Extreme Multi-Label Text Classification

New and a more realistic dataset for XTMC tasks

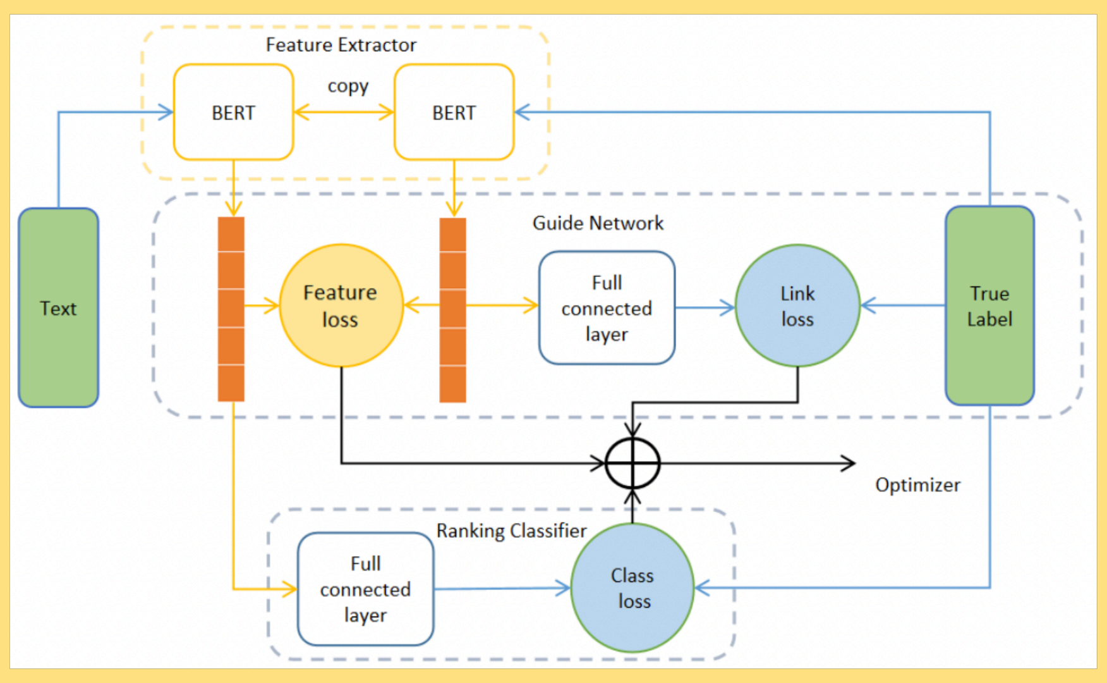
Team: Pinchas Gutter

USC Viterbi School of Engineering

Introduction

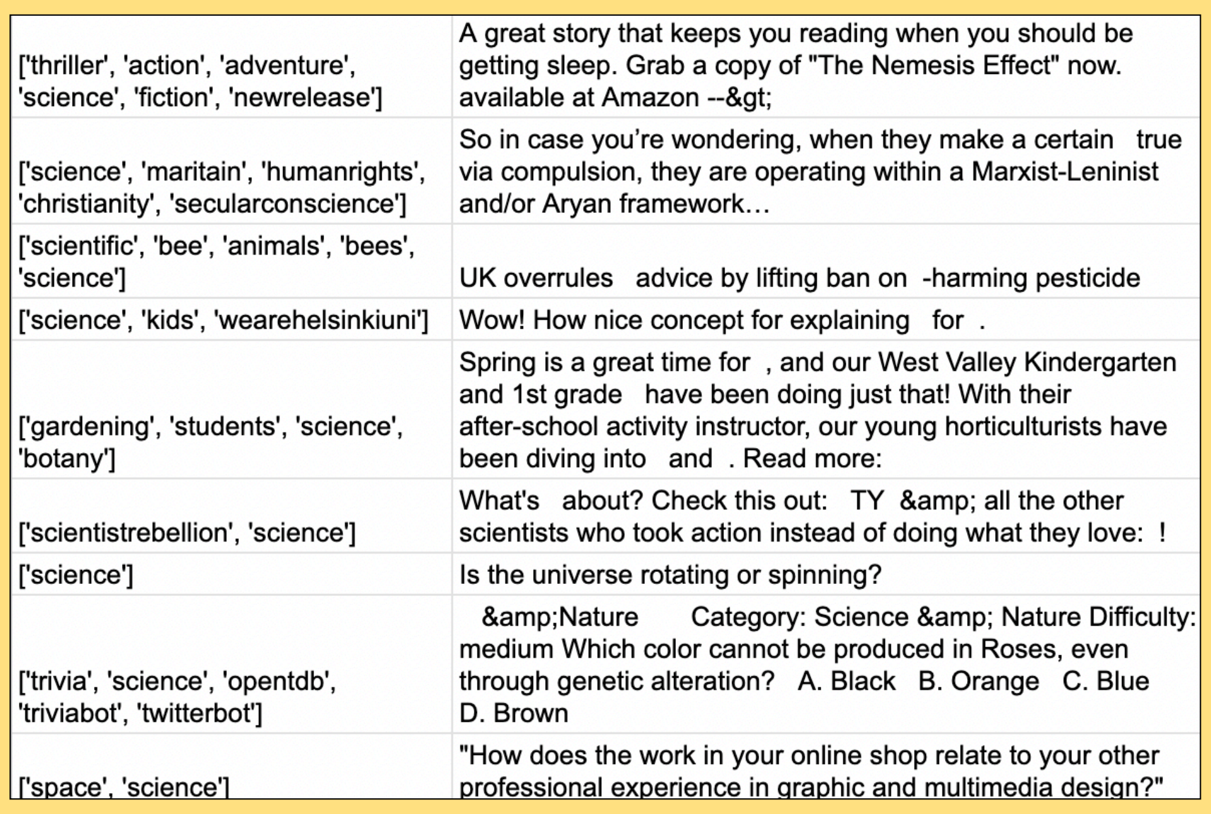
XMTC is a multi-label text classification problem in which the labels can scale up to millions. GUDN by [Wang et al., 2022] have achieved the state-of-the-art performance in XTMC task using EURLEX-4K dataset. However, this dataset is very verbose and domain-specific which results in major overlap between between the labels and the text.

