Multilingual Taxonomic Web Page Classification for Contextual Targeting at Yahoo

Eric Ye
Yahoo Research
San Jose, CA, USA
jiayunye@yahooinc.com

Xiao Bai
Yahoo Research
San Jose, CA, USA
xbai@yahooinc.com

Neil O'Hare
Yahoo Research
San Francisco, CA, USA
nohare@yahooinc.com

Eliyar Asgarieh
Yahoo Research
San Jose, CA, USA
eliyar.asgarieh@yahooinc.com

Kapil Thadani
Yahoo Research
New York, NY, USA
thadani@yahooinc.com

Francisco Perez-Sorrosal
Yahoo Research
San Francisco, CA, USA
fperez@yahooinc.com

Sujyothi Adiga
Yahoo
Bangalore, India
sujyothi@yahooinc.com

Abstract

As we move toward a cookie-less world, the ability to track users' online activities for behavior targeting will be drastically reduced, making contextual targeting an appealing alternative for advertising platforms. Category-based contextual targeting displays ads on web pages that are relevant to advertiser-targeted categories, according to a pre-defined taxonomy. Accurate web page classification is key to the success of this approach. In this paper, we use multilingual Transformer-based transfer learning models to classify web pages in five high-impact languages. We adopt multiple data sampling techniques to increase coverage for rare categories, and modify the loss using class-based re-weighting to smooth the influence of frequent versus rare categories. Offline evaluation shows that these are crucial for improving our classifiers. We leverage knowledge distillation to train accurate models that are lightweight in terms of (i) model size, and (ii) the input text used. Classifying web pages using only text from the URL addresses a unique challenge for contextual targeting in that bid requests come to ad systems as URLs without content, while crawling is time consuming and costly. We launched the proposed models for contextual targeting in the Yahoo DSP, significantly increasing its revenue.

CCS Concepts

• **Information systems → Content match advertising**; • **Computing methodologies → Neural networks**.

Keywords

contextual targeting, text classification, knowledge distillation

ACM Reference Format:
Eric Ye, Xiao Bai, Neil O'Hare, Eliyar Asgarieh, Kapil Thadani, Francisco Perez-Sorrosal, and Sujyothi Adiga. 2022. Multilingual Taxonomic Web Page Classification for Contextual Targeting at Yahoo. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539189

1 Introduction

Due to regulations such as GDPR and CCPA along with general privacy concerns, the ability to track historical user behavior for advertising purposes is shrinking. It is no longer possible to serve targeted ads through Audience Targeting (e.g., demographic targeting, behavioral targeting) to users who have opted out. Contextual targeting thus emerges as an important advertising strategy for serving ads that are relevant to the web pages on which they are displayed, providing a unique opportunity for delivering a personalized ad experience to users without tracking their identities (e.g., browser cookies, mobile device ids). The main types of contextual targeting are: *category-based* targeting, where ads targeted to web pages that are relevant to some pre-defined topics, and *keyword-based* targeting, where ads are targeted to web pages containing specific keywords. We focus on category-based contextual targeting and describe taxonomic web page classification models to support Yahoo's contextual targeting business for its global market.

The categories used in this work are drawn from the Yahoo Interest Category (YIC) taxonomy, which consists of 442 categories over 5 tiers. Figure 1 shows a few examples of YIC categories. Since a web page may be characterized by multiple categories across multiple tiers, the categorization task is inherently a multi-label classification problem with a skewed label distribution. There are two main approaches to address such problems [31]. The first approach transforms the problem into a collection of independent binary classifiers trained in a one-vs-rest formulation: this approach was used by our previous production system. The second approach trains a single multi-label model that can predict all categories simultaneously, and well-known models such as SVMs have been extended to support this approach [14]. Recently, transfer learning models [13, 19], especially pre-trained Transformers such as BERT [12] have had great success in improving many natural language processing tasks including text classification, but existing implementations of such models do not support multi-label classification.

In this work, we adapt a popular open-source transfer learning framework to the multi-label classification problem by modifying

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539189

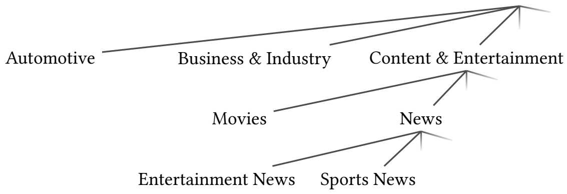


Figure 1: Example categories from the YIC taxonomy

the output layer. We rely on professional editors to annotate web pages with the taxonomic categories that are relevant to their content. We found that a naive attempt at using editorial data to train such models is sub-optimal, and that there are a number of challenges that need to be addressed in order to train accurate models.

First, web pages are highly skewed in terms of categories, with a small number of frequent categories and a long tail. The hierarchical nature of the taxonomy further increases the skew as categories at higher tiers are more frequent than their descendants. This makes it difficult to create a training dataset that represents the entire taxonomy and also leads to large class imbalance. To address this, we adopt sampling strategies that increase coverage for rare categories. We also propose a category-based re-weighting strategy that takes into account two aspects: the weight for positive labels, which are sparse relative to negative labels, as well as the weight for infrequent categories, which are disadvantaged during training.

The second challenge, given the importance of the international market, is the need to classify a large number of non-English pages. We target four high potential non-English languages, and use human annotation and machine translation to collect non-English training data in a scalable way. We train a single multilingual model with data from each language, and use knowledge distillation (KD) techniques to reduce the computational cost of these models while achieving excellent classification accuracy in all languages.

