

A Hierarchical multi-input and output Bi-GRU Model for Sentiment Analysis on Customer Reviews

Liujie Zhang^{1, a}, Yanquan Zhou^{1, 2, b}, Xiuyu Duan^{1, b}, Ruiqi Chen^{1, b}

¹School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China

²Engineering Research Center of Information Networks, Ministry of Education, Beijing, China

^aliujiezhangupt@gmail.com, ^bzhouyanquan, duanxy, chenruiqi@bupt.edu.cn

Abstract: Multi-label sentiment classification on customer reviews is a practical challenging task in Natural Language Processing. In this paper, we propose a hierarchical multi-input and output model based bi-directional recurrent neural network, which both considers the semantic and lexical information of emotional expression. Our model applies two independent Bi-GRU layer to generate part of speech and sentence representation. Then the lexical information is considered via attention over output of softmax activation on part of speech representation. In addition, we combine probability of auxiliary labels as feature with hidden layer to capturing crucial correlation between output labels. The experimental result shows that our model is computationally efficient and achieves breakthrough improvements on customer reviews dataset.

1. Introduction

Sentiment analysis is an important research field in natural language processing, text mining, and computer linguistics [[1]]. By analyzing, processing, summarizing and reasoning of the subjective text, sentiment analysis aims to identify and extract the subjective information from original materials, such as attitudes, emotions, and opinions.

With the rapid development of the Internet in recent years, the ways that information is generated and disseminated become more and more diversified, especially for user generated content (UGC). It presents an explosive growth in the social, service consumption platform and media communication platform. The vast of majority of UGC data contains a certain aspect of emotional expression from user more or less. The value of efficiently processing and digging the emotional tendencies of UGC data is significant for business customers, which can greatly help to conduct feedback research on products or services, as well as product decisions.

In this paper, we follow the research work on multi-label sentiment classification of Qi. et al. [[11]], and propose a hierarchical multi-input and output model based bi-directional recurrent neural network, which can fully consider the semantic and lexical information of emotional expression. Moreover, Bi-GRU layer over word-related part of speech is learned as attention layer for sentiment words. This paper is organized as followed. We will firstly introduce current work on multi-label sentiment classification in section 2. Then, in Section 3, our model will be introduced in detail. Finally, the experimental results and conclusion will be presented and discussed in Section 4 and 5, respectively



2. Literature Survey

Sentiment analysis as an important branch of NLP, has been studied on a lot of topics, such as news, blogs, movie and product reviews [[2]]. Traditionally, it is mainly to analyze whether the attitude negative or positive, supported or opposed. Turney [[3]] and Pang [[4]] firstly proposed the problem on academic studies in 2002. Most of the research methods are mainly rule-based, emotional dictionary or statistical machine learning method. Kim [[5]] handled sentiment analysis with convolution neural network, but still considered it as multi-classification of single tag.

In real life, the emotional expression is usually a complex combination of emotion, such as the customer reviews of a product. It is more scientific and reasonable to regard sentiment analysis as a multi-label classification, which has drawn more attention of researchers. There are two ways to solve the multi-label classification problem: problem transformation and algorithm adaptation [[6], [7]]. The former method transformed multi-label classification problem into other known problem by processing data, such as Tsoumakis, Katakis et al. proposed Binary Relevance (BR) algorithm [[8]] in 2010, which resolves multi-label task into multiple binary classification task. Based on Bayes network, Zhang et al. proposed sophisticated Bayes networks to estimate the label dependencies [[9]]. But the model is more computationally expensive and time-consuming. Rai and Kumar et al. proposed multi-output relationship learning algorithm [[10]] to capture output correlation and designed model for regression problem rather than classification problem in 2012. The latter method retrofit existing algorithms to apply multi-label classification learning. For instance, ML-kNN transforms traditional kNN algorithm by maximizing the posterior probability. Rank-SVM algorithm transforms the traditional SVM algorithm [[11]].

In this section, Qi. et al reported the approaches of sentiment analysis on UGC. [[11]] They improved ML-kNN algorithm and proposed HML-kNN model by introducing auxiliary tags, which made a little bit of promotion in the result. However, the text similarity of the model is based on the words, which means large computational complexity and ignoring the context semantic association. The generalization ability of the model is not strong.

3. Proposed Method

Currently, recurrent neural networks has been widely used in the field of text analysis, especially in the context semantics [[12]]. In this work, we propose a hierarchical multi-input and output model (HMIO) based bi-directional recurrent neural network, which considers the semantic and lexical information of emotional expression. We employ Gated Recurrent Units (GRUs) as our recurrent neurons, which share many of same properties with LSTMs. GRU is a simpler variant of LSTMs.

3.1 Gated Recurrent Units

LSTM and GRU are the two most widely used network models of RNN. There are two gates for both GRU and LSTM layer: an update gate z and a reset gate r . Respectively, the update gate judges how much of the memory of previous cell to keep alive, and the reset gate decides how to combine the input of new cell with previous memory. But there is a little difference with LSTM: Firstly, a GRU cell has two gates instead of three. Secondly, the forget and input gates of GRU cell are coupled into the update gate z , and reset gate r is directly applied to the previous hidden state. Considering that GRUs have fewer parameters and thus may train faster, so we choose GRU as our recurrent neurons.