To address this problem we came up a new dataset called USC-31k. Twitter is a potential data source for XTMC problem since the tweets are limited in length and they cover variety of domains. USC-31k is a dataset of tweets as texts and hashtags as labels.



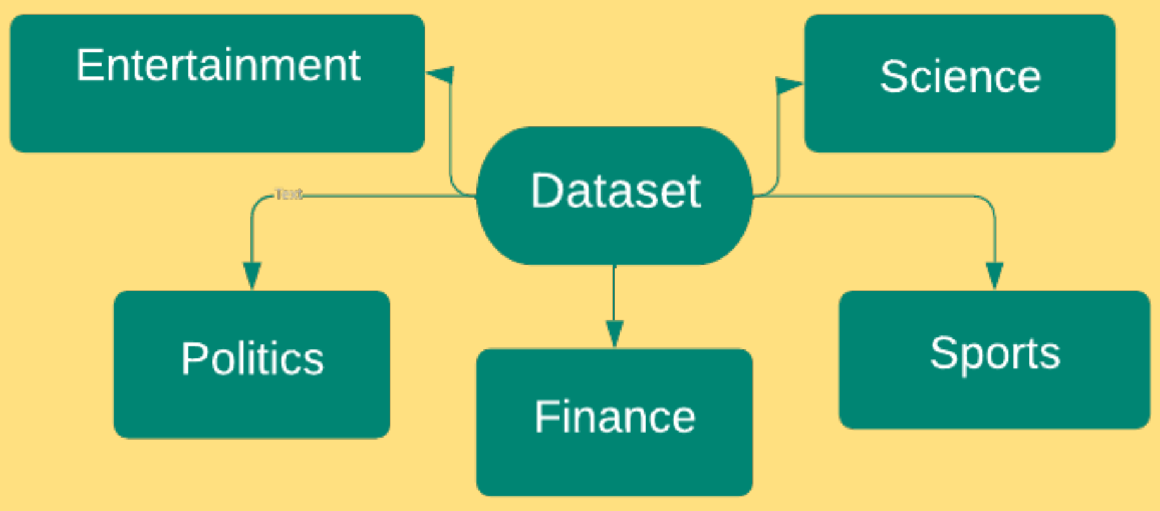
Dataset Creation

USC-31k has 50 thousand tweets. The dataset is divided into three subsets (easy, medium and hard) depending upon the amount of preprocessing done on the extracted tweets, and five subdomains.



