AGTG: Text Sentiment Analysis on Social Network Domain Based on Pre-Training and Regular Optimization Techniques

Zufeng Wu, Jingyou Peng, Ruiting Dai and Songtao Liu
University of Electronic Science and Technology of China, Chengdu, China

Keywords: Further Pre-Training, Word Embedding Aggregation, Regularization Optimization.

Abstract: The data volume of social networks is very large nowadays, and the information contained in social texts can be used in various application scenarios, especially the information of emotional features revealed in the texts, which is needed to be studied in the fields of opinion control and risk assessment, so text sentiment analysis for social networks is an important research. Many advanced modeling techniques suffer from high arithmetic requirements for large number of parameters, difficulty in extracting keyword information, lack of training data volume leading to overfitting phenomenon and catastrophic forgetting in the fine-tuning phase. Aiming to improve the sentiment classification effect of short texts in the textual task domain of social networks, we propose a new text sentiment analysis algorithm, which contains three important components: 1. Use social networks in the pre-training phase The model is further pre-trained using text data from the task domain in the pre-training phase; 2. A dynamic-static word embedding aggregation method is used to enrich the semantic representation information of the text; 3. The loss function is tuned by adding a trust domain smoothing control adversarial regular optimization method in the fine-tuning phase. Our experiments show that the proposed algorithm achieves new optimal performance in the social network domain.

1 INTRODUCTION

Social networks, as one of the most popular social communication platforms today, have become one of the main channels for people to express their emotions and opinions. Text sentiment analysis in the field of social networks is also an important research direction in natural language processing, whose main task is to analyze and understand the sentiment information contained in the text content people post on social media platforms. With the continuous development of deep learning technology, pre-trained models have become one of the most advanced natural language processing methods and have achieved great success in text sentiment analysis in the social network domain. Aiming to improve the sentiment classification effect of short texts in the textual task domain of social networks, we address the problems of high arithmetic requirements for model parametric quantities, difficulty in extracting keyword information, lack of training data volume leading to overfitting phenomenon and catastrophic forgetting in the fine-tuning phase. We make a series of improvements in the pre-training and fine-tuning phases:

1. The pre-training phase uses text data from the social network task domain to further pre-train the ALBERT pre-training model.

2. The text representation layer uses a new dynamic-static word embedding aggregation method to refine the text information.

3. The loss function is tuned in the fine-tuning phase using smoothly **G**overnance adversarial **R**egularization optimization for **A**ccreditatio**N** **D**omain, referred to as **GRAND**.

The model is fine-tuned using TextCNN structure in fine-tuning, so we propose the ALBERT-GloVe-TextCNN-GRAND model, referred to as AGTG for social network text sentiment analysis.

2 BACKGROUND

BERT[0] is a novel pre-trained language model released and introduced by Google's AI team in 2018, and its application domain covers many tasks in natural language processing, such as text question-answer discrimination and text sentiment computation. The BERT model has also evolved over time, with variants such as ALBERT[0] and RoBERTa[0]. Among

Wu, Z., Peng, J., Dai, R. and Liu, S.
AGTG: Text Sentiment Analysis on Social Network Domain Based on Pre-Training and Regular Optimization Techniques.
DOI: 10.5220/0012283100003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 338-343
ISBN: 978-989-758-677-4

them, the ALBERT pre-training model is considered as a streamlined version of BERT, which reduces the model computation by word embedding parameter factorization and cross-layer parameter sharing, thus improving the training efficiency of the model. In addition, ALBERT introduces Sentence-Order Prediction to refine its understanding of semantic relations between sentences, which in turn leads to improved model accuracy. Compared with BERT, ALBERT achieves a significant reduction in the number of parameters while maintaining model performance. After pre-training the semantic representation of the language model, the model framework is fine-tuned by adjusting it to suit the downstream tasks. This is done by replacing the top layer of the language model with a task-specific layer and then continuing the training on the downstream task.

The regular optimization method we utilize in the fine-tuning phase combines two existing techniques, namely smoothness inductive adversarial[0] and Bregman approximation point optimization[0]. Smoothed inductive adversarial regularization is a regularization framework that combines inductive and adversarial learning, exploiting the idea of adaptive smoothing in inductive learning and introducing adversarial training in order to improve the robustness and generalization of the model; Bregman approximation point optimization is an approximation algorithm commonly used for optimization problems by constructing an approximation to the Bregman function in each iteration. The purpose of introducing this method when training neural networks is to minimize the loss function by adjusting the model weights. During each iteration round, the Bregman scatter is essentially a powerful regularizer that prevents the points after the iteration from differing too much from the points in the previous iteration.