The third challenge, unique to the contextual targeting business, is that bid requests come to ad systems as URLs of the pages on which an ad may be displayed. Although using the full page content leads to higher classification accuracy, accessing page content to enable prediction at ad request time does not meet the SLA of ad systems. Therefore, we selectively crawl web pages and predict their categories off-line using a near real-time stream processing system. However, crawling incurs a considerable cost, and discovering new URLs can be time consuming. To include un-crawled pages in our contextual targeting system, we need a model that can accurately classify web pages based only on text extracted from the URL.

To address these challenges, we propose multilingual Transformer-based models that can classify crawled and uncrawled web pages with respect to the YIC taxonomy. Our main contributions are:

- We adapt Transformer models to multi-label classification by modifying the output classification layer and propose novel class-based loss re-weighting and data sampling techniques to deal with label skew, in the process achieving 37% higher mean average precision than legacy classifiers.
- We extend these models to address multilingual classification for content in 5 target languages, and use knowledge distillation to reduce their computational cost, allowing us to deploy a single small model while improving accuracy over larger language-specific models by at least 4%.
- We propose an accurate classification model solely based on the text in the URL itself, allowing us to significantly increase market share and do real-time topic classification without being restricted by crawling capacity. We distill a large content model to a small URL-only model, achieving a 26% improvement over our legacy XGBoost content model. To the best of our knowledge, this is the first use of real-time category-based contextual targeting using only URL text.
- We deploy the proposed models to support category-based contextual targeting at Yahoo, and show through online metrics how these models positively influence ad delivery. We also explore a novel application of the category-based URL profiles, which improves the revenue of behavior targeting by 0.57% through real-time user interest expansion.

2 Related Work

In this section, we summarize relevant prior literature in a number of areas that are relevant to this work.

Taxonomic Text Classification. Hierarchical Multi-label Classification (HMC) combines hierarchical classification—where categories/labels are organized into a class hierarchy [30]—and multi-label classification (MLC) [4, 31], where every document can be assigned one or more labels. There are two main approaches to HMC [31]; (i) train independent binary classifiers for each category, and (ii) train a single multi-label model that can predict all categories simultaneously, which is the approach we pursue here.

Weight Redistribution. Weight redistribution has been explored in machine learning to correct class imbalance, biased datasets or corrupted labels [7][17]. Recent research aims to extend these ideas to *online* class re-weighting, e.g., by minimizing the loss on a clean unbiased validation set using a meta-gradient descent step on the weights of the current mini-batch [25]. We are not aware of previous work in applying weight redistribution to correct extreme class imbalances in multi-label classification over large taxonomies.

Pre-trained Language Models. In recent years, models based on the Transformer architecture [32], which uses a *self-attention* mechanism, have driven significant advances on a variety of tasks such as language generation, translation, question-answering and classification. Some examples of recent models that build on this architecture include BERT [12], RoBERTa [21], GPT [6, 24] and DistilBERT [28]. These models are pre-trained on a large unsupervised document corpus, and subsequently fine-tuned on a supervised downstream task [12]. We follow this approach and fine tune pre-trained language models for hierarchical multi-label classification.

Multilingual Text Classification. Early work on multilingual document classification typically learns cross-lingual sentence representations using parallel corpora [2, 27, 29], which limits the number of languages that can be classified. Advances in multilingual masked language models, pre-trained with over 100 languages, such as Multilingual BERT [23], XLM [10], XLM-RoBERTa [9] have pushed the state-of-the-art for multilingual text classification tasks in the XNLI benchmark [11]. Different from our work on multi-label taxonomic classification that targets hundreds of classes, most previous work was either on binary classification or multi-class classification tasks with a very limited number of classes.

Knowledge Distillation. Knowledge distillation (KD) is a model compression technique for training smaller student models from larger teacher models without significant loss in accuracy. It has been extensively used across different natural language tasks to decrease the sizes of ever growing neural network models. There are three major categories of KD [15]: response-based [18], feature-based [26], and relation-based [33]. Most knowledge distillation work in the natural language domain falls into the first two categories. In response-based KD, the information from the output layer of the teacher model is used to train the student model. We apply response-based KD in our work and show that this approach significantly improves the performance of smaller models.

Contextual Targeting. Contextual targeting [34] is an advertising strategy that displays ads relevant to the content of a web page. Category-based [5] and keyword-based [3] approaches have been widely adopted in industry. While work on keyword-based approaches focuses on summarizing page content [1, 20] and matching them against ads [3, 5], category-based approaches focus on classifying web pages into a category taxonomy. Hashemi [16] summarizes the major web page classification approaches using text, images or both. To the best our knowledge, our work is the first web page classification model that solely relies on the URL itself.

3 Web Page Categorization for Contextual Targeting

Category-based contextual targeting relies on techniques to classify web pages in terms of categories that reflect user interests. In this section, we formulate the task in terms of multi-label classification into a category taxonomy and describe the adaptation of pretrained transformer encoders like BERT [12] to address it. We also incorporate knowledge distillation to make models compact and practical to serve. Finally, we extend these models to support for multiple languages and the ability to classify web pages without crawling their content, by using only information from the URL.

3.1 Taxonomic Categories

Categories for contextual targeting must be broad enough to apply to diverse web pages, but specific enough to capture meaningful user interests. A tree-structured taxonomy is a natural choice for organizing such categories, with specific or niche interests grouped under more general ones. In this work we use the Yahoo Interest Categories (YIC) taxonomy, which contains 442 interest categories: 12 tier-1 categories, 100 tier-2 categories, 259 tier-3 categories, 66 tier-4 categories and 5 tier-5 categories. Figure 1 shows a few examples of category names with their paths, and the full taxonomy is visible in the advertiser interface of the Yahoo Ad Platform.

The hierarchical structure implies that a web page assigned to any category (e.g., "Content & Entertainment/News") would also be categorized to its ancestor categories (e.g., "Content & Entertainment"). A page may also be described by multiple categories (e.g., a car blog by "Automotive" and "Content & Entertainment/News"), making the category-based contextual targeting scenario a *multi-label* taxonomic classification problem, different from the more common multi-class setting where classes are mutually exclusive.

