Improved Multi-label Classification under Temporal Concept Drift: Rethinking Group-Robust Algorithms in a Label-Wise Setting

Anonymous ACL submission

Abstract

In document classification for, e.g., legal and biomedical text, we often deal with hundreds of classes, including very infrequent ones, as well as temporal concept drift caused by the influence of real world events, e.g., policy changes, conflicts, or pandemics. Both class imbalance and drift are often approached by resampling the training data to simulate (or compensate for) a known target distribution, but what if the target distribution is determined by unknown future events? Instead of resampling uniformly to hedge our bets, we focus on the underlying optimization algorithms used to train such document classifiers and evaluate several group-robust optimization algorithms, initially proposed to mitigate group-level disparities. Reframing group-robust algorithms as adaptation algorithms under concept drift, we find that Invariant Risk Minimization and Spectral Decoupling outperform sampling-based approaches to class imbalance and concept drift, and lead to *much* better performance on minority classes. The effect is more pronounced the larger the label set.

1 Introduction

Multi-label document classification is the task of assigning a subset of labels from a large predefined set – of, say, hundreds or thousands of labels – to a given document. Common applications include labeling scientific publications with concepts from ontologies (Tsatsaronis et al., 2015), associating medical records with diagnostic and procedure labels (Johnson et al., 2017), pairing legislation with relevant legal concepts (Mencia and Fürnkranzand, 2007), or categorizing product descriptions (Lewis et al., 2004). The task in general presents interesting challenges due to the large label space and two-tiered skewed label distributions.

Class Imbalance In multi-label classification, datasets often exhibit class imbalance, i.e., skewed label distributions (Fig. 2). Common methods include resampling and reweighting based on heuristic assumptions, but methods are known to suffer from unstable performance, poor applicability, and high computational cost in complex tasks where their assumptions do not hold (Liu et al., 2020). Datasets with long-tail frequency distributions, like the ones considered below – sometimes referred to as *power-law datasets* (Rubin et al., 2012) – can be particular challenging. Also, the heuristics fix the trade-off between exploiting as much of the training data as possible and balancing the classes, instead of trying to learn the optimal trade-off.

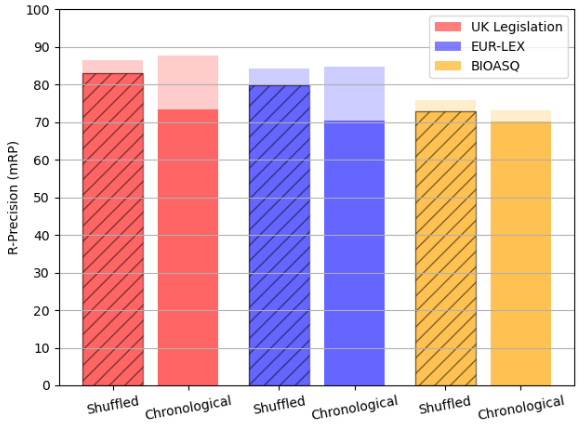

Figure 1: Model performance using *random* vs. *chronological* splits across the medium-sized datasets (Table 1). The shaded parts of the bars are the train/test discrepancy due to *over-fitting*. The performance drop from random to chronological splits demonstrates the *temporal concept drift*.

Temporal Concept Drift Moreover, class distributions may change over time. This is one dimension of the *temporal generalization* problem (Lazaridou et al., 2021). Recently, Søgaard et al. (2021) argued chronological data splits are necessary to estimate real-world performance, contrary to random splits (Gorman and Bedrick, 2019), because random splits artificially removes drift. Temporal concept drift, which we focus on here – in-

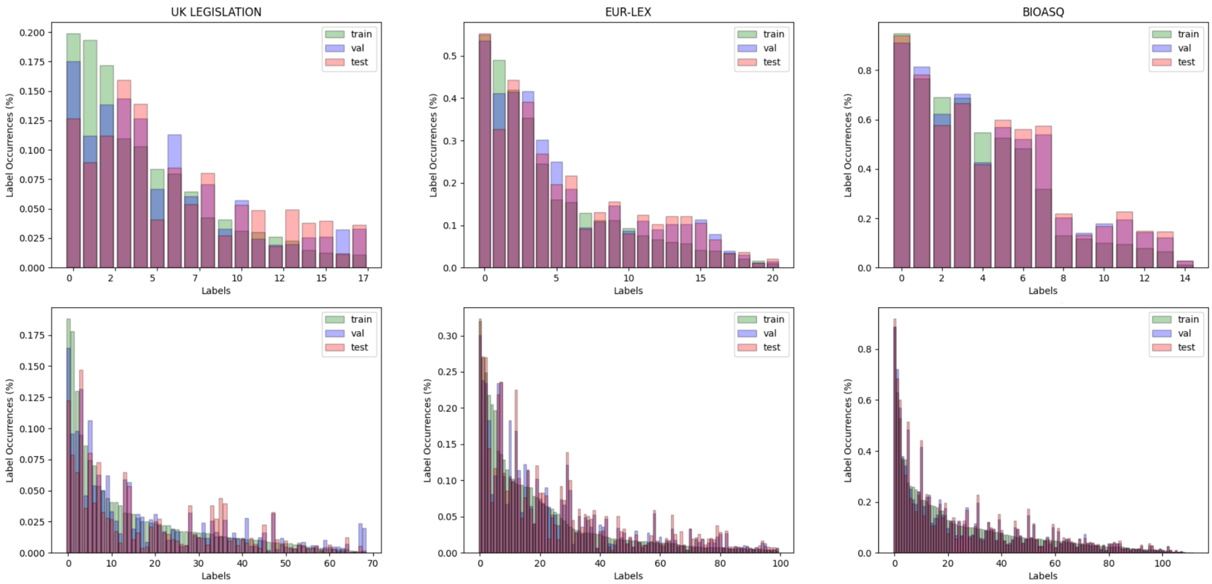

Figure 2: Label distributions of the medium-sized datasets. *Class imbalance* across labels (bars) in the x axis and *temporal concept drift* across subsets depicted with different coloured bars in the y axis.

stead of covariate shift (Shimodaira, 2000), for example – is an instance of concept drift (Gama et al., 2014), often discussed in the domain adaptation literature, e.g., Chan and Ng (2006).

