Multi-label Classification of Short Texts with Label Correlated Recurrent Neural Networks

Zhiyuan Peng
Santa Clara University
Santa Clara, CA, USA
zpeng@scu.edu

Behnoush Abdollahi
Walmart Labs
Sunnyvale, CA, USA
behnoush.abdollahi@walmartlabs.com

Min Xie
Instacart
San Francisco, CA, USA
min.xie@instacart.com

Yi Fang
Santa Clara University
Santa Clara, CA, USA
yfang@scu.edu

ABSTRACT

Short texts are commonly seen nowadays on the Internet in various forms such as tweets, queries, comments, status updates, snippets of search results, and reviews from social platforms. Accurate categorization of these short texts is critical for enhancing information services as it provides the foundation for better search and recommendation. In many real-world applications, a short text is often associated with multiple categories. Due to the sparsity of context information, traditional multi-label classification methods do not perform well on short texts. In this paper, we propose a novel Label Correlated Recurrent Neural Network (LC-RNN) for multi-label classification of short texts by exploiting correlations between categories. We utilize a tree structure to represent the relationships among labels and consequently an efficient max-product algorithm can be developed for exact inference of label prediction. We conduct experiments on four testbeds and the results demonstrate the effectiveness of the proposed model.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**.

KEYWORDS

Multi-label Classification, Short Text, RNN

ACM Reference Format:
Zhiyuan Peng, Behnoush Abdollahi, Min Xie, and Yi Fang. 2021. Multi-label Classification of Short Texts with Label Correlated Recurrent Neural Networks. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3471158.3472246

ICTIR '21, July 11, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472246

1 INTRODUCTION

Short texts have been increasingly and widely used on the Web such as tweets, comments, status updates, snippets of search results, and reviews from various social platforms. In many practical applications, a short text is often labeled with multiple labels. For instance, one comment on Reddit may be tagged with "threat" and "hate" at the same time. Multi-label classification of short texts is to assign a piece of short text to a subset of relevant categories.

Unlike ordinary documents, short texts are usually much shorter, nosier, and sparser. They may not provide sufficient and accurate word co-occurrence or shared context for traditional text classification methods to achieve a desired accuracy. To tackle the challenge of short texts, we exploit the correlations between classes, which provide a valuable source of information for the classification of short texts. For example, if a tweet is labeled with "Kung Fu", it is often also labeled with "Chinese". On the other hand, some of the prior works attempted to mitigate the sparsity of short texts by leveraging topic models to reduce the data dimensionality [2]. These topic models treat texts as a bag of words, which largely neglects the ordering and semantic information of the short texts. In recent years, considerable progress has been made in deep learning. Some prior works have demonstrated that neural network based models are useful at capturing sequential and semantic information of textual data and in particular effective for text classification with Convolutional Neural Networks [4] and Recurrent Neural Networks [8].

In this paper, we propose a novel Label Correlated Recurrent Neural Network (LC-RNN) for multi-label classification of short texts. Specifically, based on frequent label co-occurrence patterns in the underlying dataset, we generate a tree-structured undirected graph which is actually a Conditional Random Field (CRF) by using maximum spanning tree algorithm. Piecewise training can then be applied to CRF. Due to the tree structure of the label graph, we can perform exact inference for label prediction using the max-product message passing algorithm. Our major contributions can be summarized as follows.

- We propose a novel multi-label classification model for short texts. To the best of our knowledge, this is the first work that exploits the label correlations for short text classification.
- We represent the relationships among labels with a tree structure and consequently an efficient max-product algorithm can be utilized for exact inference of label prediction.

- We conduct experiments on three public testbeds and one proprietary dataset in E-Commence. The results demonstrate the effectiveness of the proposed model. We will make our code publicly available upon paper acceptance.

2 RELATED WORK

Multi-label classification models can be generally divided into two families: 1) problem transformation models which fit data to algorithms, and 2) algorithm adaptation models which fit algorithms to data. The review article [12] provides a survey of various multi-label learning models in these two families. Numerous methods that encode label correlations have been proposed [3], but most of them were designed for image data. To the best of our knowledge, no prior work has exploited label correlations for short texts.