3 PROPOSED METHOD

3.1 Within-Task Pre-Training

ALBERT is a model obtained by pre-training in the general-purpose domain, using a 40G Chinese corpus as its pre-training corpus, including more than 10 billion Chinese characters, which consists of various sources such as encyclopedias, news articles and interactive community comments. The data distribution between the generic domain and the target task domain is clearly different, and many studies have shown that within-Task pre-training pre-trained models in the target task domain can improve the performance of pre-trained models in downstream tasks[0].

To further improve the short text sentiment classification in the social network text task domain, we chose an open source text dataset, also in the social network domain, to further pre-train the ALBERT model, which is a sentiment analysis dataset of Weibo text from ChineseNlpCorpus, a Chinese natural language processing corpus project. The model is saved as a checkpoint for every 50k steps of training, and the final training ends with 300k steps of further pre-training, which will be followed by comparative experiments to verify the impact of further pre-training ALBERT on the final classification effect of the final model.

3.2 Dynamic-Static Word Embedding

We introduce the dynamic static word embedding aggregation method in the representation module of the model. The use of static word vectors alone has obvious drawbacks in the problem of multiple meanings of words, while using only dynamic word vectors slows down the convergence of the model.
To solve the above problems, we consider aggregating the ALBERT and GloVe pre-training models so that the text representation vectors output from the ALBERT pre-training model and the GloVe pre-training model are spliced and fused to form the embedding matrix of the models to obtain the final text representation matrix. The word embedding matrices obtained from the ALBERT model and the GloVe model are $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$,with d_1 and d_2 denote the word vector dimensions. The final representation matrix is $Z=[X, Y]=[x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n]$, $Z \in \mathbb{R}^{n \times (d_1+d_2)}$. The specific aggregation is shown in Figure 1.

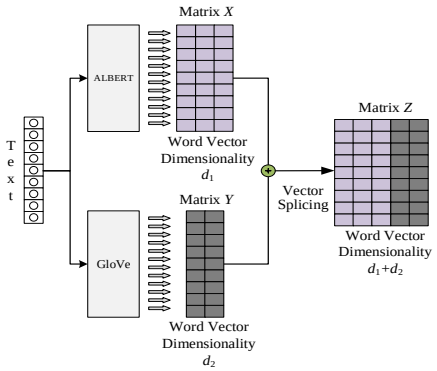

Figure 1. Schematic diagram of dynamic-static word embedding aggregation.

3.3 Grand Regular Optimization

We propose a new regular optimization algorithm for stable and efficient fine-tuning of the text sentiment classification model, which can solve the problems of poor generalization and catastrophic forgetting. The catastrophic forgetting problem in the fine-tuning phase.

The purpose of GRAND is to force the model to make similar classification predictions at the proximal points in the trust domain. First, an adversarial regularization perturbation term is imposed in the fine-tuning phase of the model, given the AGTG model denoted as $f(\cdot;\omega)$ and n data sample points in the target task domain, which solves the optimization problem of the objective function $\mathcal{F}$ in the fine-tuning phase, where ω is the model parameter, $L(\omega)$ is the overall loss function, λ is the adjustment parameter greater than 0, and $Re(\omega)$ is the adversarial regularization term, as follows:

$$\min_w F(\omega) = L(\omega) + \lambda Re(\omega) \quad (1)$$

where the specific formula for the regularization term $Re(\omega)$ is defined as follows:

$$Re(\omega) = \frac{1}{n}\sum_{i=1}^{n} \max_{\left\| \tilde{x}_i - x_i \right\|_p \leqslant \delta} \ell\left(f(x_i;\omega), f\left(\tilde{x}_i;\omega\right)\right)$$

and δ is the perturbation parameter greater than 0, and $\tilde{x}$ is the training sample to which the random perturbation is added.

The specific formula of the loss function $L(\omega)$ is defined as follows:

$$L(\omega) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{AGTG}}\left(f(x_i;\omega), y_i\right)$$

where $\ell_{\text{AGTG}}(\cdot,\cdot)$ denotes the loss function determined by the specific task of the model, x is the training sample, and y is the label corresponding to the training sample.

