *WIREs Data Mining and Knowledge Discovery*, 4(6):411–444.

A. F. Giraldo-Forero, J. A. Jaramillo-Garzón, and C. G. Castellanos-Domínguez. 2013. A comparison of multi-label techniques based on problem transformation for protein functional prediction. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2688–2691.

Rohit Gupta, Mamshad Nayeem Rizve, Jayakrishnan Unnikrishnan, Ashish Tawari, Son Tran, Mubarak Shah, Benjamin Yao, and Trishul Chilimbi. 2024. Open vocabulary multi-label video classification.

Mohammadreza Heydarian, Thomas E. Doyle, and Reza Samavi. 2022. Mlcm: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095.

Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. 2024. A comprehensive survey on contrastive learning. *Neurocomputing*, 610:128645.

Shan Huang, Wenlong Hu, Bin Lu, Qiang Fan, Xinyao Xu, Xiaolei Zhou, and Hao Yan. 2024. Application of label correlation in multi-label classification: A survey. *Applied Sciences*, 14(19).

Shu Huang, Wei Peng, Jingxuan Li, and Dongwon Lee. 2013. Sentiment and topic analysis on social media: a multi-task multi-label classification approach. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 172–181, New York, NY, USA. Association for Computing Machinery.

Nikmah Isnaini, Adiwijaya, Mohamad Syahrul Mubarok, and Muhammad Yuslan Abu Bakar. 2019. A multi-label classification on topics of indonesian news using k-nearest neighbor. *Journal of Physics: Conference Series*, 1192(1):012027.

Ling Jia, Jin Fan, Dong Sun, Qingwei Gao, and Yixiang Lu. 2022. Research on multi-label classification problems based on neural networks and label correlation. In *2022 41st Chinese Control Conference (CCC)*, pages 7298–7302.

Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2022. Unsupervised text classification with neural word embeddings. *Computational Linguistics in the Netherlands Journal*, 12:165–181.

Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2023. Advancing topical text classification: A novel distance-based method with contextual embeddings. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 586–597, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Jens Lemmens, Tess Dejaeghere, Tim Kreutz, Jens Van Nooten, Ilia Markov, and Walter Daelemans. 2021. Vaccinpraat: Monitoring vaccine skepticism in dutch twitter and facebook comments. *Computational Linguistics in the Netherlands Journal*, 11:173–188.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

Xiang Li, Jiexi Liu, Xinrui Wang, and Songcan Chen. 2024a. A survey on incomplete multi-label learning: Recent advances and future trends.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Xintong Li, Jinya Jiang, Ria Dharmani, Jayanth Srinivasa, Gaowen Liu, and Jingbo Shang. 2024b. Open-world multi-label text classification with extremely weak supervision. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15084–15096, Miami, Florida, USA. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Junyang Lin, Qi Su, Pengcheng Yang, Shuming Ma, and Xu Sun. 2018. Semantic-unit-based dilated convolution for multi-label text classification. *CoRR*, abs/1808.08561.

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently.

Zhongchen Ma and Songcan Chen. 2022. A similarity-based framework for classification task. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):5438–5443.

Usman Malik, Simon Bernard, Alexandre Pauchet, Clément Chatelain, Romain Picot-Clémente, and Jérôme Cortinovis. 2024. Pseudo-labeling with large language models for multi-label emotion classification of french tweets. *IEEE Access*, 12:15902–15916.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Ratri Mukherjee and Kishlay Jha. 2024. Context-aware contrastive representation learning for zero-shot biomedical text classification. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3611–3614, United States. IEEE.

Manjunath Mulimani and Annamaria Mesaros. 2024. Class-incremental learning for multi-label audio classification.

Ghulam Mustafa, Muhammad Usman, Lisu Yu, Muhammad Tanvir Afzal, Muhammad Sulaiman, and Abdul Shahid. 2021. Multi-label classification of research articles using word2vec and identification of similarity threshold. *Scientific Reports*, 11(1):21900.

Nikolaos Mylonas, Stamatis Karlos, and Grigorios Tsoumakas. 2020a. Zero-shot classification of

biomedical articles with emerging mesh descriptors. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 175–184, New York, NY, USA. Association for Computing Machinery.

Nikolaos Mylonas, Stamatis Karlos, and Grigorios Tsoumakas. 2020b. Zero-shot classification of biomedical articles with emerging mesh descriptors. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 175–184, New York, NY, USA. Association for Computing Machinery.

Nobal Niraula, Samet Ayhan, Balaguruna Chidambaram, and Daniel Whyatt. 2024. Multi-label classification with generative large language models. In *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, pages 1–7.

Jens Van Nooten, Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2025. One size does not fit all: Exploring variable thresholds for distance-based multi-label text classification.

OpenAI. 2024. Gpt-4 technical report.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning.

Irina Radeva, Ivan Popchev, and Miroslava Dimitrova. 2024. Similarity thresholds in retrieval-augmented generation. In *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, pages 1–7.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training. In *arxiv*.

Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. 2016. MEKA: A multi-label/multi-target extension to Weka. *Journal of Machine Learning Research*, 17(21):1–5.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91.

Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2022. Exploring universal sentence encoders for zero-shot text classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 135–147, Online only. Association for Computational Linguistics.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16218–16233, Singapore. Association for Computational Linguistics.