2 Related Work

Temporal Drift Temporal drift has been studied in several NLP tasks, including document classification (Huang and Paul, 2018, 2019), sentiment analysis (Lukes and Søgaard, 2018), Named Entity Recognition (NER) (Rijhwani and Preotiuc-Pietro, 2020), Neural Machine Translation (NMT) (Levenberg et al., 2010) and Language Modelling (Lazaridou et al., 2021). None of these papers focus on class imbalance and temporal concept drift. These papers have mainly been diagnostic, not providing technical solutions that are applicable in our case.

Multi-label Class Imbalance Class imbalance in multi-label classification has so far been studied through the lens of network *architectures*, searching for the best neural architecture for handling few- and zero-shot labels in the multi-label setting. To improve the performance for underrepresented (few-shot) classes, (Snell et al., 2017) introduced Prototypical Networks that average all instances in each class to form *prototype* label vectors (encodings), a form of inductive bias, which improved few-shot learning. In a similar direction, Mullenbach et al. (2018) developed the Label-Wise Attention Network (LWAN) architecture, in which label-wise document representations are learned by attending to the most informative words for each label, using trainable label encodings (representations). Rios and Kavuluru (2018) extended LWAN and the idea of *prototype* label encodings. They combined label descriptors with information from a graph convolutional network (Kipf and Welling, 2017) that considered the relations of the label hierarchy to improve the results in few-shot and zero-shot settings. Alternatives to LWAN were considered by Chalkidis et al. (2020a), presenting minor improvements in the few-shot setting, but harming the overall performance.

Fairness The literature on inducing approximately fair models from biased data is rapidly growing. See Mehrabi et al. (2021) for a recent survey. We rely on this literature in how we define fairness, and for the algorithms that we compare in our experiments below. The fairness-promoting learning algorithms we evaluate are discussed in detail in Section 4. Recent studies targeting fairness show that class imbalance has connections to bias (Blakeney et al., 2021; Subramanian et al., 2021), i.e., mitigating class-wise disparities has a chain effect on lowering group-wise disparities.

We focus on (large-scale) multi-label document classification and study a fundamental component of the learning process leading to performance disparities across labels, i.e., the underlying *optimization algorithm* used for training. We consider group-robust optimization algorithms initially proposed to mitigate group disparities given specific protected attributes (e.g., gender, race), but re-frame these algorithms to optimize for good performance across labels rather than across groups.

Dataset	Domain	No. of Documents	Setting	No. of Labels	Distribution Swift (WS) Random	Chronological	Diff.
UK-LEX (new)	UK Legislation	36,500	Small	18 / 18	0.002	0.016	(8×)
			Medium	69 / 69	0.001	0.005	(5×)
EUR-LEX (Chalkidis et al., 2021)	EU Legislation	65,000	Small	20 / 21	0.003	0.027	(9×)
			Medium	100 / 127	0.001	0.007	(7×)
BIOASQ (Tsatsaronis et al., 2015)	Biomedical Articles	100,000	Small	16 / 16	0.002	0.058	(29×)
			Medium	112 / 116	0.002	0.009	(5×)

Table 1: Main characteristics of the examined datasets. We report the application domain, the number of documents, the available setting and the corresponding number of labels (used / total), and the label distribution swift measured as the Wasserstein Distance (WS) between train-test label probability distributions.

3 Datasets

We experiment with three datasets (Table 1) from two domains (legal and biomedical), which support two different classification settings (label granularities), i.e., label sets including more abstract or more specialized concepts (labels).[1]

UK-LEX United Kingdom (UK) legislation is publicly available as part of the United Kingdom's National Archives.[2] Most of the laws have been categorized in thematic categories (e.g., health-care, finance, education, transportation, planing) that are presented in the document preamble and are used for archival indexing purposes. We release a new dataset, which comprises 36.5k UK laws (documents). The dataset is chronologically split in training (20k, 1975–2002), development (8.5k, 2002–2008), test (8.5k, 2008–2018) subsets. It supports two different label granularities, comprising 18, and 40 topics (labels), respectively.

EUR-LEX European Union (EU) legislation is published in EUR-Lex.[3] All EU laws are annotated by EU's Publications Office with multiple concepts from EuroVoc, a thesaurus maintained by the Publications Office.[4] EuroVoc has been used to index documents in systems of EU institutions, e.g., in web legislative databases, such as EUR-Lex and CELLAR, the EU Publications Office's common repository of metadata and content. We use the English part of the dataset of Chalkidis et al. (2021), which comprises 65k EU laws (documents). The dataset is chronologically split in training (55k, 1958–2010), development (5k, 2010–2012), test (5k, 2012–2016) subsets. It supports four different label granularities. We use the 1st and 2nd level of the EuroVoc taxonomy including 21 and 127 categories, respectively.

BIOASQ The BIOASQ (Task A: Large-Scale Online Biomedical Semantic Indexing) dataset (Tsatsaronis et al., 2015) comprises biomedical articles from PubMed,[5] annotated with concepts from the Medical Subject Headings (MeSH) taxonomy.[6] MeSH is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information, e.g., in MEDLINE/PubMed, and the NLM databases. We use a subset of 100k documents derived from the latest version (v.2021) of the dataset. We sub-sample documents in the period 2000-2021, and we consider chronologically split training (80k, 1964–2015), development (10k, 2015–2018), test (10k, 2018–2020) subsets. We use the 1st and 2nd levels of MeSH, including 16 and 116 categories.

