SEMSUP-XC: SEMANTIC SUPERVISION FOR EXTREME CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Extreme classification (XC) considers the scenario of predicting over a very large number of classes (thousands to millions), with real-world applications including serving search engine results, e-commerce product tagging, and news article classification. The zero-shot version of this task involves the addition of new categories at test time, requiring models to generalize to novel classes without additional training data (e.g. one may add a new class *"fidget spinner"* for e-commerce product tagging). In this paper, we develop SEMSUP-XC, a model that achieves state-of-the-art zero-shot (ZS) and few-shot (FS) performance on three extreme classification benchmarks spanning the domains of law, e-commerce, and Wikipedia. SEMSUP-XC builds upon the recently proposed framework of semantic supervision that uses semantic label descriptions to represent and generalize to classes (e.g., *"fidget spinner"* described as "A popular spinning toy intended as a stress reliever"). Specifically, we use a combination of contrastive learning, a hybrid lexico-semantic similarity module and automated description collection to train SEMSUP-XC efficiently over extremely large class spaces. SEMSUP-XC significantly outperforms baselines and state-of-the-art models on all three datasets, by up to 6-10 precision@1 points on zero-shot classification and >10 precision points on few-shot classification, with similar gains for recall@10 (3 for zero-shot and 2 for few-shot). Our ablation studies and qualitative analyses demonstrate the relative importance of our various improvements and show that SEMSUP-XC's automated pipeline offers a consistently efficient method for extreme classification.

1 INTRODUCTION

Extreme classification (XC) studies multi-class and multi-label classification problems with a large number of classes, ranging from thousands to millions (Bengio et al., 2019; Bhatia et al., 2015; Chang et al., 2019; Lin et al., 2014; Jiang et al., 2021). The paradigm has multiple real-world applications including movie and product recommendation, search-engines, and e-commerce product tagging. Moreover, in practical scenarios where XC is deployed, environments are constantly changing, with new classes with zero or few labeled examples being added. Recent work such as ZestXML (Gupta et al., 2021), MACLR (Xiong et al., 2022), LightXML (Jiang et al., 2021), and GROOV (Simig et al., 2022), has explored zero-shot and few-shot extreme classification (ZS-XC and FS-XC). These setups are challenging because of (1) the presence of a large number of fine-grained classes which are often not mutually exclusive, (2) limited or no labeled data per class, (3) and increased computational expense and model size because of the large label space. While the aforementioned works have tried to tackle the latter two issues, they have not attempted to build a semantically rich representation of classes for improved classification, using only class names to represent them.

A large fine-grained label space necessitates capturing the semantics of different attributes of classes. To this end, we leverage semantic supervision (SEMSUP) (Hanjie et al., 2022), a recently proposed framework that represents classes using diverse descriptions to better capture their semantics. This design choice allows SEMSUP to better generalize to novel classes by using corresponding descriptions, compared to standard classifiers. However, SEMSUP as designed in (Hanjie et al., 2022) cannot be naively applied to XC due to several reasons: (1) SEMSUP performs full cross-entropy learning which is computationally intractable for large label spaces (2) it uses only semantic similarity between the instance and label description to measure compatibility, thus ignoring lexically similar common terms like "*soccer*" and "*football*". (3) it uses a semi-automatic pipeline for collecting label descriptions for

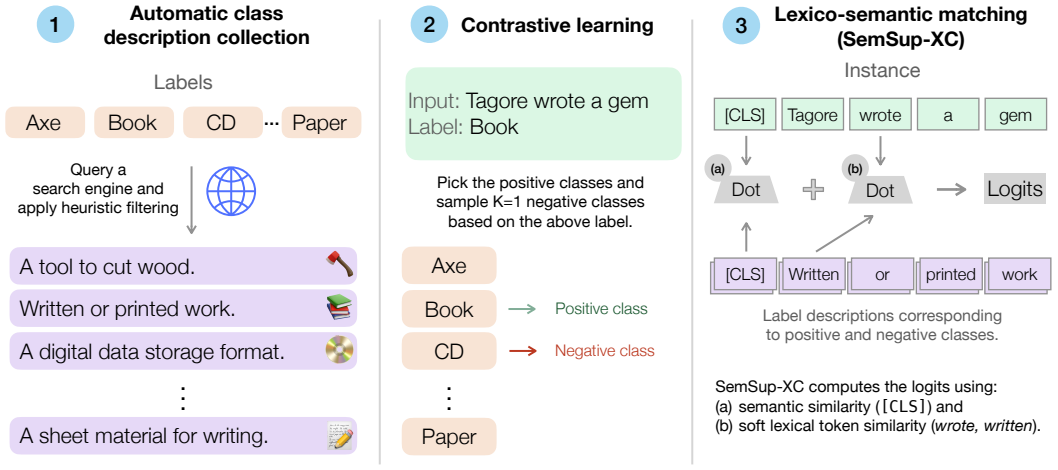


Figure 1: Our model SEMSUP-XC improves the framework of semantic supervision (Hanjie et al., 2022), by adding (1) large-scale automated class description collection with heuristic filtering, (2) contrastive learning, and (3) a novel lexico-semantic matching model building on COIL (Gao et al., 2021a). While (1) and (2) significantly improve SEMSUP's computational speed (over 99% on Wikipedia), (3) boosts performance (§ 5.1)

classes that requires a small amount of human intervention, which is expensive for large label spaces we are dealing with.

We remedy these deficiencies by developing a new model SEMSUP-XC that scales to large class spaces in XC using three innovations. First, SEMSUP-XC employs a contrastive learning objective (Hadsell et al., 2006) which samples a fixed number of negative label descriptions, improving computation speed by as much as 99.9%. Second, we use a novel hybrid lexical-semantic similarity model called Relaxed-COIL (based on COIL (Gao et al., 2021b)) that combines semantic similarity of sentences with soft matching between all token pairs. And finally, we propose SEMSUP-WEB which is a fully automatic pipeline with precise heuristics to scrape high-quality descriptions.

SEMSUP-XC achieves state-of-the-art performance on three diverse XC datasets based on legal (EURLex), e-commerce (AmazonCat), and wiki (Wikipedia) domains, across three settings (zero-shot, generalized zero-shot and few-shot extreme classification – ZS-XC, GZS-XC, FS-XC). For example, on ZS-XC, SEMSUP-XC outperforms the next best baseline by 5 to 19 Precision@1 points on different datasets, and maintains the advantage across all metrics. On FS-XC, SEMSUP-XC consistently outperforms baselines by over 10 Precision@1 points on $5, 10, 20$ shot classification. Interestingly, SEMSUP-XC also outperforms larger models like T5 and Sentence Transformers (e.g., by over 30 P@1 points on EURLex) which are pre-trained on web-scale corpora, which shows the importance of contrastive learning to adapt to a specific domain. We perform several ablation studies to dissect the importance of each component in SEMSUP-XC, and also provide a qualitative error analysis of the model.

2 Related Work

Extreme classification Extreme classification (XC) (Bengio et al., 2019) studies multi-class and multi-label classification problems over large label spaces. Traditionally, studies have used *sparse-features* extracted from the bag-of-words representation of input documents (Bhatia et al., 2015; Chang et al., 2019; Lin et al., 2014), and have also explored one-versus-all binary classifiers (Babbar & Schölkopf, 2017; Yen et al., 2017; Jain et al., 2019; Dahiya et al., 2021a) and tree-based methods which utilize the label hierarchy (Prabhu et al., 2018; Wydmuch et al., 2018; Khandagale et al., 2020). Recently, neural-network (NN) based *dense-feature* methods have demonstrated improved accuracies due to their ability to generate semantically rich and contextual representations of text. Different studies have experimented with architectures like convolutional neural networks (Liu et al., 2017), Transformers (Chang et al., 2020; Jiang et al., 2021; Zhang et al., 2021), attention-based

networks (You et al., 2019) and shallow networks (Medini et al., 2019; Mittal et al., 2021; Dahiya et al., 2021b). While the aforementioned works show impressive performance when the labels during training and evaluation are the same, they do not consider the practical zero-shot classification scenario with unseen labels during evaluation.