There exists much less work on multi-label classification of short texts than other types of data such as long documents and images. The majority of prior approaches attempt to enrich the representation of a short text using additional semantics which can be derived from the same short text collection, a collection of much longer documents in a similar domain as the short texts, or from much larger external sources such as Wikipedia. Hierarchical multi-Label classification of social text streams is tackled by extending each short document via entity linking and sentence ranking strategies [7]. A concept based approach was proposed in [11] by leveraging a large taxonomy knowledge base. Some other approaches took an opposite direction by trimming a short text representation to get a few most representative words for topical classification [9].

3 LABEL CORRELATED RECURRENT NEURAL NETWORKS

In addition to treating multi-label classification as a set of independent binary classification problems, we propose to improve it by utilizing the frequent label co-occurrence patterns in the training data. We encode a short text by a recurrent neural network such as LSTM in order to capture the semantic information of the input. We identify the informative label pairs by learning a tree-structured undirected graph in the label space as shown in Figure 2(a) which is actually a CRF model. The joint probability of a configuration of L label variables can be given by

$$P(y_1, y_2, ..., y_L|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C} \Psi_c(\mathbf{y_c}, \mathbf{x}) \quad (1)$$

where C is the set of maximum cliques of the graph, and Ψ_c is the potential function for maximum clique c, which maps the clique label configuration y_c and the input data instance $\mathbf{x}$ into a positive scalar value. $Z(\mathbf{x})$ is a partition function that ensures a valid conditional distribution and L is the total number of labels. Finally we apply the trained model to predict labels for test data using a max-product exact inference algorithm [1].

3.1 Piecewise Training of CRF

To train a tree-structured CRF, computing $Z(\mathbf{x})$ is needed at each parameter update step which can make the training process computationally expensive, especially for large and densely connected graphs and large training sets. To avoid calculating $Z(\mathbf{x})$ at each training step, we formulate the learning process as a piecewise training procedure under the framework of undirected graphical models [10]. That is, we train a set of L binary classifiers from the training data, one for each label in the original label space, as the potential functions for the node cliques. In addition, we train a set of $(L-1)$ binary classifiers independently from the data for the constructed new labels (i.e., edges in the tree) as the potential functions for the corresponding edge cliques. It has been shown in [10] that such piecewise training of an undirected graphical model can be justified as minimizing a family of upper bounds on the log partition function of the data log-likelihood.

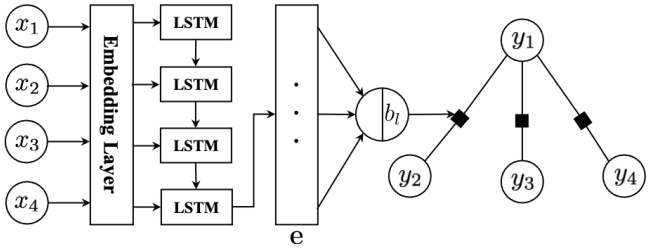

Figure 1. The Architecture of Label Correlated Recurrent Neural Networks (LC-RNN)

3.2 Model Architecture

The proposed Label Correlated Recurrent Neural Network (LC-RNN) with only one binary classifier for edge $y_{1,2}$ is shown in Figure 1. Each binary classifier consists of three layers.

- Embedding Layer: Embedding layer maps each word x_i in the input short text into a vector representation $\mathbf{e}$. In the experiments, we use pre-trained word representation by GloVe [6] for word embeddings.
- LSTM Layer: Long Short Term Memory (LSTM), a type of Recurrent Neural Network (RNN), transforms the word embeddings from the previous layer to capture the sequential and semantic information in the short text. It outputs a semantic vector.
- Dense Layer: A fully connected layer takes as the input the LSTM encoding of the input text, and output probability b_l for a) a class label $y \in \{0, 1\}$ or b) an edge in the tree between two nodes y_i and y_j. It uses the Sigmoid activation function to produce the probability after the fully connected layer.