Tim Schopf, Daniel Braun, and Florian Matthes. 2023. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR)*, NLPIR 2022, New York, NY, USA. Association for Computing Machinery.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

Dezhao Song, Andrew Vold, Kanika Madan, and Frank Schilder. 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, 106:101718.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. 2023. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1579–1585. AAAI Press.

Yangqiu Song, Shyam Upadhyay, Haoruo Peng, and Dan Roth. 2016. Cross-lingual dataless classification for many languages. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2901–2907. AAAI Press.

Balázs Szalkai and Vince Grolmusz. 2018. Near perfect protein multi-label classification with deep neural networks. *Methods*, 132:50–56. Comparison and Visualization Methods for High-Dimensional Biological Data.

Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Adane Nega Tarekegn, Mohib Ullah, and Faouzi Alaya Cheikh. 2024. Deep learning for multi-label learning: A comprehensive survey.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, and Garrett Tanzer. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Gemma Team. 2025. Gemma 3 technical report.

Simon Chi Lok U, Jie He, Víctor Gutiérrez-Basulto, and Jeff Z. Pan. 2023. Instances and labels: Hierarchy-aware joint supervised contrastive learning for hierarchical multi-label text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8858–8875, Singapore. Association for Computational Linguistics.

Jens Van Nooten and Walter Daelemans. 2023. Improving Dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270, Toronto, Canada. Association for Computational Linguistics.

Jens Van Nooten and Walter Daelemans. 2025. Jump to hyperspace: Comparing Euclidean and hyperbolic loss functions for hierarchical multi-label text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4260–4273, Abu Dhabi, UAE. Association for Computational Linguistics.

Jens Van Nooten and Andriy Kosar. 2024. Advancing CSR theme and topic classification: LLMs and training enhancement insights. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 292–305, Torino, Italia. Association for Computational Linguistics.

Josefien Van Olmen, Jens Van Nooten, Hilde Philips, Annet Sollie, and Walter Daelemans. 2022. Predicting covid-19 symptoms from free text in medical records using artificial intelligence: Feasibility study. *JMIR Med Inform*, 10(4):e37771.

Bram Vanroy. 2024a. Fietje: An open, efficient llm for dutch.

Bram Vanroy. 2024b. Geitje 7b ultra: A conversational model for dutch.

Sai Pavan Kumar Veeranki, Akhila Abdulnazar, Diether Kramer, Markus Kreuzthaler, and David Benjamin Lumenta. 2024. Multi-label text classification via secondary use of large clinical real-world data sets. *Scientific Reports*, 14(1):26972.

Sappadla Prateek Veeranna, Jinseok Nam, Eneldo Loza Mencıa, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *Proceeding of european symposium on artificial neural networks, computational intelligence and machine learning. bruges, belgium: Elsevier*, pages 423–428.

Fengjun Wang, Moran Beladev, Ofri Kleinfeld, Elina Frayerman, Tal Shachar, Eran Fainman, Karen Lastmann Assaraf, Sarai Mizrachi, and Benjamin Wang. 2023. Text2Topic: Multi-label text classification system for efficient topic detection in user generated content with zero-shot capabilities. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing: Industry Track, pages 93–103, Singapore. Association for Computational Linguistics.

Qian Wang, Ning Jia, and Toby P. Breckon. 2019. A baseline for multi-label image classification using an ensemble of deep convolutional neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 644–648.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packaged resources to advance general chinese embedding.

Changzhen Xiong and Yanmei Shan. 2018. Subject features and hash codes for multi-label image retrieval. In *2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, pages 808–812.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

Pingyue Zhang and Mengyue Wu. 2024. Multi-label supervised contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16786–16793.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022a. Contrastive data and learning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47, Seattle, United States. Association for Computational Linguistics.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. 2022b. Use all the labels: A hierarchical multi-label contrastive learning framework.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words.

A. Appendix A: Adjusted Label Names

Dataset	Label Name	Adjusted Label Name
BioTech	foundation	Foundation Establishment
	closing	Company Closing Operations
Reuters	acquisition	Acquisition and Mergers
	consumer price index	Consumer Price Index (CPI)
	cornglutenfeed	Corn Gluten Feed
	f-cattle	Feeder Cattle
LitCovid	Mechanism	Biological Mechanism
	Diagnosis	Diagnostic Methods

Table 9: Examples of adjusted label names.

B. Appendix B: LLM Prompts

B.1. Zero-shot Example (LitCovid)

You are an expert annotator with expertise in analyzing abstracts from academic papers related to COVID-19.

Classify each of the following texts with all relevant topics expressed in the text. ONLY use the following topic labels: [LABEL_LIST]. It is possible that multiple labels are relevant to the text or that NO label is relevant!!!

The final output should be a JSON object that looks like this:
[OUTPUT_EXAMPLE]
Respond ONLY with the JSON that contains your predictions for each text!!!
Texts: [NUMBERED_TEXTS]

B.2. Few-shot Example (LitCovid)

You are an expert annotator with expertise in analyzing abstracts from academic papers related to COVID-19.