4 Fine-tuning Algorithms

In our experiments, we rely on pre-trained English language models (Devlin et al., 2019) and fine-tune these using different learning objectives. Our main goal during fine-tuning is to find a hypothesis (h) for which the risk $R(h)$ is minimal:

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \tag{1}$$

$$R(h) = \mathbf{E}[\mathcal{L}(h(x), y)] \tag{2}$$

where y are the targets (*ground truth*) and $h(x) = \hat{y}$ is the system hypothesis (model's predictions).

Similar to previous studies, $R(h)$ is an expectation of the selected loss function ($\mathcal{L}$). In this work, we study multi-label text classification (Section 3), thus we aim to minimize the binary cross-entropy loss across L classes:

$$\mathcal{L}(x) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \tag{3}$$

[1] We originally also considered the MIMIC-III dataset of Johnson et al. (2017) including discharge summaries fro US hospitals annotated with ICD-9 medical codes, but the publication date of the documents has been "counterfeited" as part of the anonymization process. Experimental results with random splits are presented in Appendix A.

[2] https://www.legislation.gov.uk/

[3] http://eur-lex.europa.eu/

[4] http://eurovoc.europa.eu/

[5] https://pubmed.ncbi.nlm.nih.gov

[6] https://www.nlm.nih.gov/mesh/

ERM (Vapnik, 1992), which stands for Empirical Risk Minimization, is the most standard and widely used optimization technique to train neural methods. The loss is calculated as follows:

$$\mathcal{L}_{ERM} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(x_i) \tag{4}$$

where N is the number of instances (training examples) in a batch, and $\mathcal{L}_i$ is the loss per instance.

Furthermore, we consider a representative selection of group-robust fine-tuning algorithms that try to mitigate performance disparities with respect to a given attribute (A), e.g., in a standard scenario that could be the gender of a document's author in sentiment analysis, or the background landscape in image classification. In our case, the attribute of interest is the labeling of the documents. The attribute is split into G groups, which in our case are the classes ($G = L$). All algorithms rely on a balanced group sampler, i.e., an equal number(N_{g_i}) of instances (samples) per group (g_i) are included at each batch. Most of the algorithms are built upon group-wise losses ($\mathcal{L}_{g_i}$), computed as follows:

$$\mathcal{L}(g_i) = \frac{1}{N_{g_i}} \sum_{j=1}^{N_{g_i}} \mathcal{L}(x_j) \tag{5}$$

In our case, contrary to previous applications of group-robust algorithms, the groups (classes) are not mutually exclusive (documents are tagged with multiple labels). Hence, the group sampler can only guarantee that *at least* N groups (labels) will be considered at each step, but most probably even more. In this work, we examine the following group-robust algorithms in a label-wise fashion:

Group Uniform is the more naive group robust algorithm that uses the average of the group-wise (label-wise) losses -all groups (labels) are considered equally important-, instead of the standard sample-wise average, as follows:

$$\mathcal{L}_{GM} = \frac{1}{G} \sum_{i=1}^{G} \mathcal{L}(g_i) \tag{6}$$

Group DRO (Sagawa et al., 2020), stands for Group Distributionally Robust Optimization (DRO). Group DRO is an extension of the Group Uniform algorithm, where the group-wise (label-wise) losses are weighted inversely proportional to the group (label) performance. The total loss is calculated as follows:

$$\mathcal{L}_{DRO} = \sum_{i=1}^{G} w_{g_i} * \mathcal{L}(g_i), \text{ where} \tag{7}$$

$$w_{g_i} = \frac{1}{W}(\hat{w}_{g_i} * e^{L(g_i)}) \quad \text{and} \quad W = \sum_{i=1}^{G} w_{g_i} \tag{8}$$

where G is the number of groups (labels), $\mathcal{L}_g$ are the averaged group-wise (label-wise) losses, w_g are the group (label) weights, $\hat{w}_g$ are the group (label) weights as computed in the previous update step.

V-REx (Krueger et al., 2020), which stands for Risk Extrapolation, is yet another proposed group-robust optimization algorithm. Krueger et al. (2020) hypothesize that variation across training groups is representative of the variation later encountered at test time, so they also consider the variance across the group-wise (label-wise) losses. In V-REx the total loss is calculated as follows:

$$\mathcal{L}_{REX} = \mathcal{L}_{ERM} + \lambda * \text{Var}([\mathcal{L}_{g_1}, \ldots, \mathcal{L}_{g_G}]) \tag{9}$$

where Var is the variance among the group-wise (label-wise) losses, and λ, a weighting hyper-parameter scalar.

IRM (Arjovsky et al., 2020), which stands for Invariant Risk Minimization, mainly aims to penalize variance across multiple training dummy estimators across groups, i.e., performance cannot vary in samples that correspond to the same group. The total loss is computed as follows:

$$\mathcal{L}_{IRM} = \frac{1}{G} \sum_{i=1}^{G} [\mathcal{L}(g_i) + \lambda * P(g_i)] \tag{10}$$

$$P_{g_i} = \nabla[\mathcal{L}_{g_i=1,3,\ldots}^{N_{g_i}} \mid 1\,] * \nabla[\mathcal{L}_{g_i=2,4,\ldots}^{N_{g_i}} \mid 1\,] \tag{11}$$

where $\mathcal{L}_{gi}$ is the loss of the i_{th} instance, which is part of the g_{th} group (label). Refer to Arjovsky et al. (2020) for a more detailed introduction of the group penalty terms (P_g).

Deep CORAL (Sun and Saenko, 2016), minimizes the difference in second-order statistics (covariances) between the source and target feature activations. In practice, it introduces group-pair penalties:

$$\mathcal{L}_{CORAL} = \mathcal{L}_{ERM} + \lambda * \frac{1}{G}\left(\sum_{i=1}^{G} P(g_i, g_{i+1})\right) \tag{12}$$

$$P(g_i, g_{i+1}) = [\overline{C_{g_i}} - \overline{C_{g_{i+1}}}]^2 + [\overline{X_{g_i}} - \overline{X_{g_{i+1}}}]^2 \tag{13}$$

where $\overline{C_{g_i}}$ are the averaged covariances of the ith group and $\overline{X_{g_i}}$ are the averaged features (document representations) of the ith group, respectively. Refer to Sun and Saenko (2016) for a more detailed introduction of the group penalty terms (P_g).