Zero-shot extreme classification (ZS-XC) Zero-shot classification (ZS) (Larochelle et al., 2008) aims to predict unseen classes not encountered during training by utilizing auxiliary information like the class name or a prototype. Multiple works have attempted to improve performance for the text domain (Dauphin et al., 2014; Nam et al., 2016; Wang et al., 2018; Pappas & Henderson, 2019; Hanjie et al., 2022), however, given the large label space of XC, these cannot be easily be extended because of computational expense and performance degradation. ZestXML (Gupta et al., 2021) was the first study to attempt ZS extreme classification by projecting bag-of-words input features close to corresponding label features using a sparsified linear transformation, but this limits them to using non-contextual text representations. Subsequent works have used neural-networks to generate contextual text representations (Xiong et al., 2022; Simig et al., 2022; Zhang et al., 2022; Rios & Kavuluru, 2018), with MACLR (Xiong et al., 2022) using an inverse cloze pre-training step and GROOV (Simig et al., 2022) using a sequence-to-sequence generative model to predict novel labels. However, all these works have the shortcoming that they use only label names (e.g., the word "cat") which lack semantic information to represent classes. In this work, we adapt the recently proposed method of semantic supervision (Hanjie et al., 2022) that uses semantically rich and diverse descriptions to represent classes. SEMSUP underperforms out-of-the-box, and we propose several training and modeling changes (§ 3) to achieve state-of-the-art performance on ZS XC tasks.

3 METHODOLOGY

3.1 BACKGROUND

Zero and Few-shot Extreme Classification Extreme classification (XC) contains classification problems with label spaces (thousands to millions classes). Zero-shot extreme classification (ZS-XC) is a version of XC where a model is evaluated on unseen classes not encountered during training. We consider two settings, (1) Zero-shot (ZS), where the model is tested only on unseen classes not containing any train classes, and (2) Generalized zero-shot (G-ZS), where the model is tested on a combined set of train and unseen classes.

Background: Semantic supervision Semantic supervision (SEMSUP) (Hanjie et al., 2022) is a framework for zero-shot classification that represents classes using rich textual descriptions (e.g., "A form of competitive physical activity or game" for the class *sports*), instead of discrete IDs (e.g., *Class-1*, *Class-2*). This allows a trained model to generalize to new classes as long as their corresponding descriptions are provided. In addition to using an input encoder (f_{IE}), SEMSUP also has an output encoder (g_{OE}) to encode label descriptions, and makes a class prediction by measuring the compatibility of the input representation of the instance and output representation of the label description corresponding to a class.

Formally, let C be the number of classes, d be the dimensionality of the input representation, x_i be the input document, $\mathcal{D} = (d_1, \ldots, d_C)$ be descriptions corresponding to the classes, $f_{\text{IE}}(x_i) \in \mathbb{R}^d$ be the input representation, and $g_{\text{OE}}(d_j) \in \mathbb{R}^d$ be the output representation of the j^{th} class. We operate under the multi-label classification setting, which is the default for the extreme classification benchmarks we consider. Then, we have the probability of picking the j^{th} class as:

$$\text{SEMSUP} \coloneqq P(y_j = 1|x_i) = \sigma\left(g_{\text{OE}}(d_j)^{\intercal} \cdot f_{\text{IE}}(x_i)\right) \tag{1}$$

SEMSUP is trained using the binary cross-entropy (BCE) loss between the predicted probability and gold answer, where N is the number of instances in the dataset.

$$\mathcal{L}_{\text{SEMSUP}} = \frac{1}{N \cdot |C|} \sum_{(x_i, Y_i)} \sum_{j=1}^{C} \mathcal{L}_{\text{BCE}}\left(P\left(y_{ij} = 1|x_i\right), y_{ij}\right) \tag{2}$$

where $Y_i = \{y_{i1}, \ldots, y_{iC} | y_{ij} \in \{0, 1\}\}$ is the set of the labels for instance x_i, with x_i belonging to class j if and only if $y_{ij} = 1$.

SEMSUP relies on multiple high-quality label descriptions of classes that contain semantic information, which are semi-automatically scraped from the web and filtered by an expert in (Hanjie et al., 2022). During training, diverse descriptions are randomly sampled, endowing the model with a semantically rich representation since they contain information on different class attributes. For example, the class *sports* can have a definition ("A form of competitive physical activity or game"), examples ("Examples are football, cricket, and hockey"), or etymology ("Derived from a French word meaning leisure") in each description, among other attributes.

Shortcomings for large class spaces While (Hanjie et al., 2022) show improved performance of SEMSUP on zero-shot classification, their vanilla method cannot be directly applied to extreme classification due to several reasons. First, they use the binary cross-entropy loss over all the classes in the dataset (Eq 2), which involves encoding C label descriptions for each batch. This is computationally intractable for large label spaces because of GPU memory constraints. Second, they use a simple bi-encoder model which measures the semantic similarity between the input instance and the label description. However, instances and descriptions often share lexical terms with the same or similar lemma (e.g., *wrote* and *written*), which are not directly exploited by their method. And third, although their label description collection pipeline is semi-automatic, human intervention of any form is not feasible for extreme classification datasets, especially ones with labels in the order of millions. Our method SEMSUP-XC (Section 3.2) addresses the above contraints by using: (1) contrastive learning with negative samples for improved computational speed, (2) a novel hybrid lexical-semantic similarity model for improved performance, and (3) a completely automatic description scraping pipeline with accurate heuristics for filtering poor descriptions.

3.2 SEMSUP-XC: EFFICIENT GENERALIZATION FOR ZERO-SHOT EXTREME CLASSIFICATION

We provide an overview of our method in Figure 1 and explain it in detail below.

Training using contrastive learning For datasets with a large label space (large $|C|$), we improve SEMSUP's computational speed by sampling negative classes for each instance rather than encoding the label descriptions of all classes. For instance x_i, consider two partitions of the labels $Y_i = \{y_{i1}, \ldots, y_{iC} | y_{ij} \in \{0, 1\}\}$, with Y_i^+ containing the positive classes ($y_{ij} = 1$) and Y_i^- containing the negative classes ($y_{ij} = 0$). SEMSUP-XC is trained using all positive classes ($|Y_i^+|$), but drawing inspiration from contrastive learning (Hadsell et al., 2006), we sample $K - |Y_i^+|$ negative classes from Y_i^- instead of $C - |Y_i^+|$, with $K \ll |C|$. We refer readers to appendix C for exact details. Intuitively, our training objective incentivizes the representations of the instance and label descriptions of the positive classes to be similar while simultaneously increasing the distance with respect to representations of the negative classes. Furthermore, rather than picking negative labels at random, we sample hard negatives that are lexically similar to positive labels. This allows for more explicit separation of embeddings of closely related labels. A typical dataset we consider (AmazonCat) contains $|C| = 13,000$ and $K \approx 1000$, which leads to SEMSUP-XC being $\frac{12000}{13000} = 92.3\%$ faster than SEMSUP. Mathematically, the following is the training objective:

$$\mathcal{L}_{\text{SEMSUP-XC}} = \frac{1}{N \cdot K} \sum_i \left(\sum_{y_k \in Y_i^+} \mathcal{L}_{\text{BCE}} \left(P\left(y_k = 1 | x_i\right), y_k \right) + \sum_{l=1, y_l \ Y_i^-}^{K - |Y_i^+|} \mathcal{L}_{\text{BCE}} \left(P(y_l = 1 | x_i), y_l \right) \right) \quad (3)$$

We follow a similar procedure for inference and we refer readers to Appendix C for further details.

