



Hierarchical Sentiment Estimation Model for Potential Topics of Individual Tweets

Qian Ji , Yilin Dai , Yinghua Ma , Gongshen Liu , Quanhai Zhang ,
and Xiang Lin 

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 800 Dongchuan RD, Shanghai 200240, China

{jeicy_good, lydai1108, ma-yinghua, lgshen, qhzhang, lionel}@sjtu.edu.cn

Abstract. Twitter has gradually become a valuable source of people’s opinions and sentiments. Although tremendous progress has been made in sentiment analysis, mainstream methods hardly leverage user information. Besides, most methods strongly rely on sentiment lexicons in tweets, thus ignoring other non-sentiment words that imply rich topic information. This paper aims to predict individuals’ sentiment towards potential topics on a two-point scale: positive or negative. The analysis is conducted based on their past tweets for the precise topic recommendation. We propose a hierarchical model of individuals’ tweets (HMIT) to explore the relationship between individual sentiments and different topics. HMIT extracts token representations from fine-tuned Bidirectional Encoder Representations from Transformer (BERT). Then it incorporates topic information in context-aware token representations through a topic-level attention mechanism. The Convolutional Neural Network (CNN) serves as a final binary classifier. Unlike conventional sentiment classification in the Twitter task, HMIT extracts topic phrases through Single-Pass and feeds tweets without sentiment words into the whole model. We build six user models from one benchmark and our collected datasets. Experimental results demonstrate the superior performance of the proposed method against multiple baselines on both classification and quantification tasks.

Keywords: Sentiment analysis in Twitter · Hierarchical model · Deep learning

1 Introduction

Nowadays, the information exploited from tweets is abundant and useful, thus receiving great attention from researchers. The task in this paper is to predict individuals’ sentiment towards potential topics on a two-point scale: positive or negative based on their past tweets. Generally, a user’s attitudes towards different topics are closely related and won’t change dramatically in a short time, so building models of individuals’ tweets and estimating sentiment polarities towards potential topics are beneficial for precise topic recommendations for

individuals, including related topics, advertisements and social circles. Earlier researchers [5] use a Support Vector Machines with part-of-speech features to categorize tweets. An adaptive recursive neural network for target-dependent classification is proposed, which propagates sentiment signals from sentiment-bearing words to specific targets on a dependence tree [4].

All the methods mentioned above ignore potential sentiment relations within individuals' tweets but rely heavily on sentiment lexicons. Besides, the models mentioned above only focus on the classification task while more practical applications require a combination of extraction and classification. In this work, we propose a hierarchical model of individuals' tweets, which extracts topics with Single-Pass algorithm and models the relationship between individual sentiments and different topics. The main contributions of this paper are three-fold:

- Models are built on individuals' tweets and the topic phrase of each tweet is obtained through Single-Pass. Individuals' tweets without sentiment words, along with extracted topic and gold labels are inputs of HMIT. Based on the approach, it's possible to provide precise topic recommendations for individuals.
- We propose a novel topic-dependent hierarchical model, which extracts features from fine-tuned BERT and incorporates topic information through topic-level attention. CNN categorizes sentence representations into positive or negative.
- We build models on six users separately from one Twitter benchmark dataset and dataset collected by ourselves. We also create new test dataset, collecting neutral sentences from three general topics. In our experiments, the proposed method is able to outperform multiple baselines on both datasets in terms of classification and quantification.

2 Related Work

Target-based sentiment analysis aims to judge the sentiment polarity expressed for each target being discussed. To capture semantic relations flexibly, a target-dependent Long Short-Term Memory (TD-LSTM) is proposed [10]. As attention mechanism has been successfully applied to many tasks, a variety of attention-based RNN models have proven to be effective ways [12]. To our knowledge, we are the first to exploit target-individual relation for target-based sentiment analysis. Hierarchical models have been used predominantly for representing sentences. A hierarchical ConvNet to extract salient sentences from reviews is employed in [2]. Along with the wide use of pre-trained language models, there is a recent trend of incorporating extra knowledge to pre-trained language models as a different hierarchical model [1]. For BERT, it is difficult to be applied to downstream tasks which need to put emphasis on several specific words. We propose a hierarchical model that extracts overall information from fine-tuned BERT and then incorporates topic information. CNN categorizes the whole sentence representation into positive and negative.

3 Proposed Method

The HMIT architecture is shown in Fig. 1. We describe each component and how it is used in learning and inference in detail. For one user, $\{s_1, s_2, \dots, s_m\}$ is a collection of his/her tweets, containing m tweets of various topics. A tweet s_i composed of n words is denoted as $s_i = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}$ with a gold sentiment label $y_i = \{POSITIVE, NEGATIVE\}$.