Spectral Decoupling (Pezeshki et al., 2020) relies on the idea of *Gradient Starvation*. Pezeshki et al. state that a network could become over-confident in its predictions by capturing only one or a few dominant features. Thus, adding an L2 penalty on the network's logits ($\hat{y}_i$) provably decouples the fixed points of the dynamics. The total loss is computed as follows:

$$\mathcal{L}_{SD} = \mathcal{L}_{ERM} + \lambda * \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i^2 \tag{14}$$

In our work, we consider the aforementioned algorithms in a label-wise setting, instead of a group-wise setting given a protected attribute. In our case, $G = L$, where L is the number of labels.

5 Experimental SetUp

Baseline Models For both legal datasets (UK-LEX, EUR-LEX), we use the small LEGAL-BERT model of Chalkidis et al. (2020b), a BERT (Devlin et al., 2019) model pre-trained on English legal corpora. For BIOASQ, we use the small English BERT model of Turc et al. (2019). Following Devlin et al. (2019), we feed each document to the pre-trained model and obtain the top-level representation $h_{\texttt{[cls]}}$ of the special `[cls]` token as the document representation. The latter goes through a dense layer of L output units, one per label, followed by a sigmoid activation.

We also experiment with the Label-Wise Attention Network (LWAN) relying on a BERT encoder (Chalkidis et al., 2020a), dubbed BERT-LWAN.[7] Chalkidis et al. reported state-of-art results in EUR-LEX and AMAZON-13K using BERT-LWAN compared to several baselines. BERT-LWAN uses one attention head per label to generate L document representations d_l:

$$a_{lt} = \frac{\exp(K(h_t)Q_l)}{\sum_{t'} \exp(K(h_{t'})Q_l)} \tag{15}$$

$$d_l = \frac{1}{T} \sum_{t=1}^{T} a_{lt} V(h_t) \tag{16}$$

T is the document length in tokens, h_t the context-aware representation of the t-th token, K, V are linear transformations of h_t, and Q_l a trainable vector used to compute the attention scores of the l-th attention head; u_l can also be viewed as a label representation. Intuitively, each head focuses on possibly different tokens of the document to decide if the corresponding label should be assigned. BERT-LWAN employs L linear layers (o_l) with sigmoid activations, each operating on a different label-wise document representation d_l, to produce the probability of the corresponding label p_l:

$$p_l = \text{sigmoid}(d_l \cdot o_l) \tag{17}$$

Training and Evaluation Details We fine-tune all models using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5. We use a batch size of 64 and train models for up to 20 epochs using early stopping on the development set. Across experiments, we use BERT models following a small configuration (6 transformer blocks, 512 hidden units and 8 attention heads), which allows us to increase the batch size up to 64 and consider samples with multiple labels (groups) in the group robust algorithms. In practice, this enables us to sample at least 4 samples per group (label) for all labels in the small label sets, and at least 1 sample per group (label) for 64 labels in the medium-sized label sets (69-112 labels).

Given the large number and skewed distribution of labels, retrieval measures have been favored in large-scale multi-label text classification literature (Mullenbach et al., 2018; You et al., 2019; Chalkidis et al., 2020a). Following Chalkidis et al. (2020a), we report *mean R-Precision* (m-RP) (Manning et al., 2009), while we also report the standard *micro-F1* (μ-F_1) and *macro-F1* (m-F_1) to better estimate the class-wise performance disparity.

In our experiments, we use and extend the WILDs (Koh et al., 2021) library, which provides an experimental framework for experimenting with group-robust algorithms. We effectively rewrote all parts of code to consider label-wise groups and losses, while we also implemented the unsupported methods (Group Uniform, V-REx, and Spectral Decoupling). For reproducibility and further exploration with new group-robust methods, we release our code on Github.[8]

[7]The original model was proposed by Mullenbach et al. (2018), with a CNN encoder.

[8]The Github repository will be released upon acceptance. Meanwhile, reviewers have access to the internally submitted code (.zip).

Algorithm	EUR-LEX						UK-LEX						BIO-ASQ					
	Small			*Medium*			*Small*			*Medium*			*Small*			*Medium*		
	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP
ERM	**79.3**	64.4	**84.2**	68.4	40.4	70.5	80.2	75.2	83.6	66.5	35.8	73.3	**85.9**	75.7	**87.6**	68.6	46.7	70.3
ERM+GS	79.2	65.7	83.1	69.0	42.8	70.9	80.1	75.4	83.9	67.8	41.4	73.8	85.3	75.9	86.3	68.4	48.2	69.8
Group Uniform	78.4	67.9	81.9	68.6	50.2	70.0	79.1	74.5	84.1	69.1	56.2	**75.0**	85.2	76.3	86.8	68.6	51.5	69.5
Group DRO	77.8	62.6	79.0	67.5	43.8	67.4	78.8	73.4	83.5	60.9	29.3	68.9	84.3	72.8	84.9	43.9	13.9	43.8
Deep CORAL	78.7	68.1	82.3	67.7	44.1	70.5	79.6	75.2	83.6	67.2	53.1	74.7	85.1	75.4	86.1	68.8	53.2	69.9
V-REx	78.6	68.0	82.6	69.0	49.4	69.7	80.2	75.8	84.6	68.4	52.1	74.7	85.1	76.3	86.8	69.3	51.8	71.4
IRM	78.5	67.7	81.1	69.9	**54.8**	70.7	79.4	74.9	84.2	69.4	58.9	**75.0**	85.2	76.4	86.8	69.5	**54.7**	70.0
SD	**79.3**	**69.2**	79.5	**70.7**	52.4	**72.2**	**80.3**	**76.8**	**84.8**	**70.0**	**59.1**	74.8	85.5	**76.8**	86.9	**71.0**	53.4	**72.2**

Table 2: **Overall** results of the **group-robust (label-robust) algorithms** across all datasets (UK-LEX, EUR-LEX, BIOASQ) and settings (small and medium sized label sets).