Hybrid lexico-semantic similarity model SEMSUP uses a bi-encoder architecture with two different BERT (Devlin et al., 2019b) models as the input and output encoder respectively. They use the final-layer representation corresponding to the `[CLS]` token to encode the instance and label description, with the inner product measuring the semantic similarity of the input instance and output class. Drawing inspiration from recent IR models like COIL (Gao et al., 2021b) and ColBERT (Khattab & Zaharia, 2020), we note that BERT's semantic similarity can ignore lexical matching of words present in the input and output text which exhibit strong evidence of compatibility, thus leading to

performance degradation. COIL is a bi-encoder architecture which alleviates this by incorporating both semantic and lexical similarity. Apart from the dot product between [CLS] vectors, they also include an exact lexical match scoring function which is based on the dot product of representations corresponding to tokens with exact matches in the two pieces of text considered (e.g., input text: "Capture the best momemts in high quality pictures." and label description "A camera is used to take photos." If there are multiple occurrences of a common token type, the maximum similarity score is chosen, and the scores are then aggregated over all token types that are present in both sentences. Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ be the input instance with n tokens, $d_j = (d_{j1}, d_{j2}, \ldots, d_{jm})$ be a label description of class j with m tokens, $v_{\text{cls}}^{x_i}$ and $v_{\text{cls}}^{d_j}$ be the [CLS] representations of the input and label description, and $v_k^{x_i}$ and $v_l^{d_j}$ be the token representation of the k^{th} and l^{th} token of x_i and d_j respectively. Mathematically, the following would be the probability of picking class j for both vanilla SEMSUP and SEMSUP + COIL:

$$\begin{aligned} \text{SEMSUP} &:= P(y_j = 1|x_i) = \sigma\left(v_{\text{cls}}^{x_i\intercal} \cdot v_{\text{cls}}^{d_j}\right) \\ \text{SEMSUP + COIL} &:= P(y_j = 1|x_i) = \sigma\left(v_{\text{cls}}^{x_i\intercal} \cdot v_{\text{cls}}^{d_j} + \sum_{w \in x_i \cap d_j} \max_{w = x_{ik} = d_{jl}} \left(v_k^{x_i\intercal} v_l^{d_j}\right)\right) \end{aligned} \tag{4}$$

where exact lexical match is used to get the list of common tokens – $w \in x_i \cap d_j$. As a result of the exact match, COIL has the drawback that semantically similar tokens (e.g., pictures and photos in the above sentence pair) and words with the same lemma (e.g., walk and walking) are treated as dissimilar tokens. To avoid this, we propose the use of soft lexical matching based on token clustering and lemmatization in our model Relaxed-COIL. We create clusters of tokens based on two characteristics: 1) the BERT token-embedding similarity (Reimers & Gurevych, 2019) and 2) the lemma of the word. Let $CL(w)$ denote the cluster membership of the token w, with $CL(w_i) = CL(w_j)$ if and only if they have a high token-embedding similarity or if they share the same lemma. Instead of using the exact lexical match in COIL, we use a soft lexical match which only checks if two tokens belong to the same cluster, thus allowing the model to exploit semantic similarity of tokens. Mathematically, Relaxed-COIL computes the probability as follows, where $CL(x_i) = \{CL(x_{i1}), \ldots, CL(x_{in})\}$ denotes the set of cluster memberships of tokens in x_i.

$$\begin{aligned} \text{Relaxed-COIL} :=& P(y_j = 1|x_i) = \\ & \sigma\left(v_{\text{cls}}^{x_i\intercal} \cdot v_{\text{cls}}^{d_j} + \sum_{\text{idx} \in CL(x_i) \cap CL(d_j)} \max_{\text{idx} = CL(x_{ik}) = CL(d_{jl})} \left(v_k^{x_i\intercal} v_l^{d_j}\right)\right) \end{aligned} \tag{5}$$

Automatically collecting high-quality descriptions SEMSUP uses a semi-automatic pipeline for collecting multiple descriptions for classes, with an expert required for filtering irrelevant ones. However, in our case, large label spaces (e.g., 1 million for Wikipedia) make any degree of human involvment infeasible. To alleviate this, we create a completely automatic pipeline for collecting descriptions which includes heuristics for removing spam, advertisements, and irrelevant descriptions, and we detail the list of heuristics used in Appendix B. In addition to web-scraped label descriptions, we utilize label-hierarchy information if provided by the dataset(EURLex and AmazonCat), which allows us to encode properties about parent and children classes wherever present. Further details are present in Appendix B.2 As we show in the ablation study (§ 5.3), both label descriptions that we collect and the label hierarchy provide significant performance boosts.

4 EXPERIMENTAL SETUP

Datasets We evaluate our model on three diverse public benchmark datasets. They are, **EURLex-4.3K** (Chalkidis et al., 2019) which is legal document classification dataset with 4.3K classes, **AmazonCat-13K** (McAuley & Leskovec, 2013) which is an e-commerce product tagging dataset including Amazon product descriptions and titles with 13K categories, and **Wikipedia-1M** (Gupta et al., 2021) which is an article classification dataset made up of over 5 million Wikipedia articles

Dataset	Documents				Labels	
	N_{train}	N_{test}	N_{testzsl}	$\lvert Y_{avg} \rvert$	$\lvert Y_{\text{seen}} \rvert$	$\lvert Y_{\text{unseen}} \rvert$
EURLex-4.3K	45 K	6 K	5.3 K	547.5	3,136	1,057
AmazonCat-13K	1.1M	307K	268 K	210.4	6,830	6,500
Wikipedia-1M	2.3M	2.7M	2.2M	275.6	495,107	776,612

Table 1: Dataset statistics along with information about zero-shot (ZS-XC) splits.

with 1 million categories. We provide detailed statistics about the number of instances and classes in train and test set in Table 1. We refer readers to Appendix A.2 for additional details.

Baselines We perform extensive experiments with six diverse baselines. **1) TF-IDF** performs a nearest neighbour match between the sparse tf-idf features of the input and label description. **2) T5** (Raffel et al., 2019) is a large sequence-to-sequence model which has been pre-trained on 750GB unsupervised data and further fine-tuned on MNLI (Williams et al., 2018) to allow us to check if a label description entails an input instance. Ranking is done on top 50 labels predicted by TF-IDF **3) Sentence Transformer** (Reimers & Gurevych, 2019) is a semantic text similarity model fine-tuned using a contrastive learning objective on over 1 billion sentence pairs. We rank the labels based on similarity of their output embeddings with document's embeddings. The latter two baselines use significantly more data than SEMSUP-XC and T5 has $9\times$ the parameters. The aforementioned baselines are unsupervised and not fine-tuned on our datasets. The following baselines are previously proposed supervised models which are fine-tuned on the datasets we consider. **4) ZestXML** (Gupta et al., 2021) learns a highly sparsified linear transformation between sparse input and label features. **5) MACLR** (Xiong et al., 2022) is a bi-encoder based model pre-trained on two self-supervised learning tasks to improve extreme classification—Inverse Cloze Task (Lee et al., 2019) and SimCSE (Gao et al., 2021c), and we fine-tune it on the datasets considered. **6) GROOV** (Simig et al., 2022) is a T5 model that learns to generate both seen and unseen labels given an input document. **7) SPLADE** (Formal et al., 2021) is a sparse neural retreival model that learns label/document sparse expansion via a Bert masked language modelling head. It is one of the current state of the art in information retrieval in out-of-domain tasks. To make comparisons with SEMSUP-XC fair, we fine-tune and re-evaluate the above models on the datasets we consider while including label descriptions and label hierarchy information. We refer readers to Appendix A.2 for additional details.

SEMSUP-XC implementation details We use the Bert-base model (Devlin et al., 2019a) as the backbone for the input encoder and Bert-small model (Turc et al., 2019) for the output encoder. SEMSUP-XC follows the model architecture described in Section 3.2 and we use contrastive learning (Hadsell et al., 2006) to train our models. During training, we randomly sample $1000-p$ negatives for each instance, where p is the number of positive labels for the instance. At inference, to improve computational efficiency, we precompute the output representations of label descriptions. We use the AdamW optimizer (Loshchilov & Hutter, 2019) and tune our hyperparameters using grid search on the respective validation set. We provide further details in Appendix A.1.

