USC-31K: Extreme Multi-Label Text Classification Dataset

Astha Mishra, Manav Jain, Megha Bagri, Siddhant Gupta, Shicuo Xie
asthamis@usc.edu, manavpra@usc.edu, mbagri@usc.edu, sgupta86@usc.edu, shicuoxi@usc.edu

1 Introduction

Extreme multi-label text classification (XMTC) aims to match the most relevant labels for a text from an extremely large number of labels. For example, Figure 1, shows the Wikipedia article on NBA which contains many related labels such as player names, team names, team city, and even other related sports. It is emphasized that extreme multi-label text classification is a challenging task that differs from multi-class or usual multi-label text classification because the labels in XMTC tasks may reach hundreds of thousands or even more. This task has attracted many research interests due to the wide downstream applications, such as advertisement recommendation, user profiling, or web search. As the number of labels in the multi-label text classification can scale up to millions, this problem is intractable and challenging.

Many works have been proposed to solve the XMTC problem, including [Prabhu and Varma, 2014; Bhatia et al., 2015; Babbar and Shoelkopf, 2016]. Deep learning methods have started to shine in XMTC tasks in recent years. X-Transformer [Chang et al., 2019] used the pre-trained model, e.g., BERT [Devlin et al., 2018] to effectively extract features from raw texts so that accuracy can be improved significantly. LightXML [Jiang et al., 2021] later improved the X-Transformer to make it lighter and faster. However, the easily accessible but critical label semantics are not taken into account by LightXML.

National Basketball Association

From Wikipedia, the free encyclopedia

"NBA" redirects here. For other uses, see NBA (disambiguation).

The **National Basketball Association** (**NBA**) is a professional basketball league in North America. The league is composed of 30 teams (29 in the United States and 1 in Canada) and is one of the major professional sports leagues in the United States and Canada. It is the premier men's professional basketball league in the world.[3]

The league was founded in New York City on June 6, 1946, as the Basketball Association of America (BAA).[1] It changed its name to the National Basketball Association on August 3, 1949, after merging with the competing National Basketball League (NBL).[4] In 1976, the NBA and the American Basketball Association (ABA) merged, adding four franchises to the NBA. The NBA's regular season runs from October to April, with each team playing 82 games. The league's playoff tournament extends into June. As of 2020, NBA players are the world's best paid athletes by average annual salary per player.[5][6][7]

The NBA is an active member of USA Basketball (USAB),[8] which is recognized by the FIBA (International Basketball Federation) as the national governing body for basketball in the United States. The league's several international as well as individual team offices are directed out of its head offices in Midtown Manhattan, while its NBA Entertainment and NBA TV studios are directed out of offices located in Secaucus, New Jersey. In North America, the NBA is the third wealthiest professional sport league after the National Football League (NFL) and Major League Baseball (MLB) by revenue, and among the top four in the world.[9]

The Milwaukee Bucks are the defending league champions, as they defeated the Phoenix Suns 4–2 in the 2021 NBA Finals.

Figure 1: An example of XMTC is extracted from https://en.wikipedia.org/wiki/National_Basketball_Association.

The model proposed by [Wang et al., 2022], which utilizes label semantics through a guided network (GUDN), has achieved state-of-the-art accuracy. Some of the datasets in which the existing models have trained are Eurlex-4K, AmazonCat-13K, Wiki10-31K, and Wiki-500K. Most of these datasets are extremely verbose, domain-specific, and contain overlapping text and labels.

We propose a new dataset based on Twitter that overcomes these limitations of the existing datasets. Twitter seems like a potential data source for this problem statement because the number of hashtags (labels) associated with the tweets accounts for a huge label space, which creates a case for XMTC.

2 Related Work

Many novel methods have been proposed to improve accuracy while controlling computational complexity and model sizes in XMTC. These methods can be broadly categorized into two directions according to the input: One is traditional machine learning methods that use the sparse features of text like Bag of Words features as input, and the other is deep learning methods that use raw text. Traditional machine learning methods can be further divided into three directions: one-vs-all methods, tree-based methods and embedding-based methods.

One-vs-all methods treat each label as a binary classification problem and classification tasks are independent of each other. Although many one-vs-all methods like DiSMEC [Babbar and Scholkopf 2017], and PPDSparse [Yen et al. 2017] focus on improving model efficiency, one-vs-all methods still suffer from expensive computational complexity and large model size.

Embedding-based methods project high dimensional label space into a low dimensional space to simplify the XMTC problem. However, no matter how the label compression part is designed, label compression will always lose a part of information. It makes these methods achieve worse accuracy compared with one-vs-all methods and methods.