Algorithm	EUR-LEX						UK-LEX						BIOASQ					
	Head			*Tail*			*Head*			*Tail*			*Head*			*Tail*		
	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP
ERM	73.7	61.8	74.5	26.6	19.0	51.8	71.8	55.3	77.4	36.4	15.8	76.4	71.8	60.6	73.2	45.9	32.9	57.7
ERM+GS	74.1	62.4	74.7	30.3	23.1	52.5	72.7	57.4	77.8	40.2	28.3	77.5	72.2	61.2	72.7	47.6	38.7	57.4
Group Uniform	73.0	62.0	73.5	43.1	38.5	54.5	70.9	60.2	77.8	62.2	52.1	79.5	71.5	60.4	72.5	50.0	42.5	57.1
Group DRO	70.1	50.9	70.3	8.4	5.5	28.3	66.7	46.6	73.4	29.4	11.6	69.4	58.9	26.9	59.6	1.1	0.9	0.8
Deep CORAL	72.4	59.7	73.7	35.0	28.5	57.0	69.7	53.3	75.1	61.2	43.4	80.0	72.7	63.1	73.6	55.2	49.3	63.1
V-REx	73.2	61.7	73.1	43.1	37.1	55.1	70.4	56.6	76.7	60.6	47.6	80.2	71.3	59.5	72.4	47.1	37.4	56.7
IRM	73.8	**64.3**	74.1	**48.8**	**45.2**	57.0	71.3	62.6	77.8	62.6	55.2	80.7	72.0	62.5	72.7	**53.3**	**47.0**	59.2
SD	**74.7**	63.8	**75.2**	47.0	41.0	**59.1**	**71.7**	**62.4**	**77.1**	**64.0**	**55.8**	**82.2**	**73.6**	**64.0**	**74.7**	52.7	42.8	**62.9**

Table 3: Results of group-robust algorithms in ***head* and *tail* classes** in the medium-sized datasets. *Head* are the 50% most represented (frequent) classes in the training set, and *tail* are the bottom 50%.

6 Results

Main Results To highlight the temporal concept drift, we initially fine-tune BERT in all datasets with the standard ERM optimization algorithm using both *random* and *chronological* splits. Table 4 shows that the real-world performance achieved using the chronological split is severely overestimated using the random split (approx. +10% across evaluation measures) in two out of threee datasets. While all datasets have inherently skewed distributions (class imbalance), which is naturally demonstrated by the performance discrepancy between μ-F$_1$ and m-F$_1$ scores (especially when we consider the larger label sets), the temporal dimension further exacerbate the performance discrepancy as label distributions also vary across subsets (Fig 2).

Dataset	Random			Chronological		
	μ-F$_1$	m-F$_1$	m-RP	μ-F$_1$	m-F$_1$	m-RP
UK-LEX-SM	**89.3**	**87.5**	**92.9**	80.2	75.2	83.6
UK-LEX-MD	**78.2**	**45.6**	**85.0**	66.5	35.8	73.3
EUR-LEX-SM	**86.8**	**76.5**	**89.5**	79.3	64.4	84.2
EUR-LEX-MD	**77.6**	**49.8**	**79.8**	68.4	40.4	70.5
BIOASQ-SM	**86.5**	**75.9**	**88.8**	85.9	75.7	87.6
BIOASQ-MD	**71.9**	**48.2**	**72.3**	68.6	46.7	70.3

Table 4: Overall results across all datasets and settings using **random vs. chronological splits** with ERM.

In Table 2, we present the overall results for the different optimization algorithms considering the baseline model, BERT. We observe that using a *group sampler* (ERM+GS), which equals standard oversampling of minority classes, slightly improve the results in m-F$_1$ (+1-4%) in many cases, while the performance is comparable in μ-F$_1$ and m-RP. Considering the results of group-robust algorithms, we observe that most of them improve m-F$_1$ across datasets compared to ERM and ERM+GS, +1-4% for small-sized datasets and +5-12% in medium-sized datasets. Again the performance in μ-F$_1$ and m-RP is mostly comparable or a bit lower, as sample-wise averaged measures are dominated by frequent classes due to class imbalance.

Contrary, Group DRO is consistently outperformed even by the standard ERM. Recall that Group DRO uses a weighted average of the group-wise (label-wise) losses (Eq. 7-8), where the group weights rely on the momentum of the group-wise (label-wise) losses (Eq. 8). In our case, this regularization acts counter-intuitively, as weights for the infrequent classes, which are rarely present across batches, are not updated (increased) constantly. This leads to an asymmetry, where some weights are frequently updated, while others not, and in time the latter are almost zeroed-out and not affect the training objective (loss).

