A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing

Sophie Henning[1,2] **William Beluch**[1] **Alexander Fraser**[2] **Annemarie Friedrich**[1]
[1] Bosch Center for Artificial Intelligence, Renningen, Germany
[2] Center for Information and Language Processing, LMU Munich, Germany
sophieelisabeth.henning|william.beluch@de.bosch.com
fraser@cis.lmu.de
annemarie.friedrich@de.bosch.com

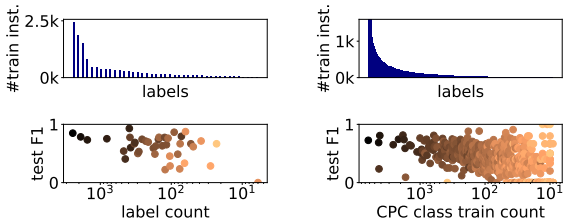
Abstract

Many natural language processing (NLP) tasks are naturally imbalanced, as some target categories occur much more frequently than others in the real world. In such scenarios, current NLP models tend to perform poorly on less frequent classes. Addressing class imbalance in NLP is an active research topic, yet, finding a good approach for a particular task and imbalance scenario is difficult.

In this survey, the first overview on class imbalance in deep-learning based NLP, we first discuss various types of controlled and real-world class imbalance. Our survey then covers approaches that have been explicitly proposed for class-imbalanced NLP tasks or, originating in the computer vision community, have been evaluated on them. We organize the methods by whether they are based on sampling, data augmentation, choice of loss function, staged learning, or model design. Finally, we discuss open problems and how to move forward.

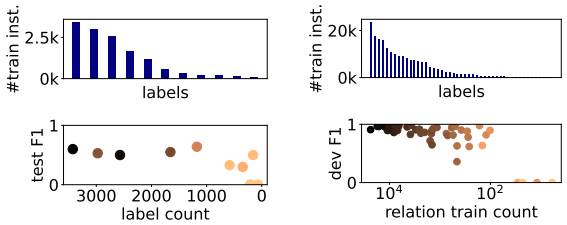
1 Introduction

Class imbalance is a major problem in natural language processing (NLP), because target category distributions are almost always skewed in NLP tasks. As illustrated by Figure 1, this often leads to poor performance on minority classes. Which categories matter is highly task-specific and may even depend on the intended downstream use. Developing methods that improve model performance in imbalanced data settings has been an active area for decades (e.g., Bruzzone and Serpico, 1997; Japkowicz et al., 2000; Estabrooks and Japkowicz, 2001; Park and Zhang, 2002; Tan, 2005), and is recently gaining momentum in the context of maturing neural approaches (e.g., Buda et al., 2018; Kang et al., 2020; Li et al., 2020; Yang et al., 2020; Jiang et al., 2021; Spangher et al., 2021). The problem is exacerbated when classes overlap in the feature space (Lin et al., 2019; Tian et al., 2020). For example, in patent classification, technical categories differ largely in frequency, and the concepts mentioned in the different categories can be very similar.



(a) Single-label **relation classification** on TACRED (Zhou and Chen, 2021)

(b) Hierarchical multi-label **patent classification** (Pujari et al., 2021)



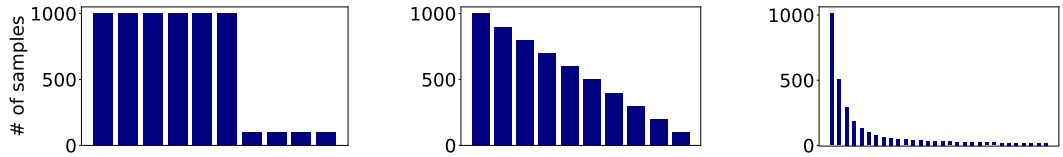
(c) Implicit **discourse relation** classification (PDTB) (Shi and Demberg, 2019)

(d) UD **dependency parsing** using RoBERTa on EWT (Grünewald et al., 2021)

Figure 1: **Class imbalance** has a negative effect on **performance** especially for minority classes in a variety of NLP tasks. Upper charts show label count distributions, lower part show test/dev F1 by training instance count (lighter colors indicate fewer test/dev instances). All models are based on transformers.

On a large variety of NLP tasks, transformer models such as BERT (Vaswani et al., 2017; Devlin et al., 2019) outperform both their neural predecessors and traditional models (Liu et al., 2019; Xie et al., 2020; Mathew et al., 2021). Performance for minority classes is also often higher when using self-supervised pre-trained models (e.g., Li and Scarton, 2020; Niklaus et al., 2021), which parallels findings from computer vision (Liu et al., 2022). However, the advent of BERT has not solved the class imbalance problem in NLP, as illustrated by Figure 1. Tänzer et al. (2022) find that on synthetically imbalanced named entity datasets with

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 523–540
May 2-6, 2023



(a) Step imbalance, $\mu = 0.4$, $\rho = 10$ (b) Linear imbalance, $\rho = 10$ (c) Long-tailed distribution

Figure 2: Instance counts per label follow different distributions: examples of **class imbalance types**.

majority classes having thousands of examples, at least 25 instances are required to predict a class at all, and 100 examples to learn to predict it with some accuracy.

Despite the relevance of class imbalance to NLP, related surveys only exist in the computer vision domain (Johnson and Khoshgoftaar, 2019b; Zhang et al., 2021b). Incorporating methods addressing class imbalance can lead to performance gains of up to 20%. Yet, NLP research often overlooks how important this is in practical applications, where minority classes may be of special interest.

Our contribution is to draw a clear landscape of approaches applicable to deep-learning (DL) based NLP. We set out with a problem definition (Sec. 2), and then organize approaches by whether they are based on sampling, data augmentation, choice of loss function, staged learning, or model design (Sec. 3). Our extensive survey finds that re-sampling, data augmentation, and changing the loss function can be relatively simple ways to increase performance in class-imbalanced settings and are thus straightforward choices for NLP practitioners.[1] While promising research directions, staged learning or model modifications often are implementation-wise and/or computationally costlier. Moreover, we discuss particular challenges of non-standard classification settings, e.g., imbalanced multi-label classification and catch-all classes, and provide useful connections to related computer vision work. Finally, we outline promising directions for future research (Sec. 4).

Scope of this survey. We focus on approaches evaluated on or developed for neural methods. Work from "traditional" NLP (e.g., Tomanek and Hahn, 2009; Li et al., 2011; Li and Nenkova, 2014; Kunchukuttan and Bhattacharyya, 2015) as well as Natural Language Generation (e.g., Nishino et al., 2020) and Automatic Speech Recognition (e.g., Winata et al., 2020; Deng et al., 2022) are not addressed in this survey. Other types of imbalances such as differently sized data sets of subtasks in continual learning (Ahrens et al., 2021) or imbalanced regression (Yang et al., 2021) are also beyond the scope of this survey. In Sec. 3.5, we briefly touch upon the related area of few-shot learning (Wang et al., 2020c).