['thriller', 'action', 'adventure', 'science', 'fiction', 'newrelease']	A great story that keeps you reading when you should be getting sleep. Grab a copy of "The Nemesis Effect" now. available at Amazon --&gt;
['science', 'maritain', 'humanrights', 'christianity', 'secularconscience']	So in case you're wondering, when they make a certain true via compulsion, they are operating within a Marxist-Leninist and/or Aryan framework…
['scientific', 'bee', 'animals', 'bees', 'science']	UK overrules advice by lifting ban on -harming pesticide
['science', 'kids', 'wearehelsinkiuni']	Wow! How nice concept for explaining for .
['gardening', 'students', 'science', 'botany']	Spring is a great time for , and our West Valley Kindergarten and 1st grade have been doing just that! With their after-school activity instructor, our young horticulturists have been diving into and . Read more:
['scientistrebellion', 'science']	What's about? Check this out: TY &amp; all the other scientists who took action instead of doing what they love: !
['science']	Is the universe rotating or spinning?
['trivia', 'science', 'opentdb', 'triviabot', 'twitterbot']	&amp;Nature Category: Science &amp; Nature Difficulty: medium Which color cannot be produced in Roses, even through genetic alteration? A. Black B. Orange C. Blue D. Brown
['space', 'science']	"How does the work in your online shop relate to your other professional experience in graphic and multimedia design?"