The effect of group-robust algorithms in relation to the size of the label set. In Tables 2, we can also observe that the performance gains of group-robust algorithms compared to ERM are greater when we use the larger label sets. This is also as the class imbalance and temporal concept drift are more severe when we consider more

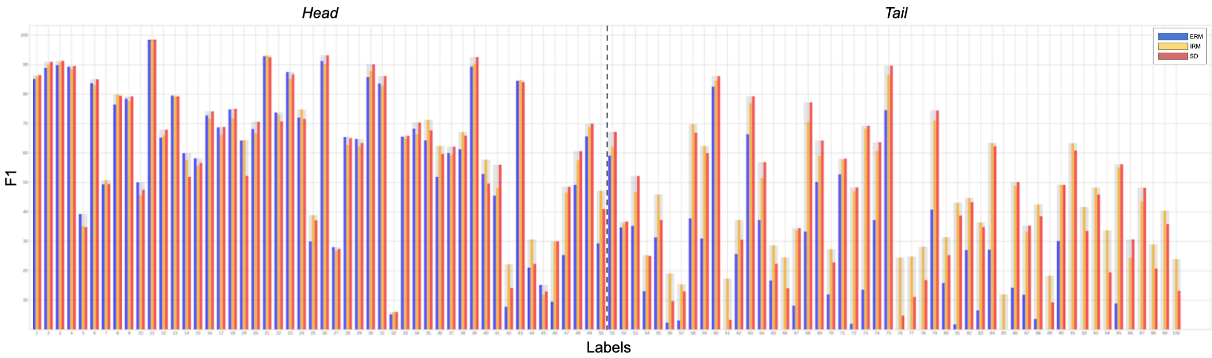

Figure 3: ***Class-wise F1-score*** results for ERM, IRM and Spectral Decoupling on medium-sized EUR-LEX. The classes have been ordered (left-to-right) based on the label distribution in the training subset.

Algorithm	**BERT**									**BERT-LWAN**								
	Overall			*Head*			*Tail*			*Overall*			*Head*			*Tail*		
	μ-F_1	m-F_1	m-RP	μ-F_1	m-F_1	m-RP	μ-F_1	m-F_1	m-RP	μ-F_1	m-F_1	m-RP	μ-F_1	m-F_1	m-RP	μ-F_1	m-F_1	m-RP
ERM	68.4	40.4	70.5	73.7	61.8	74.5	26.6	19.0	51.8	70.2	50.3	**71.8**	74.4	64.3	**75.6**	44.2	36.2	54.8
ERM+GS	69.0	42.8	70.9	74.1	62.4	74.7	30.3	23.1	52.5	69.1	54.1	69.9	73.1	63.6	73.4	47.6	44.5	56.6
Group Uniform	68.6	50.2	70.0	73.0	62.0	73.5	43.1	38.5	54.5	68.9	54.7	69.7	73.2	63.9	73.7	47.7	**45.4**	56.8
Group DRO	63.5	28.2	63.4	70.1	50.9	70.3	8.4	5.5	28.3	66.8	39.8	65.9	72.1	59.4	70.7	31.0	20.2	43.6
Deep CORAL	67.7	44.1	70.5	72.4	59.7	73.7	35.0	28.5	57.0		n/a			n/a			n/a	
V-REx	69.0	49.4	69.7	73.2	61.7	73.1	43.1	37.1	55.1	69.2	**54.6**	70.3	73.0	63.8	74.2	48.1	**45.4**	56.8
IRM	69.9	**54.8**	70.7	73.8	**64.3**	74.1	**48.8**	**45.2**	57.0	69.1	54.2	70.1	73.3	63.7	73.9	47.8	44.7	56.3
SD	70.7	52.4	**72.2**	74.7	63.9	**75.1**	47.0	41.0	**59.1**	**70.3**	54.2	70.6	74.4	64.4	73.6	47.8	44.1	**58.4**

Table 5: Results of group-robust algorithms with **different models** (BERT, and BERT-LWAN) in the medium-sized version of EUR-LEX. Deep CORAL is not applicable (n/a) in LWAN -there is not a universal featurizer-.

refined labels, especially considering m-F_1.

The effect of group-robust algorithms in relation to class frequency. In Table 3, we present results for the different optimization algorithms considering two groups of classes based on their frequency. *Head* classes are the 50% most frequent classes in the training set, while *tail* are the bottom 50%. As expected, the performance in head classes is much better compared to tail ones across datasets (approx. +20-40% in m-F_1). We observe that the performance gains of group-robust algorithms compared to ERM are greater in the tail classes (+10-20% in m-F_1). This is further highlighted in Figure 3, where we observe that IRM and Spectral Decoupling have larger gains in the right part (tail labels). This is highly expected as the goal of the group-robust algorithms is to minimize the group-wise (in our case, label-wise) disparity. Group DRO is severely out-performed in both head and tail, especially in the tail classes (whose weights have been zeroed-out, as previously noticed).

The effect of group-robust algorithms using BERT-LWAN. In this part, we compare the effect of the group-robust algorithms in between standard BERT and BERT-LWAN. In Table 5, we observe that BERT-LWAN closes the gap between ERM and the best-of group-robust algorithms. The results of ERM when we use BERT-LWAN are improved across measures, especially when we consider m-F_1 with a 10% improvement over the standard BERT. Both IRM and Spectral Decoupling seem quite insensitive to the underlying model (Fig. 4). Similarly, the results for the rest of the group-robust algorithms are improved. Nonetheless, there are still benefits in m-F_1 and less represented (*rare*) labels in general. Interestingly, Spectral Decoupling improves results in both F_1 scores. Although, we observe a mild performance drop (-1-2%) in m-RP when we consider overall and head classes. We hypothesize that group-robust algorithms negatively affect the ability of the model to correctly rank labels, as they force the model to consider all labels and be less confident (discriminatory) with one way or another.