3.3 Tree-structured Graph with Label Correlation

In this section, we construct a tree-structured graph in the label space by identifying a set of correlated label combination pairs. We take all possible label pairs as candidates by forming a fully connected graph over the L label variables. Then we measure the correlation strength of each label pair as the weight of the corresponding edge using an appropriate criterion. In the experiments, three criteria are investigated to calculate the weights between labels: Normalized Co-occurrence [3], cosine similarity, and Pearson Correlation. In Normalized Co-occurrence, for each pair of the

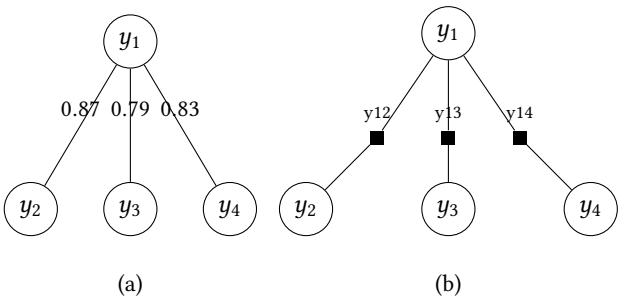

Figure 2. An Example of (a) Maximum Spanning Tree and (b) the Corresponding Factor Graph based on Label Correlation

labels (i, j), we calculate the co-occurrence weight as follows:

$$NC(i,j) = \frac{count(i,j)}{min(count(i), count(j))}$$

where $count(i, j)$ is the number of the instances tagged with both label i and label j and $count(i)$ is the number of instances with label i. In Cosine similarity, we represent each label/class by a binary vector with size N which is the total number of documents in the training set. If a document belongs to the class, the corresponding dimension is 1 and otherwise 0. The correlation between labels can then be computed based on the Cosine similarity between the label vectors. Similarly, Pearson correlation can also be calculated based on such binary vector representations of labels.

Given the correlation criterion, we can compute the weights for all edges between the label variables. Then we use a maximum spanning tree algorithm to select $(L-1)$ edges according to the computed weights, which produces a tree with the maximum strength of connection. An example of the derived tree structure is shown in Figure 2(a).

3.4 Inference/Label Prediction

After the tree-structured model is trained, the inference process on a test instance $\mathbf{x}$ (i.e., a sequence of words $x_1, x_2, ...$) is to find the maximum a posteriori (MAP) label assignment $\mathbf{y} = y_1, y_2, ..., y_L$ by solving

$$\mathbf{y}^* = \underset{\mathbf{y}}{\text{argmax}}\, P(\mathbf{y}|\mathbf{x})$$

where $\mathbf{y}^*$ is the predicted labels and $P(\cdot)$ is given in Eqn.(1). Since the label graph is tree-structured, the multi-label prediction can be performed using the max-product inference algorithm [1]. The max-product algorithm predict labels through message passing on the factor graph. Given the trained pairwise graphical model, we first translate it into a factor graph by keeping all variable nodes and adding a factor node for each edge clique. For example, the factor graph in Figure 2(b) is constructed for the tree-structured graph in Figure 2(a), where each variable node is represented as a circle and each factor node is represented as a rectangle. For a given test instance $\mathbf{x_i}$, the potentials of the two types of nodes in the factor graph are computed using the probabilistic binary classifiers obtained in the training step, i.e., $\Psi(\mathbf{y_i}) = P(\mathbf{y_i}|\mathbf{x_i})$ and $\Psi(\mathbf{y_i}, \mathbf{y_j}) = P(\mathbf{y_i}, \mathbf{y_j}|\mathbf{x_i})$.