The function of each gate is described as followed:

$$\begin{aligned}
 z &= \sigma(x_t U^z + s_{t-1} W^z) \\
 r &= \sigma(x_t U^r + s_{t-1} W^r) \\
 h &= \tanh(x_t U^h + (s_{t-1} \odot r) W^h) \\
 s_t &= (1 - z) \odot h + z \odot s_{t-1}
 \end{aligned} \tag{1}$$

where $\odot$ stands for element-wise multiplication.

3.2 Model

Unlike the most common deep learning network with only one input and output layer, our model has two input and output layer, respectively. The auxiliary labels are placed in the middle output layer to constraint the learning of the part of the speech, while the middle outputs are also regarded as part of hidden feature for predicting final sentiment labels.

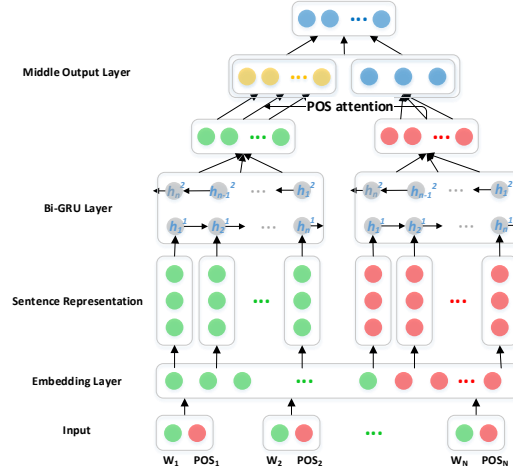


Fig.1 The architecture of hierarchical multi-input and output model based Bi-GRU

In Feature 1, we feed paired word and part of speech (POS) into the embedding layer, which embeds word and POS into fixed dimension semantic space [[13]]. The embedding layer consists of two parts: the red units represent POS embedding layer, and the green units represent word embedding layer. Each word w_i and POS_i is mapped to $\square^d$. We combine the two embedding layer into the same layer for updateable training.

Meanwhile, two independent Bi-GRU layers are applied to study the representation of the text words and POS. We use the product of the corresponding hidden unit of forward and backward layers as a semantic representation [[14], [15]].

$$h_i^w = h_i^{w,1} \square h_i^{w,2} \quad (2)$$

$$h_i^{pos} = h_i^{pos,1} \square h_i^{pos,2} \quad (3)$$

where $h_i^{w,1}$ and $h_i^{w,2}$ donate the output of the i -th hidden unit of the word forward and backward networks, with similar to $h_i^{pos,1}$ and $h_i^{pos,2}$.

In the hierarchical output structure, the lexical feature is used to predict the probability distribution of the auxiliary labels:

$$y_a = \sigma(W_a h^{pos} + b_a) \quad (4)$$

where σ is sigmoid function, y_a is the probability of the prediction on the auxiliary labels. In addition, we applied softmax activation function on the lexical feature to weight the output of word layer, which can be seen as a kind of attention mechanism for sentiment words.

$$h_{at}^w = \alpha \square h^w \quad (5)$$

$$\alpha_i = \frac{\exp(h_i^{pos})}{\sum_{i=1}^C \exp(h_i^{pos})} \quad (6)$$

h_{at}^w donates the output of word hidden layer after applying attention weighted with POS. α_i is i -th weight from softmax activation, C is number of hidden unit cells. $\square$ stands for element-wise multiplication. we observe the POS of the emotional word usually has a relatively pronounced lexical distribution, such as adjectives, adverbs and gerunds, which are more likely to be related to sentiment

expression. The emotional semantic of text can be modeled by weighted representation of word. Finally, the probability distribution of the auxiliary tag is also taken as a feature, which is combined with weighted word feature to predict the final sentiment labels.

$$y_c = \sigma(W_c h_c^w + b_c) \quad (7)$$

where h_c^w is combined feature with probability of the auxiliary tag and the word representation. y_c is final predicted probability of sentiment labels. In our model, we jointly minimize the cross-entropy of the auxiliary labels and the final sentiment labels:

$$L = - \sum_{x \in X} (\sum_{c=1}^C p_c^g(x) \log(p_c(x)) + \sum_{m=1}^M p_m^g(x) \log(p_m(x))) \quad (8)$$

where p^g is the ground truth of auxiliary labels or final labels. X is dataset samples. C and M are the numbers of auxiliary class and the final sentiment classes, respectively.