Why IRM and Spectral Decoupling are a better fit compared to the rest of the algorithms? To answer this question, we need to identify the main differentiation between IRM, Spectral Decoupling and the rest of the methods. Both IRM and Spectral Decoupling follow similar incentives. IRM penalizes variance across losses in the same group (Eq. 10), i.e., in our case, the network is penalized if there is a performance disparity between samples labeled with the same classes using as a reference a dummy classifier. Spectral Decoupling penalizes the variance across label predictions (Eq. 14), i.e.,

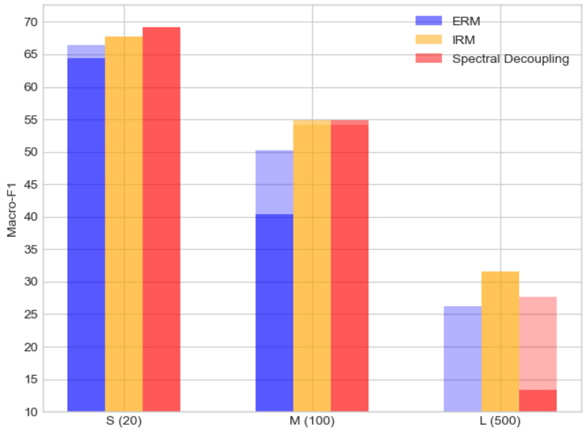

Figure 4: LWAN-BERT performance using ERM, IRM, and Spectral Decoupling across all EUR-LEX settings. The shaded part denotes the performance improvement compared to the standard BERT.

the network is penalized for being over-confident. The rest of the algorithms mainly rely on an equal consideration of the group-wise (in our case, label-wise) losses (Eq. 6), i.e., in our case, all classes are equally important for the training objective.

The latter incentive (averaging across group-wise losses) seems very intuitive, although in practice the groups (labels) co-occur (are not mutually exclusive) in a multi-label setting, thus frequent labels remain "first class citizens" in the optimization process, biasing parameter updates in their favor.

Contrary, both IRM and Spectral Decoupling use a learning component (loss term), which penalizes *label degeneration*. This is particularly important in multi-label classification, especially when we consider large label sets, as networks tend to over-fit (specialize) in few dominant (frequent) labels that shape the training loss and finally ignore (zero-out) the rest of the labels. This is quite different from the concept of *Gradient Starvation*, introduced by Pezeshki et al. (2020), where a network becomes over-confident in its predictions by capturing only few dominant features, as in our case the main issue is the label degeneration rather than possible spurious correlations learned by the network. Moreover, Spectral Decoupling does not rely on group-wise losses, similar to the rest.

In Figure 4, we compare the performance of ERM, IRM, and Spectral Decoupling across three EUR-LEX settings, small-sized, medium-sized, and one extra large-sized considering the 3rd level of EuroVoc including 500 concepts (labels). In the small label set, we observe that the use of LWAN-BERT slightly improves the performance when trained with ERM compared to standard BERT (shaded part of the bars). In the medium label set, as already discussed, we observe a 10% improvement with ERM, while in case of the large label set, using LWAN-BERT leads to a 25% improvement with ERM, and 15% with Spectral Decoupling, while IRM proves to be robust across all settings and both neural methods.

7 Conclusions & Future Work

We considered one of the main challenges in large-scale multi-label text classification, which comes from the fact that not all labels are well represented in the training set due to the class imbalance and the effect of temporal concept drift. To mitigate label disparities, we considered several group-robust optimization algorithms initially proposed to mitigate group disparities given specific attributes. Experimenting with three datasets in two different settings, we empirically find that group-robust algorithms vastly improve performance considering macro-averaged measures, while two of the group-robust algorithms (Invariant Risk Minimization and Spectral Decoupling) improve performance across all measures. Considering a more well-suited neural method (LWAN-BERT), we observe a vast performance improvement using ERM, which is still outperformed by both group-robust algorithms.

In the future, we would like to further investigate the two-tier anomaly (class imbalance and temporal concept drift). In this direction, we would like to directly take into consideration the time dimension by utilizing this information in group sampling and algorithms (e.g., groups over period of time). We would also like to consider data augmentation techniques (e.g., paraphrasing via masked-language modeling (Ng et al., 2020), and teacher forcing exploiting unlabeled data (Eisenschlos et al., 2019)) to improve the data (feature) sampling variability, as the group sampler used in group-robust algorithms over-sample minority classes with the same limited instances. Further on, we would like to investigate the use of zero-shot LWAN methods (Rios and Kavuluru, 2018; Chalkidis et al., 2020a), which currently harm averaged performance in favor of improved worst case performance.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*.

Cody Blakeney, Gentry Atkinson, Nathaniel Huish, Yan Yan, Vangelis Metris, and Ziliang Zong. 2021. Measure twice, cut once: Quantifying bias and fairness in deep neural networks.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020a. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020b. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online.

Yee Seng Chan and Hwee Tou Ng. 2006. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 89–96, Sydney, Australia. Association for Computational Linguistics.

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.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China. Association for Computational Linguistics.

João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4).

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Alistair EW Johnson, David J. Stone, Leo A. Celi, and Tom J. Pollard. 2017. MIMIC-III, a freely accessible critical care database. *Nature*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Rémi Le Priol, and Aaron C. Courville. 2020. Out-of-Distribution Generalization via Risk Extrapolation (REx). *CoRR*.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomás Kociský, Susannah Young, and Phil Blunsom. 2021. Pitfalls of static language modelling. *CoRR*, abs/2102.01951.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 394–402, Los Angeles, California. Association for Computational Linguistics.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res.*, 5:361–397.

Zhining Liu, Pengfei Wei, Jing Jiang, Wei Cao, Jiang Bian, and Yi Chang. 2020. MESA: boost ensemble imbalanced learning with meta-sampler. In *Advances in Neural Information Processing Systems*

33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6).

Eneldo Loza Mencia and Johannes Fürnkranzand. 2007. An Evaluation of Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain. In *Proceedings of the LWA 2007*, pages 126–132, Halle, Germany.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1101–1111.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. 2020. Gradient starvation: A learning proclivity in neural networks. *arXiv preprint arXiv:2011.09468*.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, Brussels, Belgium. Association for Computational Linguistics.

Timothy Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine Learning*, 88:157–208.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175.

Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Fairness-aware class imbalanced learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2045–2051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Baochen Sun and Kate Saenko. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham. Springer International Publishing.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 2021 Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

V. Vapnik. 1992. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification. In *Advances in Neural Information Processing Systems*, pages 5812–5822.