Table 1. Statistics of the datasets. "cardinality" is the average number of labels per instance

Dataset	# instances	# labels	avg length	cardinality
Comment	16,225	6	51	2.16
MedWeb	2,560	8	13	1.38
EC1	7,725	16	11	2.30
E-Commerce	56,306	5	3.2	1.49

4 EXPERIMENTS

4.1 Experimental Setup

We use the following three public datasets with various characteristics for evaluation, and statistics of the three datasets are presented in Table 1. 1) *Comment*[1]. 2) *Medical Natural Language Processing for Web Document (MedWeb)*[2] which are tweets and annotated with 8 labels such as cold, fever, etc. 3) *SemEval-2018 Task 1: Affect in Tweets (EC1)*[3], which are also tweet texts. All these three datasets are short texts with the average lengths much shorter than normal documents as shown in Table 1.

Additionally, we test our proposed approach on a query categorization dataset on fashion from the Walmart E-commerce website. User queries on E-commerce are usually short in length, and given the ambiguity of the query, many of them can be associated with more than one categories. For example, the query "red dress" is labeled as both "juniors" and "women".

Each dataset is split into three parts: 80% for training, 10% for validation and the left 10% is for testing. The following data preprocessing is done on each dataset: 1) convert all the words to lowercase; 2) remove all the characters other than the spelling alphabet; 3) remove all the redundant blanks. The dimensionality of the pre-trained GloVe word embeddings is set to 100 and the dimensionality of the LSTM embeddings b is 32. All the experiments were done on a server with 2 Intel E5-2630 CPUs and 4 GeForce GTX TITAN X GPUs. The proposed deep models were implemented in TensorFlow 2.0.

Naive Bayes and a simple LSTM-based classifier are used as two baselines for comparison, which are well-known methods for text classification. The LSTM baseline has a similar architecture with LC-RNN except the output layer. The embeddings learned from LSTM is fed to Sigmoid functions to make independent classifications on the corresponding labels, without utilizing the label correlations as LC-RNN does. Three evaluation metrics are used in the experiments: Accuracy, F1, and Area Under Curve (AUC) [5].

4.2 Experimental Results

Table 2 shows the Accuracy of the baselines and the proposed model with different label correlation criteria on four datasets. As we can see, the proposed model consistently outperformed the LSTM baseline on all the datasets with noticeable margins. Since LC-RNN only differs from the LSTM baseline in the output layer where label correlations are utilized, these results demonstrated the effectiveness

[1]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview
[2]http://research.nii.ac.jp/ntcir/permission/ntcir-13/perm-en-MedWeb.html
[3]https://competitions.codalab.org/competitions/17751#learn_the_details-datasets

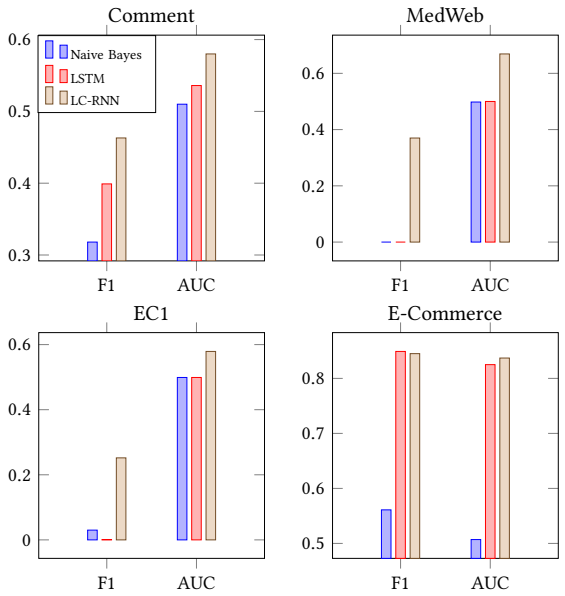

Figure 3. The F1 and AUC scores of Naive Bayes, LSTM and LC-RNN (NC) on four datasets