Classify each of the following texts with all relevant topics expressed in the text. ONLY use the following topic labels: [LABEL_LIST]. It is possible that multiple labels are relevant to the text or that NO label is relevant!!!

The final output should be a JSON object that looks like this:
[OUTPUT_EXAMPLE]
Respond ONLY with the JSON that contains your predictions for each text!!!
Here are some annotated examples from the data: [ANNOTATED_EXAMPLES]
Texts: [NUMBERED_TEXTS]

C. Appendix C: Model Hyperparameters

Model	Loss	SemEval					BioTech					Reuters					AAPD					LitCovid				
		LR	Ep	BS	NTK	mgn	LR	Ep	BS	NTK	mgn	LR	Ep	BS	NTK	mgn	LR	Ep	BS	NTK	mgn	LR	Ep	BS	NTK	mgn
GIST-Large	L_LM	2e-5	10	128	3	0.5	2e-5	10	128	7	0.5	2e-5	10	128	7	0.5	2e-5	15	128	7	0.5	2e-5	5	128	3	0.5
	L_{TM}	2e-5	10	128	3	0.5	2e-5	10	128	7	0.5	2e-5	15	128	7	0.5	2e-5	15	128	7	0.5	2e-5	5	128	3	0.5
	L_P	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5
BGE	L_LM	2e-5	10	128	3	0.5	2e-5	10	128	10	0.5	2e-5	10	128	7	0.5	2e-5	10	128	7	0.5	2e-5	5	128	3	0.5
	L_{TM}	2e-5	10	128	3	0.5	2e-5	10	128	10	0.5	2e-5	10	128	7	0.5	2e-5	10	128	7	0.5	2e-5	5	128	3	0.5
	L_P	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	128	N/A	0.5	2e-5	15	128	N/A	0.5
Stella	L_LM	2e-5	5	128	3	0.5	2e-5	10	128	7	0.5	2e-5	10	128	7	0.5	2e-5	15	128	7	0.5	2e-5	5	128	3	0.5
	L_{TM}	2e-5	10	128	3	0.5	2e-5	10	128	7	0.5	2e-5	10	128	7	0.5	2e-5	15	128	7	0.5	2e-5	5	128	3	0.5
	L_P	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	16	N/A	0.5	2e-5	15	128	N/A	0.5
RoBERTa	N/A	2e-5	10	8	N/A	N/A	2e-5	20	8	N/A	N/A	2e-5	10	8	N/A	N/A	2e-5	10	8	N/A	N/A	2e-5	20	8	N/A	N/A
RoBERTa (sample)	N/A	2e-5	10	8	N/A	N/A	2e-5	20	8	N/A	N/A	2e-5	10	8	N/A	N/A	2e-5	20	8	N/A	N/A	2e-5	10	8	N/A	N/A

Table 10: Hyperparameters for the best models per setup. *LR* = Learning Rate, *Ep* = number of epochs, *BS* = Batch Size, *NTK* = negative top k; number of hard negatives, *mgn* = margin.

D. Appendix D: DBC Results with Uniform Thresholds

Model		SemEval			BioTech			Reuters			AAPD			LitCovid			Avg. Metrics			
		EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	Clf. Rank	Avg EMR	Avg. MaF1	Avg. MiF1
GIST	Base	7.24	25.27	29.18	0.52	16.13	18.61	5.82	12.26	15.81	4.2	29.94	33.95	12.07	22.47	26.93	12.7	5.97 (4.23)	21.214 (7.07)	24.896 (7.53)
	L_{LM}	17.77	51.53	65.22	11.02	9.04	40.54	59.31	33.94	61.98	17.1	36.72	53.72	30.21	63.44	73.34	5.73	27.082 (19.32)	38.934 (20.51)	58.96 (12.47)
	L_{TM}	14.08	50.23	65.9	3.15	24.67	40.1	20.11	22.58	37.46	3.8	36.17	50.08	26.99	62.72	70.73	7.6	13.626 (10.33)	39.274 (17.11)	52.854 (14.98)
	L_P	22.61	57.25	68.92	21.26	40.15	55.79	70.33	37.42	72.82	21.8	46.19	60.33	46.55	69.26	75.73	2.13	34.69 (18.29)	49.808 (13.47)	65.312 (7.7)
BGE	Base	1.99	40.35	43.83	0.52	15.13	17.3	1.52	8.36	9.95	1	23.39	24.09	15.32	35.4	39.3	13.3	4.07 (6.31)	24.526 (13.41)	26.894 (14.39)
	L_{LM}	17.8	47.84	62.79	4.72	13.69	40.8	50.25	14.42	32.79	13.2	17.09	21.3	49.85	60.48	73.62	8.43	27.164 (21.41)	30.704 (21.91)	46.26 (21.54)
	L_{TM}	14.15	49.5	62.37	1.31	21.95	33.97	16.61	29.42	50.48	11.1	26.9	36.17	38.05	62.18	74.34	7.5	16.244 (13.5)	37.99 (17.11)	51.466 (17.2)
	L_P	19.61	55.39	67.81	12.86	34.01	50.49	40.65	22.36	48.11	16.2	42.52	57.04	72.3	80.71	87.03	3.87	32.324 (24.84)	46.998 (22.38)	62.096 (15.89)
Stella	Base	4.79	15.17	17.44	1.57	17.32	19.94	19.12	20.94	29.83	4.9	31.81	35.07	8.03	21.01	19.52	11.87	7.682 (6.79)	21.25 (6.4)	24.36 (7.67)
	L_{LM}	14.7	50.64	59.67	8.66	18.08	41.16	9.59	34.83	56.85	27.1	42.58	61.79	68.62	69.89	84.67	4.6	25.734 (25.07)	43.204 (19.16)	60.828 (15.6)
	L_{TM}	12.86	51.5	57.7	3.67	21.81	33.09	0.99	21.02	32.85	22.9	43.38	57.79	32.81	64.57	72.87	7.13	14.646 (13.31)	40.456 (18.96)	50.86 (17.46)
	L_P	13.75	57.07	66.12	19.42	42.04	56.76	69.4	38	71.75	23.1	44.36	59.2	73.7	81.23	87.56	2.13	38.078 (31.23)	52.1 (16.89)	66.34 (13.56)