Related surveys. We review imbalance-specific data augmentation approaches in Sec. 3.2. Feng et al. (2021) give a broader overview of data augmentation in NLP, Hedderich et al. (2021) provide an overview of low-resource NLP, and Ramponi and Plank (2020) discuss neural domain adaptation.

2 Problem Definition

Class imbalance refers to a classification setting in which one or multiple classes (**minority classes**) are considerably less frequent than others (**majority classes**). More concrete definitions, e.g., regarding the relative share up to which a class is seen as a minority class, depend on the task, dataset and labelset size. Much research focuses on improving all minority classes equally while maintaining or at least monitoring majority class performance (e.g., Huang et al., 2021; Yang et al., 2020; Spangher et al., 2021). We next discuss prototypical types of imbalance (Sec. 2.1) and then compare controlled and real-world settings (Sec. 2.2).

2.1 Types of Imbalance

To systematically investigate the effect of imbalance, Buda et al. (2018) define two prototypical types of label distributions, which we explain next.

Step imbalance is characterized by the fraction of minority classes, μ, and the size ratio between majority and minority classes, ρ. Larger ρ values indicate more imbalanced data sets. In prototypical step imbalance, if there are multiple minority classes, all of them are equally sized; if there are several majority classes, they also have equal size. Figure 2a shows a step-imbalanced distribution with 40% of the classes being minority classes

[1] We provide practical advice on identifying potentially applicable class imbalance methods in the Appendix (Figure 3).

and an imbalance ratio of $\rho = 10$. NLP datasets with a large catch-all class as they often arise in sequence tagging (see Sec. 2.2) or in relevance judgments in retrieval models frequently resemble step-imbalanced distributions. The ρ ratio has also been reported in NLP, e.g., by Li et al. (2020), although more task-specific imbalance measures have been proposed, e.g., for single-label text classification (Tian et al., 2020). In **linear imbalance**, class size grows linearly with imbalance ratio ρ (see Figure 2b), as, e.g., in the naturally imbalanced SICK dataset for natural language inference (Marelli et al., 2014).

Long-tailed label distributions (Figure 2c) are conceptually similar to linear imbalance. They contain many data points for a small number of classes (*head classes*), but only very few for the rest of the classes (*tail classes*). These distributions are common in computer vision tasks like instance segmentation (e.g., Gupta et al., 2019a), but also in multi-label text classification, for example with the goal of assigning clinical codes (Mullenbach et al., 2018), patent categories (Pujari et al., 2021), or news and research topics (Huang et al., 2021).

2.2 Controlled vs. Real-World Class Imbalance

Most real-world label distributions in NLP tasks do not perfectly match the prototypical distributions proposed by Buda et al. (2018). Yet, awareness of these settings helps practitioners to select appropriate methods for their data set or problem by comparing distribution plots. Using synthetically imbalanced data sets, researchers can control for more experimental factors and investigate several scenarios at once. However, evaluating on naturally imbalanced data provides evidence of a method's real-world effectiveness. Some recent studies combine both types of evaluation (e.g., Tian et al., 2021; Subramanian et al., 2021; Jang et al., 2021).

Many NLP tasks require treating a large, often heterogenous **catch-all class** that contains all instances that are not of interest to the task, while the remaining (minority) classes are approximately same-sized. Examples include the "Outside" label in IOB sequence tagging, or tweets that mention products in contexts that are irrelevant to the annotated categories (Adel et al., 2017). Such real-world settings often roughly follow a step imbalance distribution, with the additional difficulty of the catch-all class.

2.3 Evaluation

As *accuracy* and *micro*-averages mostly reflect majority class performance, choosing a good evaluation setting and metric is non-trivial. It is also highly task-dependent: in many NLP tasks, recognizing one or all minority classes well is at least equally important as majority class performance. For instance, non-hateful tweets are much more frequent in Twitter (Waseem and Hovy, 2016), but recognizing hateful content is the key motivation of hate speech detection. Which classes matter may even depend on downstream considerations, i.e., the same named entity tagger might be used in one application where a majority class matters, and another where minority classes matter more. Several evaluation metrics exist that have been designed to account for class-imbalanced settings, but no de facto standard exists. For example, *balanced accuracy* (Brodersen et al., 2010) corresponds to the average of per-class recall scores. It is often useful to record performance on *all* classes and to report *macro*-averages, which treat all classes equally.

3 Methods for Addressing Class Imbalance in NLP

In this section, we survey methods that either have been explicitly proposed to address class-imbalance issues in NLP or that have been empirically shown to be applicable for NLP problems. We provide an overview of which methods are applicable to a selection of NLP tasks in Appendix A.

3.1 Re-Sampling

To increase the importance of minority instances in training, the label distribution can be changed by various sampling strategies. Sampling can either be executed once or repeatedly during training (Pouyanfar et al., 2018). In *random oversampling* (**ROS**), a random choice of minority instances are duplicated, whereas in *random undersampling* (**RUS**), a random choice of majority instances are removed from the dataset. ROS can lead to overfitting and increases training times. RUS, however, discards potentially valuable data, but has been shown to work well in language-modeling objectives (Mikolov et al., 2013).

When applied in DL, ROS outperforms RUS both in synthetic step and linear imbalance (Buda et al., 2018) and in binary and multi-class English and Korean text classification (Juuti et al., 2020; Akhbardeh et al., 2021; Jang et al., 2021). More

flexible variants, e.g., re-sampling only a tunable share of classes (Tepper et al., 2020) or interpolating between the (imbalanced) data distribution and an almost perfectly balanced distribution (Arivazhagan et al., 2019), can also further improve results. *Class-aware sampling* (**CAS**, Shen et al., 2016), also referred to as *class-balanced sampling*, first chooses a class, and then an instance from this class. Performance-based re-sampling during training, following the idea of Pouyanfar et al. (2018), works well in multi-class text classification (Akhbardeh et al., 2021).

Issues in multi-label classification. In multi-label classification, label dependencies between majority and minority classes complicate sampling approaches, as over-sampling an instance with a minority label may simultaneously amplify the majority class count (Charte et al., 2015; Huang et al., 2021). CAS also suffers from this issue, and additionally introduces within-class imbalance, as instances of one class are selected with different probabilities depending on the co-assigned labels (Wu et al., 2020). Effective sampling in such settings is still an open issue. Existing approaches monitor the class distributions during sampling (Charte et al., 2015) or assign instance-based sampling probabilities (Gupta et al., 2019b; Wu et al., 2020).

