A Label-Enhanced Text Classification Model

Yingjie Liu¹, Xueyang Liu²,*, Wenhui Hu², Long Zhang¹ and Minghui Zhang³

¹School of Software and Microelectronics, Peking University
²National Engineering Research Centre for Software Engineering, Peking University
³Handan Institute of Innovation, Peking University

*Corresponding author email: liuxueyang@pku.edu.cn

Abstract. In the field of text classification, most of the previous work only uses one-hot labels, ignoring the correlations between labels. The paper proposes a novel label-enhanced text classification model, which utilizes the semantic correlation between sentences and category labels in the Natural Language Processing (NLP) classification task to integrate label information. We measure the similarity between instance and the category with the correlations among the labels. We test the proposed model on text classification tasks in two levels: text classification (document level) and sentiment analysis (sentence level). Experimental results show that the label-enhanced text classification model achieves great performance in multiple text classification tasks. In addition, experiment results on the unbalanced data sets show that our model is able to mitigate the impact of unbalanced data in classification tasks.

Keywords: Label-enhanced text classification; Sentence level; Document level.

1. Introduction

In the field of text classification[1], most of existing work uses one-hot labels for training, but in actual scenarios, there is a certain correlation between labels. For example, in sentiment analysis, a sentence may have two different sentiments, but the traditional one-hot label forcibly labels it as a single sentiment, ignoring the potential correlation between the labels, which may affect the model learning ability and hinder the performance of text classification[2-5].

Since natural language contains rich semantic information, the semantic information in the sentences plays an important role on the performance of classification. Inspired by this, we use the correlation between category labels and instance sentences and propose a simple and easy-to-use label-enhanced text classification model (LETC), which expands the training goal from simple category labels to the semantic correlation between category labels and instance sentences. Therefore, our proposed model changes the category labels from one-hot discrete values to continuous values. The experimental results show the efficiency of our model on two different levels of text classification tasks: text classification (document level) and sentiment analysis (sentence level). The results show that our label-enhanced text classification model has achieved a great improvement.

The rest of the paper is organized as follows: Section 2 describes our approach in detail. Section 3 shows our experimental setup. Our experiments are discussed in Section 4. The conclusion of our work is in Section 5.

2. Label-enhanced Text Classification Model

In traditional binary supervised datasets, each instance \( x \) is assigned in \( l^y \in \{0,1\} \), with an one-hot label \( y \) to describe. In this paper, we call \( l^y \) a normal label, because \( l^y \) denotes the relevance between the
category label and the sentence. But the one-hot labels neglect the relative importance of each label. To solve the defect, we use a way to label instances \( x \) is to assign a real value \( \in [0,1] \) to all possible label \( y \), representing the degree to which \( y \) describes \( x \). We assume that all the labels in the label set can always represent the instance completely, thus \( \sum_y d_{xy} = 1 \). The \( d_{xy} \) is called the relevance degree of \( y \) to \( x \). For a specific instance, the relevance degree of all labels constitutes a label vector called label-enhanced. Compared with ordinary one-hot labels, label-enhanced contains more semantic information to describe the instance more comprehensively.

Let \( X \in \mathbb{R}^q \) represents an input space with \( q \) dimension. The specific \( i^{th} \) instance is represented by \( x_i \), which is a sequence of words \( x_i = \{v_1, v_2, \ldots, v_n\} \), where \( n \) is the maximal-length of the sentence, \( v_n \) is the corresponding \( q \) dimension embedding vector for each word in \( x_i \). Let \( Y = \{y_1, y_2, \ldots, y_p\} \) represent a limited set of labels, where \( p \) is the value of possible labels and \( y_i \) is the label of \( x_i \). The ordinary label vector of \( x_i \) is represented by \( \in \mathbb{R}^{p} \), where \{0, 1\} \( \subset \mathbb{R}^{p} \). The label-enhanced of \( x_i \) is represented by \( \in \mathbb{R}^{p} \). Then, the construction process of the label-enhanced text classification model is as follows:

Given a training dataset \( \mathcal{S} = \{(x_i, l_i) | \forall i \leq N\} \), where \( N \) is the size of training dataset and \( x_i \in X \), we should recover the enhanced label \( d_i \) of \( x_i \) from the ordinary label vector \( l_i \), and thus we transform \( \mathcal{S} \) into an enhanced label training set \( \mathcal{P} = \{(x_i, d_i) | \forall i \leq N\} \).