Algorithm	*Small* m-F_1	*Small* μ-F_1	*Medium* m-F_1	*Medium* μ-F_1
ERM	71.8	60.2	47.4	10.3
ERM+GS	71.7	62.4	47.5	12.6
Group Uniform	71.9	66.1	48.2	13.3
Group DRO	65.2	47.4	14.0	3.8
Deep CORAL	72.1	67.1	47.1	12.3
V-REx	71.9	65.9	47.6	11.3
IRM	72.0	66.6	**53.3**	**18.3**
Spectral Decoupling	**72.3**	**67.2**	53.1	16.1

Table 6: **Overall** results of the **group-robust algorithms** across all datasets.

A Additional Results

MIMIC-III dataset (Johnson et al., 2017) contains approx. 50k discharge summaries from US hospitals. Each summary is annotated with one or more codes (labels) from the ICD-9 hierarchy, which has 8 levels.[9]. The International Classification of Diseases, Ninth Revision (ICD-9) is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States and is maintained by the World Health Organization (WHO).

MIMIC-III has been anonymized to protect patients privacy, including chronological information (e.g., entry/discharge dates). We split the dataset randomly in training (30k), development (10k), test (10k) subsets. We use the 1st and 2nd level of ICD-9 including 19 and 184 categories, respectively. In Table 6, we present the results, which lead to the very same observations discussed for the rest of the datasets.

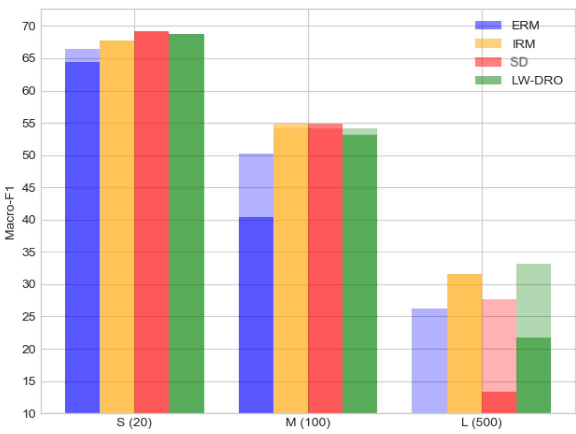

Figure 5: LWAN-BERT performance using ERM, IRM, Spectral Decoupling, and LW-DRO across all EUR-LEX settings. The shaded part denotes the performance improvement over standard BERT.

[9]www.who.int/classifications/icd/en/

B Alternative Combined Algorithm

Having a clear understanding of what IRM and Spectral Decoupling offer, it seems that we could combine both to leverage all features: (a) rely on group-wise (label-wise) losses as the main driver of the optimization process (Eq. 6); (b) penalize the classifier if there is a performance disparity between samples labeled with the same classes (Eq. 10–11); and (c) penalize the classifier for being over-confident (Eq. 14). We name the new algorithm Label-Wise Distributional Robust Optimization (LW-DRO), where the total loss term ($\mathcal{L}_{LW-DRO}$), is computed as follows:

$$\frac{1}{G}\left(\sum_{i=1}^{G} \mathcal{L}(g_i) + \lambda_1 P(g_i)\right) + \lambda_2 \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i^2 \qquad (18)$$

In Fig. 5, we present the results of the 3 overall best group-robust algorithms (IRM, Spectral Decoupling, and LW-DRO) across all EUR-LEX settings. LW-DRO has comparable perfomance in the first two setting (small, medium), while being the best in the large-sized setting.

C Measuring class-wise bias

Blakeney et al. (2021) recently introduced two evaluation measures to estimate class-wise bias of two models in comparison to one another in a multi-class setting, and show that these metrics can be also used to measure fairness and bias with respect to protected attributes.

Following Blakeney et al. (2021), in Figure 6 we present the normalized Combined Error Variance (CEV) in-between algorithms. CEV estimates the class-wise bias of a model A relative to another model B has increased of the change between model A and a random predictor.[10] In our case, as different models, we consider BERT trained with a different algorithm. In both UK-LEX and EUR-LEX, swapping Group Uniform, IRM, or Spectral Decoupling with ERM, or Group DRO leads to a higher class-wise bias, which is highly expected given the aforementioned performance analysis, i.e., improved m-F_1 scores.

[10]For a detailed analysis of the CEV metric, please refer to Blakeney et al. (2021).

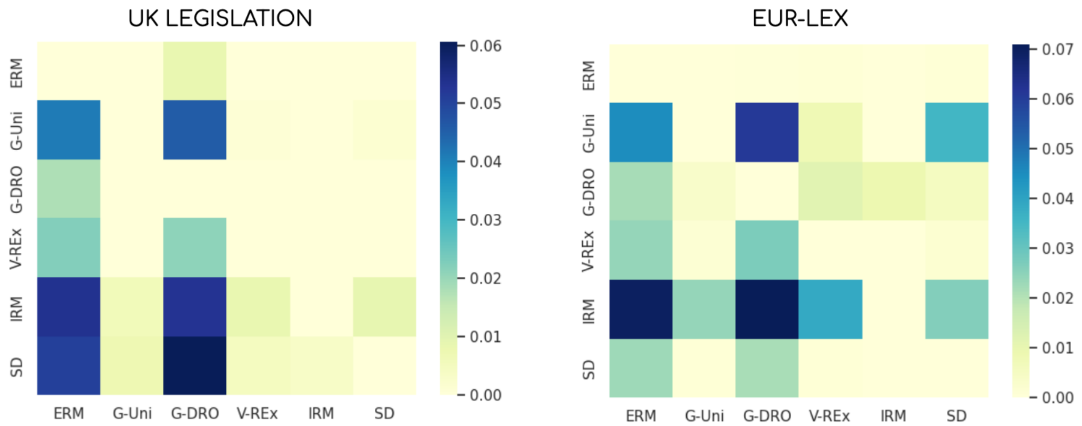

Figure 6: ***Class-wise bias*** in-between algorithms across datasets, measured with the normalized Combined Error Variance (CEV) as defined by Blakeney et al. (2021).