3.2 Hierarchical Multi-label Classification

In recent years, text classification benchmarks have been dominated by Transformer-based architectures pretrained on large text corpora, such as BERT [12] and RoBERTa [21]. These models typically have a softmax output layer and are fine-tuned using cross-entropy loss, which is well-suited for multi-class classification. For multi-label classification we replace the output layer with a sigmoid activation for each category, allowing output units to learn binary classifiers independently of other units, although all units share their input representation from preceding Transformer layers.

Given a single category c, the model is effectively trained as a standard binary classifier. A binary cross-entropy loss can be defined over all N training examples $x_1, \ldots, x_N$, where each x_i has a corresponding binary label $y_{i,c} \in \{0, 1\}$ indicating whether it belongs to category c.

$$L_c = \frac{1}{N} \sum_{i=1}^{N} y_{i,c} \log \hat{y}_{i,c} + (1 - y_{i,c}) \log(1 - \hat{y}_{i,c}) \tag{1}$$

where $\hat{y}_{i,c} \in [0, 1]$ indicates the real-valued activation of the sigmoid corresponding to category c when x_i is provided as input.

When training multiple categories simultaneously, we introduce category-specific weights w_c for one of the terms in the loss. The final loss of the network is thus a weighted average of N per-instance losses summed over all C categories.

$$L = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N} w_c \left(y_{i,c} \log \hat{y}_{i,c} \right) + (1 - y_{i,c}) \log(1 - \hat{y}_{i,c}) \tag{2}$$

These category weights w_c serve an important role. For hierarchical multi-label classification, many rare categories may have very few positive examples in the data. This imbalance would mean that, without re-weighting, rare classes will have very little influence on the loss. Additionally, the loss function incorporates a separate binary cross entropy loss for every category as shown above in equation (2). Since, for a given example, most of these categories will be negative (i.e., $y_{i,c} = 0$), it is easy for this loss to become dominated by the negative labels and for the classifier to converge on a trivial classifier which makes negative predictions for *all* categories.

We propose a re-weighting strategy that allows us to simultaneously balance the loss between classes (i.e., amplify the influence of rare classes) and change the global influence of positive versus negative labels. For negative labels (i.e., $y_{i,c} = 0$), the weight is implicitly always 1, while for positive labels we propose a weighting function that increases the influence of rare classes and limits that of frequent classes. We do not strictly enforce *equal* influence but define a smoothing factor to control the amount of re-weighting that is applied. Specifically, the weight w_c is defined as

$$w_c = \mu \frac{\max_k f_k + \alpha}{f_c + \alpha} \tag{3}$$

where f_c denotes the frequency of category c in the training data,

$$f_c = \sum_{i=1}^{N} y_{i,c} \tag{4}$$

α determines the degree of class-based re-weighting, and μ is a constant multiplier that controls the overall influence of positive versus negative labels. As α approaches ∞, all categories will have the same w_c, and class-based weighting will not be in effect. If α is 0, the loss will be perfectly balanced to ensure equal influence from each category. Finally, since the influence of α is sensitive to the corpus size N, we compute it as a function of N:

$$\alpha = \gamma \times N \tag{5}$$

where γ is a scale-free smoothing factor that can be tuned.

3.3 Knowledge Distillation

The accuracy of BERT-like models typically improves significantly as the model size—inner dimensionality, number of layers, number of self-attention heads—is increased. Larger models, however, require significantly greater computational resources to serve and are consequently impractical for many large-scale classification tasks. We address this limitation through knowledge distillation [18], a framework in which the predictions of a large model (the *teacher*) are used to train a lightweight distilled model (a *student*).

Distillation can be accomplished using a variety of techniques. In our work, knowledge is transferred to the student model through a large dataset of unlabeled examples. Given a trained teacher model and an example x_i from this dataset, the real-valued predictions $\hat{y}_{i,c} \in [0, 1]$ from the teacher for each category c are recorded as *soft labels* for training the student model. These soft labels communicate additional information to the student about the richer teacher model and contribute to improved accuracy and generalization [22]. The loss function for training the student follows the description in Section 3.2, with the change that the labels $y_{i,c}$ in equations (2) and (4) are now real-valued and lie within $[0, 1]$.

Most previous work on knowledge distillation has been in the multi-class setting, where temperature scaling is applied to smooth the predicted label distribution from the teacher to ensure that predictions for classes outside of the sole positive class have non-trivial values. In the multi-label classification setting —where predictions are not normalized over all classes—the benefit of this smoothing is not clear. In preliminary experiments, we found that temperature of 1 (i.e., no scaling) performed best, and consequently we use the teacher predictions directly without any scaling.

We use this process to distill large models like XLM-RoBERTa-Large (355M parameters) into models like XLM-RoBERTa-Base (125M parameters) or smaller, resulting in more computationally efficient inference at scale. In addition, we also use knowledge distillation to reduce inference latency by classifying web pages *without* first crawling them, as described in the following section.

3.4 Categorizing without Crawling

BERT-based classifiers require a sequence of tokens as input and typically support two segments of input text separated by a [SEP] token. To classify a web page given its URL, we crawl the HTML and extract the page `title` and `body`, stripping HTML tags, white space and special characters. In addition, we parse the URL and extract the domain and path as additional tokens to use in the input. The URL domain, path and page title are designated as the first segment while the page body constitutes the second segment.

However, crawling to extract page content can add significant latency, in addition to consuming a lot of resources, when running contextual targeting at scale. Moreover, we observe that there is often sufficient information in URL domains (e.g., `news.yahoo.com`) and paths (e.g., `/sports/football`) to produce reasonable categories. We use the knowledge distillation approach from Section 3.3 to distill a teacher model trained on page content and URL text into a student model that is provided only a URL. This has the effect of establishing an association between URL tokens with categories in the taxonomy, so even URL tokens that are not clearly linked to a category (e.g., `vox.com`) can be predictive after distillation.