3.2 Data Augmentation

Increasing the amount of minority class data during corpus construction, e.g., by writing additional examples or selecting examples to be labeled using Active Learning, can mitigate the class imbalance problem to some extent (Cho et al., 2020; Ein-Dor et al., 2020). However, this is particularly laborious in naturally imbalanced settings as it may require finding "the needle in the haystack," or may lead to biased minority class examples, e.g., due to collection via keyword queries. Synthetically generating additional minority instances thus is a promising direction. In this section, we survey data augmentation methods that have been explicitly proposed to mitigate class imbalance and that have been evaluated in combination with DL.

Text augmentation generates new natural language instances of minority classes, ranging from simple string-based manipulations such as synonym replacements to Transformer-based generation. *Easy Data Augmentation* (**EDA**, Wei and Zou, 2019), which uses dictionary-based synonym replacements, random insertion, random swap, and random deletion, has been shown to work well in class-imbalanced settings (Jiang et al., 2021; Jang et al., 2021; Juuti et al., 2020). Juuti et al. (2020) generate new minority class instances for English binary text classification using EDA and embedding-based synonym replacements, and by adding a random majority class sentence to a minority class document. They also prompt the pre-trained language model GPT-2 (Radford et al., 2019) with a minority class instance to generate new minority class samples. Tepper et al. (2020) evaluate generation with GPT-2 on English multi-class text classification datasets, coupled with a flexible balancing policy (see Sec. 3.1).

Similarly, Gaspers et al. (2020) combine machine-translation based text augmentation with dataset balancing to build a multi-task model. Both the main and auxiliary tasks are German intent classification. Only the training data for the latter is balanced and enriched with synthetic minority instances. In a long-tailed multi-label setting, Zhang et al. (2022) learn an attention-based text augmentation that augments instances with text segments that are relevant to tail classes, leading to small improvements. In general, transferring methods such as EDA or backtranslation to multi-label settings is difficult (Zhang et al., 2022, 2020; Tang et al., 2020).

Hidden space augmentation generates new instance vectors that are not directly associated with a particular natural language string, leveraging the representations of real examples. Using representation-based augmentations to tackle class imbalance is not tied to DL. SMOTE (Chawla et al., 2002), which interpolates minority instances with randomly chosen examples from their K-nearest neighbours, is popular in traditional machine learning (Fernández et al., 2018), but leads to mixed results in DL-based NLP (Ek and Ghanimifard, 2019; Tran and Litman, 2021; Wei et al., 2022). Inspired by CutMix (Yun et al., 2019), which cuts and pastes a single pixel region in an image, **TextCut** (Jiang et al., 2021) randomly replaces small parts of the BERT representation of one instance with those of the other. In binary and multi-class text classification experiments, TextCut improves over non-augmented BERT and EDA.

Good-enough example extrapolation (**GE3**, Wei, 2021) and **REPRINT** (Wei et al., 2022) also operate in the original representation space. To synthesize a new minority instance, GE3 adds the vec-

tor representing the difference between a majority class instance and the centroid of the respective majority class to the mean of a minority class. Evaluations on synthetically step-imbalanced English multi-class text classification datasets show improvements over oversampling and hidden space augmentation baselines. GE3 assumes that the distribution of data points of a class around its mean can be extrapolated to other classes, an assumption potentially hurting performance if the minority class distribution differs. To account for this when subtracting out majority characteristics, REPRINT performs a principal component analysis (PCA) for each class, leveraging the information on relevant dimensions during sample generation. This method usually outperforms GE3, with the cost of an additional hyperparameter (subspace dimensionality).

MISO (Tian et al., 2021) generates new instances by transforming the representations of minority class instances that are located nearby majority class instances. They learn a mapping from minority instance vectors to "disentangled" representations, making use of mutual information estimators (Belghazi et al., 2018) to push these representations away from the majority class and closer to the minority class. An adversarially-trained generator then generates minority instances using these disentangled representations. Tian et al. apply MISO in naturally and synthetically imbalanced English and Chinese binary and multi-class text classification with a single minority class.

ECRT (Chen et al., 2021) learns to map encoder representations (feature space) to a new space (source space) whose components are independent of each other given the class, assuming an invariant causal mechanism from source to feature space. The independence enables them to generate new meaningful minority examples by permuting or sampling components in the source space, resulting in medium improvements on a large multi-label text classification dataset with many labels.

Further related work exists in the area of **transfer learning** (Ruder et al., 2019), e.g., from additional datasets that provide complementary information on minority classes. For instance, Spangher et al. (2021) achieve small gains by manually selecting auxiliary datasets to improve imbalanced sentence-based discourse classification. However, complementary datasets have to be retrieved for each application, and task loss coefficients have to be tuned. Adapting methods to predict useful transfer sources (Lange et al., 2021) might help alleviate these problems.

3.3 Loss Functions

Standard *cross-entropy loss* (**CE**) is composed from the predictions for instances that carry the label in the gold standard, which is why the resulting classifiers fit the minority classes less well. In this section, we summarize loss functions designed for imbalanced scenarios. They either re-weight instances by class membership or prediction difficulty, or explicitly model class margins to change the decision boundary. Throughout this section, we use the variables and terms as shown in Table 1.

Losses for Single-Label Scenarios. *Weighted cross-entropy* (**WCE**) uses class-specific weights α_j that are tuned as hyperparameters or set to the inverse class frequency (e.g., Adel et al., 2017; Tayyar Madabushi et al., 2019; Li and Xiao, 2020). While WCE treats all instances of one class in the same way, *focal loss* (**FL**, Lin et al., 2017) down-weights instances for which the model is already confident (implemented with the $(1-p_j)^\beta$ coefficient). FL can of course also be used with class weights. Instead of mimicking accuracy like CE, *dice loss* (**Dice**, Milletari et al., 2016) tries to capture class-wise F1 score, with the predicted probability p_j proxying precision and the ground truth indicator y_j proxying recall. *Self-adjusting dice loss* (**ADL**, Li et al., 2020) combines confidence-based down-weighting via $1-p_j$ with Dice loss. For sequence labeling, QA and matching on English and Chinese datasets, Dice performs better than FL and ADL.

Rather than re-weighting instances, *label-distribution-aware margin loss* (**LDAM**, Cao et al., 2019), essentially a smooth hinge loss with label-dependent margins Δ_j, aims to increase the distance of the minority class instances to the decision boundary with the aim of better generalization for these classes. Cao et al.'s evaluation largely focuses on computer vision, but they also report results for LDAM on a synthetically imbalanced version of the IMDB review dataset (Maas et al., 2011), achieving a much lower error on the minority class than vanilla CE or CE with re-sampling or re-weighting. Subramanian et al. (2021) propose LDAM variants that consider bias related to socially salient groups (e.g., gender-based bias) in addition to class imbalance, evaluating them on binary text classification.