Evaluation setting and metrics We evaluate all models on three different settings: Zero-shot classification (ZS) on a set of unseen classes, generalized zero-shot classification (G-ZS) on a combined set of seen and unseen classes, and few-shot classification (FS) on a set of classes with minimal amounts of supervised data (1 to 20 examples per class). We use Precision@K and Recall@K (with multiple values of K) as our evaluation metrics, as is standard practice. Precision@K measures how accurate the top-K predictions of the model are, and Recall@K measures what fraction of correct labels are present in the top-K predictions, and they are mathematically defined as $P@k = \frac{1}{k}\sum_{i\in \text{rank}_k(\hat{y})} y_i$ and $R@k = \frac{1}{\sum_i y_i}\sum_{i\in \text{rank}_k(\hat{y})} y_i$, where $\text{rank}_k(\hat{y})$ is the set of top-K predictions.

Model	EURLex-4.3K				AmazonCat-13K				Wikipedia-1M			
	ZS-XC		GZS-XC		ZS-XC		GZS-XC		ZS-XC		GZS-XC	
	P@1	R@10	P@1	R@10	P@1	R@10	P@1	R@10	P@1	R@10	P@1	R@10
w/o Descriptions												
TF-IDF	44.0	55.8	53.4	41.2	18.7	21.0	14.7	21.5	14.5	18.3	14.4	14.7
T5	7.2	29.2	10.4	23.0	2.5	10.5	3.2	10.2	8.2	23.6	4.2	15.1
Sent. Transformer	16.6	23.2	20.9	42.0	18.2	25.0	21.1	17.9	7.8	13.3	5.2	9.1
ZestXML (Gupta et al., 2021)	9.6	25.7	84.8	54.8	12.7	21.2	87.9	52.5	12.9	20.0	26.7	25.7
ZestXML + TF-IDF	24.7	46.4	84.9	54.2	15.6	24.4	87.6	54.2	15.8	20.8	26.3	17.2
SPLADE (Formal et al., 2021)	20.2	24.4	52.3	34.2	17.2	28.7	75.8	41.3	14.3	17.8	25.3	26.4
MACLR (Xiong et al., 2022)	25.3	41.7	59.8	54.2	35.5	55.0	42.9	44.3	28.6	40.1	26.7	30.7
GROOV (Simig et al., 2022)	1.2	7.0	84.1	49.4	0.0	2.4	87.1	47.6	5.9	15.4	**31.4**	29.1
with Descriptions												
TF-IDF	43.7	50.4	57.2	39.5	17.4	20.8	21.1	15.0	9.2	12.5	9.1	10.3
T5	5.0	24.8	3.3	8.1	2.8	7.7	3.2	4.2	3.7	13.4	3.4	13.2
Sent. Transformer	15.9	31.1	18.8	25.5	15.2	22.2	16.0	18.4	19.6	22.5	14.2	16.6
ZestXML (Gupta et al., 2021)	9.7	25.8	83.8	55.2	3.8	20.5	71.1	49.7	9.1	13.9	19.2	17.8
ZestXML + TF-IDF	22.6	44.6	84.2	60.7	5.4	24.8	76.9	50.7	10.6	14.1	20.9	17.9
MACLR (Xiong et al., 2022)	20.9	37.9	60.3	53.8	18.4	22.3	36.5	23.8	30.7	**41.9**	28.1	33.6
GROOV (Simig et al., 2022)	0.3	0.6	80.2	18.1	0.0	0.0	84.5	23.5	0.5	0.2	7.0	1.5
SEMSUP-XC	**49.7**	**62.0**	**87.0**	**59.0**	**48.5**	**73.1**	**88.5**	**71.7**	**36.5**	38.5	**33.7**	**34.1**

Table 2: Results for zero-shot (ZS-XC) and generalized zero-shot (GZS-XC) for all models on three XC benchmarks. SEMSUP-XC significantly outperforms state-of-the-art models on both precision (P@) and recall (R@) metrics across the board.

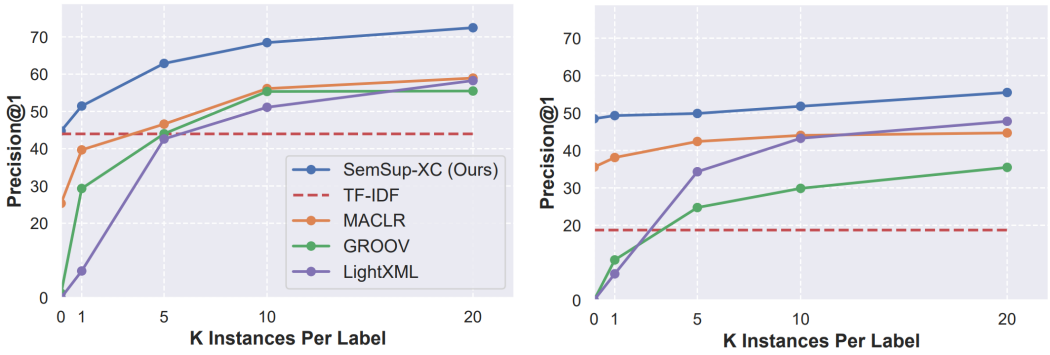
5 RESULTS

5.1 ZERO-SHOT EXTREME CLASSIFICATION

We consider two variants of baselines: with and without descriptions. We provide label hierarchy as output supervision in both cases. Table 2 shows that SEMSUP-XC significantly outperforms baselines on all datasets and metrics, under both zero-shot (ZS-XC) and generalized zero-shot (GZS-XC) settings. On ZS-XC, SEMSUP-XC outperforms MACLR by over 20, 13, and 15 P@1 points on the three datasets, respectively, even though MACLR uses XC specific pre-training (Inverse Cloze Task and SimCSE) , while SEMSUP-XC does not. SEMSUP-XC also outperforms GROOV (e.g., over 45 P@1 points on EURLex) which uses a T5 seq2seq model pre-trained on significantly more data than BERT, and this is likely because GROOV's output space is unconstrained. SEMSUP-XC's semantic understanding of instances and labels stands out against ZestXML which uses sparse non-contextual features with the former consistently scoring twice as higher compared to the latter. Interestingly TF-IDF performs better than all other baselines for EURLex ZS. This is because sparse methods often perform better than dense bi-encoders in zero-shot settings (Thakur et al., 2021), as the latter fail to capture fine-grained information. However, due to the introduction of Relaxed COIL in our method, SEMSUP-XC can perform fine-grained lexical matching in descriptions along with capturing deep semantic information, thus resulting in superior ZS and GZS scores. SEMSUP-XC also outperforms the other unsupervised baselines of T5, and Sentence-Transformer, even though the latter two are pre-trained on significantly larger amounts of data than BERT (T5 use $50\times$ compared our base model).

In addition, SEMSUP-XC achieves higher recall on all datasets, beating the best performing baselines by 6, 18, and 5.2R@10 points on the three datasets, respectively. Since GZS-XC includes labels seen during training while evaluation, all methods have higher scores than ZS-XC and the gaps between different models are smaller, but we see both on precision and recall metrics that SEMSUP-XC again outperforms all the baselines considered by margins of 1-2 precision@1 points. Table 5 in Appendix D contains additional results with more methods and metrics.

Further, while SEMSUP-XC improves from the inclusion of descriptions, other methods have little to no advantage. This is because web-scraped descriptions are often noisy and need suitable architecture to make use of them. Unlike other methods, SEMSUP-XC has a hybrid lexical matching module, which improves from the inclusion of descriptions. This demonstrates the combined advantage of our proposed architectural changes and the use of web-scraped descriptions.