3.5 Multilingual Classification

When considering pages crawled on the web, their content may be in languages other than English. Neural language models can be extended to support multiple languages with modifications to pretraining [9, 10], allowing contextual targeting to be scaled to new regions with no additional considerations for modeling other than obtaining multilingual data for fine-tuning. Using translations of web page content to expand our dataset, we investigate the use of multilingual models that support four additional languages: Spanish, French, Portuguese and Traditional Chinese.

4 Methodology

In this section we describe our approach to building a corpus for taxonomic multi-label classification, in addition to describing the methodology that we use for offline evaluation of these models.

4.1 Corpus Development

4.1.1 *English Language Corpus.*

For our initial work on corpus development, we collected a traffic-based stratified sample of English language bid request URLs from the Yahoo Demand-side Platform (DSP) during a 6-month period from Jan 2020 to July 2020. All data labelling was carried out internally by an in-house editorial team. The long tail category distribution, however, means that a random sample of bid request pages has very low coverage of torso and tail categories. For example, 27 categories have no labelled pages in a 15k random sample, while 114 categories have less than 5 labelled pages in the same sample. For this reason, we adopt two targeted sampling approaches to address this problem: (i) URL collection, and (ii) active learning.

URL Collection. Since many of the target categories have zero or very few labelled samples in our initial random sample, we did not have data to bootstrap a model that could be used to assist with data collection. For this reason, in collaboration with the Yahoo editorial team, we used a method that we refer to as *URL Collection* or *UC*, where the editorial team is given a set of categories and asked to find URLs from diverse websites that are relevant to those categories. After these candidate URLs have been collected, they are then fully annotated with respect to additional taxonomy categories that they are relevant to. Although this is clearly a biased form of data collection, since data does not come from the population of bid request URLs, our results will demonstrate that this is nevertheless a very useful way to bootstrap models for rare categories.

Active Learning. Active learning is a method for using model predictions to sample documents for annotation. After first bootstrapping data with URL Collection, we can train initial models for tail and torso categories. To gather additional candidate pages for these rare categories, we adopt the simple approach of sampling

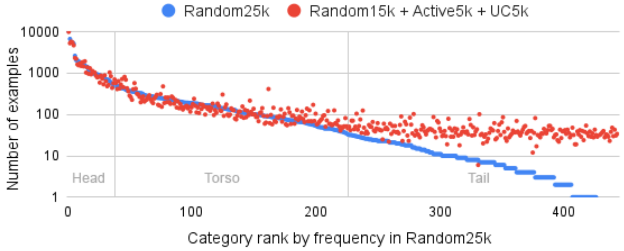


Figure 2: Pages per category in a random sample and in a combined sample with URL Collection and active learning.

pages for which the model score is higher than a threshold, and these pages are then manually labelled by our editorial team.

During corpus construction, this process of random sampling followed by targeted sampling was iterated a number of times, each time based on a recent 6-month stratified sample. Figure 2 shows how this improves coverage for tail categories.

4.1.2 Non-English Corpus.

In addition to English, we identified 4 other target languages for bid request page classification: Spanish, French, Portuguese and Traditional Chinese. 28k of the English language documents were automatically translated into each of the 4 target languages using the Google Translate API.[1] In addition, 29.2k bid request pages per language were sampled using a mix of stratified random sampling, active learning and URL Collection per language for each of the 4 non-English target languages. These documents were annotated with YIC taxonomy labels by an editorial team.

4.1.3 Corpus Partitioning.

The corpus was partitioned for each target language as follows:

- *Training Set,* containing a mix of data sampled randomly, by URL Collection, and by active learning, as described above.
- *Development Set,* used to make initial decisions on optimal hyperparameters and model selection via early stopping. This data is a random subset of the stratified random sample of DSP bid request URLs.
- *Test Set:* held-out dataset, which serves as a gatekeeper that determines whether the model can be deployed in production. If the model passes the overall quality requirements agreed upon, it is put into production. Like the development set, this is a random subset of the stratified random sample.

Corpus statistics for each target language can be found in Table 1.

4.2 Implementation & Evaluation

Evaluation Metrics and Data. We primarily rely on mAP (Mean Average Precision)—which is computed as the mean of the average precision for each category—as a single metric that allows us to compare the fine-tuned models in a threshold-independent manner. All models were trained over the entire set of 442 categories. Due to the skewed category distribution in the random test set, we did not have labelled examples available for all categories. For our evaluation, we only calculate mAP for categories that are *testable,* defined as any category with at least one positive test example. Table 1 shows the number of testable categories for each language.

Table 1: Corpus statistics per language considered.

Language	Train docs	Dev docs	Test docs	Testable categories
English	56k	5k	13k	391
Spanish	48k	1.2k	7k	378
French	48k	1.2k	8k	384
Portuguese	48k	1.2k	8k	380
Traditional Chinese	48k	1.2k	8k	387
Total Non-English	192k	-	-	-
Total	248k	-	-	-

Table 2: mAP for various weighting strategies for English web page classification with RoBERTa-Large.

Model	5 epochs	80 epochs
XGBoost	0.337	
Transformer models		
No re-weighting	0.326	0.450
Positive-class weighting	0.435	0.460
Class-based weighting	0.440	0.462

Hyperparameter Optimization and Model selection. All hyperparameters were optimized using grid search by selecting the values that corresponded to the optimal mAP on the development set. For each training run, early stopping was used to select the best model, again according to mAP on the development set. Unless otherwise stated, models were trained for 80 epochs.