Single-label	**CE**	$-\sum_{j=1}^{C} y_j \log p_j$	
	WCE	$-\sum_{j=1}^{C} \alpha_j y_j \log p_j$	
	FL	$-\sum_{j=1}^{C} y_j (1-p_j)^{\beta} \log p_j$	
	Dice	$\sum_{j=1}^{C} 1 - \frac{2p_j y_j + \gamma}{p_j^2 + y_j^2 + \gamma}$	
	ADL	$\sum_{j=1}^{C} 1 - \frac{2(1-p_j)p_j y_j + \gamma}{(1-p_j)p_j + y_j + \gamma}$	
	LDAM	$-\sum_{j=1}^{C} y_j \log \frac{\exp(z_j - \Delta_j)}{\exp(z_j - \Delta_j) + \sum_{l \neq j} \exp(z_l)}$	with $\Delta_j = K / n_j^{1/4}$
	RL	$\mathbb{1}(gt \neq A) \log(1 + \exp(\rho(m^+ - z_{gt}))) + \log(1 + \exp(\rho(m^- - z_{c^-})))$	
Multi-label	**BCE**	$-\sum_{j=1}^{C} [y_j \log p_j + (1-y_j) \log(1-p_j)]$	
	WBCE	$-\sum_{j=1}^{C} \alpha_j [y_j \log p_j + (1-y_j) \log(1-p_j)]$	
	FL	$-\sum_{j=1}^{C} [y_j (1-p_j)^{\beta} \log p_j + (1-y_j) p_j^{\beta} \log(1-p_j)]$	
	DB	$-\sum_{j=1}^{C} [y_j \hat{\alpha}_j (1-q_j)^{\beta} \log q_j + (1-y_j) \hat{\alpha}_j \frac{1}{\lambda} q_j^{\beta} \log(1-q_j)]$	with $q_j = y_j \sigma(z_j - v_j) + (1-y_j)\sigma(\lambda(z_j - v_j))$

C	number of classes
y	target vector
p	model prediction vector
α	class weights
β	tunable focusing parameter
z	model logits vector
γ	smoothing constant
n_j	size of class j
K	label-independent constant
gt	index of ground-truth class
m^+	margin to correct class
m^-	... to most competitive incorrect class
A	special catch-all class
c^-	index of largest non-gt logit
λ	scaling factor
$\hat{\alpha}_j$	instance-specific class weights
v_j	class-specific bias

Table 1: Overview of **loss functions** formulated for one instance. See Appendix A for references/implementations.

In settings with a large artificial and potentially heterogeneous **catch-all class** (see Sec. 2.2), many areas of the space contain representations of the catch-all class. Here, vanilla LDAM might be an appropriate loss function as it encourages larger margins for minority classes. In such cases, *ranking losses* (**RL**) can also be effective to incentivize the model to only pay attention to "real" classes. On an imbalanced English multi-class dataset with a large catch-all class, Adel et al. (2017) find a ranking loss introduced by dos Santos et al. (2015) improves over CE and WCE. For minority classes, this loss function maximizes the score of the correct label z_{gt} while at the same time minimizing the score of the highest-scoring incorrect label z_{c^-}. For the catch-all class A, only z_{c^-} is minimized; z_{gt} is ignored. Similarly, Hu et al. (2022) apply class weights only to non-catch-all classes.

Losses for Multi-Label Scenarios. In multi-label classification, each label assignment can be viewed as a binary decision, hence *binary cross-entropy* (**BCE**) is often used here. Under imbalance, two issues arise. First, although class-specific weights have been used with BCE (e.g., Yang et al., 2020), their effect on minority classes is less clear than in the single-label case. For each instance, all classes contribute to BCE, with the labels *not* assigned to the instance (called *negative classes*) included via $(1-y_j)\log(1-p_j)$. Thus, if *weighted binary cross-entropy* (**WBCE**) uses a high weight for a class, it also increases the importance of negative instances for a minority class, which may further encourage the model to *not* predict this minority class.

To leverage class weights more effectively in BCE, one option is to only apply them to the loss of positive instances as proposed for multi-label image classification (Kumar et al., 2018). Related work includes uniformly upweighting positive instances of *all* classes in hierarchical multi-label text classification (e.g., Rathnayaka et al., 2019). An approach to multi-label emotion classification by Yilmaz et al. (2021) performs training time balancing by adapting FL such that for a given mini-batch the loss over all instances in this mini-batch has exactly the same value for every class.

If a classifier already correctly predicts a negative class for an instance, the loss can be further decreased by reducing the respective label's logits. In CE, due to the softmax that uses the logits of all classes, the impact of this effect becomes minor once the logit for the correct class is much larger than those of the other classes. However, the problem is more severe in BCE (Wu et al., 2020), as logits are treated independently. As minority labels mostly occur as negative classes, this logit suppression leads to a bias in the decision boundary, making it less likely for minority classes to be predicted. To tackle this issue and based on a multi-label version of FL, Wu et al. (2020) propose *distribution-balanced loss* (**DB**) for object detection, adding *Negative Tolerant Regularization* for the loss for negative classes by transforming the logits of positive and negative classes differently (see q_j in Table 1). This regularization imposes a sharp drop in the loss function for negative classes once the respective logit is below a threshold. Moreover, DB introduces instance-specific class weights $\hat{\alpha}$ to account for imbalances caused by class-aware sampling (see Sec. 3.1) in multi-label scenarios. These weights reflect the frequency of a class and the quantity and frequency of the positive labels of the instance. Huang et al. (2021) have shown large

improvements of DB over BCE even when using uniform sampling on two long-tailed multi-label English text classification datasets.

While Cao et al. (2019) propose and theoretically justify LDAM for single-label classification only, it has been successfully applied to multi-label text classification as well (Biswas et al., 2021). Ferreira and Vlachos (2019) show that applying a cross-label dependency loss (Yeh et al., 2017; Zhang and Zhou, 2006) can be helpful for multi-label stance classification. Similarly, Lin et al. (2019) introduce a label-confusion aware cost factor into their loss function. The adaptive loss of Suresh and Ong (2021) integrates inter-label relationships into a contrastive loss (Khosla et al., 2020), which compares the score of a positive example with the distance to that of other positive and negative examples in order to push its representation closer to the correct class and further away from the wrong class(es). The resulting loss function learns how to increase the weight of confusable negative labels relative to other negative labels. Combining label-confusion aware loss functions with class weighting techniques is a promising research direction.

Re-Sampling vs. Loss Functions. Re-sampling and loss functions that are specifically designed for class-imbalanced settings are based on the same idea of increasing the importance of minority instances. Re-sampling is conceptually simpler and has a direct impact on training time, e.g., oversampling may cause a considerable increase. By contrast, the loss functions explained above are more flexible, e.g., by modeling desirable properties of margins, but also mostly harder to interpret.