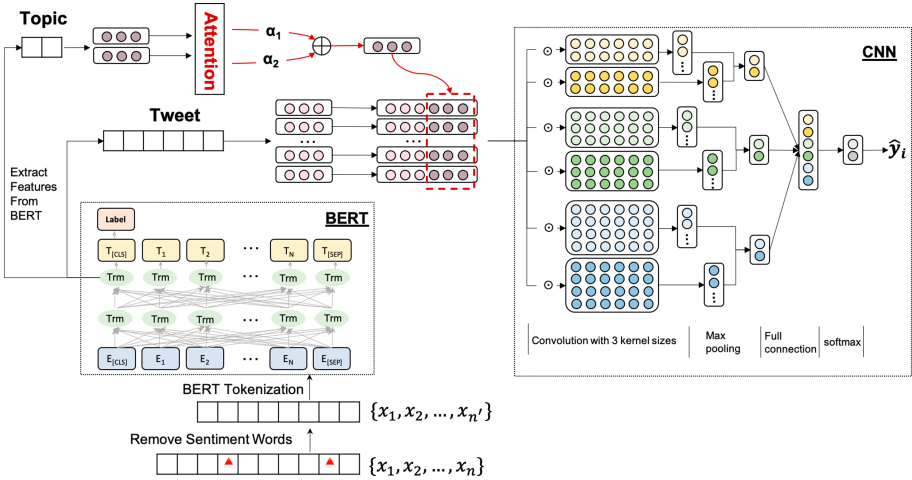


Fig. 1. The overall architecture of HMIT

3.1 Topic Phrases Extraction

To extract several topic phrases in each tweet, we employ Single-Pass algorithm. The core idea is to input texts continuously to determine the matching degree between the input text and an existing cluster. Texts whose maximum similarity to the cluster core are greater than the given threshold p_0 will be clustered as one category. After all texts are clustered, we set the most frequently occurred bi-gram as topic phrases for this cluster, so that every tweet is associated with a topic phrase $\{TP_1, TP_2\}$.

3.2 Fine-Tuned BERT with Non-sentiment Words

All sentiment words are removed from tweets first according to sentiment lexicons [6]. $\{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}$ represents a tweet without sentiment words and is further fed into BERT tokenization. Each sentence is tokenized and padded to length N by inserting padding tokens. The embedding layer of BERT integrates word, position and token type embeddings where $E_j \in \mathbb{R}^K$ is the K -dimensional vector of the j -th word in the tweet. BERT is a multi-layer bidirectional Transformer

encoder. In text classification, the decoder applies first token pooling to a full connection layer with softmax activation, returning a probability distribution on two categories. After fine-tuning BERT on our own dataset, we extract one layer of the latent vector from the encoder of fine-tuned BERT.

3.3 Topic-Level Attention

We use a topic-level attention mechanism over a topic phrase to produce a single representation. Since different tokens in a topic phrase may contribute to its semantics differently, we calculate an attention vector for a topic phrase. The hidden outputs corresponding to $\{TP_1^{(i)}, TP_2^{(i)}\}$ is denoted as $H^{(i)} = \{h_{TP_1}^{(i)}, h_{TP_2}^{(i)}\}$. We compute the aggregated representation of a topic phrase as

$$H^{(i)'} = \alpha^{(i)\top} H^{(i)} = \sum_{o \in \{1,2\}} \alpha_j^{(i)} h_{TP_o}^{(i)} \quad (1)$$

where the topic attention vector $\alpha^{(i)} = \{\alpha_1^{(i)}, \alpha_2^{(i)}\}$ is distributed over topic phrase $H^{(i)}$. The attention vector $\alpha^{(i)}$ is a self-attention vector that takes the hidden outputs of a topic phrase as input and feeds them into a bi-layer perceptron. We concatenate each token representation and the aggregated topic representation $H^{(i)'}$ to obtain the final context-aware representation for each word.

3.4 CNN Classification

CNN has grabbed increasing attention in text classification tasks recently due to its strong ability to capture local contextual dependencies. Based on that, we propose to apply CNN to the final layer of classification. As shown in Fig. 1, convolution operation involves kernels with three different sizes. Suppose $w \in \mathbb{R}^{q \times 2K}$ is a filter of q tokens, a feature c_j is generated by:

$$c_j = f(w \circ h_{j:j+q-1}^{(i)'} + b) \quad (2)$$

Here \circ denotes convolution, while $b \in \mathbb{R}$ is a bias term and f is ReLU activation function. This filter applies to whole possible tokens in the sentence to produce a feature map:

$$c = [c_1, c_2, \dots, c_{N-q+1}] \in \mathbb{R}^{N-q+1} \quad (3)$$

Max-pooling layer take the maximum value $\hat{c} = \max\{c\}$ of c as the feature corresponding to filter w . \hat{y}_i denotes the predicted label for the i -th tweet.