Tree-based methods aim to overcome high computational complexity in one-vs-all methods. These methods will construct a hierarchical tree structure by partitioning labels like Parabel

[Prabhu et al. 2018] or sparse features like FastXML [Prabhu and Varma 2014]. Though the tree structure reduces the prediction time, accuracy is affected as the tree grows deep.

With the development of NLP, deep learning methods such as XML-CNN [Liu et al. 2017], AttentionXML [You et al. 2018] and XTransformers [Chang et al. 2019] have shown great improvement in XMTC. But the main challenge to these methods is how to couple with millions of labels with limited GPU resources.

Considering the X-Transformer's computational complexity and the model's size, LightXML [Jiang et al., 2021] intends to improve it to obtain a light and fast model. The guided network (GUDN) model proposed by [Wang et al., 2022], uses BERT to extract specific label semantic features from raw labels and combines a deep pre-trained model to extract features, which is more effective than previous work for finding the latent space between texts and labels. This model has achieved state-of-the-art accuracy.

3 Motivation

Eurlex-4K is text data about European Union law, containing nearly four thousand labels formed according to EU-ROVOC descriptors. Amazon-13K is a product-to-product recommendation dataset, where the labels are product categories. Wiki10-31K and Wiki-500K are excerpts from Wikipedia articles with thirty-one thousand and five hundred thousand labels, respectively.

Upon exploration, we realized that the text in these datasets is verbose. Wiki10-31K and Wiki-500K are full-length articles ranging from 5 to 10 pages in length. Eurlex-4K also shares this problem. Moreover, Amazon-13K and Eurlex-4K datasets are domain-specific and have an overlapping text and label set. AmazonCat-13K, Wiki10-31K, and Wiki-500K contain representative symbols such as '++++-' and '08P01B'. These symbols do not provide any semantic information and are difficult to predict.

The state-of-the-art model, GUDN, performs best on the Eurolex-4k dataset. Since GUDN explores the latent space between the label and text, a strong relation between text and the labels does not provide the right opportunity for the guide network to show its contributions.

Considering the ubiquitous application scope of the XMTC problem we find that these characteristics of a dataset are limiting. For tasks like user profiling or social media tagging, these datasets are not representative of the real world. Our motivation is to produce a new dataset that addresses these problems.

4 Dataset

We created a new dataset which consists of tweets as the text and hashtags as the labels. The length of each tweet is limited to 280 characters and the overlap between the tweet text and the label is almost negligible. The labels are human readable and contain semantic information. These tweets also contain special symbols, emojis and tags unlike the existing dataset that are comparatively less noisy. We collect tweets from five broad domains and provide the user the flexibility to test the generalizability of their models.

4.1 Data Collection

Due to the limitation of Twitter API, we plan to use Tweepy, an advanced Twitter scraping tool without any rate limitation for data scraping. We scraped about 50k tweets in total as our dataset.

We collected tweets from five broad domains, namely "Science", "Sports", "Finance", "Entertainment", and "Politics". We did this by manually defining ten sub-domains or gold class labels, and for each sub-domain, we scraped 1000 tweets containing the sub-domain as a hashtag. The rest of the labels that came along with the tweets became the label set.

Science	science, physics, chemistry, biology, astronomy, neurobiology, technology, bioinformatics, neuroscience, microbiology, cellbiology

Fig 2: Sample domain and its sub-domain

4.2 Data Cleaning

Data preprocessing is done by removing non-English tweets, special characters, hashtags, hyperlinks, username mentions, and emojis. To understand how models perform on our dataset, we divided it into three categories 1) Easy 2) Moderate 3) Hard.

Twitter data can be very noisy. Tweets can be short, sometimes just a few words. Hashtags don't have to be valid English words or hashtags

can be formed by combining multiple English words. To compare the performance of state-of-the-models on this dataset we had to come up with a way to reduce this noise. Hence we decided to divide the dataset into the above-mentioned three classes.

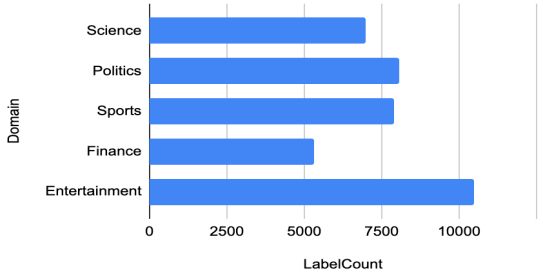

Figure 3 :Number of Lables per domain in USC-31k.

Apart from the basic pre-processing we performed some class-specific cleaning tasks. In the easy mode, we kept a minimum length restriction of 128 characters on the text of the tweets. We created a frequency mapping to count occurrences of different hashtags for a particular domain. We kept only the topmost relevant labels associated with the text and neglected the ones which had lesser frequencies. Since hashtags could also be embedded in the tweet's body it can increase the overlap percentage between the labels and the text. But for the easy mode of the dataset, we decided to keep these embedded hashtags as a part of the body.