3.4 Staged Learning

One approach to finding a good trade-off between learning features that are representative of the underlying data distribution and reducing the classifier's bias towards the majority class(es) is to perform the training in several stages. *Two-staged training* is common in imbalanced or data-scarce computer vision tasks (e.g., Wang et al., 2020b,a; Zhang et al., 2021a). The first stage usually performs standard training in order to train or fine-tune the feature extraction network. Later stages may freeze the feature extractor and re-train the classifier layers using special methods to address class imbalance, e.g., using more balanced data distributions or specific losses. For example, Cao et al. (2019) find their LDAM loss to be most effective when the training happens in two stages. In NLP, deep-learning models are usually based on pre-trained neural text encoders or word embeddings. Further domain-specific pre-training before starting the fine-tuning stage(s) can also be effective (Gururangan et al., 2020).

Several NLP approaches that fall under *staged learning* are directly inspired by computer vision research. In the context of long-tailed image classification, Kang et al. (2020) find that class-balanced sampling (see Sec. 3.1) helps when performing single-stage training, but that in their two-stage *classifier re-training* (**cRT**) method, using the original distribution in the first stage is more effective than class-balanced sampling. cRT employs the latter only in the second stage after freezing the representation weights. Yu et al. (2020) perform a similar decoupling analysis on long-tailed relation classification, essentially confirming Kang et al. (2020)'s results on this NLP task with respect to the re-sampling strategies. Additionally, they find that loss re-weighting under this analysis behaves similar to re-sampling, i.e., it leads to worse performance when applied during representation learning, but boosts performance when re-training the classifier. Hu et al. (2022) successfully leverage Kang et al.'s ideas for event detection, where both trigger detection and trigger classification suffer from class imbalance.

Jang et al. (2021) model imbalanced classification as a continual learning task with k stages where the data gradually becomes more balanced (*sequential targeting*, **ST**). The first stage contains the most imbalanced subset, and then the degree of imbalance decreases, with the last stage presenting the most balanced subset. The training objective encourages both good performance on the current stage and keeping information learnt in previous stages. Their experiments include binary and ternary English and Korean text classification. Active Learning (AL), which contains several stages by definition, has also been shown to boost performance of BERT models for minority classes (Ein-Dor et al., 2020). For a discussion about AL and deep learning, see Schröder and Niekler (2020).

3.5 Model Design

The methods described so far are largely independent of model architecture. This section describes model modifications that aim at improving performance in imbalanced settings.

Observing that the weight vectors for smaller classes have smaller norms in standard joint training compared to staged-learning based cRT (see Sec. 3.4), Kang et al. (2020) normalize the classifier weights directly in one-staged training using a hyperparameter τ to control the normalization "temperature" (τ**-norm**). τ-norm achieves similar or better performance than cRT in long-tailed image classification and outperforms cRT in relation extraction, but cRT works better for named entity recognition and event detection (Nan et al., 2021).

SetConv (Gao et al., 2020) and **ProtoBERT** (Tänzer et al., 2022) learn representatives for each class using *support sets* and classify an input (the *query*) based on its similarity to these representatives. SetConv applies convolution kernels that capture intra- and inter-class correlations to extract class representatives. ProtoBERT uses class centroids in a learned BERT-based feature space, treating the distance of any instance to the catch-all class as just another learnable parameter. At each training step, SetConv uses only one instance per class in the query set, but preserves the original class imbalance in the support set, whereas ProtoBERT uses fixed ratios. In the respective experimental studies, ProtoBERT performs better than using a standard classification layer on top of BERT for minority classes in NER if less than 100 examples are seen by the model, while SetConv excels in binary text classification with higher degrees of imbalance and in multi-class text classification.

The **HSCNN** model (Yang et al., 2020) uses class representatives only for the classification of tail classes, while head classes are assigned using a standard text CNN (Kim, 2014). HSCNN learns label-specific similarity functions, extracting instance representations from the pre-final layers of two copies of the original CNN, and assigns a tail class if the similarity to the class representative (computed as the mean of 5 random support instances) exceeds 0.5. On tail classes, HSCNN consistently improves over the vanilla CNN.

In addition, there exist a number of **task-specific solutions**. Prange et al. (2021) propose to construct CCG supertags from predicted tree structures rather than treating the problem as a standard classification task. In order to recognize implicit positive interpretations in negated statements in a class-imbalanced dataset, van Son et al. (2018) argue that leveraging information structure could be one way to improve inference. *Structural causal models* (**SCM**s) have been applied to imbalanced NLP tasks, encoding task-specific causal graphs (e.g., Nan et al., 2021). Similarly, Wu et al. (2021) causally model how bias in long-tailed corpora affects topic modeling (Blei et al., 2003) and use this to improve training of a variational autoencoder.

A research area closely related to class imbalance is **few-shot learning** (FSL, Wang et al., 2020c), which aims to learn classes based on only very few training examples. Model ideas from FSL can be leveraged for long-tailed settings, e.g., by making use of relational information about class labels in the form of knowledge graph embeddings or other forms of embedding hierarchical relationships between labels (Han et al., 2018; Zhang et al., 2019), or computing label-specific representations (Mullenbach et al., 2018).

4 Insights and Future Directions

We have provided a comprehensive, concise and structured overview of current approaches to dealing with class imbalance in DL-based NLP.

What works (best)? As there is no established benchmark for class-imbalanced settings, evaluation results are hard to compare across papers. In general, re-sampling or changing the loss function may lead to small to moderate gains. For data augmentation approaches, the reported performance increases tend to be larger than for re-sampling or new loss functions. The effects of staged training or modifications of the model vary drastically, ranging from detrimental to very large performance gains.

Hence, re-sampling, data augmentation, and changing the loss function are straightforward choices in class-imbalanced settings. Approaches based on staged learning or model design may sometimes outperform them, but often come with a higher implementation or computational cost. For a practical decision aid and potential application settings of some class imbalance methods, see Figure 3 in Appendix B and Table 3 in Appendix A.

How should we report results? Much NLP research only reports aggregate statistics (Harbecke et al., 2022), making it hard to judge the impact on improvements by class, which is often important in practice. We thus argue that NLP researchers should *always* report per-class statistics, e.g., as in Figure 1. Open-sourcing spreadsheets with the exact numbers would enable the community to compare systems more flexibly from multiple angles, i.e., with respect to whichever class(es) matter in

a particular application scenario, and to re-use this data in research on class imbalance. Reviewers should also value works that analyze performance for relevant minority classes rather than focusing largely only on overall accuracy improvements.

A main hindrance to making progress on class imbalance in computer vision and NLP alike is that experimental results are often hard to compare (Johnson and Khoshgoftaar, 2019a, 2020). A first important step would be to not restrict baselines to methods of the same type, e.g., a new data augmentation approach should not only compare to other data augmentation methods, but also to using loss functions for class imbalance. Establishing a shared and systematic benchmark of a diverse set of class-imbalanced NLP tasks would be highly beneficial for both researchers and practitioners.