(a) EURLex-4.3K (b) AmazonCat-13K

Figure 2: Few-Shot Precision@5 for different Values of K on EurLex and AmazonCat. SEMSUP-XC starts off significantly higher for K=0 and maintains the gap for larger K to the second best model, MACLR (Xiong et al., 2022).

Method	Components				EURLex-4.3K			AmazonCat-13K		
	Descri-ptions	Hierarchy	COIL	Relaxed COIL	P@1	P@5	R@10	P@1	P@5	R@10
Label description ablation										
SEMSUP-XC	Web	✓	✓	✓	44.8	**21.1**	**58.1**	**48.5**	**27.4**	**73.1**
Replace Web Descriptions with Names	Names	✓	✓	✓	**45.3**	20.4	56.9	44.1	25.2	69.3
Remove Hierarchy	Names	✗	✓	✓	31.0	14.8	42.4	22.4	13.1	38.1
Relaxed-COIL ablation										
SEMSUP-XC	Web	✓	✓	✓	**44.8**	**21.1**	**58.1**	**48.5**	**27.4**	**73.1**
Replace Relaxed with Exact Lexical Matching	Web	✓	✓	✗	42.4	19.3	53.4	44.8	24.9	67.4
Remove all lexical matching	Web	✓	✗	✗	13.4	9.1	30.1	37.1	21.9	60.2

Table 3: Component-wise Model Analysis of SEMSUP-XC for ZS-XCon EURLex and AmazonCat. Each component contributes to the final performance, with lexical-matching playing an important role.

5.2 FEW-SHOT EXTREME CLASSIFICATION

We now consider the FS-XC setup, where new classes added at evaluation time have a small number of labeled instances each ($K \in \{1, 5, 10, 20\}$). For the sake of completeness, we also include zero-shot performance (ZS-XC, $K = 0$) and report results in Figure 2. Detailed results for other metrics are in appendix E (showing the same trend as P@1) and implementation details regarding creation of the few-shot splits are in appendix E. Similar to the ZS-XC case, SEMSUP-XC outperforms the baselines for all values of K considered. As expected, SEMSUP-XC's performance increases with K because of access to more labeled data, but crucially, it continues to outperform baselines by the same margins. Interestingly, SEMSUP-XC's zero-shot performance is higher than even the few-shot scores of baselines that have access up to $K = 20$ labeled samples on AmazonCat, which further strengthens the model's applicability to the XC paradigm. We also note that adding a few labeled examples seems to be more effective in EURLex than AmazonCat, with the performance difference between $K = 1$ and $K = 20$ being 21 and 6 P@1 points respectively. Combined with the fact that performance seems to plateau for both datasets, we believe that the larger label space with rich descriptions for AmazonCat has allowed SEMSUP-XC to learn label semantics better than for EURLex.

5.3 ABLATIONS

We analyze the performance of SEMSUP-XC by conducting ablation studies and qualitative analysis on EURLex and AmazonCat for the zero-shot extreme classification setting (ZS-XC) in the following sections.

Method	EURLex-4.3K			AmazonCat-13K		
	P@1	P@5	R@10	P@1	P@5	R@10
SEMSUP-XC	44.8	21.1	58.1	**48.5**	**27.4**	**73.1**
+ Augmentation	**46.6**	**22.3**	**60.3**	47.7	26.3	72.4

Table 4: Description augmentation helps boost performance for ZS-XC on EURLex, but not as much on AmazonCat, which is a significantly larger dataset ($4\times$ the number of labels).

Analyzing components of SEMSUP-XC SEMSUP-XC's use of the Relaxed-COIL model and semantically rich descriptions enables it to outperform all baselines considered, and we analyze the importance of each component in Table 3. As our base model (first row) we consider the SEMSUP-XC without ensembling it with TF-IDF. We note that the SEMSUP-XC base model is the best performing variant for both datasets and on all metrics other than P@1, for which it is only 0.5 points lower. Web scraped label descriptions are important because removing them decreases both precision and recall scores (e.g., P@1 is lower by 4 points on AmazonCat) on all settings considered. We see bigger improvements with AmazonCat, which is the dataset with larger number of classes (13K), which substantiates the need for semantically rich descriptions when dealing with fine-grained classes. Label hierarchy information is similarly crucial, with large performance drops on both datasets in its absence (e.g., 26 P@1 points on AmazonCat), thus showing that access to structured hierarchy information leads to better semantic representations of labels.

On the modeling side, we observe that both exact and soft lexical matching are important for Relaxed-COIL, with their absence leading to 11 and 4 point P@1 degradation on AmazonCat. While exact lexical-matching is significantly more important for EURLex which is the smaller dataset, we see that both types of matching are important for AmazonCat which tends to have classes which are more related.

Automatically Augmenting Label Descriptions The previous result showed the importance of web scraped descriptions, and we explore the effect of augmenting label descriptions to increase their number, and hence SEMSUP-XC's understanding of the class, and report the results in Table 4. We use the widely used Easy Data Augmentation(EDA) (Wei & Zou, 2019) method for descriptions augmentations. Specifically, we apply random word deletion, random word swapping, random insertion, and synonym replacement each with a probabilty 0.5 on the description. We notice that augmentation improves performance on EURLex by 2, 1, and 2 P@1, P@5, and R@10 points respectively, suggesting that augmentation can be a viable way to increase the quantity of descriptions. But results on AmazonCat show that augmentation does not improve and actually slightly hurts the performance (e.g., 0.8 P@1 points). Given that AmazonCat has $3\times$ the number of labels compared to EURLex, we believe that this shows SEMSUP-XC's effectiveness in capturing the label semantics in the presence of larger number of classes, thus making data augmentation redundant. However, we believe that data augmentation might be a simple tool to boost performance on smaller datasets with lesser labels or descriptions.

6 CONCLUSION

We tackle the task of extreme classification (XC) (Bengio et al., 2019), which involves very large label spaces, using the framework of semantic supervision (Hanjie et al., 2022) that uses class descriptions instead of label IDs. Our method SEMSUP-XC innovates using a combination of contrastive learning, hybrid lexico-semantic matching and automated description collection to train effectively for XC. We achieve state-of-the-art results on three standard XC benchmarks and significantly outperform prior work, while also providing several ablation studies and qualitative analyses demonstrate the relative importance of our various modeling choices. Future work can further improve description quality and use stronger models for input-output similarity to further push the boundaries on this practical task with real-world applications.

REFERENCES

Rohit Babbar and Bernhard Schölkopf. Dismec: Distributed sparse machines for extreme multi-label classification. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.

Samy Bengio, Krzysztof Dembczynski, Thorsten Joachims, Marius Kloft, and Manik Varma. Extreme classification (dagstuhl seminar 18291). In Dagstuhl Reports, volume 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In NIPS, 2015.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. In ACL, 2019.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. A modular deep learning approach for extreme multi-label text classification. ArXiv, abs/1905.02331, 2019.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. Taming pretrained transformers for extreme multi-label text classification. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3163–3171, 2020.

Kunal Dahiya, Ananye Agarwal, Deepak Saini, K Gururaj, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, and Manik Varma. Siamesexml: Siamese networks meet extreme classifiers with 100m labels. In International Conference on Machine Learning, pp. 2330–2340. PMLR, 2021a.

Kunal Dahiya, Deepak Saini, Anshul Mittal, Ankush Shaw, Kushal Dave, Akshay Soni, Himanshu Jain, Sumeet Agarwal, and Manik Varma. Deepxml: A deep extreme multi-label learning framework applied to short text documents. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 31–39, 2021b.

Yann N Dauphin, Gökhan Tür, Dilek Hakkani-Tür, and Larry P Heck. Zero-shot learning and clustering for semantic utterance classification. In ICLR (Poster), 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, 2019b. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.

Ben Friedland. profanity: A python library to check for (and clean) profanity in strings, 2013. URL https://github.com/ben174/profanity.

Luyu Gao, Zhuyun Dai, and Jamie Callan. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. In NAACL, 2021a.