In order to incorporate the correlation between category labels and instance sentences, we assume that each category label can be encoded by using its instance sentences in combination:

\[
\psi_m = \frac{1}{|X_n|} \sum_{X_n \in X_m} \sum_{Y_k \in Y} \theta^{m}_{j} v_j, 1 \leq m \leq |Y|, X_n = \{x_i | x_i \in X, L^m_n = 1\}
\]

(1)

where \( \theta^{m}_{j} \) denotes the weight of the relevance between \( y_m \) with \( v_j \), \( |X_n| \) denotes the size of \( X_n \), and \( |Y| \) denotes the size of \( Y \). \( \psi_m \) is the embedding representation of \( y_m \) whose instance is encoded. We use Term Frequency Inverse Document Frequency[6][7], an efficient and simple algorithm for matching words in a query to documents related to the query, to calculate the relative weight between the label and the corresponding word:

\[
\theta^{m}_{j} = \frac{f_{j, X_n} X_m \log \frac{|Y|}{F_{j, Y} + 1}}{1 \leq k \leq C}
\]

(2)

where \( f_{j, X_n} \) equals the counts that \( v_j \) appears in \( X_m \), and \( F_{j, Y} \) equals the number of labels in which \( v_j \) appears in \( Y \). \( C \) is total number of words in the corpus.

As \( x_i \) is a sequence of words, we encode \( x_i \) with the word embeddings of its words:

\[
X_i = \sum_{v_j \in x_i} u^j_i v_j
\]

(3)

where \( u^j_i \) denotes the weight of the relevance between \( x_i \) and \( v_j \). \( X_i \) is the embedding representation of \( x_i \). According to Jones, \( u^j_i \) can be formulated as follows:

\[
u^j_i = \frac{f_{j, X_n} \log \frac{|S|}{F_{j, X} + 1}}{1 \leq i \leq |S|, 1 \leq k \leq |x_i|}
\]

(4)
Where $f_{j|x_i}$ is the counts that $v_j$ appears in $x_i$, and $F_{j,S}$ is the number of instances where $v_j$ appears in $S$. $|x_i|$ is the length of $x_i$, and $|S|$ is the size of the training set.

The correlation between $y^m$ and $x_i$ can be formulated as:

$$r^m = \text{Corr}(\psi^m, \chi_i)$$

where $\text{Corr}$ is the correlation function, which represents the correlation between the $i$-th instance and the $c$-th label. For example, we can formulate $\text{Corr}$ with Spearman[8] correlation as:

$$\text{Corr}(\psi^m, \chi_i) = \frac{\sum_{j=1}^{q} (\chi_i^j - \bar{\chi}_i)(\psi^m_i - \bar{\psi}_m)}{\sqrt{\sum_{j=1}^{q} (\chi_i^j - \bar{\chi}_i)^2 \sum_{j=1}^{q} (\psi^m_i - \bar{\psi}_m)^2}}$$

where $\bar{\psi}_m$ represents the mean value of $\psi^m$ and $\bar{\chi}_i$ is the mean value of $\chi_i$. To further eliminate the impacts of the uncorrelated labels to an instance, we subtract the min element of the $r^i$:

$$\sigma_i = r_i - \min(r_1, r_2, ..., r^n_i)$$

Finally, we can obtain the enhanced text classification label for the instance:

$$d_i = (1 - \alpha)l_i + \alpha \sigma_i$$

The parameter $\alpha \in [0, 1]$ is the weight to control the degree of correlation between category labels and instance sentences. When $\alpha$ is set to be 0, the model will degenerate into a standard multi-class classifier. When $\alpha$ is set to be 1, the rough label-enhanced we get may not match the problem. Therefore, we could obtain a suitable enhanced label-enhanced with a proper value of $\alpha$ by experimental results.

### 3. Experiment Setup

To investigate the effect of our proposed label-enhanced text classification model, we conduct an experimental comparison between the benchmark model and the label-enhanced text classification model. Considering recent contextualized representations such as ELMo[9] and BERT[10] has been proven to be effective for improving the performances in various NLP tasks, we use BERT-Base as the benchmark model. Two different levels of text classification tasks are used as datasets in the experiment: text classification (document level) and sentiment analysis (sentence level).

#### 3.1. Text Classification

We use the text classification task (document level) to evaluate the effect of the label-enhanced text classification model on document-level text classification task. The following two datasets are used in the experiment.

**AG News**

AG News Corpus[11] contains news articles from the AG Corpus network related to the 4 largest categories. The dataset contains 30,000 training and 1,900 test examples for each class.

**DBpedia**

The DBpedia[12] ontology dataset contains 560,000 training sentences and 70,000 test sentences, each from 14 non-overlapping classes of DBpedia.

#### 3.2. Sentiment Analysis

Since most instances of sentiment analysis task consists of one or two sentences, we use two datasets to evaluate the effect of the label-enhanced text classification model on sentence-level text classification tasks.

**IMDb**

The IMDb[13] dataset is a binary sentiment analysis dataset, marked by 50,000 comments from the Internet Movie Database (IMDb) as positive or negative, respectively.
Yelp The Yelp review dataset contains more than 500,000 Yelp reviews. We use a fine-grained (five types) version of the dataset, which was obtained from the 2015 Yelp DataSet Challenge.

3.3. Imbalanced Datasets
Unbalanced dataset has been a problem in text classification tasks, which greatly hinders the performance of existing methods. The long-tail effect is widely seen in many text classification corpus. To verify the effect of LETC in the unbalanced classification dataset, for each dataset of text classification and sentiment analysis described above, we under sample one class in the dataset at a ratio of 1: 100. The under-sampled datasets are further used in the experiment.