Implementation. We use the Hugging Face `Transformers`[2] library, which contains open-source implementations of a large number of models including BERT, RoBERTa, DistilBERT and XLM-RoBERTa. We modified the original code to support the multi-label output layer and loss re-weighting scheme described in Section 3.2. Models based on full-content input use a maximum sequence length of 512 tokens; URL-only models use a maximum sequence length of 128 tokens. Experiments were conducted using eight NVidia V100 GPUs with 32GB of RAM/GPU and two Intel Xeon 2.6GHz CPUs, with a total of 64 cores and 640GB of RAM.

5 Results

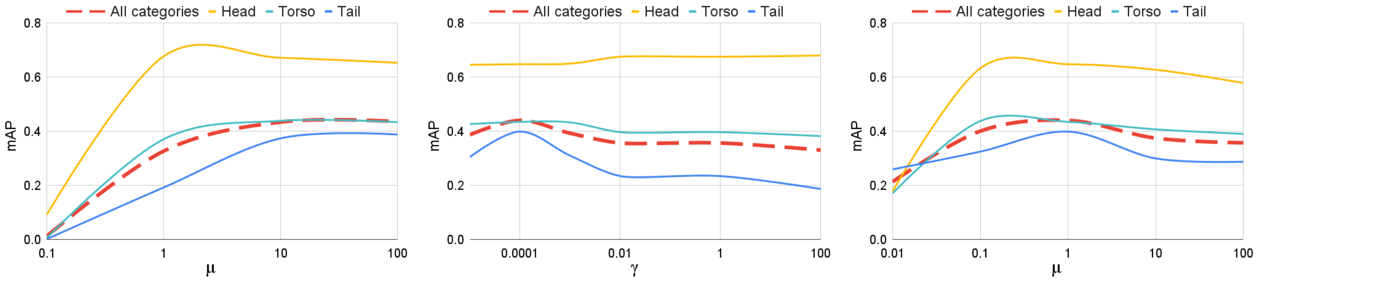
In this section, we present offline evaluation results for our models, using the test corpus described in the previous section.

5.1 Loss Re-Weighting

Since preliminary results showed that RoBERTa-Large [21] models outperformed similarly sized BERT models, consistent with published results, we use RoBERTa-Large as our English language in this section. Table 2 shows baseline results for RoBERTa-Large models evaluated on the entire set of 391 *testable categories* on the English language test set. These models are compared against a unigram+bigram XGBoost [8] model trained on the same data. We use XGBoost as a baseline here as this was used as our legacy production model, and XGBoost has been shown to perform well on a variety

[1] https://translate.google.com

[2] https://huggingface.co/docs/transformers/index



(a) Varying μ with no per-class re-weighting ($\gamma = \infty$) (b) Varying γ with no positive weight ($\mu = 1$) (c) Varying μ at optimal $\gamma = 0.0001$

Figure 3: Analysis of the effect of positive class weight μ and smoothing factor γ with RoBERTa-Large.

of tasks compared with other decision tree models [8]. Optimal values for μ and γ are selected based on mAP on the development set. The results for models trained for 5 epochs show the benefit of loss re-weighting, with a vanilla implementation achieving an mAP of 0.326, which is worse than the baseline XGBoost model, and far behind that of the best mAP of 0.44, when re-weighting is used. Interestingly, *positive class weighting* achieves results almost on par with the more sophisticated *class-based weighting*. Although the early work on BERT models reported task-specific fine tuning results for a small number of epochs [12], our preliminary results suggested that training for longer can lead to significant improvements. The results in Table 2 show that the mAP for the best model improves significantly, from 0.440 to 0.462, with longer training. Longer training also appears to make these models more robust to the choice of loss, as the difference between the unweighted vs weighted models is much smaller in this case. Overall, apart from the 5-epoch RoBERTa model with no re-weighting, all models show dramatic improvements over the baseline XGBoost model.

Figure 3 examines the impact of the re-weighting hyperparameters μ and γ in more detail for models trained for 5 epochs. In addition to reporting mAP averaged over all categories, for this analysis we split the taxonomy into head, torso and tail categories based on the number of training samples per category in a 25k stratified random sample, as follows: (i) *head*: categories with >=500 examples, (ii) *torso*: categories with >=30 and <500 examples, (iii) *tail*: categories with <30 examples. Figure 3a examines the effect of the positive weight factor μ without per-class re-weighting: while a default value of 1 is optimal for the head categories, this leads to poor performance on tail categories and sub-optimal performance on torso categories, with an optimal value for torso/tail at around $\mu = 10$. Figure 3b examines the effect of the smoothing factor when the positive labels have no additional weight ($\mu = 1$). For larger values of the γ hyperparameter (i.e.*less* re-weighting), we see that torso/tail categories perform poorly, with torso/tail performance improving as re-weighting is applied by reducing γ. Figure 3c shows the effect of changing μ when an approximately optimal smoothing factor is used: while the trend is the same, the interaction of the two hyperparameters leads to a lower optimal value for μ.

5.2 Evaluation of Targeted Data Collection

In this section, we examine the effect of the targeted data sampling approaches described in Section 4. For these experiments, we train RoBERTa-Large models using various subsets of the training set, comparing 15k, 20k and 25k documents composed of various combinations of random, URL Collection and active learning data. The results in Table 3 show a moderate improvement across all segments when adding additional random data to the 15k dataset, while URL Collection data leads to a massive improvement for tail categories and a modest improvement for torso categories. Although this is expected, as URL Collection specifically targets these categories, the results demonstrate that this strategy, although biased, is an effective way of bootstrapping a model for rare categories.

Table 3: mAP for RoBERTa-Large on English web page classification using various sampling methods. Random - Random sample. UC - URL Collection. Active - Active learning.