Luyu Gao, Zhuyun Dai, and Jamie Callan. COIL: revisit exact lexical match in information retrieval with contextualized inverted list. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pp. 3030–3042. Association for Computational Linguistics, 2021b. doi: 10.18653/v1/2021.naacl-main.241. URL https://doi.org/10.18653/v1/2021.naacl-main.241.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. ArXiv, abs/2104.08821, 2021c.

María Grandury. roberta-base-finetuned-sms-spam-detection, 2021. URL https://huggingface.co/mariagrandury/roberta-base-finetuned-sms-spam-detection.

Nilesh Gupta, Sakina Bohra, Yashoteja Prabhu, Saurabh.S. Purohit, and Manik Varma. Generalized zero-shot extreme multi-label learning. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pp. 1735–1742. IEEE, 2006.

Austin W. Hanjie, A. Deshpande, and Karthik Narasimhan. Semantic supervision: Enabling generalization over output spaces. ArXiv, abs/2202.13100, 2022.

Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019.

Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In AAAI, 2021.

Sujay Khandagale, Han Xiao, and Rohit Babbar. Bonsai: diverse and shallow trees for extreme multi-label classification. Machine Learning, pp. 1 – 21, 2020.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In AAAI, volume 1, pp. 3, 2008.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In ACL, 2022.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. ArXiv, abs/1906.00300, 2019.

Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Multi-label classification via feature-aware implicit label space encoding. In ICML, 2014.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR, 2019.

Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. Proceedings of the 7th ACM conference on Recommender systems, 2013.

Tharun Kumar Reddy Medini, Qixuan Huang, Yiqiu Wang, Vijai Mohan, and Anshumali Shrivastava. Extreme classification in log memory using count-min sketch: A case study of amazon search with 50m products. Advances in Neural Information Processing Systems, 32, 2019.

Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. Decaf: Deep extreme classification with label features. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 49–57, 2021.

Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. All-in text: Learning document, label, and word representations jointly. In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

Nikolaos Pappas and James Henderson. Gile: A generalized input-label embedding for text classification. Transactions of the Association for Computational Linguistics, 7:139–155, 2019.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. pp. 993–1002, 04 2018. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3185998.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019. URL https://arxiv.org/abs/1910.10683.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. ArXiv, abs/1908.10084, 2019.

Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2018, pp. 3132. NIH Public Access, 2018.

Daniel Simig, Fabio Petroni, Pouya Yanki, Kashyap Popat, Christina Du, Sebastian Riedel, and Majid Yazdani. Open vocabulary extreme classification using generative models. ArXiv, abs/2205.05812, 2022.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. ArXiv, abs/2104.08663, 2021.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. arXiv: Computation and Language, 2019.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2321–2331, 2018.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019. URL https://arxiv.org/abs/1901.11196.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Marek Wydmuch, Kalina Jasinska, Mikhail Kuznetsov, Róbert Busa-Fekete, and Krzysztof Dembczynski. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In NeurIPS, 2018.

Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Inderjit S. Dhillon. Extreme zero-shot learning for extreme text classification. ArXiv, abs/2112.08652, 2022.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In NAACL, 2021.

Ian En-Hsu Yen, Xiangru Huang, Wei Dai, Pradeep Ravikumar, Inderjit S. Dhillon, and Eric P. Xing. Ppdsparse: A parallel primal-dual sparse method for extreme classification. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In NeurIPS, 2019.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. Advances in Neural Information Processing Systems, 34:7267–7280, 2021.

Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. Metadata-induced contrastive learning for zero-shot multi-label text classification. In Proceedings of the ACM Web Conference 2022, pp. 3162–3173, 2022.

APPENDICES

A TRAINING DETAILS

A.1 HYPERPARAMETER TUNING

We tune the learning rate, batch_size using grid search. For the EURLex dataset, we use the standard validation split for choosing the best parameters. We set the input and output encoder's learning rate at $5e^{-5}$ and $1e^{-4}$, respectively. We use the same learning rate for the other two datasets. We use batch_size of 16 on EURLex and 32 on AmazonCat and Wikipedia. For Eurlex, we train our zero-shot model for fixed 2 epochs and the generalized zero-shot model for 10 epochs. For the other 2 datasets, we train for a fixed 1 epoch. For baselines, we use the default settings as used in respective papers.

Training

All of our models are trained end-to-end. We use the pretrained BERT model (Devlin et al., 2019b) for encoding input documents, and Bert-Small model (Turc et al., 2019) for encoding output descriptions. For efficiency in training, we freeze the first two layers of the output encoder. We use contrastive learning to train our models and sample hard negatives based on TF-IDF features. All implementation was done in PyTorch and Huggingface transformer and experiments were run NVIDIA RTX2080 and NVIDIA RTX3090 gpus.

A.2 BASELINES

We use the code provided by ZestXML, MACLR and GROOV for running the supervised baselines. We employ the exact implementation of TF-IDF as used in ZestXML. We evaluate T5 as an NLI task (Xue et al., 2021). We separately pass the names of each of the top 100 labels predicted by TF-IDF, and rank labels based on the likelihood of entailment. We evaluate Sentence-Transformer by comparing the similarity between the emeddings of input document and the names of the top 100 labels predicted by TF-IDF. Splade is a sparse neural retreival model that learns label/document sparse expansion via a Bert masked language modelling head. We use the code provided by authors for running the baselines. We experiment with various variations and pretrained models, and find splade_max_CoCodenser pretrained model with low sparsity($\lambda_d = 1e-6$ & $\lambda_q = 1e-6$) to be performing the best.

B LABEL DESCRIPTIONS FROM THE WEB

B.1 AUTOMATICALLY SCRAPING LABEL DESCRIPTIONS FROM THE WEB

We mine label descriptions from web in an automated end-to-end pipeline. We make query of the form 'what is <class_name>'(or component name in case of Wikipedia) on duckduckgo search engine. Region is set to United States(English), and advertisements are turned off, with safe search set to moderate. We set time range from 1990 uptil June 2019. On average top 50 descriptions are scraped for each query. To further improve the scraped descriptions, we apply a series of heuristics:

- We remove any incomplete sentences. Incomplete sentences do not end in a period or do not have more than one noun, verb or auxiliary verb in them.
  Eg: Label = **Adhesives** ; Removed Sentence = *What is the best glue or gel for applying*
- Statements with lot of punctuation such as semi-colon were found to be non-informative. Descriptions with more than 10 non-period punctuations were removed.
  Eg: Label = **Plant Cages & Supports** ; Removed Description = *Plant Cages & Supports. My Account; Register; Login; Wish List (0) Shopping Cart; Checkout $ USD $ AUD THB; R$ BRL $ CAD $ CLP $ ...*
- We used regex search to identify urls and currencies in the text. Most of such descriptions were spam and were removed.
  Eg: Label = **Accordion Accessories** ; Removed Description = *Buy Accordion Accessories*

Online, with Buy Now & Pay Later and Rental Options. Free Shipping on most orders over $250. Start Playing Accordion Accessories Today!