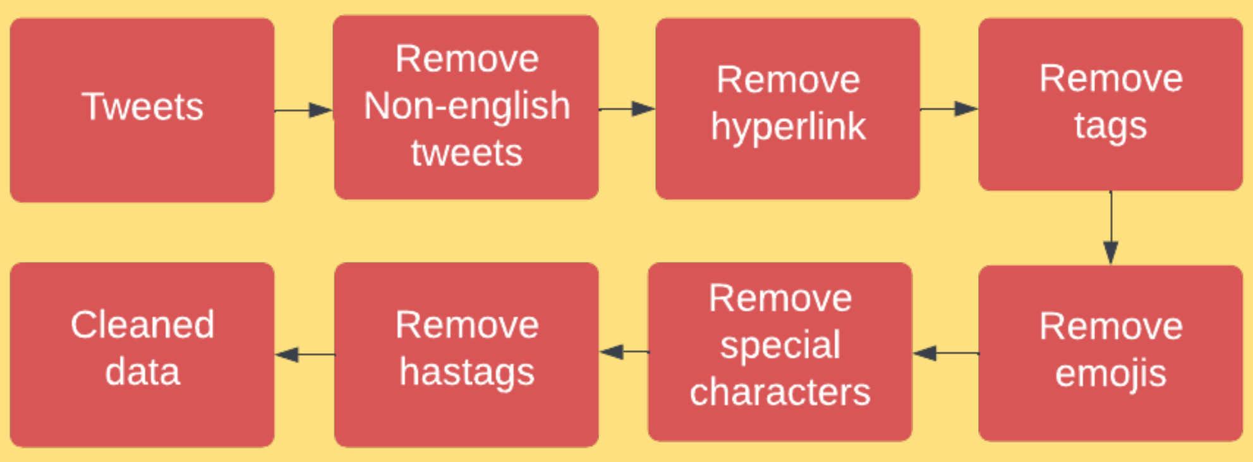
Sample of USC-31k



Dataset Preprocessing

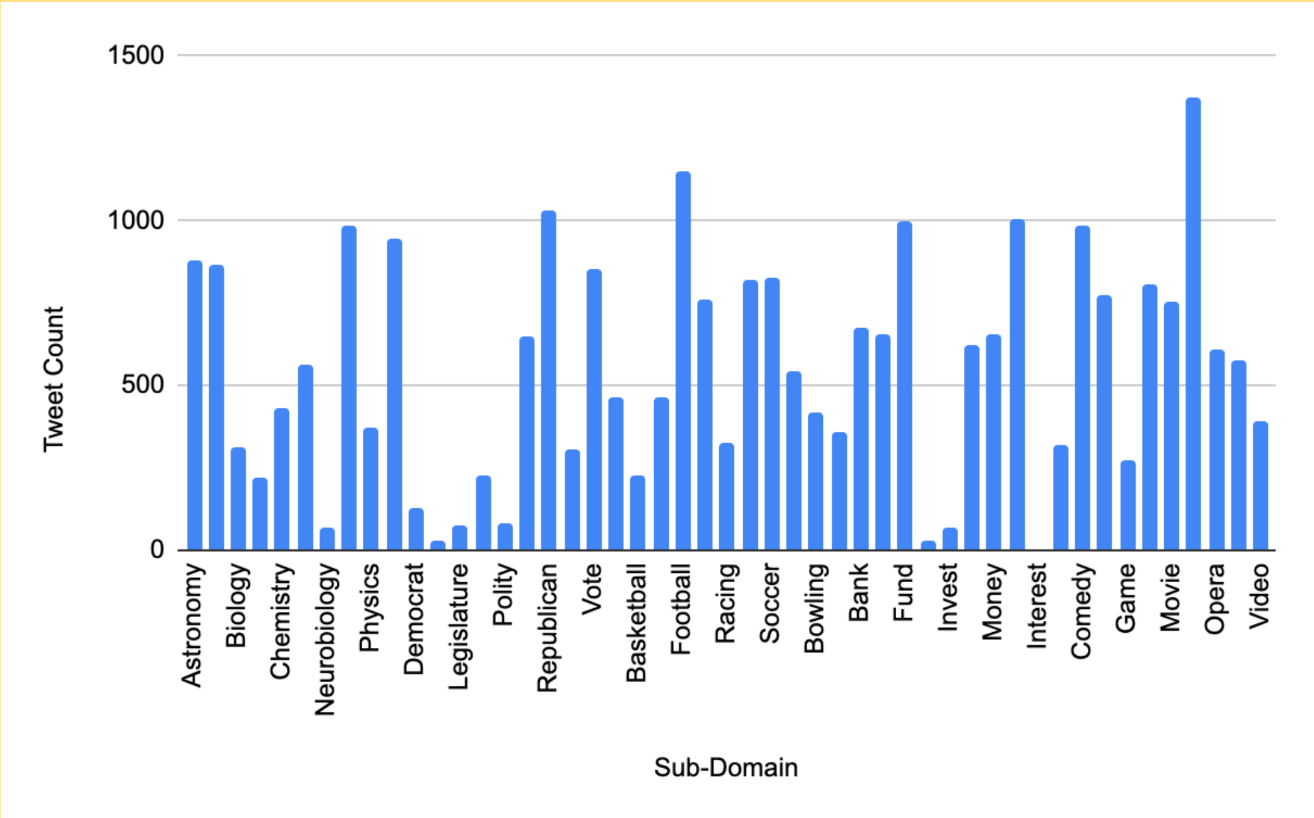
The tweets are collected using Tweepy, which is a python module that provides a convenient way to access the twitter API.