In moderate cleaning, we keep the emojis because emojis do convey some meaning and context in the tweets. In hard cleaning, we keep the data as it is without doing any preprocessing. The data was changed as per the model input requirements.

5 Experiments

5.1 Benchmarking

We evaluated our dataset against two state-of-the-art neural networks for the Extreme Multi-Label Text Classification task - LightXML and GUDN.

['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 -->
['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 & 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']	&Nature Category: Science & 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?"

Figure 4: A sample of twitter dataset

5.1.1 LightXML

LightXML by [Jiang et al., 2021] is a deep learning model which fine-tunes a single transformer model with dynamic negative label sampling. LightXML consists of three parts: text representing, label recalling, and label ranking. For text representation, the model uses multi-layer features of the transformer model, which can prove rich text information for the other two parts. With the advantage of dynamic negative sampling, the model proposes generative cooperative networks to recall and rank labels. For the label recalling part, it uses the generator network based on label clusters, which is used to recall labels. Moreover, for the label ranking part, a discriminator network to distinguish positive labels from recalled labels is used.

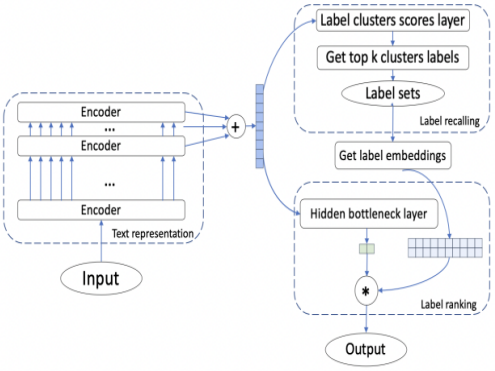

Figure 5: LightXML Model

5.1.2 Guided Network

Guided Network proposed by [Wang et al., 2022] has three parts - a feature extractor, guide network and a ranking classifier. First, the feature extractor extracts the features of the texts and the labels, and then the features of texts and labels are input into the guide network. A close relationship is established through the guide network for texts and labels, and the relationship is fed back to the feature extractor for continuous optimization.

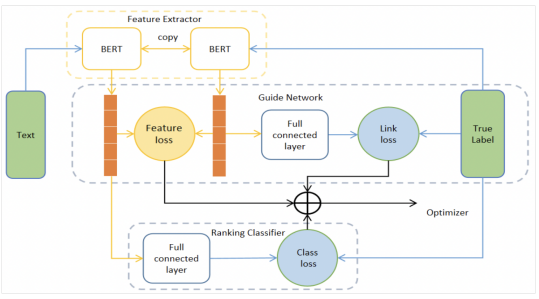

Figure 6: Guided Network Model

5.2 Evaluation

In XMTC, the label space is huge and there are only a few relevant labels for each text. The model calculates a probability for each of the labels based on relevance. For this reason, rank-based evaluation metrics are widely used in XMTC tasks. These metrics emphasize the top k most relevant labels of a text. We used P@k which represents the precision of the top k labels as our evaluation criteria.

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

We implemented these two models to benchmark our dataset. The results of our experiments showed that although the performance of both the model fell significantly as compared to that on EURLex-4K, the performance of the Guided Network model on the Twitter dataset fell comparatively less.

5.3 Results and Discussion

In [Wang et al., 2022], the author concluded that GUDN is sensitive to label semantics. While Eurex-4k provided label semantics, the latent space between the text and label is very minimal for the dataset. Hence the guiding network does not get a chance to add value and the model performs similarly to LightXML. Other datasets like AmazonCat-13K, Wiki10-31K, and Wiki-500K contain codes and symbols for labels and make it difficult for the model [Wang et al., 2022]. The new dataset USC-31k provides the right environment to test the potency of the model by providing a dataset where the labels are semantically rich and the latent space is large.

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

Table 1: P@k values for EURLex & USCUSC-31k.

As expected the Guide Network helps the GUDN model to perform better on USC-31k as compared to LightXML. The performance for both the models fall as the dataset is simply more vast and more difficult to learn. However, the accuracy for LightXML falls more sharply compared to GUDN and this is because the guiding network helps bridge the latent space enabling us to get better performance.