4. Experimental Results
We conduct a comparison between the benchmark model and the label-enhanced text classification model on all the above tasks. We report classification accuracy, precision, recall and F1 in test set. From the table 1, we can see:

| Data Set | Model    | Accuracy | Precision | Recall | F1   |
|----------|----------|----------|-----------|--------|------|
| AG News  | baseline | 0.9425   | 0.9424    | 0.9425 | 0.9424 |
|          | LETC     | 0.9442   | 0.9440    | 0.9442 | 0.9441 |
| DBpedia  | baseline | 0.9919   | 0.9919    | 0.9918 | 0.9919 |
|          | LETC     | 0.9921   | 0.9921    | 0.9918 | 0.9919 |
| IMDb     | baseline | 0.9321   | -         | -      | -    |
|          | LETC     | 0.9345   | -         | -      | -    |
| Yelp     | baseline | 0.6953   | 0.7047    | 0.6953 | 0.6986 |
|          | LETC     | 0.6981   | 0.7052    | 0.6964 | 0.6997 |
| AG News  | baseline | 0.7347   | 0.8190    | 0.7349 | 0.6378 |
|          | LETC     | 0.7609   | 0.8424    | 0.7613 | 0.6703 |
| DBpedia  | baseline | 0.9727   | 0.9714    | 0.9731 | 0.9736 |
|          | LETC     | 0.9782   | 0.9789    | 0.9782 | 0.9776 |
| IMDb     | baseline | 0.5002   | 0.2503    | 0.5002 | 0.3333 |
|          | LETC     | 0.5026   | 0.2651    | 0.5023 | 0.3340 |
| Yelp     | baseline | 0.5145   | 0.4497    | 0.5181 | 0.4734 |
|          | LETC     | 0.5257   | 0.4580    | 0.5292 | 0.4835 |

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)
\]

\[
\text{precision} = \frac{TP}{TP + FP} \quad (10)
\]

\[
\text{recall} = \frac{TP}{TP + FN} \quad (11)
\]

\[
F_1 = 2 * \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (12)
\]

Our proposed label-enhanced text classification model outperforms the baseline model in all datasets, which verifies that incorporation of labels correlation is effective enhancement.

For unbalanced datasets, our methods gain significant improvements over baseline model in all datasets. For example, compared with baseline model, LETC gains 3.25% F1-score improvements in under-sampled AG News dataset. In the IMDB dataset, when the dataset becomes unbalanced, the performance of the baseline model will drop sharply. The reason is that if a category lacks training data, our model
can use the semantic correlation between labels and instance sentences to provide other supervised information of other labels to help training models for categories with few samples.

5. Conclusion
In this article, we propose a label-enhanced text classification model. Traditional text classification models usually use one-hot labels purely, but in actual scenarios, there is a certain correlation between labels. In order to capture the additional relevance information, we use not only the semantic information in the feature space, but also the correlation between instance sentences and category labels to improve the category label information, improving the classification performance of the text classification model. We evaluated our proposed label-enhanced text classification model on different levels of text classification task: text classification (document level) and sentiment analysis (sentence level). Experimental results show that incorporation of enhanced label achieves good improvement on both document-level and sentence-level text classification tasks. In further work, we will explore our model in other areas such as relation extraction, named entity recognition.

References
[1] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
[2] Geng X, Yin C, Zhou Z H. Facial age estimation by learning from label distributions[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(10): 2401-2412.
[3] T. Wang, L. Liu, N. Liu, H. Zhang, L. Zhang, and S. Feng, “A multi-label text classification method via dynamic semantic representation model and deep neural network,” Appl. Intell., pp. 1–13, Mar. 2020.
[4] J. Gong, H. Ma, Z. Teng, Q. Teng, H. Zhang, L. Du, S. Chen, M. Z. A. Bhuiyan, J. Li, and M. Liu, “Hierarchical Graph Transformer-Based Deep Learning Model for Large-Scale Multi-Label Text Classification,” IEEE Access, vol. 8, pp. 30885–30896, 2020.
[5] H. Azarbonyad, M. Dehghani, M. Marx, and J. Kamps, “Learning to rank for multi-label text classification: Combining different sources of information,” Nat. Lang. Eng., pp. 1–23, Feb. 2020.
[6] Jones K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1988, 60(5): 132-142.
[7] Jones K S. IDF term weighting and IR research lessons[J]. Journal of Documentation, 2004, 60(5): 521-523.
[8] Spearman C. The proof and measurement of association between two things. By C. Spearman, 1904.[J]. American Journal of Psychology, 1987.
[9] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.
[10] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[11] Del Corso G M, Gulli A, Romani F. Ranking a stream of news[C]//Proceedings of the 14th international conference on World Wide Web. 2005: 97-106.
[12] Lehmann J, Isele R, Jakob M, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.
[13] Maas A L, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. Association for Computational Linguistics, 2011: 142-150.