of incorporating label correlations for multi-label classification of short texts. The three variants of the proposed model produced similar results in general. The proposed model also outperformed Naive Bayes with the only exception on the E-Commerce dataset where LC-RNN (NC) with Normalized Co-occurrence yielded better results than Naive Bayes while the other two variants of LC-RNN did not. Figure 3 shows the F1 and AUC metrics of the proposed model LC-RNN (NC) and two baselines on the four datasets. Interestingly, on the E-Commerce dataset both LSTM and LC-RNN outperformed Naive Bayes with a very large margin. The reason might be due to the class imbalance of the dataset, Naive Bayes may under-predict the minority categories which can still lead to a high Accuracy overall but poor performance in other metrics. In general, LC-RNN outperformed the two baselines with substantial margins in F1 and AUC on the three public dataset while it delivered comparable performance with LSTM on the E-Commence dataset.

Since our proposed model relies on the predicted connectivity (i.e., strength of edges) between labels to utilize label correlations, we investigate the impact of the predicted connectivity on the model. Table 3 shows the accuracy of our proposed methods when we assume using the true edge information for inference. Specifically, if two labels appear in the same test instance based on the ground truth, we set $\Psi(y_i, y_j) = P(y_i, y_j|x_i) = 1$; otherwise, it is zero. The results in Table 3 would be the upper-bound performance of the proposed methods because it is almost impossible for a model to make perfect predictions on all the edges. Table 4 shows the F1 score of the predicted edges by LC-RNN (NC). As we can see, the predicted results are still far from perfect which leaves room for further improvement in this component in future work.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a multi-label classification model for short texts by taking advantage of the correlations among labels. The experimental results show that the correlations do help improve the accuracy of the classifier. In future work, we plan to conduct more comprehensive experiments to evaluate the effectiveness of the proposed model with more baselines. The proposed use of label correlations can also be applied to other text classification methods such as Convolutional Neural Networks.

Table 2. Accuracy of the proposed methods and two baselines on four testbeds

	Dataset			
Method	Comment	MedWeb	EC1	E-Commerce
Naive Bayes	0.820	0.830	0.790	0.843
LSTM	0.822	0.834	0.778	0.822
LC-RNN (NC)	0.843	0.872	0.819	0.865
LC-RNN (Cosine)	0.852	0.869	0.822	0.823
LC-RNN (Pearson)	0.846	0.866	0.823	0.827

Table 3. The upper-bound Accuracy of our proposed methods based on the true edge information

	Dataset			
Method	Comment	MedWeb	EC1	E-Commerce
LC-RNN (NC)	0.919	0.904	0.953	0.987
LC-RNN (Cosine)	0.942	0.907	0.951	0.989
LC-RNN (Pearson)	0.913	0.880	0.936	0.989

Table 4. F1 scores of LC-RNN on the predicted edges

	Dataset			
Method	Comment	MedWeb	EC1	E-Commerce
LC-RNN (NC)	0.587	0.637	0.561	0.786

REFERENCES

[1] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory* 47, 2 (2001), 498–519.

[2] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *ACM SIGIR*. 165–174.

[3] Xin Li, Feipeng Zhao, and Yuhong Guo. 2014. Multi-label Image Classification with A Probabilistic Label Enhancement Model.. In *UAI*, Vol. 1. 3.

[4] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *ACM SIGIR*. 115–124.

[5] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[7] Zhaochun Ren, Maria-Hendrike Peetz, Shangsong Liang, Willemijn Van Dolen, and Maarten De Rijke. 2014. Hierarchical multi-label classification of social text streams. In *ACM SIGIR*. 213–222.

[8] Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In *EMNLP*. 811–816.

[9] Aixin Sun. 2012. Short text classification using very few words. In *ACM SIGIR*. 1145–1146.

[10] Charles Sutton and Andrew McCallum. 2012. Piecewise training for undirected models. *arXiv preprint arXiv:1207.1409* (2012).

[11] Fang Wang, Zhongyuan Wang, Zhoujun Li, and Ji-Rong Wen. 2014. Concept-based short text classification and ranking. In *ACM CIKM*. 1069–1078.

[12] Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE TKDE* 26, 8 (2013), 1819–1837.