Sampling Strategy	#samples	All	Head	Torso	Tail
Random15k	15k	0.390	0.652	0.439	0.269
Random20k	20k	0.397	0.659	0.450	0.271
Random15k + UC5k	20k	0.447	0.652	0.452	0.394
Random25k	25k	0.401	0.655	0.451	0.282
Random20k + UC5k	25k	0.445	0.647	0.448	0.393
Random15k + Active5k + UC5k	25k	0.452	0.649	0.448	0.413

Having trained a model with URL Collection data, we have enough data to bootstrap models for torso/tail categories, enabling the use of active learning for sampling those categories. The final 3 rows of Table 3 show the impact of further increasing the training corpus size, demonstrating that active learning sampling gives additional improvements over URL Collection for tail categories.

5.3 Evaluation of Multilingual Models

For multilingual classification, we train models on the multilingual dataset described in Section 4.1.2 using XLM-RoBERTa-Large, which has been shown to be state-of-the-art for multilingual NLP tasks [9]. We then use knowledge distillation to reduce the model size. The results in Table 4, which compare XLM-RoBERTa-Large models trained with human annotated editorial data, translated data and both, show that training a non-English classifier purely based on translated data performs competitively with training data directly annotated in the target language. These results are very encouraging as using automatically translated data not only increases our training data without additional human effort, but also allows us to benefit from English-language URL Collection and

Table 4: mAP for multi-lingual models, with various sources of training data. E - Editorial data. T - Translated data.

Model	Train Language	en	es	fr	pt	zh-tw
Monolingual						
RoBERTa-Large	E:en	0.460	-	-	-	-
XLM-R-Large	E:en	0.458	-	-	-	-
Multilingual						
XLM-R-Large	E:en + T:es/fr/pt/zh-tw	0.447	0.544	0.519	0.517	0.485
XLM-R-Large	E:en/es/fr/pt/zh-tw	0.468	0.555	0.552	0.536	0.532
XLM-R-Large	E+T:es/fr/pt/zh-tw	-	0.550	0.529	0.536	0.520
XLM-R-Large	E+T:en/es/fr/pt/zh-tw	0.474	0.577	0.557	0.560	0.543

active learning strategies for increasing coverage of tail categories. Combining translated and directly annotated data gives the best results for all languages, including English, improving mAP by 1.3% for English, 4.0% for Spanish, 0.9% for French, 4.5% for Portuguese and 2.1% for Traditional Chinese.

Surprisingly, these results also show that the mAP of the multilingual model trained with data from 5 languages evaluated on English not only matches the English-only models, it achieves a relative improvement of 3.0% over the English-only RoBERTa-Large model and 3.5% over an English-only XLM-RoBERTa-Large model. Similarly, Table 4 indicates that adding English data while fine-tuning the XLM-RoBERTa-Large multilingual model significantly improves mAP for all the non-English languages.

Knowledge Distillation. The results above show that it is possible to achieve promising results using large Transformer models such as XLM-RoBERTa-Large. These models are too computationally expensive to directly deploy in our production environment, however, so it is necessary to train smaller models. We choose XLM-RoBERTa-Base as a student model as it is the smaller sibling of the multilingual XLM-RoBERTa-Large teacher model.

The results of our distillation experiments are presented in Table 5. Standard training of smaller models with editorial labels (XLM-R-Base) leads to performance that is much worse than the large model (XLM-R-Large). We compare this model against XLM-R-Base KD models trained for 1M steps using soft labels from editorial data, randomly sampled data, or a combination of both. First, training directly with soft labels from the editorial data does, as expected, lead to an increase in mAP. Introducing random data labelled by the teacher model for distillation leads to further performance gains, matching or exceeding the teacher model, likely because the label distribution is better represented by this data. However, when mixing editorial and unlabelled random training data, increasing the proportion of random data does not appear to significantly affect performance. This is also observed when we distill with *only* the larger random dataset, removing editorial data entirely.

5.4 Classification of Uncrawled URLs

Due to the limited capacity of our web crawlers, there is a significant number of web pages for which content is not available: we rely solely on tokens from the URL to classify these pages. Similarly to content-based models, we train DistilBERT and XLM-R-Base models using URL tokens as the input. We also train an XLM-R-Large URL-only model to be used as a teacher model for knowledge distillation. In addition, we train a distilled model that uses the best content-based teacher model as the teacher, enabling us to investigate if a model trained with impoverished input (URL-only) can benefit from the knowledge contained in a model trained with richer input (URLs and page content).

Table 5: Results for knowledge distillation of multilingual models. mAP for various KD training datasets with soft labels extracted from editorial labeled data (E) and additional random unlabeled data (R). Numbers in the #R column are per language: e.g. 50k × 5 indicates 50k each for 5 languages.

Model	#E	#R	en	es	fr	pt	zh-tw
Trained with binary editorial labels							
XLM-R-Base	248k	-	0.440	0.542	0.520	0.529	0.502
XLM-R-Large	248k	-	0.474	0.577	0.557	0.560	0.543
Distilled from XLM-R-Large							
XLM-R-Base KD	248k	-	0.468	0.585	0.556	0.566	0.551
XLM-R-Base KD	248k	50k × 5	0.480	0.588	0.555	0.569	0.552
XLM-R-Base KD	248k	600k × 5	0.483	0.581	0.564	0.570	0.548
XLM-R-Base KD	0	600k × 5	0.485	0.585	0.564	0.569	0.549

The results in Table 6 show that URL-only inference using a model trained with content yields poor performance in terms of mAP, motivating the need for separate URL-only model for classifying uncrawled pages. And, indeed, a model trained specifically with URL-only input shows significant improvement. We can also see that for URL-only models trained with editorial data, models of larger size lead to higher mAP, although these models are significantly outperformed by the best content models. More importantly, when using knowledge distillation to train a URL model with soft labels generated by an XLM-RoBERTa-Large model (either using page content or URL only), a base-sized model can achieve higher mAP than a large model trained with editorial data. We again observe that soft labels from a randomly sampled dataset play a critical role in improving the mAP. Finally, the teacher model trained with content trains a much better student model compared with a URL-only trained teacher model. A possible reason is that this model makes more accurate classifications and thus generates higher-quality soft labels. Overall, the best distilled multilingual URL-only model achieves a relative improvement in mAP of 13.1% for English, 8.1% for Spanish, 10.7% for French, 11.2% for Portuguese, and 11.2% for Traditional Chinese, compared with a traditionally trained XLM-R-Large model, while making the model much less computationally expensive for deployment. It also achieves a 26% improvement in mAP over our previously-deployed XGBoost categorization model.