Table 11: All classification results with calibrated uniform thresholds.

E. Appendix E: DBC Ranking Results

Model		SemEval			BioTech			Reuters			AAPD			LitCovid			Avg. Metrics			
		NDCG@3	NDCG@5	P@1	NDCG@3	NDCG@5	P@1	NDCG@3	NDCG@5	P@1	NDCG@3	NDCG@5	P@1	NDCG@3	NDCG@5	P@1	Ranking Rank	Avg. NDCG@3	Avg. NDCG@5	Avg. P@1
GIST	Base	65.97	73.02	68.18	25.91	31.73	20.73	59.74	63.22	50.55	60.45	66.24	64.3	65.27	72.86	44.96	13.8	55.468 (16.76)	72.94 (18.3)	42.64 (34.06)
	L_{LM}	77.48	82.32	81.1	54.18	59.14	43.83	83.84	84.66	82.3	69.5	74.03	75.7	93.24	94.74	89.93	5.73	75.648 (14.82)	88.53 (15.23)	68.552 (47.14)
	L_{TM}	76.85	82.19	80.64	50.96	58.72	44.09	83.69	84.57	82.67	68.3	72.86	74.5	92.5	94.13	89.53	6.87	74.46 (15.86)	88.16 (15.28)	60.19 (35.24)
	L_P	79.31	83.67	81.71	69.21	72.9	64.3	87.29	87.63	86.27	75.14	79.11	79.8	95.59	96.42	92.43	2.07	81.178 (10.26)	90.045 (10.18)	72.272 (28.04)
BGE	Base	56.7	64.58	57.32	22.4	27.34	17.85	60.2	64.44	51.04	48.13	54.51	49.6	68.49	75.01	53.55	14.4	51.184 (17.67)	69.795 (20.45)	37.076 (24.37)
	L_{LM}	75.84	80.69	77.05	49.06	57.07	43.31	83.72	84.73	82.04	67.02	71.77	71.4	92.38	93.84	87.71	8.73	73.604 (16.63)	87.265 (16)	60.7 (40.81)
	L_{TM}	76.19	80.85	77.48	49.46	56.33	44.36	82.84	83.92	80.91	66.24	70.86	68.6	91.79	93.38	87.15	8.97	73.304 (16.27)	87.115 (16.1)	55.348 (32.99)
	L_P	79.52	83.76	82.39	67	71.32	61.42	87.27	87.67	86.14	74.51	78.46	77.9	96.44	97.04	93.83	2	80.948 (11.38)	90.4 (11.17)	67.458 (31.03)
Stella	Base	62.44	70.23	65.91	28.49	32.83	23.1	76.34	77.88	69.04	57.2	62.47	56.8	44.47	54.18	18.67	13.8	53.788 (18.18)	62.205 (18.78)	38.462 (27.92)
	L_{LM}	74.66	80.3	79.38	49.45	55.46	48.56	85.95	86.49	85.38	70.85	75.14	76.2	94.51	95.7	91.49	5.97	75.084 (17.11)	88 (17.33)	59.848 (41.1)
	L_{TM}	74.06	78.19	79.96	49.26	54.51	51.71	84.92	85.61	85.05	72.53	76.33	78.7	92.78	94.72	91.6	6.27	74.71 (16.46)	86.455 (17.24)	58.926 (40.23)
	L_P	76.57	81.86	77.54	71.15	75.49	66.4	87.83	88.14	86.8	73.48	76.82	77.4	96.3	96.96	93.68	2.67	80.98 (10.46)	89.7 (10.06)	73.184 (25)

Table 12: Ranking results (NDCG@3, NDCG@5 and P@1) on all datasets from all models trained contrastive losses.