4. Experimental Results

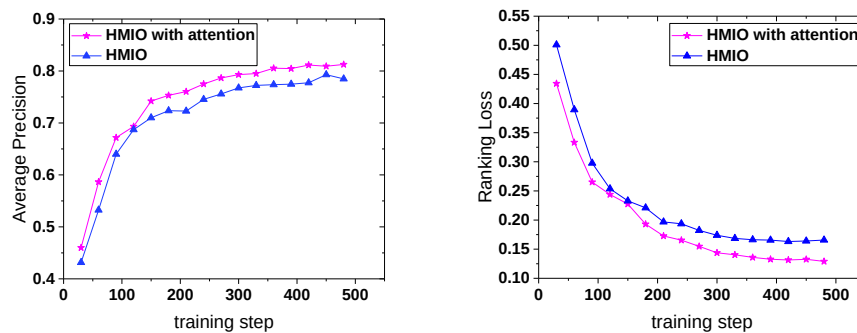
We selected 34799 customer reviews of mobile phones from www.jd.com website, which is same as Qi et al. These reviews data have been labeled to 6 major categories: satisfaction, disappointment, admiration, reproach, like and dislike. Furthermore, each sample is tagged with three different categories by different object: (1) reviews for purchase experience; (2) reviews for service providers; (3) reviews for goods (mobile phone). We use the three different object as middle layer output, six kinds of emotional tags as the final predicted labels.

we select traditional ML-kNN and HML-kNN algorithm as our baseline. Moreover, we cut off the softmax activation of lexical output connected with word feature to show the performance of attention layer clearly. That means, lexical feature is only used to predict auxiliary labels in this case.

Table.1 multi-label sentiment classification results on customer reviews

Models	Ranking Loss	Hamming Loss	One Error	Coverage	Average Precision
ML-kNN	0.2913	0.0913	0.4863	1.8512	0.4664
HML-kNN	0.2894	0.0908	0.4859	1.8436	0.4676
HMIO	0.1373	0.0844	0.3150	1.8141	0.7918
attention + HMIO	0.1291	0.0823	0.2343	1.6123	0.8126

Five kinds of multi-label indicators are applied to evaluate the performance of our models: Ranking loss, Hamming loss, One-error, Coverage, Average precision. Except for the last indicator is the bigger for better, the rest of the evaluation indicators are all the smaller for better. Table 1 demonstrates our model has a great improvement on the performance by the various indicators on customer reviews, indicating that our model can better capture the emotional semantic of the text and relationship between labels. The network structure that takes into account the part of speech network can make full use of pragmatic and grammatical information.



(a) Average precision over attention layer (b) Ranking loss over attention layer

Fig.2 Comparative experimental results of HMIO model in training process.

We compare the changes in the curve of ranking loss and average precision whether there is a POS attention layer in the model. Figure 2, the pink line shows the performance of attention mechanism with information of POS. From the result, we can observe that the ranking loss and average precision of model with lexical attention layer converge faster with better results in the training process. It demonstrates that additional lexical information contributes to help model easier capture the complex emotional semantics than only considering word.

5. Conclusion

In this paper, we have studied the multi-label sentiment classification for customer reviews based on Bi-GRU, which can accurately capture expression of emotional word semantic. The hierarchical multi-input and output model based on recurrent neural network can both effectively reduce the computational complexity and greatly improve the accuracy of prediction. In addition, softmax activation applied to lexical layer can be considered as an attention mechanism for word feature, which can make full use of the emotional semantic and lexical information. The middle layer of the auxiliary labels not only can well constrain the embedding layer of learning, but also can serve as effective feature for predicting the final sentiment labels.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China with grant No.71231002.

References

- [1] Chen, Huimin, et al. "Neural Sentiment Classification with User and Product Attention." EMNLP, 2016
- [2] Thakkar, Harsh, and Dhiren Patel. "Approaches for sentiment analysis on twitter: A state-of-art study." arXiv preprint arXiv:1512.01043, 2015
- [3] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
- [4] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
- [5] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882, 2014
- [6] Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." International Journal of Data Warehousing and Mining, 2006
- [7] Zhang, Min-Ling, and Zhi-Hua Zhou. "A review on multi-label learning algorithms." IEEE transactions on knowledge and data engineering, August, 2014, PP-1819-1837

- [8] Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Mining multi-label data." Data mining and knowledge discovery handbook. Springer US, 2009, pp-667-685
- [9] Zhang, Min-Ling, and Kun Zhang. "Multi-label learning by exploiting label dependency." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp-999-1008
- [10] Rai, Piyush, Abhishek Kumar, and Hal Daume. "Simultaneously leveraging output and task structures for multiple-output regression." Advances in Neural Information Processing Systems. 2012.
- [11] Qi, Hongwei, Yanquan Zhou, and Qi Guo. "A Hierarchical ML-kNN Method for Complex Emotion Analysis on Customer Reviews.", 2016
- [12] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." Acoustics, speech and signal processing (icassp), 2013 iee international conference on. IEEE, 2013
- [13] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014
- [14] Tai, Kai Sheng, Richard Socher, and Christopher D. Manning. "Improved semantic representations from tree-structured long short-term memory networks." arXiv preprint arXiv:1503.00075, 2015
- [15] Tang, Duyu, et al. "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." ACL (1). 2014