How can we move forward? Most work on class-imbalanced NLP has focused on single-label text classification. Finding good solutions for multi-label settings is still an open research challenge. Class imbalance also poses problems in NLP tasks such as sequence labeling or parsing, and we believe that the interaction of structured prediction models with methods to address class imbalance is a promising area for future research. Moreover, we need to study how class imbalance methods affect prediction calibration in order to provide reliable confidence estimates.

In general, inspiration for new model architectures could for example be drawn from approaches developed for few-shot learning (Wang et al., 2020c). Recently, prompting (Radford et al., 2019) has emerged as a new paradigm in NLP, which performs strongly in real-world few-shot settings (Schick and Schütze, 2022). Methods that improve worst-case performance under distribution shift (e.g., Sagawa et al., 2020) might also be applied to improve minority-class performance.

Acknowledgements

We thank the anonymous reviewers for their valuable comments, and Heike Adel, Stefan Grünewald, Subhash Pujari, and Timo Schrader for providing data for our teaser image. We thank them and Talita Anthonio, Mohamed Gad-Elrab, Lukas Lange, Stefan Ott, Robert Schmier, Hendrik Schuff, Daria Stepanova, Jannik Strötgen, Thang Vu, and Dan Zhang for helpful discussions and feedback on the writing. We also thank Jason Wei and Jiaqi Zeng for answering questions about their work.

Limitations

This paper is a survey, structuring, organizing and describing works and concepts to address *class imbalance* including *long-tailed learning*. While we touch upon *data augmentation* and *few-shot learning*, we do not comprehensively review those areas. Details on the scope of this review have also been described in Sec. 1.

The search process for the survey included searching for the keywords *class imbalance* and *long tail* in Google Scholar and the ACL Anthology, as well as carefully checking the papers that cite relevant papers.

Finally, the paper only constitutes a literature review, it does not yet provide a comprehensive empirical study which is much needed in this research area, but it will be of use in carrying out such a study.

References

Heike Adel, Francine Chen, and Yan-Ying Chen. 2017. Ranking convolutional recurrent neural networks for purchase stage identification on imbalanced Twitter data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 592–598, Valencia, Spain. Association for Computational Linguistics.

Kyra Ahrens, Fares Abawi, and Stefan Wermter. 2021. Drill: Dynamic representations for imbalanced lifelong learning. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 409–420, Cham. Springer International Publishing.

Farhad Akhbardeh, Cecilia Ovesdotter Alm, Marcos Zampieri, and Travis Desell. 2021. Handling extreme class imbalance in technical logbook datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4034–4045, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the 35th*

International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR.

Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. 2021. Transicd: Transformer based code-wise attention model for explainable ICD coding. In *Artificial Intelligence in Medicine - 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15-18, 2021, Proceedings*, volume 12721 of *Lecture Notes in Computer Science*, pages 469–478. Springer.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124.

L. Bruzzone and S.B. Serpico. 1997. Classification of imbalanced remote-sensing data by neural networks. *Pattern Recognition Letters*, 18(11):1323–1328.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems Progress in Intelligent Systems Mining Humanistic Data.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Junya Chen, Zidi Xiu, Benjamin Goldstein, Ricardo Henao, Lawrence Carin, and Chenyang Tao. 2021. Supercharging imbalanced data learning with energy-based contrastive representation transfer. *Advances in Neural Information Processing Systems*, 34.

Won Ik Cho, Youngki Moon, Sangwhan Moon, Seok Min Kim, and Nam Soo Kim. 2020. Machines getting with the program: Understanding intent arguments of non-canonical directives. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 329–339, Online. Association for Computational Linguistics.

Keqi Deng, Gaofeng Cheng, Runyan Yang, and Yonghong Yan. 2022. Alleviating asr long-tailed problem by decoupling the learning of representation and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:340–354.

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.

Cícero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 626–634, Beijing, China. Association for Computational Linguistics.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Adam Ek and Mehdi Ghanimifard. 2019. Synthetic propaganda embeddings to train a linear projection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 155–161, Hong Kong, China. Association for Computational Linguistics.

Andrew Estabrooks and Nathalie Japkowicz. 2001. A mixture-of-experts framework for text classification. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. 2018. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Researc*, 61:863–905.

William Ferreira and Andreas Vlachos. 2019. Incorporating label dependencies in multilabel stance detection. 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 6350–6354, Hong Kong, China. Association for Computational Linguistics.

Yang Gao, Yi-Fan Li, Yu Lin, Charu Aggarwal, and Latifur Khan. 2020. SetConv: A New Approach for Learning from Imbalanced Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1284–1294, Online. Association for Computational Linguistics.

Judith Gaspers, Quynh Do, and Fabian Triefenbach. 2020. Data Balancing for Boosting Performance of Low-Frequency Classes in Spoken Language Understanding. In *Proc. Interspeech 2020*, pages 1560–1564.

Stefan Grünewald, Annemarie Friedrich, and Jonas Kuhn. 2021. Applying occam's razor to transformer-based dependency parsing: What works, what doesn't, and what is really necessary. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 131–144, Online. Association for Computational Linguistics.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019a. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. 2019b. LVIS: A dataset for large vocabulary instance segmentation. *CoRR*, abs/1908.03195.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245, Brussels, Belgium. Association for Computational Linguistics.

David Harbecke, Yuxuan Chen, Leonhard Hennig, and Christoph Alt. 2022. Why only micro-f1? class weighting of measures for relation classification. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 32–41, Dublin, Ireland. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Bo Hu, Yun Liu, Naiyue Chen, Lifu Wang, Ning Liu, and Xing Cao. 2022. Segcn-dcr: A syntax-enhanced event detection framework with decoupled classification rebalance. *Neurocomputing*, 481:55–66.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Jang, Yoonjeon Kim, Kyoungho Choi, and Sungho Suh. 2021. Sequential targeting: A continual learning approach for data imbalance in text classification. *Expert Systems with Applications*, 179:115067.

Nathalie Japkowicz et al. 2000. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. AAAI Press Menlo Park, CA.

Wanrong Jiang, Ya Chen, Hao Fu, and Guiquan Liu. 2021. Textcut: A multi-region replacement data augmentation approach for text imbalance classification. In *Neural Information Processing*, pages 427–439, Cham. Springer International Publishing.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019a. Deep learning and thresholding with class-imbalanced big data. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 755–762.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019b. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1).

Justin M. Johnson and Taghi M. Khoshgoftaar. 2020. The Effects of Data Sampling with Deep Learning and Highly Imbalanced Big Data. *Information Systems Frontiers*, 22(5):1113–1131.

Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N. Asokan. 2020. A little goes a long way: Improving toxic language classification despite data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2991–3009, Online. Association for Computational Linguistics.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation

and classifier for long-tailed recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Pulkit Kumar, Monika Grewal, and Muktabh Mayank Srivastava. 2018. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In *International conference image analysis and recognition*, pages 546–552. Springer.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2015. Addressing class imbalance in grammatical error detection with evaluation metric optimization. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 2–10, Trivandrum, India. NLP Association of India.

Lukas Lange, Jannik Strötgen, Heike Adel, and Dietrich Klakow. 2021. To share or not to share: Predicting sets of sources for model transfer learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8744–8753, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinfen Li and Lu Xiao. 2020. syrapropa at SemEval-2020 task 11: BERT-based models design for propagandistic technique and span detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1808–1816, Barcelona (online). International Committee for Computational Linguistics.

Junyi Jessy Li and Ani Nenkova. 2014. Addressing class imbalance for improved recognition of implicit discourse relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 142–150, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Shoushan Li, Guodong Zhou, Zhongqing Wang, Sophia Yat Mei Lee, and Rangyang Wang. 2011. Imbalanced sentiment classification. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 2469–2472, New York, NY, USA. Association for Computing Machinery.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Yue Li and Carolina Scarton. 2020. Revisiting rumour stance classification: Dealing with imbalanced data. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 38–44, Barcelona, Spain (Online). Association for Computational Linguistics.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Cost-sensitive regularization for label confusion-aware event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5278–5283, Florence, Italy. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. 2022. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*.

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. *CoRR*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571.

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, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. 2020. Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2223–2236, Online. Association for Computational Linguistics.

Seong-Bae Park and Byoung-Tak Zhang. 2002. A boosted maximum entropy model for learning text chunking. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 482–489.

Samira Pouyanfar, Yudong Tao, Anup Mohan, Haiman Tian, Ahmed S. Kaseb, Kent Gauen, Ryan Dailey, Sarah Aghajanzadeh, Yung-Hsiang Lu, Shu-Ching Chen, and Mei-Ling Shyu. 2018. Dynamic sampling in convolutional neural networks for imbalanced data classification. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 112–117.

Jakob Prange, Nathan Schneider, and Vivek Srikumar. 2021. Supertagging the long tail with tree-structured decoding of complex categories. *Transactions of the Association for Computational Linguistics*, 9:243–260.

Subhash Chandra Pujari, Annemarie Friedrich, and Jannik Strötgen. 2021. A multi-task approach to neural multi-label hierarchical patent classification using transformers. In *European Conference on Information Retrieval*, pages 513–528. Springer.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Prabod Rathnayaka, Supun Abeysinghe, Chamod Samarajeewa, Isura Manchanayake, Malaka J. Walpola, Rashmika Nawaratne, Tharindu R. Bandaragoda, and Damminda Alahakoon. 2019. Gated recurrent neural network approach for multilabel emotion detection in microblogs. *CoRR*, abs/1907.07653.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *CoRR*, abs/2008.07267.

Li Shen, Zhouchen Lin, and Qingming Huang. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision – ECCV 2016*, pages 467–482, Cham. Springer International Publishing.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. 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 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 498–517, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Songbo Tan. 2005. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, 28(4):667–671.

Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan. 2020. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8:193248–193256.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China. Association for Computational Linguistics.

Naama Tepper, Esther Goldbraich, Naama Zwerdling, George Kour, Ateret Anaby Tavor, and Boaz Carmeli. 2020. Balancing via generation for multi-class text classification improvement. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1440–1452, Online. Association for Computational Linguistics.

Jiachen Tian, Shizhan Chen, Xiaowang Zhang, and Zhiyong Feng. 2020. A graph-based measurement for text imbalance classification. In *ECAI 2020*, pages 2188–2195. IOS Press.

Jiachen Tian, Shizhan Chen, Xiaowang Zhang, Zhiyong Feng, Deyi Xiong, Shaojuan Wu, and Chunliu Dou. 2021. Re-embedding difficult samples via mutual information constrained semantically oversampling for imbalanced text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3148–3161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katrin Tomanek and Udo Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, page 105–112, New York, NY, USA. Association for Computing Machinery.

Nhat Tran and Diane Litman. 2021. Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*, pages 148–153, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chantal van Son, Roser Morante, Lora Aroyo, and Piek Vossen. 2018. Scoring and classifying implicit positive interpretations: A challenge of class imbalance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2253–2264, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. 2020a. The devil is in classification: A simple framework for long-tail instance segmentation. In *Computer Vision – ECCV 2020*, pages 728–744, Cham. Springer International Publishing.

Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. 2020b. Frustratingly simple few-shot object detection. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9919–9928. PMLR.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020c. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jason Wei. 2021. Good-enough example extrapolation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5923–5929, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Jiale Wei, Qiyuan Chen, Pai Peng, Benjamin Guedj, and Le Li. 2022. Reprint: a randomized extrapolation based on principal components for data augmentation. *CoRR*, abs/2204.12024.

Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven C. H. Hoi. 2020. Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition. *CoRR*, abs/2012.01687.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision – ECCV 2020*, pages 162–178, Cham. Springer International Publishing.

Xiaobao Wu, Chunping Li, and Yishu Miao. 2021. Discovering topics in long-tailed corpora with causal intervention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 175–185, Online. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Wenshuo Yang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2020. HSCNN: A hybrid-Siamese convolutional neural network for extremely imbalanced multi-label text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6716–6722, Online. Association for Computational Linguistics.

Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. 2021. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11842–11851. PMLR.

Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. 2017. Learning deep latent space for multi-label classification. In *Thirty-first AAAI conference on artificial intelligence*.

Selim F. Yilmaz, E. Batuhan Kaynak, Aykut Koç, Hamdi Dibeklioğlu, and Suleyman Serdar Kozat. 2021. Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Haiyang Yu, Ningyu Zhang, Shumin Deng, Zonggang Yuan, Yantao Jia, and Huajun Chen. 2020. The devil is the classifier: Investigating long tail relation classification with decoupling analysis. *CoRR*, abs/2009.07022.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020. On data augmentation for extreme multi-label classification. *CoRR*, abs/2009.10778.

Jiaxin Zhang, Jie Liu, Shaowei Chen, Shaoxin Lin, Bingquan Wang, and Shanpeng Wang. 2022. Adam: An attentional data augmentation method for extreme multi-label text classification. In *Advances in Knowledge Discovery and Data Mining*, pages 131–142, Cham. Springer International Publishing.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. 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 3016–3025, Minneapolis, Minnesota. Association for Computational Linguistics.

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. 2021a. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2361–2370.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2021b. Deep long-tailed learning: A survey. *CoRR*, abs/2110.04596.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *CoRR*, abs/2102.01373.

Appendix

A Method Overview

We here provide details on a selection of methods surveyed in this paper. Table 3 shows whether they have been applied respectively whether they are applicable in binary, multi-class, and multi-label classification. Moreover, it contains information on whether authors open-sourced their implementation. For links to open-sourced code, see Table 2.