F. Appendix F: LLM Results

Model	Method	SemEval			BioTech			Reuters			AAPD			LitCovid			Average Results		
		EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	EMR	MaF1	MiF1	Avg. EMR	Avg. MaF1	Avg. MiF1
Qwen3:1.7b	zero-shot	9.73	44.96	52.94	**16.01**	34.64	38.34	13.96	29.78	38.16	0.3	10.76	12.79	10.23	40.98	42.38	10.05	32.22	36.92
	few-shot	8.93	48.63	56.15	8.39	30.91	39.09	19.48	31.63	38.47	4.2	16.07	26.23	23.16	50.89	48.57	12.83	35.63	41.70
Qwen3:14b	zero-shot	**16.26**	**49.65**	**60.41**	13.91	45.88	45.52	57.26	**64.45**	72.59	4.8	42.98	42.52	33.03	61.7	64.48	25.05	52.93	57.10
	few-shot	15.71	48.97	59.46	14.44	**46.84**	**47.2**	**59.78**	58.86	**73.13**	**9.8**	**44.17**	**46.76**	**39.25**	**63.47**	**66.31**	**27.79**	**52.46**	**58.57**
Gemma3:12b	zero-shot	13.07	49.85	57.92	3.94	35.97	40.44	36.72	55.98	60.94	2.3	37.31	34.85	35.04	60.97	64.96	18.21	48.02	51.82
	few-shot	12.64	49.06	56.78	8.14	39.16	43.26	44.26	56.77	63.49	5.6	39.77	40.65	36.13	59.72	63.87	21.35	48.89	53.61

Table 13: Zero-shot and few-shot classification results from LLMs on all English datasets. The best results per dataset are marked in bold.

G. Appendix G: Classification Reports and Confusion Matrices

G.1. SemEval

	Stella	Stella (L_P)	Qwen3:14b	RoBERTa	support
anger	51.70	**79.89**	76.46	75.03	1101
anticipation	23.09	27.41	32.00	**32.94**	425
disgust	56.99	**77.12**	50.92	75.72	1099
fear	45.79	71.50	60.36	**77.16**	485
joy	66.22	82.87	80.02	**85.43**	1442
love	38.86	**64.77**	41.38	46.88	516
optimism	60.99	73.75	50.09	**74.80**	1143
pessimism	32.70	**43.31**	37.65	31.91	375
sadness	52.33	**72.55**	70.07	72.31	960
surprise	12.53	27.82	**29.26**	18.72	170
trust	12.02	8.21	**10.42**	1.28	153
micro avg	48.96	66.83	59.46	**70.30**	7869
macro avg	41.20	**57.20**	48.97	53.83	7869

Table 14: Class-wise F1-scores per label and approach.

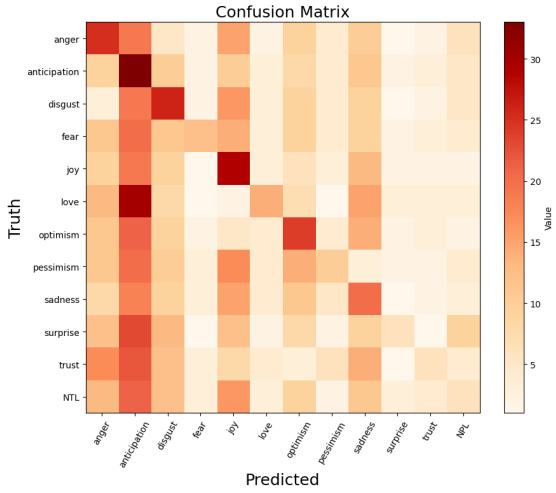

Figure 5: Confusion matrix for the SemEval dataset, obtained from Stella (DBC).

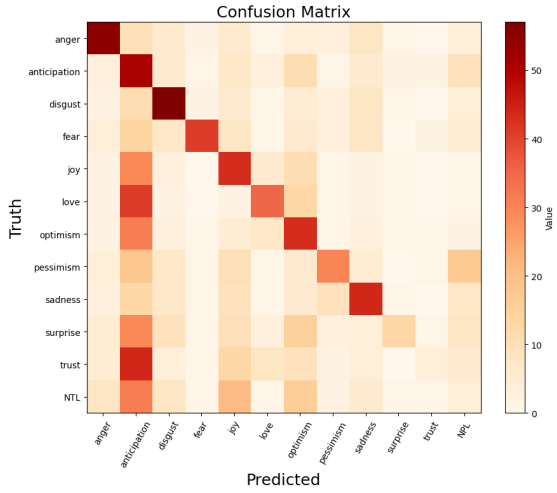

Figure 6: Confusion matrix for the SemEval dataset, obtained from Stella (L_P, DBC).

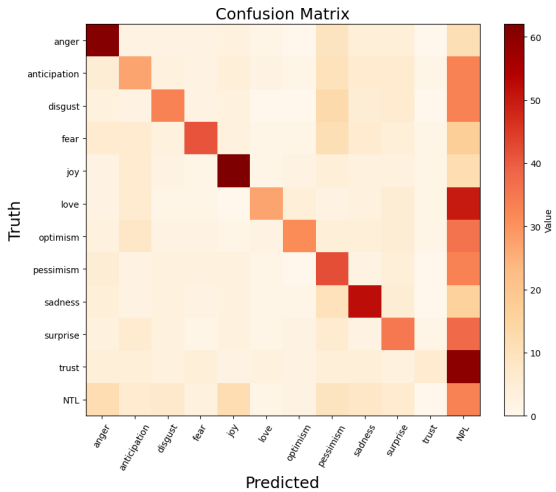

Figure 7: Confusion matrix for the SemEval dataset, obtained from *Qwen3:14b*.