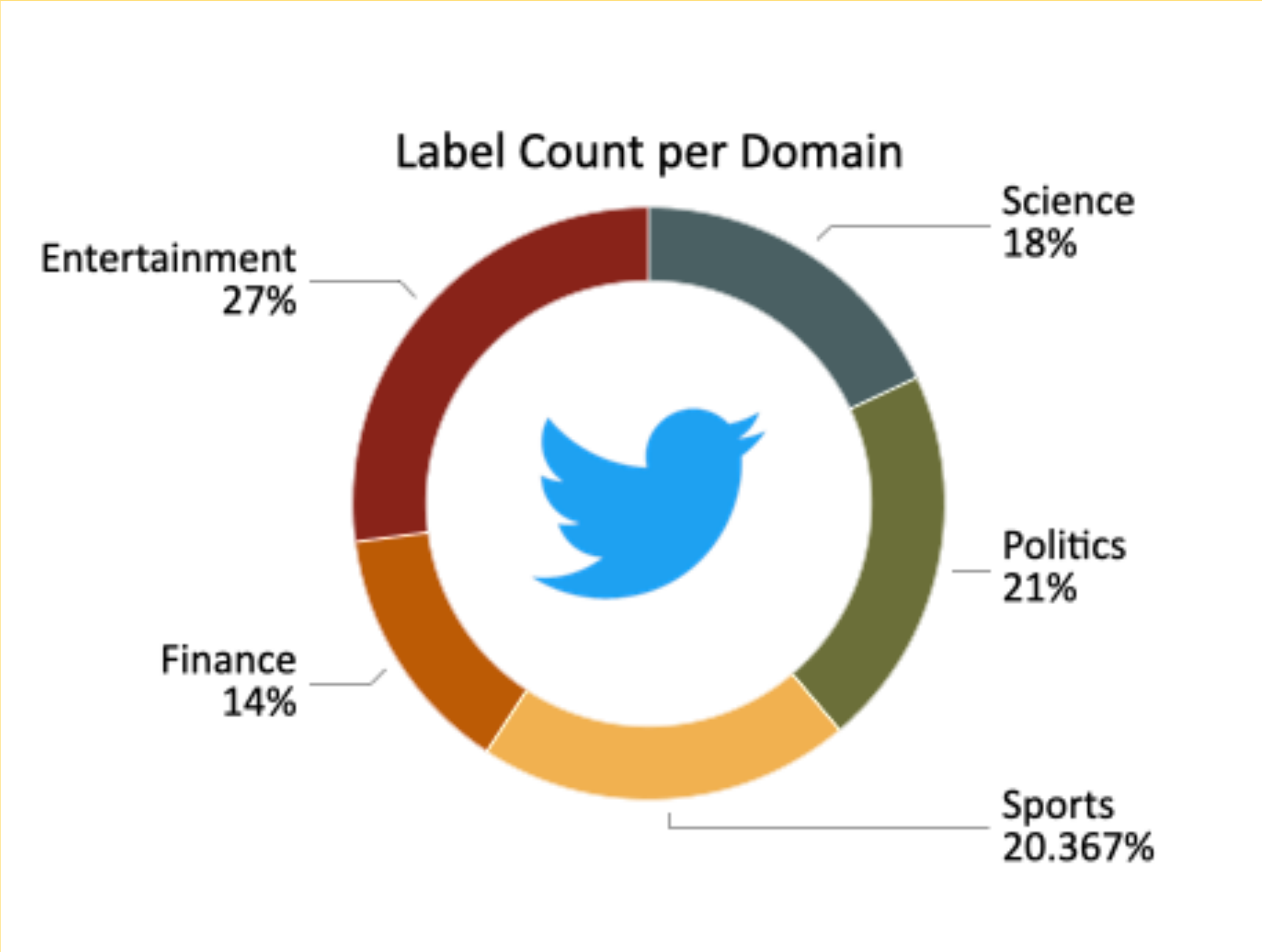
Data preprocessing is done by removing non English tweets and special characters and symbols such as hashtag, hyperlink, tags, and emojis.



Data Statistics

Number of tweets collected per sub-domain





Experiments

The performance of GUDN by [Wang et al., 2022] and LightXML by [Jiang et al., 2021] is analyzed with the EURLEX-4k and USC-31k dataset. Eurlex-4K is text data about European Union law, containing nearly four thousand labels formed according to EUROVOC descriptors. EURLEX-4k has an overlap between the texts and labels. Hence, we have measured the performance of both models with USC-31k, which is a real world dataset in which the text and labels are far apart.

Analysis

We use a simple but intuitive evaluation metric used extensively in XMTC tasks called P@k. The calculation formula of P@k is as follows:

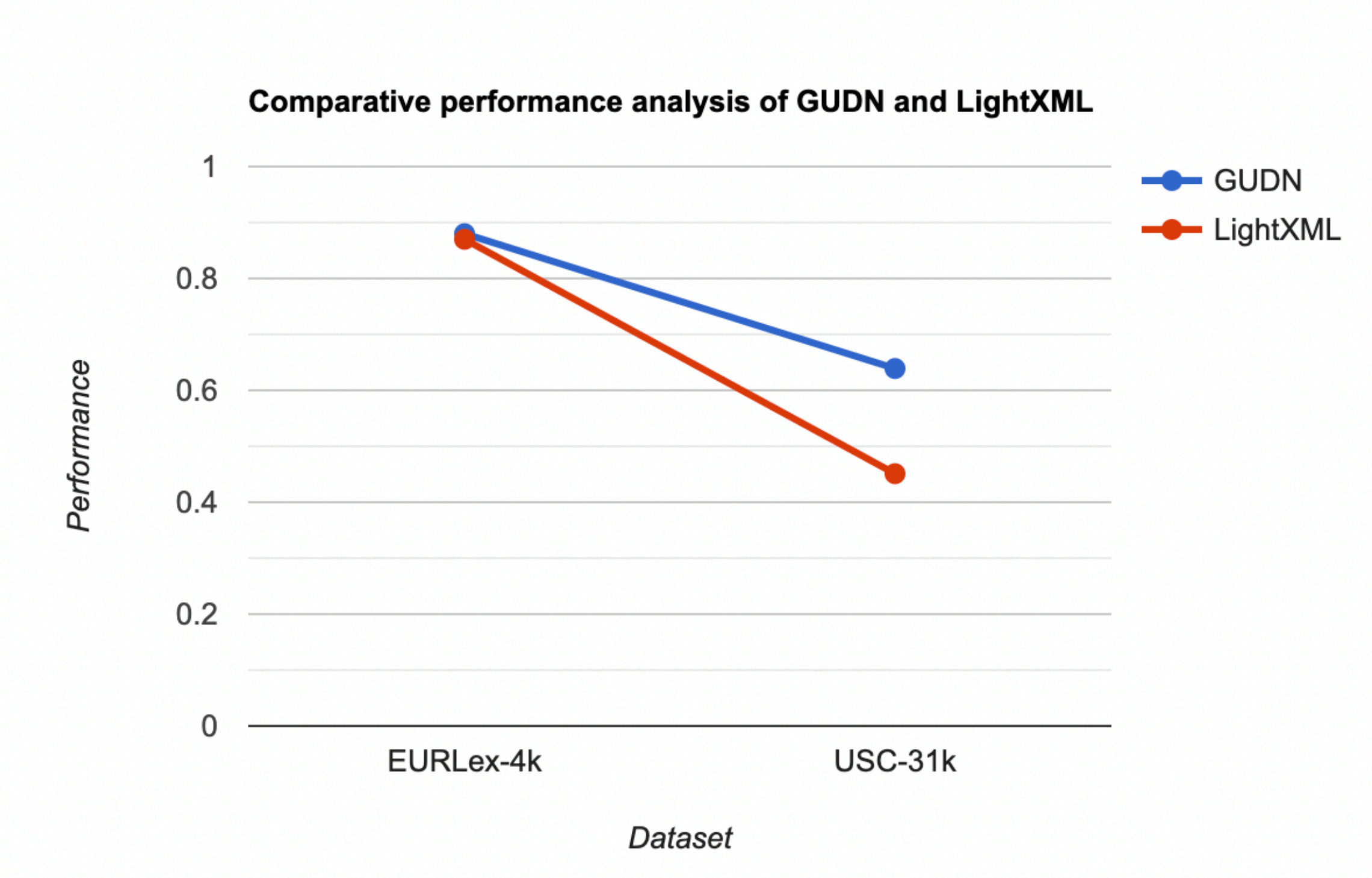
$$P@k = \frac{1}{k} \sum_{i \in \mathrm{rank}_k(\hat{y})} y_i,$$

where k is a given constant is usually 1, 3, 5.

Dataset	P@k	GUDN	LightXML
Eurlex-4k	P@1	*0.88*	*0.87*
	P@3	*0.77*	*0.75*
	P@5	*0.65*	*0.63*
USC-31k	P@1	*0.639*	*0.451*
	P@3	*0.42*	*0.312*
	P@5	*0.324*	*0.219*

Results

In [Wang et al., 2022], author proposed a Guided network to bridge the gap between the text and the labels. According to the results the two models, LightXML and GUDN performed very similarly for the Eurolex-4K data.



We hypothesized that this is because the text and the tables are very similar for the Eurolex dataset. To test this hypothesis we designed the USC-31k data set. This dataset makes it harder for the model to predict labels as the text is very small and the labels are not very similar to the text. The results show that when the text and labels are more far apart the guided network kicks in and helps the GUDN model out perform Lightxml by a larger margin.