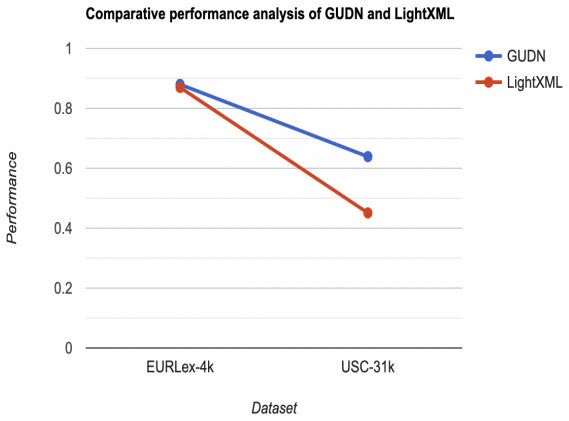

Figure 7: Comparative analysis of performance of GUDN and LightXML for EURLex-4k and USC-31k.

6 Conclusion and Future Work

In this paper, we have created a new dataset called USC-31k using twitter data. The dataset contains tweets as text and the respective hashtags as the labels. Both the text and the labels are semantically meaningful. We have collected data for five domains namely Entertainment, Science, Sports, Politics and Finance. Each of these domains are divided into ten sub-domains. We have collected ten thousand tweets for each subdomain. Each tweet can have multiple hashtags associated with it, and all of these hashtags combined makes the label set for USC-31k which has a total of thirty one thousand labels.

The dataset generated by us is tested with two state of the art models proposed by [Wang et al., 2022] and [Jiang et al., 2021]. The GUDN model proposed by [Wang et al., 2022] has shown a better performance as compared to the LightXML model proposed by [Jiang et al., 2021]. However, both of the models have shown a significant decrease in performance when tested with the USC-31k dataset.

The dataset generated by us, however, has some limitations that can be improved. Currently, we are collecting random data from different domains, however the tweets can be repeated among different domains. The USC-31k dataset can be improved to eliminate redundancy. The dataset can also be further extended for other domains and subdomains.

7 Team responsibilities

Astha Mishra	Implementing the LightXML model. Evaluate and document the results.
Manav Jain	Collection and cleaning of Twitter data.
Megha Bagri	Implementing the LightXML model. Evaluate and document the results.
Siddhant Gupta	Implementing the GUDN model. Evaluate and document the results.
Shicuo Xie	Collection and cleaning of Twitter data.

References

[Chang et al., 2019] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. Taming pretrained transformers for extreme multi-label text classification, arxiv:1905.02331, 2019.

[Mittal et al., 2021a] Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. Decaf: Deep extreme classification with label features, arxiv:2108.00368. arXiv e-prints, 2021.

[Yeh et al., 2017] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent spaces for multi-label classification, arxiv:1707.00418, 2017.

[Jiang et al., 2021] 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, arxiv:2101.03305, 2021.

[Prabhu and Varma, 2014] Yashoteja Prabhu and Manik Varma. 2014. FastXML: a fast, accurate and stable tree-classifier for extreme multi-label learning. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). Association for Computing Machinery, New York, NY, USA, 263–272. https://doi.org/10.1145/2623330.2623651

[Bhatia et al., 2015] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In Proceedings of the 28th Annual Conference on Advances in Neural Information Processing Systems (NIPS), December 2015

[Babbar and Shoelkopf, 2016] Rohit Babbar and Bernhard Shoelkopf. Dismec - distributed sparse machines for extreme multi-label classification, arxiv:1609.02521, 2016.

[Devlin et al., 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, arxiv:1810.04805, 2018.

[Wang et al., 2022] Qing Wang, Hongji Shu, and Jia Zhu. 2022. *GUDN A novel guide network for extreme multi-label text classification*. arXiv. DOI:https://doi.org/10.48550/ARXIV.2201.11582

[Yen et al. 2017] Ian E.H. Yen, Xiangru Huang, Wei Dai, Pradeep Ravikumar, Inderjit Dhillon, and Eric Xing. 2017. PPDsparse: A Parallel Primal-Dual Sparse Method for Extreme Classification. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '17*). Association for Computing Machinery, New York, NY, USA, 545–553. https://doi.org/10.1145/3097983.3098083

[Babbar and Scholkopf 2017] Rohit Babbar and Bernhard Schölkopf. 2017. DiSMEC: Distributed Sparse Machines for Extreme Multi-Label Classification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (WSDM '17), Association for Computing Machinery, Cambridge, United Kingdom, 721–729. DOI:https://doi.org/10.1145/3018661.3018741

[Prabhu et al. 2018] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising. In *Proceedings of the 2018 World Wide Web Conference* (WWW '18), International World Wide Web Conferences Steering Committee, Lyon, France, 993–1002. DOI:https://doi.org/10.1145/3178876.3185998

[Liu et al. 2017] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multilabel text classification. In International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017.

[You et al. 2018] 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, arxiv:1811.01727, 2018.