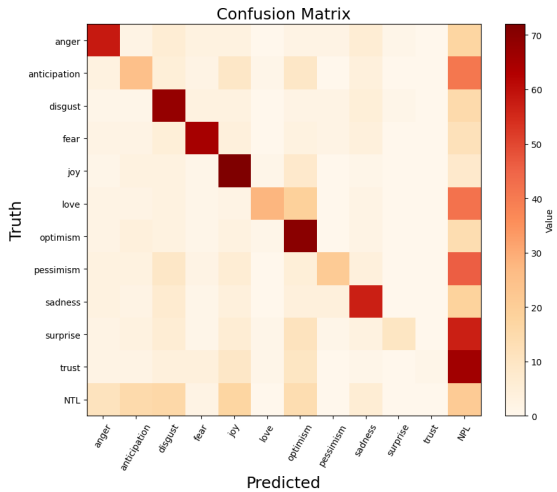

Figure 8: Confusion matrix for the SemEval dataset, obtained from *RoBERTa*.

G.2. BioTech

	Stella	Stella (L_P)	Qwen3:14b	RoBERTa	support
Company Description	48.59	47.31	47.62	**79.49**	103
Event Organization	17.39	18.18	38.46	**59.46**	9
Product Launch and Presentation	23.75	46.58	34.92	**50.00**	32
Industry Expansion	**8.33**	0.00	6.78	0.00	4
Service and Product Provision	9.88	33.33	7.81	**71.15**	8
Partnerships and Alliances	0.00	0.00	0.00	0.00	0.0
New Initiatives and Programs	23.53	30.77	30.09	**66.67**	34
Investment in Public Company	38.10	55.56	57.14	**86.67**	8
Funding Round	60.00	68.57	78.57	**82.82**	14
Support and Philanthropy	21.43	45.16	**57.14**	0.00	20
Mergers and Acquisitions	58.54	78.79	66.67	**81.25**	19
Executive Statement	64.67	77.35	75.00	**78.95**	162
Executive Appointment	54.55	73.33	**80.00**	**80.00**	15
Product Updates	9.30	0.00	30.30	**66.67**	9
Article Publication	0.00	25.81	21.82	**40.00**	18
Participation in an Event	8.33	11.76	25.81	**58.82**	10
Hiring	27.59	58.82	54.55	**97.30**	12
IPO Exit	28.57	**66.67**	10	51.52	2
Department Establishment	0.00	0.00	0.00	**70.45**	1
Foundation Establishment	1.30	40.00	0.00	**63.16**	2
Other Activities	32.46	**42.86**	35.00	0.00	101
Alliance and Partnership	41.90	**73.24**	52.46	0.00	37
Geographic Expansion	22.22	26.09	56.67	**70.00**	19
Clinical Trial Sponsorship	33.33	**66.67**	44.44	42.11	2
Company Closing Operations	33.33	40.00	**57.14**	54.55	4
Regulatory Approval	33.33	53.33	**66.67**	53.33	8
Patent Publication	5.48	6.15	**80.00**	0.00	2
Subsidiary Establishment	25.00	66.67	50.00	**69.57**	1
micro avg	34.45	53.43	47.20	**73.16**	656
macro avg	26.10	41.18	44.82	**52.64**	656

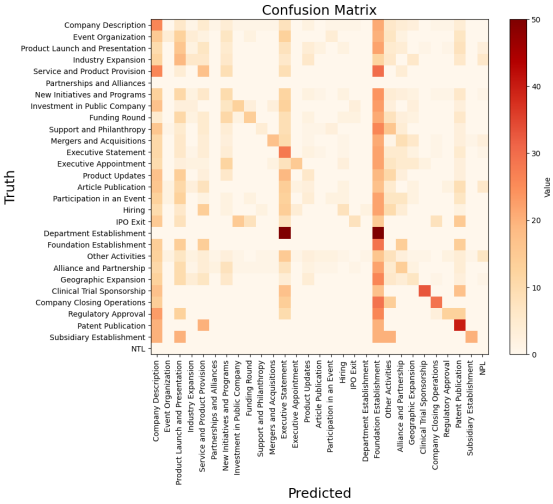

Figure 9: Confusion matrix for the BioTech dataset, obtained from Stella (DBC).

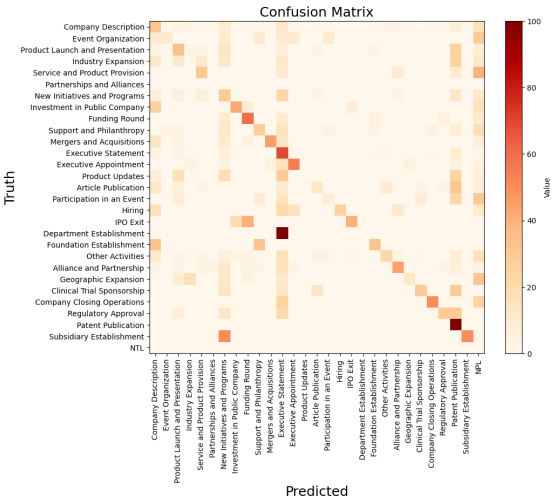

Figure 10: Confusion matrix for the BioTech dataset, obtained from Stella (L_P, DBC).