6 System Implementation and Product Impact

6.1 System Implementation

We built a Spark Streaming inference pipeline on AWS to classify the incoming ad request URLs in Yahoo's DSP using the proposed multilingual distilled XML-RoBERTa base models for crawled and uncrawled web pages. The URL profiles that consist of a list of contextual targeting segments corresponding to the predicted categories for each URL are written to a key-value store for real time lookup at ad serving time. The Spark Streaming component runs in an AWS EMR cluster and consumes input data from AWS Kafka. The EMR cluster hosts rely on docker images to run model inference for the pages that newly appear in AWS Kafka's streaming

Table 6: mAP for URL models. E - Editorial corpus. R - Randomly sampled unlabeled data (600k). KD$_u$ - distillation from teacher model trained with URL input. KD$_c$ - distillation from teacher model trained with URL + page content.

Model	Training Data	en	es	fr	pt	zh-tw
Content Model (Tested with URL+Content input) - binary editorial labels						
XLM-R-Large	E	0.474	0.577	0.557	0.560	0.543
Content Model (Tested with URL-only input) - binary editorial labels						
XLM-R-Large	E	0.309	0.380	0.381	0.368	0.306
URL Models trained with binary editorial labels						
DistilBERT (en)	E	0.353	-	-	-	-
XLM-R-Base	E	0.345	0.433	0.415	0.430	0.351
XLM-R-Large	E	0.374	0.470	0.449	0.455	0.376
URL Models distilled from XLM-R-Large						
XLM-R-Base KD$_u$	E	0.363	0.450	0.438	0.442	0.363
XLM-R-Base KD$_c$	E	0.385	0.477	0.453	0.467	0.386
XLM-R-Base KD$_u$	E+R	0.382	0.479	0.460	0.464	0.374
XLM-R-Base KD$_c$	E+R	0.423	0.508	0.497	0.506	0.418

source. The categories assigned to a web page are filtered based on pre-defined per-category thresholds to only store categories with high confidence. The category list for a web page is then extended to include all the ancestors of each predicted category according to the YIC taxonomy structure. The final results are then written to the key-value store using AWS Kafka.

6.2 Impact on Contextual Targeting

The taxonomic web page classification models described above are developed to support contextual targeting in the Yahoo DSP. As described in Section 6.1, a URL and its predicted categories are precomputed using one of our multilingual distilled XLM-RoBERTa-Base models, and the results are stored in a key-value store for real-time lookup at ad serving time. Once a bid request, requesting an ad on a publisher web page, comes to our DSP, any ads that target at least one of the URL's categories through contextual targeting (in addition to the ads that are eligible through other targeting strategies) are eligible for the ad auction.

In this section, we report the performance of three representative model launches in production for contextual targeting during the course of building the product. For each model, we apply a per-category confidence threshold so that the expected precision is at least 0.8. We measure the contribution of contextual targeting to the entire Yahoo DSP before and after a model launch for impressions, clicks and revenue. These contributions are measured 15 days before and 15 days after each launch date, and the relative improvement for each metric is summarized in Table 7. Relative changes are computed to de-emphasize temporal effects.

The first launch replaced the production XGBoost model with our Transformer-based model for crawled English web pages. The XGBoost model consists of one binary classifier for each category, trained using words and Wiki entities from the web page content as features. We observe that the Transformer-based model increased the contribution of Contextual Targeting to DSP by 56% for impressions, 17% for clicks, and 53% for revenue.

One major contribution of this work is a distilled Transformer-based model that can accurately classify uncrawled web pages solely based on tokens from the URLs. When we launched this model into production for English pages, in addition to the model that only classifies crawled web pages, the contribution of contextual targeting to DSP impressions increased by 257%. Given that a significant fraction of web pages do not have their content available for analysis, and therefore could not previously be classified, this launch enabled classification of a far greater number of documents, and so greatly increased the impact of contextual targeting.

Table 7: Post launch metrics for contextual targeting. The relative improvement for the contribution percentage of contextual targeting to DSP before and after launch are reported.

Post launch coverage		Relative change w.r.t. pre-launch		
Market	Uncrawled	Impression	Click	Revenue
en	No	+56%	+17%	+77%
en	Yes	+257%	+194%	+353%
en/es/fr/pt/zh-tw	Yes	+37%	+31%	+33%

The third launch involved the two distilled multilingual models that extend our contextual targeting solution to crawled and uncrawled web pages in Spanish, French, Portuguese, and Traditional Chinese. As expected, compared to the earlier models that only classify English pages, this launch increased the contribution of Contextual Targeting to DSP by 37% for impressions, 31% for clicks, and 33% for revenue. Together, these post-launch metrics show that each of these launches, based on model variations described in this paper, contributed significantly to the growth of contextual targeting within the DSP platform at Yahoo.