3.5 Inference and Learning

The objective to train topic-level attention and CNN classifier is defined as minimizing the sum of the cross-entropy losses of prediction on each tweet as follows:

$$\mathcal{L}_s = - \sum_{i=1}^m \hat{y}_i \log(y_i) + (1 - \hat{y}_i) \log(1 - y_i) \quad (4)$$

For inference, test news is first passed to the fine-tuned BERT to obtain its hidden vector. According to the topic-level attention mechanism, context-aware representations are incorporated with topic information and then fed to the CNN classifier. Finally, a prediction is obtained.

4 Experiments

4.1 Experimental Settings

Datasets. Table 1 shows the statistics of datasets. We select three users from Sentiment140 [9] to build models separately as \mathbb{D}_{s1} , \mathbb{D}_{s2} and \mathbb{D}_{s3} . We also collect tweets from three talkative users and label them manually as \mathbb{D}_{t1} , \mathbb{D}_{t2} and \mathbb{D}_{t3} . To verify the feasibility of the method in practical application, we collect 100 news for each of three topics: *health care*, *climate change*, *social security* as \mathbb{T}_h , \mathbb{T}_c and \mathbb{T}_s .

Table 1. Dataset statistics

Datasets	Train			Test		
	Positive	Negative	Total	\mathbb{T}_h	\mathbb{T}_c	\mathbb{T}_s
\mathbb{D}_{t1}	1243	1038	2281	100	100	100
\mathbb{D}_{t2}	882	1044	1926	100	100	100
\mathbb{D}_{t3}	772	496	1268	100	100	100
\mathbb{D}_{s1}	119	119	238	—	—	—
\mathbb{D}_{s2}	98	183	281	—	—	—
\mathbb{D}_{s3}	66	210	276	—	—	—

Network Details. For Single-Pass, we set the threshold p_0 to 0.4. We tune pre-trained base uncased BERT which sets hidden size K as 768 with 12 hidden layers and 12 attention heads. Max sequence length N , batch size and learning rate are set to 128, 32 and 5×10^{-5} respectively. For the CNN classifier, we adopt three filter sizes: 2, 3 and 4 separately. 64 filters are used for each filter size and three pooling sizes are set to 4 in the task. We train the fine-tuned BERT for 3 epochs and the CNN classifier for 20 epochs.

Evaluation Metrics. We employ accuracy and F1 score as evaluation for classification. We regard evaluation of test sets as a quantification task, which estimates the distribution of tweets across two classes. We adopt Mean Absolute Error based on a predicted distribution \hat{p} , its true distribution p and the set \mathcal{C} of classes. It's computed separately for each topic, and the results are averaged across three topics to yield the final score.

4.2 Models Under Comparison

We compare our proposed method with the methods that have been proposed for sentiment analysis (SA) and target-based sentiment analysis (TBSA).

- BERT [3]: BERT achieves state-of-the-art results in sentence classification, including sentiment classification.
- mem_absa [11]: Mem_absa adopts a multi-hop attention mechanism over an external memory to focus on the importance level of the context words and the given target.
- IAN [8]: IAN considers both attention mechanisms on the target and the full context. It uses two attention-based LSTMs to interactively capture the keywords of the target and its content.
- Cabasc [7]: Cabasc takes into account the correlation between the given target and each context word, composed of sentence-level content attention mechanism and content attention mechanism.

BERT^{-s}, mem_absa^{-s}, IAN^{-s} and Cabasc^{-s} are variant models of BERT, mem_absa, IAN and Cabasc respectively, removing sentiment words in training and testing.

4.3 Results and Analysis

Main Results. From Table 2, we observe that HMIT is able to significantly outperform other baselines in both classification and quantification tasks on our own dataset, which suggests that our proposed method is effective to capture the relationship between individual sentiments and different topics and succeed in sentiment estimation towards potential topics. We also find that BERT performs reasonably well on validation sets, which confirms its strong ability to represent a whole sentence and its feasibility as the first layer of our model. We compare HMIT with one SA model, three TBSA models and their variants. Table 2 shows that TBSA models display little advantage compared with SA models, which implies that current sentiment classification is mostly decided by sentiment lexicons or opinion words around the target instead of the target itself. Furthermore, the superior performance of the variants on both datasets indicates that removing sentiment words from tweets enables models to pay more attention to the topic in a tweet, thus constructing the relationship between topics and individual sentiments.