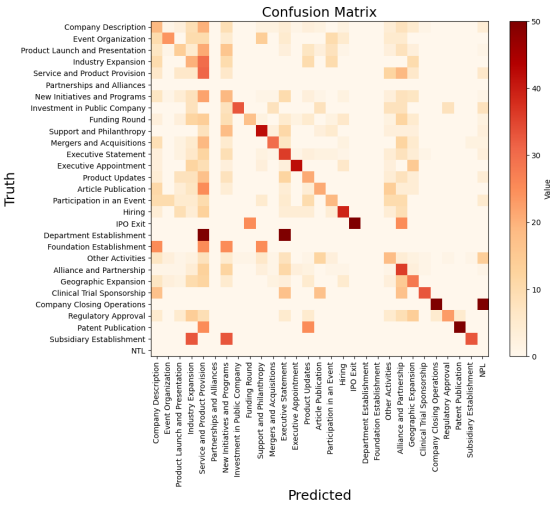

Figure 11: Confusion matrix for the Biotech dataset, obtained from *Qwen3:14b*.

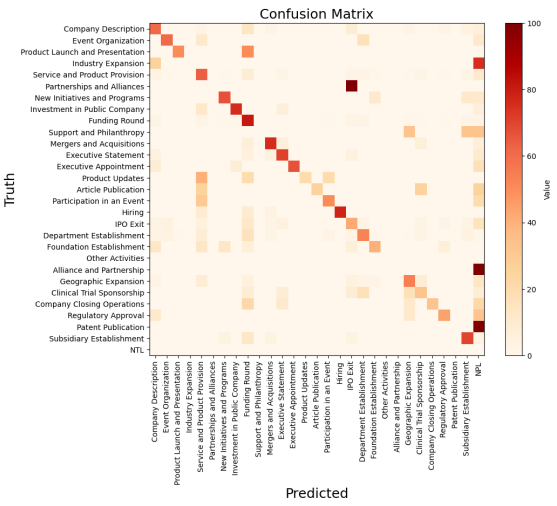

Figure 12: Confusion matrix for the BioTech dataset, obtained from *RoBERTa*.

G.3. LitCovid

	Stella	Stella (L_P)	Qwen3:14b	RoBERTa	support
Case Report	51.75	**88.14**	80.36	87.36	482
Diagnostic Methods	65.09	85.68	53.80	**86.83**	1546
Epidemic Forecasting	66.21	71.66	57.93	**85.30**	192
Biological Mechanisms	64.41	**85.41**	72.26	69.97	1073
Prevention Strategies	73.31	**93.42**	70.67	88.71	2750
Transmission Dynamics	33.63	62.30	40.07	**93.13**	256
Medical Treatments	53.72	**88.63**	69.18	61.09	2207
micro avg	61.68	88.01	66.31	**88.20**	8506
macro avg	58.30	**82.18**	63.47	81.77	8506

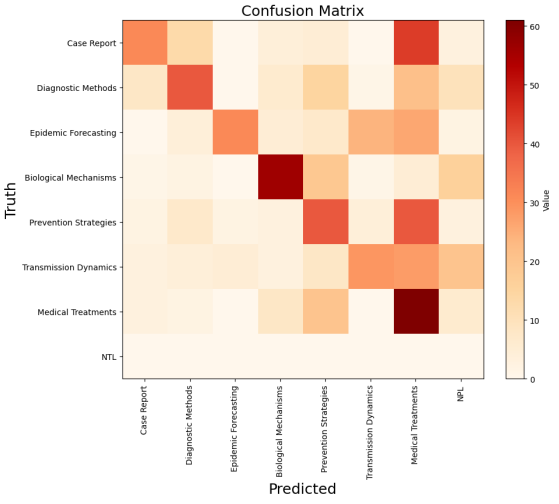

Figure 13: Confusion matrix for the LitCovid dataset, obtained from *Stella* (DBC).

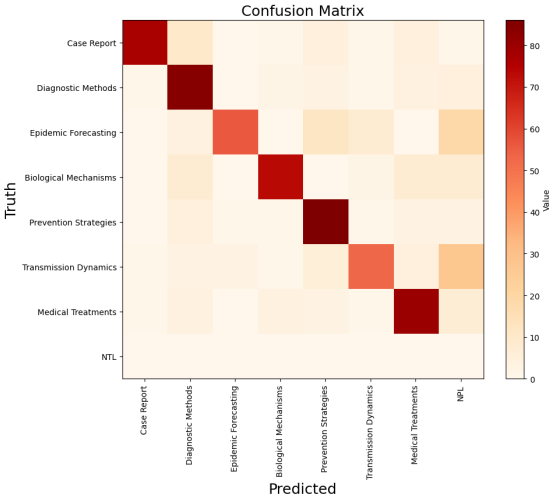

Figure 14: Confusion matrix for the LitCovid dataset, obtained from *Stella* (L_P, DBC).