6.3 Impact beyond Contextual Targeting

In addition to contextual targeting, interest-based audience targeting is another prevalent way of showing personalized ads to users. Different from contextual targeting, which targets users solely based on the interest categories derived from the current web page being viewed, interest-based targeting usually derives users' interest categories based on their historical interactions with the web (e.g., browsing, search, purchase, etc.). Regardless of the content of the current page being viewed, ads that are relevant to a user's historical interest categories are eligible to be shown on the page. If a user is viewing a web page whose categories are not yet part of her historical interests, and an advertiser is not opted in for contextual targeting, we may miss the opportunity to show this user relevant ads that target these categories. Therefore, we use the categories of the current web page for *real-time user interest expansion* for interest-based Yahoo Audience Targeting.

We run an A/B test to evaluate this new feature in Yahoo Native. In the control bucket, only the ads targeting a user's historical interest categories are eligible for the ad auction. In the test bucket, at ad serving time, if some of the contextual categories assigned by our model to the web page being viewed do not belong to a user's historical interest categories, these categories are also considered as the user's interest categories and used to qualify additional ads for the auction. Although these categories eventually become the user's historical interest categories when the user returns to our system in the future, the real-time user interest expansion feature allows user interests to be captured in real time. Both the control and test buckets ran at 20% of the entire traffic for 10 days. We observe that using the real-time user interest expansion increases CPM (Cost-Per-Mille, i.e., ad platform's revenue for one thousand ad impressions) by 0.57%. The CPM increase is more significant (1.30%) for users without any historical interest categories. In addition to increasing Yahoo Native's revenue, CPA (Cost-per-Acquisition), i.e., advertiser spend for a pre-defined conversion such as purchase or subscription, reduces by 1.27%. This implies that ads in the test bucket are better aligned with user's instant interests, leading to better return on investment for advertisers. After this successful A/B test, the real-time user interest expansion for interest-based Yahoo Audience Targeting was launched in production.

7 Conclusion

In this paper, we addressed the problem of hierarchical multi-label classification as used in category-based contextual targeting at Yahoo. We proposed for the first time a multilingual model that can accurately classify web pages into a hierarchical taxonomy (specifically, the Yahoo Interest Categories taxonomy) without crawling their content. The proposed models are fully launched to support contextual targeting in the Yahoo DSP, in addition to supporting real-time user interest targeting. We discussed a number of practical lessons learned from our experimentation: (i) URL Collection (i.e. tasking editors with actively searching for web pages relevant to rare categories) is critical to bootstrap models for torso/tail categories, which further enables the use of active learning sampling, to address the skewed category distribution; (ii) class-based loss re-weighting is important to improve classification accuracy for rare categories; (iii) knowledge distillation allows us to train lightweight and more accurate models, especially when page content is not crawled for URLs; (iv) augmenting multilingual data through machine translation significantly improves the classification accuracy for both English and non-English pages.

ACKNOWLEDGMENTS

We would like to thank all the anonymous editors from the Yahoo editorial team, led by Samantha Giordano, Katy Kuo and Karen Wu, to evaluate the datasets used in this work.

References

[1] Aris Anagnostopoulos, Andrei Z. Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. 2011. Web Page Summarization for Just-in-Time Contextual Advertising. *ACM Transactions on Intelligent Systems and Technology* 3, 1, Article 14 (2011).

[2] Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics* 7 (2019), 597–610.

[3] Aishwary Bahirat. 2022. Contextual Recommendations Using NLP in Digital Marketing. In *Proceedings of Sixth International Congress on Information and Communication Technology*. 655–664.

[4] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems*. 730–738.

[5] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. 2007. A semantic approach to contextual advertising. In *Proceedings of SIGIR*. 559–566.

[6] Tom B. Brown, Benjamin Pickman Mann, Nick Ryder, Melanie Subbiah, Jean Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krüger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric J Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165. Retrieved from https://arxiv.org/abs/2005.14165.

[7] 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 (2002), 321–357.

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*. 8440–8451.

[10] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, Vol. 32.

[11] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of EMNLP*.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*. 4171–4186.

[13] Chuong B. Do and Andrew Y. Ng. 2005. Transfer learning for text classification.. In *Advances in Neural Information Processing Systems*. 299–306.

[14] Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative Methods for Multi-labeled Classification. *Advances in Knowledge Discovery and Data Mining* (2004), 22–30.

[15] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

[16] Mahdi Hashemi. 2020. Web page classification: a survey of perspectives, gaps, and future directions. *Multimedia Tools and Applications* 79 (2020), 11921–11945.

[17] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise. In *Advances in Neural Information Processing Systems*, Vol. 31.

[18] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. arXiv:1503.02531. Retrieved from https://arxiv.org/abs/1503.02531.

[19] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL*.

[20] Pengqi Liu, Javad Azimi, and Ruofei Zhang. 2014. Automatic Keywords Generation for Contextual Advertising. In *Proceedings of the 23rd International Conference on World Wide Web*. 345–346.

[21] 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. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

[22] Mary Phuong and Christoph H. Lampert. 2021. Towards Understanding Knowledge Distillation. arXiv:2105.13093. Retrieved from https://arxiv.org/abs/2105.13093.

[23] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *Proceedings of ACL*. 4996–5001.

[24] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[25] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to Reweight Examples for Robust Deep Learning. In *Proceedings of ICML*.

[26] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. arXiv:1412.6550. Retrieved from https://arxiv.org/abs/1412.6550.

[27] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research* 65 (2019), 569–630.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. Retrieved from https://arxiv.org/abs/1910.01108.

[29] Holger Schwenk and Xian Li. 2018. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of LREC*.

[30] Carlos N Silla and Alex A Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22, 1-2 (2011), 31–72.

[31] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3, 3 (2007), 1–13.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[33] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1285–1294.

[34] Kaifu Zhang and Zsolt Katona. 2012. Contextual advertising. *Marketing Science* 31, 6 (2012), 980–994.