- Descriptions with small sentences(<5 words) were removed.
  Eg: Label = **Boats** ; Removed Description = *Boats for Sale. Buy A Boat; Sell A Boat; Boat Buyers Guide; Boat Insurance; Boat Financing ...*
- Descriptions with more than 2 interrogative sentences were filtered out.
  Eg: Label = **Shower Curtains** ; Removed Description = *So you're interested...why? you're starting a company that makes shower curtains? or are you just fooling around? Wiki User 2010-04*
- We mined top frequent n-grams from a sample of scraped descriptions, and based on it identified n-grams which were commonly used in advertisements. Examples include: *'find great deals', 'shipped by'*.
  Label = **Boat compasses** ; Removed Description = **Shop and read reviews** *about Compasses at West Marine.* **Get free shipping** *on all orders to any West Marine Store* **near you today**.
- We further remove obscene words from the datasets using an open-source library (Friedland, 2013).
- We also run a spam detection model (Grandury, 2021) on the descriptions and remove those with a confidence threshold above 0.9.
  Eg: Label = **Phones** ; Removed Description = *Check out the Phones page at <company_name> — the world's leading music technology and instrument retailer!*
- Additionally, most of the sentences in first person, were found to be advertisements, and undetected by previous model. We remove descriptions with more than 3 first person words (such as I, me, mine) were removed.
  Eg: Label = **Alarm Clocks** ; Removed Description = *We selected the best alarm clocks by taking the necessary, well, time. We tested products with our families, waded our way through expert and real-world user opinions, and determined what models lived up to manufacturers' claims. . . .*

B.2 POST-PROCESSING

We further add hierarchy information in a natural language format to the label descriptions for AmazonCat and EURLex datasets. Precisely, we follow the format of 'key is value.' with each key, value pair represented in new line. Here key belongs to the set { 'Description', 'Label', 'Alternate Label Names', 'Parents', 'Children' }, and the value corresponds to comma separated list of corresponding information from the hierarchy or scraped web description. For example, consider the label 'video surveillance' from EURLex dataset. We pass the text:
'Label is video surveillance.
Description is <web_scraped_description>.
Parents are video communications.
Alternate Label Names are camera surveillance, security camera surveillance.'
to the output encoder.
For Wikipedia, label hierarchy is not present, so we only pass the description along with the name of label.

B.3 WIKIPEDIA DESCRIPTIONS

When labels are fine-grained, as in the Wikipedia dataset, making queries for the full label name is not possible. For example, consider the label 'Fencers at the 1984 Summer Olympics' from Wikipedia categories; querying for it would link to the same category on Wikipedia itself. Instead, we break the label names into separate constituents using a dependency parser. Then for each constituent('Fencers' and 'Summer Olympics'), we scrape descriptions. No descriptions are scraped for constituents labelled by Named-Entity Recognition('1984'), and their NER tag is directly used. Finally, all the scraped descriptions are concatenated in a proper format and passed to the output encoder.

B.4 DE-DUPLICATION

To ensure no overlap between our descriptions and input documents, we used SuffixArray-based exact match algorithm (Lee et al., 2022) with a minimum threshold of 60 characters and removed the matched descriptions.

C CONTRASTIVE LEARNING

During training, for both EURLex and AmazonCat, we randomly sample $1000 - |Y_i^+|$ negative labels for each input document. For Wikipedia, we precompute the top 1000 labels for each input based on TF-IDF scores. We then randomly sample $1000 - |Y_i^+|$ negative labels for each document. At inference time, we evaluate our models on all labels for both EURLex and AmazonCat. However, even evaluation on millions of labels in Wikipedia is not computationally tractable. Therefore, we evaluate only on top 1000 labels predicted by TF-IDF for each input.

D FULL RESULTS FOR ZERO-SHOT CLASSIFICATION

D.1 SPLIT CREATION

For EURLex, and AmazonCat, we follow the same procedure as detailed in GROOV (Simig et al., 2022). We randomly sample k labels from all the labels present in train set, and consider the remaining labels as unseen. For EURLex we have roughly 25%(1057 labels) and for AmazonCat roughly 50%(6500 labels) as unseen. For Wikipedia, we use the standard splits as proposed in ZestXML (Gupta et al., 2021).

D.2 RESULTS

Table 5 contains complete results for ZS-XCacross the three datasets, including additional baselines and metrics.

E FULL RESULTS FOR FEW-SHOT CLASSIFICATION

E.1 SPLIT CREATION

We iteratively select k instances of each label in train documents. If a label has more than k documents associated with it, we drop the label from training(such labels are not sampled as either positives or negatives) for the extra documents. We refer to these labels as neutral labels for convenience.

E.2 MODELS

We use MACLR, GROOV, Light XML as baselines. We initialize the weights from the corresponding pre-trained models in the GZSL setting. We use the default hyperparameters for baselines and SEMSUP models. As discussed in the previous section, neutral labels are not provided at train time for MACLR and GROOV baselines. However, since Light XML uses a final fully-connected classification layer, we cannot selectively remove them for a particular input. Therefore, we mask the loss for labels which are neutral to the documents. We additionally include scores for TF-IDF, but since it is a fully unsupervised method, only zero-shot numbers are included.

E.3 RESULTS

The full results for few-shot classification are present in Table 6.

F COMPUTATIONAL EFFICIENCY

Extreme Classification necessitates that the models scale well in terms of time and memory efficiency with labels at both train and test times. SEMSUP-XC uses contrastive learning for efficiency at train

Method	Precision - ZSL			Recall - ZSL / GZSL		Precision - GZSL		
	@1	@3	@5	R@10	R@10	@1	@3	@5
Eurlex-4.3K								
TF-IDF	44.0	26.9	19.6	55.8	41.2	53.4	35.2	28.0
T5	7.2	7.1	7.0	29.2	23.0	10.4	11.0	11.2
Sentence Transformer	15.9	10.8	9.1	31.1	25.5	18.8	15.7	11.9
ZestXML	9.6	7.3	6.5	25.7	54.8	84.8	64.8	48.9
ZestXML + TF-IDF	24.7	17.7	14.4	46.4	54.2	84.9	65.7	50.3
MACLR	25.3	16.8	13.3	41.7	54.2	59.8	47.8	40.2
GROOV	1.2	2.6	2.6	7.0	49.4	84.1	61.5	45.3
SemSup-Hier	45.3	28.1	20.4	56.9	64.9	86.3	68.9	53.9
SemSup	44.8	28.0	21.1	58.1	**65.0**	**87.1**	**68.5**	**53.6**
SemSup + TF-IDF	**49.7**	**32.1**	**23.3**	**62.0**	59.0	87.0	67.5	51.6
Amazon-13K								
TF-IDF	18.7	11.5	8.5	21.0	14.7	21.5	14.4	11.1
T5	2.5	2.8	3	10.5	10.2	3.2	4.2	4.9
Sentence Transformer	15.2	10.5	8.3	22.2	16.0	18.4	13.4	11.0
ZestXML	12.7	8.9	7.1	21.2	52.5	87.9	58.6	41.5
ZestXML + TF-IDF	15.6	11.1	8.8	24.4	54.2	87.6	59.0	42.3
MACLR	35.5	23.7	18.3	55.0	44.3	42.9	31.4	25.4
GROOV	0.0	0.3	0.5	2.4	47.6	87.1	55.2	38.4
SemSup-Hier	44.1	31	25.2	69.3	54.3	88.7	64.6	50.0
SemSup	48.5	33.8	27.4	73.1	71.7	88.5	65.5	51.2
Wikipedia-1M								
TF-IDF	14.5	7.7	5.5	18.3	14.7	14.4	8.5	6.5
T5	8.2	7.6	6.7	23.6	15.1	4.2	4.5	4.4
Sentence Transformer	19.6	11.1	7.9	22.5	16.6	14.2	9.1	7.0
ZestXML	12.9	8.0	6.0	20.0	25.7	26.7	18.8	14.6
ZestXML + TF-IDF	15.8	8.9	6.4	20.8	26.3	30.6	22.2	17.2
MACLR	28.6	17.0	12.7	40.1	30.7	26.7	17.4	13.6
GROOV	5.9	5.8	4.9	15.4	29.1	31.4	24.9	19.1
SemSup	**36.5**	**19.5**	**13.4**	38.5	**34.1**	**33.7**	23.4	17.7

Table 5: Comparison of SemSup with other supervised and unsupervised baselines.

Method	**EURLex-4.3K**			**AmazonCat-13K**		
	P@1	**P@5**	**R@10**	**P@1**	**P@5**	**R@10**
5-shot						
SEMSUP-XC	62.9	25.9	67.4	49.9	26.2	67.4
MACLR	44.8	21.1	63.2	42.4	21.6	61.3
GROOV	44.0	17.5	41.5	24.7	10.2	25.8
Light XML	42.6	18.9	52.3	34.3	18.2	53.1

Table 6: Detailed table for few-shot results.