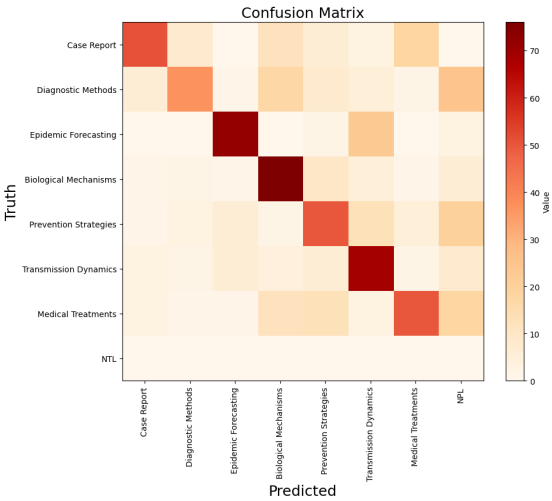

Figure 15: Confusion matrix for the LitCovid dataset, obtained from *Qwen3:14b*.

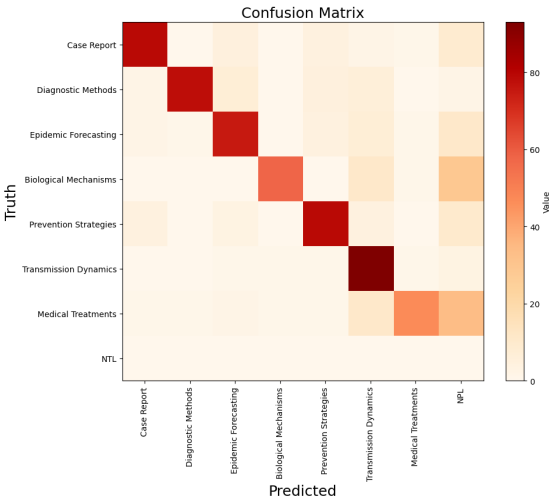

Figure 16: Confusion matrix for the LitCovid dataset, obtained from *RoBERTa*.

H. Appendix H: Statistical Significance Tests

		L_{LM} vs Base		L_{TM} vs Base		L_P vs Base	
Dataset	**Model**	**OR**	p	**OR**	p	**OR**	p
AAPD	GIST-Large	1.15	<0.001	1.06	0.002	1.75	<0.001
AAPD	BGE	1.16	<0.001	1.31	<0.001	1.73	<0.001
AAPD	Stella	1.38	<0.001	1.50	<0.001	2.41	<0.001
BioTech	GIST-Large	1.60	<0.001	1.19	<0.001	2.49	<0.001
BioTech	BGE	1.33	<0.001	1.41	<0.001	2.76	<0.001
BioTech	Stella	1.92	<0.001	1.58	<0.001	3.41	<0.001
LitCovid	GIST-Large	4.45	<0.001	4.16	<0.001	5.66	<0.001
LitCovid	BGE	2.89	<0.001	3.26	<0.001	5.16	<0.001
LitCovid	Stella	3.91	<0.001	4.09	<0.001	5.06	<0.001
Reuters	GIST-Large	1.51	<0.001	2.84	<0.001	20.18	<0.001
Reuters	BGE	2.20	<0.001	2.75	<0.001	2024	<0.001
Reuters	Stella	1.15	<0.001	1.28	<0.001	838	<0.001
SemEval	GIST-Large	1.88	<0.001	1.64	<0.001	2.05	<0.001
SemEval	BGE	1.69	<0.001	1.81	<0.001	2.75	<0.001
SemEval	Stella	2.27	<0.001	2.09	<0.001	2.82	<0.001

Table 15: GLMM odds ratios for each contrastive loss vs. the base threshold. OR > 1 indicates improved odds of a correct per-label prediction. P-values are Holm-Bonferroni corrected.

		L_P **vs** L_{LM}		L_P **vs** L_{TM}		L_{LM} **vs** L_{TM}	
Dataset	**Model**	**OR**	p	**OR**	p	**OR**	p
AAPD	GIST-Large	1.52	<0.001	1.65	<0.001	1.08	0.006
AAPD	BGE	1.49	<0.001	1.32	<0.001	1.13	<0.001
AAPD	Stella	1.74	<0.001	1.61	<0.001	1.08	0.007
BioTech	GIST-Large	1.55	<0.001	2.09	<0.001	1.34	<0.001
BioTech	BGE	2.07	<0.001	1.95	<0.001	1.06	0.185
BioTech	Stella	1.78	<0.001	2.15	<0.001	1.21	<0.001
LitCovid	GIST-Large	1.27	<0.001	1.36	<0.001	1.07	0.016
LitCovid	BGE	1.79	<0.001	1.58	<0.001	1.13	<0.001
LitCovid	Stella	1.30	<0.001	1.24	<0.001	1.05	0.115
Reuters	GIST-Large	84.41	<0.001	26.08	<0.001	3.24	<0.001
Reuters	BGE	24475	<0.001	19208	<0.001	1.27	<0.001
Reuters	Stella	82.07	<0.001	130.47	<0.001	1.59	<0.001
SemEval	GIST-Large	1.09	<0.001	1.25	<0.001	1.14	<0.001
SemEval	BGE	1.63	<0.001	1.52	<0.001	1.07	<0.001
SemEval	Stella	1.24	<0.001	1.35	<0.001	1.09	<0.001

Table 16: GLMM pairwise odds ratios between contrastive losses. OR > 1 indicates the first method has better odds of a correct prediction. P-values are Holm-Bonferroni corrected.