B Practical advice

In Figure 3, we provide practical advice which class imbalance methods might be beneficial under which circumstances. Due to the lack of an established benchmark, we can only give rough guidance.

Method	**Link**
Data Augmentation	
EDA (Wei and Zou, 2019)	GitHub
GE3 (Wei, 2021)	ACL Anthology
ECRT (Chen et al., 2021)	GitHub
Loss Functions	
FL (Lin et al., 2017)	GitHub
ADL (Li et al., 2020)	GitHub
LDAM (Cao et al., 2019)	GitHub
DB (Wu et al., 2020)	GitHub
Staged Learning	
cRT (Kang et al., 2020)	GitHub
ST (Jang et al., 2021)	GitHub
Model Design	
τ-norm (Kang et al., 2020)	GitHub
ProtoBERT (Tänzer et al., 2022)	GitHub

Table 2: **Open-sourced implementations** of examples of class imbalance methods.

Method	Binary classification	Multi-class classification	Multi-label classification	Code
Re-Sampling				
ROS/RUS (Sec. 3.1)	✓	✓	?	N/A
CAS (Shen et al., 2016)	✓	✓	?	×
Data Augmentation				
EDA (Wei and Zou, 2019)	Juuti et al. (2020) Jiang et al. (2021)	Jiang et al. (2021)	Zhang et al. (2022, 2020)	✓
TextCut (Jiang et al., 2021)	Jiang et al. (2021)	Jiang et al. (2021)	✓	×
GE3 (Wei, 2021)	✓	Wei (2021) Wei et al. (2022)	?	✓
MISO (Tian et al., 2021)	Tian et al. (2021)	Tian et al. (2021)*	?	×
ECRT (Chen et al., 2021)	✓	✓	Chen et al. (2021)	✓
Loss Functions				
WCE (Sec. 3.3)	Tayyar Madabushi et al. (2019)	Adel et al. (2017) Li and Xiao (2020)	N/A	N/A
FL (Lin et al., 2017)	✓	Li et al. (2020) Nan et al. (2021)	✓	✓
ADL (Li et al., 2020)	✓	Li et al. (2020) Spangher et al. (2021)	✓	✓
LDAM (Cao et al., 2019)	Cao et al. (2019) Subramanian et al. (2021)	✓	Biswas et al. (2021)	✓
WBCE (Sec. 3.3)	✓	×	Yang et al. (2020)	N/A
RL (dos Santos et al., 2015)	✓	Adel et al. (2017)	×	×
DB (Wu et al., 2020)	×	×	Huang et al. (2021)	✓
Staged Learning				
cRT (Kang et al., 2020)	✓	Nan et al. (2021) Hu et al. (2022)	✓	✓
ST (Jang et al., 2021)	Jang et al. (2021)	Jang et al. (2021)	✓	✓
Model Design				
τ-norm (Kang et al., 2020)	✓	Nan et al. (2021)	✓	✓
SetConv (Gao et al., 2020)	Gao et al. (2020)	Gao et al. (2020)	✓	×
ProtoBERT (Tänzer et al., 2022)	✓	Tänzer et al. (2022)	✓	✓
HSCNN (Yang et al., 2020)	✓	✓	Yang et al. (2020)	×

Table 3: **Examples of class imbalance methods and NLP application settings.** ✓: **method** applicable (but no particular reference reporting experimental results exists)/**code**: authors open-sourced their implementation, ?: application not straightforward / open research issues. *: The authors select only one class as the minority class in their experiments. For links to open-sourced code, see Table 2. Methods for binary and multi-class classification are in general applicable to classification-based **relation extraction** approaches; applying class-imbalance techniques to **sequence labeling** methods in general is similar to the case of multi-label classification. For example, if sampling for a particular category, the whole sequence sample may contain additional annotations for other categories.

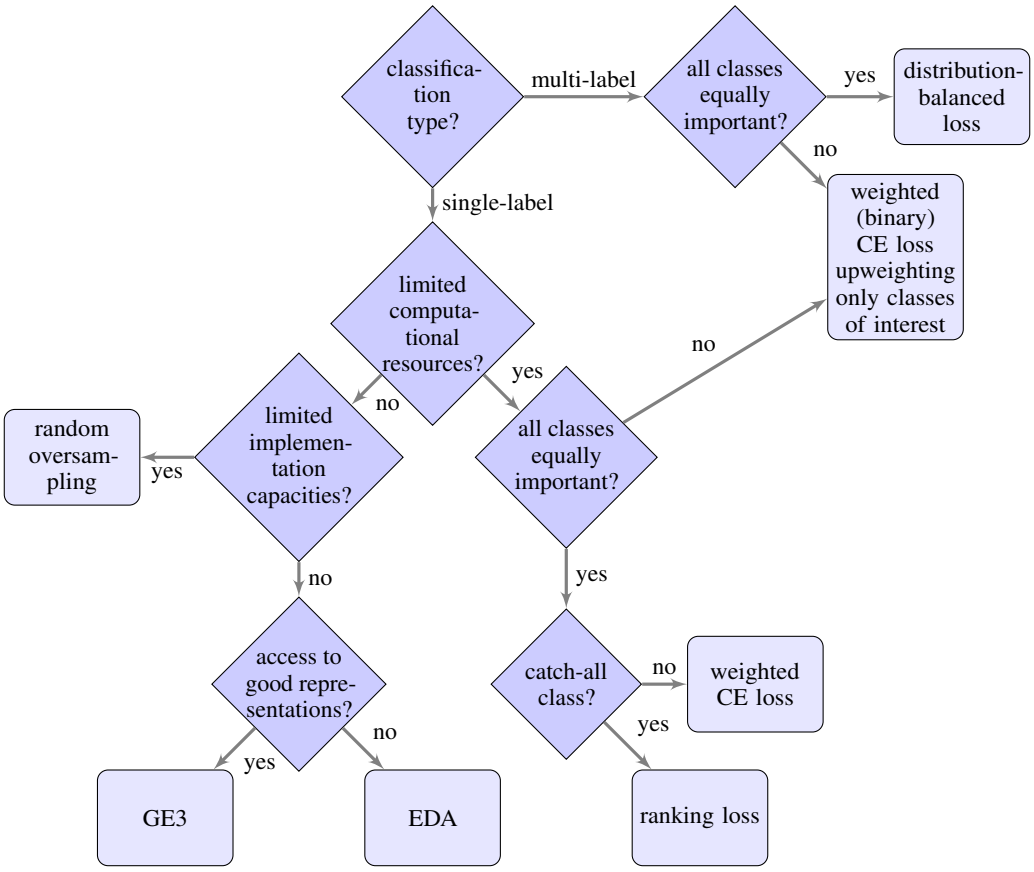


Figure 3: **Practical advice** which methods to try under which circumstances.