Model	Device	Throughput (Inputs/s)	Storage (GB)	P@1 (ZSL)
SEMSUP-XC	1 GPU	46.2	17.9	**36.5**
MACLR	1 GPU	77.8	4.6	29.8
GROOV	1 GPU	8.9	**0.4**	6.0
ZestXML	16 CPUs	**2371**	1.8	15.8

Table 7: Computational Efficiency of SEMSUP-XC and baselines on Wikipedia dataset.

time. During inference, SEMSUP-XC predicts on top 1000 shortlists by TF-IDF, thereby achieving sub-linear time. Further, contextualized tokens for label descriptions are computed only once and stored in memory-mapped files, thus decreasing computational time significantly. Overall, our computational complexity can be represented by $\mathcal{O}(T_{IE} * N + T_{OE} * |Y| + k * N * T_{lex})$, where T_{IE}, T_{OE} represent the time taken by input encoder and output encoder respectively, N is the total number of input documents, |Y| is the number of all labels, k indicates the shortlist size and $|T_{lex}|$ denotes the time in soft-lexical computation between contextualized tokens of documents and labels. In our experiments, $T_{IE} * N >> T_{OE} * |Y|$ and $T_{IE} \approx T_{lex} * k$. Thus effectively, computational complexity is approximately equal to $\mathcal{O}(\mathcal{T}_{\mathcal{IE}} * \mathcal{N})$, which is in comparison to other SOTA extreme classification methods.

Table 7 shows that SEMSUP-XC when compared to other XC state-of-the-art baselines, is computationally efficient in terms of speed while demonstrating much better performance. In terms of storage we utilize almost 4 times storage as compared to MACLR, as we need to store contextualized token embeddings of each label. However the overall storage overhead($\approx$ 17.9GB) is small in comparison to significant improvement in performance and comparable speed. We provide a more detailed analysis of our method in Appendix F

G Qualitative Analysis

We now perform a qualitative analysis of SEMSUP-XC's predictions and present representative examples in Table 8, and compare them to MACLR which is the next best performing model. Correct predictions are in bold. In the first example, even from the short text in the document SEMSUP-XC is able to figure out that it is not just a book, but a *textbook*. While MACLR predicts five labels which are all similar, SEMSUP-XC is able to predict diverse labels while getting the correct label in five predictions. In the second example, SEMSUP-XC smartly realizes the content of the document is a story and hence predictions *literature & fiction*, whereas MACLR tries to predict labels for the contents of the story instead. This shows the nuanced understanding of the label space that SEMSUP-XC has learned. The third example portrays the semantic understanding of the SEMSUP-XC's label space. While MACLR tries to predict labels like *powered mixers* because of the presence of the word *mixer*, SEMSUP-XC is able to understand the text at a high level and predict labels like *studio recording equipment* even though the document has no explicit mention of the words studio, recording or equipment). These qualitative examples show that SEMSUP-XC's understanding of how different fine-grained classes are related and how documents refer to them is better than the baselines considered. We list more such examples in Appendix G.

Input Document	Top 5 Predictions	
	SEMSUP-XC	**MACLR**
Start-Up: A Technician's Guide. In addition to being an excellent stand-alone self-instructional guide, ISA recommends this book to prepare for the Start-Up Domain of CCST Level I, II, and III examinations.	test preparation schools & teaching **new** **used and rental textbooks** software	vocational tests graduate preparation test prep & study guides testing vocational
Homecoming (High Risk Books). When Katey Bruscke's bus arrives in her unnamed hometown, she finds the scenery blurred, "as if my hometown were itself surfacing from beneath a black ocean." At the conclusion of new novelist Gussoff's "day-in-the-life-of" first-person narrative, the reader feels equally blurred by the relentless . . .	**literature & fiction** thriller & suspense thrillers genre fiction general	friendship mothers & children drugs coming of age braille
Rolls RM65 MixMax 6x4 Mixer. The new RM65b HexMix is a single rack space unit featuring 6 channels of audio mixing, each with an XLR Microphone Input and 1/4ünbalanced Line Input. A unique feature of the 1/4l̈ine inputs is they may be internally reconfigured to operate as Inserts for the Microphone Input. Each channel, in . . .	**studio recording equipment** powered mixers home audio **musical instruments** speaker parts & components	powered mixers hand mixers mixers & accessories mixers mixer parts
Political Business in East Asia (Politics in Asia). The book offers a valuable analysis of the ties between politics and business in various East Asian countries..Pacific Affairs, Fall 2003	**international & world politics** **business & investing** **politics & social sciences** **asian** **economics**	international relations policy & current events practical politics international law
Chicago Latrobe 550 Series Cobalt Steel Jobber Length Drill Bit Set with Metal Case, Gold Oxide Finish, 135 Degree Split Point, Wire Size, 60-piece, #60 - #1. This Chicago-Latrobe 550 series jobber length drill bit set contains 60 cobalt steel drill bits, including one each of wire gauge sizes #60 through #1, with a gold oxide finish and a . . .	**drill bits** **twist drill bits** **power & hand tools** **jobber drill bits** **power tool accessories**	industrial drill bits step drill bits long length drill bits reduced shank drill bits installer drill bits
Raggedy Ann and Johnny Gruelle: A Bibliography of Published Works. Patricia Hall has written and lectured extensively on Gruelle and his contributions to American culture. Her collection of Gruelle's books, dolls, correspondence, original artwork, business records and photographs is one of the most comprehensive in the world. Many . . .	**reference** history & criticism humor & entertainment **publishing & books** **research & publishing guides**	art bibliographies & indexes art & photography arts children's literature
Harmonic Analysis and Applications (Studies in Advanced Mathematics). The present book may definitely be useful for anyone looking for particular results, examples, applications, exercises, or for a book that provides the skeleton for a good course on harmonic analysis. R. Brger; Monatsheft fr Mathematik; 127.1999.3	statistics **science & math** professional science **new** **used & rental textbooks**	pure mathematics algebra applied algebra & trigonometry calculus
Soldier Spies: Israeli Military Intelligence. After its brilliant successes in the Six-Day War, the War of Attrition and the campaign against Black September, A'MAN, Israel's oldest intelligence agency, fell prey to institutional hubris. A'MAN's dangerous overconfidence only deepened, Katz here reveals, after spectacular coups such as . . .	israel middle east international & world politics politics & social sciences **history**	intelligence & espionage espionage national & international security middle eastern arms control
SF Signature White Chocolate Fondue, 4-Pound (Pack of 2). Smooth and creamy SF Signature White Chocolate Fondue provides unparalleled flavor in an incredibly easy-to-use product. This white chocolate fondue works better than other fountain chocolates due to its low viscosity and great taste. Packaged in two pound . . .	chocolate breads & bakery pantry staples canned & jarred food **kitchen & dining**	baking cocoa chocolate truffles chocolate assortments baking chocolates chocolate
Little Monsters: Monster Friends and Family (1000). Kind of a demented Monsters, Inc. meets Oliver Twist, the 1989 live-action film Little Monsters takes every child's nightmare of a monster under the bed and spins it into a dark tale of a secret underworld where children and adults turn into monsters and run wild without rules or any . . .	**movies & tv** film musical genres rock **tv**	movies **movies & tv** **tv** film theater
adidas Women's Ayuna Sandal,Newnavy/Wht/Altitude,5 M. adidas is a name that stands for excellence in all sectors of sport around the globe. The vision of company founder Adolf Dassler has become a reality, and his corporate philosophy has been the guiding principle for successor generations. The idea was as simple as it was . . .	girls clothing **sandals** sneakers outdoor	**sport sandals** shoes athletic mountaineering boots **sandals**

Table 8: Examples of class predictions from SEMSUP-XC (our model) compared to MACLR (Xiong et al., 2022). Bold represents correct predictions.