Table 2. Comparison results

Model	\mathbb{D}_{t1}		\mathbb{D}_{t2}		\mathbb{D}_{t3}		\mathbb{D}_{s1}		\mathbb{D}_{s2}		\mathbb{D}_{s3}	
	F1	MAE	F1	MAE	F1	MAE	Acc	F1	Acc	F1	Acc	F1
mem_absa	81.23	0.203	81.49	0.215	81.52	0.239	75.35	74.91	72.46	72.12	73.94	73.45
mem_absa ^{-s}	80.28	0.183	80.72	0.186	78.37	0.192	73.91	73.92	71.93	71.47	72.03	72.01
IMN	83.42	0.168	82.04	0.167	82.26	0.170	76.89	75.94	74.92	74.32	74.78	74.50
IMN ^{-s}	81.74	0.161	81.72	0.159	80.64	0.162	76.34	75.82	74.18	74.24	74.39	74.12
Cabasc	83.99	0.169	83.74	0.165	82.02	0.174	76.53	76.47	73.41	73.28	73.86	73.89
Cabasc ^{-s}	82.31	0.163	82.66	0.160	81.32	0.169	75.99	75.61	73.99	73.84	73.35	73.70
BERT	81.79	0.161	83.11	0.153	82.65	0.172	77.35	77.11	76.34	75.89	75.32	74.29
BERT ^{-s}	79.76	0.156	82.70	0.148	81.64	0.163	76.83	76.55	75.02	74.21	74.86	74.04
HMTI	85.30	0.112	86.61	0.109	83.91	0.124	78.29	77.13	75.96	74.08	75.50	74.84

Extract Features from BERT. We discover which encoding layer extracted from BERT is the most appropriate for further modification and classification. We extract features from -3, -2 and -1 encoding layer of BERT and simply add a CNN classifier after that. In Fig. 2, we report the accuracy and F1 score of cross-validation on \mathbb{D}_{t2} . It turns out that the penultimate layer is the most appropriate to make changes or incorporate external information. The last layer is too close to the target and the previous layers may not have been fully learned semantically. Therefore, we extract the penultimate layer in the method.

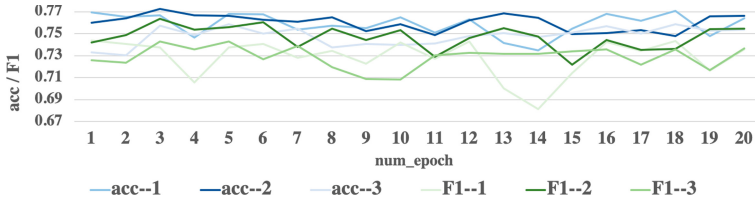


Fig. 2. Accuracy and F1 score on \mathbb{D}_{t2} with different BERT layer

5 Conclusion

We have proposed a hierarchical model to make individual sentiment estimation of potential topics. The approach extracts topics automatically and models the relationship between individual sentiments and different topics. It takes as input tweets without sentiment words, extracts features first from fine-tuned BERT and then incorporates topic information in context-aware token representation through the topic-level attention mechanism. CNN further classifies the representation into positive or negative. The proposed architecture can potentially be applied for a precise individual recommendation or group sentiment estimation towards one topic.

Acknowledgment. This research work has been funded by the National Natural Science Foundation of China (Grant No. 61772337), the National Key Research and Development Program of China NO. 2016QY03D0604.

References

1. Bao, X., Qiao, Q.: Transfer learning from pre-trained bert for pronoun resolution. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp. 82–88 (2019)
2. Denil, M., Demiraj, A., De Freitas, N.: Extraction of salient sentences from labelled documents. arXiv preprint [arXiv:1412.6815](https://arxiv.org/abs/1412.6815) (2014)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)
4. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: Adaptive recursive neural network for target-dependent Twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 49–54 (2014)
5. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project report, Stanford 1(12), 2009 (2009)
6. Liu, B., Hu, M.: Opinion lexicon (or sentiment lexicon) (2004). <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>
7. Liu, Q., Zhang, H., Zeng, Y., Huang, Z., Wu, Z.: Content attention model for aspect based sentiment analysis. In: Proceedings of the 2018 World Wide Web Conference, pp. 1023–1032. International World Wide Web Conferences Steering Committee (2018)
8. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 4068–4074. AAAI Press (2017)
9. Mohammad, S., Kiritchenko, S., Zhu, X.: NRC-Canada: building the state-of-the-art in sentiment analysis of Tweets. In: Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA, June 2013
10. Tang, D., Qin, B., Feng, X., Liu, T.: Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers, pp. 3298–3307 (2016)
11. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 214–224 (2016)
12. Wang, J., et al.: Aspect sentiment classification with both word-level and clause-level attention networks. In: IJCAI, pp. 4439–4445 (2018)