In addition to imposing the regularized perturbation term described above, a Bregman approximation point optimization method is used to add an anti-jitter mechanism in each model training iteration to prevent large updates of the model. For AGTG, during the course of the (t+1)-th iteration, the model parameters change as follows:

$$\omega_{t+1} = \text{argmin}_w F(\omega) + \gamma \nabla_{\text{Bre}}\left(\omega, \omega_t\right) \quad (2)$$

where γ is the adjustment parameter greater than 0 and $\nabla_{\text{Bre}}(\cdot,\cdot)$ is the Bregman scatter.

The specific calculation of the Bregman scatter in the above equation is as follows, where ℓ denotes the cross-entropy loss function

$$\nabla_{\text{Bre}}\left(\omega, \omega_t\right) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f\left(x_i, \omega\right), f\left(x_i, \omega_t\right)\right)$$

The Bregman scatter of each iteration essentially acts as a strong regularization term, which is intended to prevent deviations from the previous iteration too much, and is therefore called a trust domain iteration. By imposing such an anti-dithering mechanism, the knowledge learned by the model before fine-tuning can be effectively retained.

Also to accelerate the approximation point optimization method, additional momentum is added to the process of updating the model parameters in equation (2) above. The resulting model parameter iteration formula during the (t+1)-th iteration can be transformed into the following equation:

$$\omega_{t+1} = \text{argmin}_w F(\omega) + \gamma \nabla_{\text{Bre}}\left(\omega, \tilde{\omega}_t\right)$$

where $\tilde{\omega}_t$ is the sliding average of the model parameters, also called exponentially weighted average, making the update of the parameter variables related to the historical values taken over time, according to the mean teacher method proposed by Tarvainen[0] et al, which is calculated as follows:

$$\tilde{\omega}_t = (1-\alpha)\omega_t + \alpha\tilde{\omega}_{t-1},\ \alpha \in (0,1)$$

3.4 Architecture

The model first uses the ALBERT-GloVe word embedding aggregation method to splice and fuse the dynamic word vectors obtained by the ALBERT pre-training model and the static word vectors obtained by the GloVe pre-training model into dynamic and static word vectors, which effectively solves the problem of multiple meanings of a word while improving the convergence speed and classification effect. Then, the fused dynamic-static word vectors are used as the input of the convolutional layer network TextCNN in the downstream fine-tuning task, and multiple convolutional kernels are used to extract the local semantic information features in the text representation. Finally, the feature vectors obtained after the maximum pooling process and fully-

connected splicing operation are sent to the softmax classifier to calculate the final sentiment polarity classification results, where the GRAND regular optimization algorithm is used to add perturbations during the training process to improve the model robustness and generalizability. The overall network architecture of our proposed AGTG text sentiment classification model is shown in Figure 2.

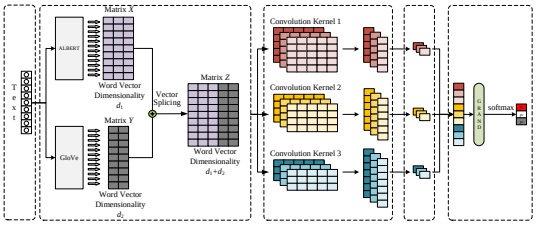

Figure 2. AGTG architecture diagram

4 EXPERIMENT

4.1 Parameters

The experiment further pre-trained the adopted ALBERT pre-training model in the target task domain by performing the ITPT task with the following key parameters: train_batch_size was set to 4096, eval_batch_size was set to 64, max_seq_length was set to 256, max_predictions_per_seq was set to 20, num_train_steps was set to 300000, num_warmup_steps was set to 1500, learning_rate was set to 1e-3, save_checkpoints_save was set to 1e-3, predictions_per_seq was set to 20, num_train_steps was set to 300000, num_warmup_steps was set to 1500, learning_rate was set to 1e-3, save_checkpoints_steps was set to 50000.

With the other parameters fixed, the model parameters involved in the experiments were changed over several experiments to obtain the parameter configuration that allowed the experiments to classify better, and the key parameters were set as follows: train_epochs was set to 8, batch_size was set to 64, max_seq_len was set to 256, learning_rate was set to 3e-5, activation function is ReLU, dropout_rate was set to 0.1, convolution kernel size is 3,4,5, and the number of convolution kernels was set to 128.

For GRAND, the adjustment parameters λ and γ in both adversarial regularization and approximate point optimization were set to 0.5, and the regularization perturbation parameter δ was set to 1e-5. To simplify the iterations, the number of regular optimization iterations T_b and T_x were both set to 2, the update parameter μ of $\tilde{x}$ was set to 1e-3, the initialized standard deviation σ of the perturbed samples was set to 1e-5, and the acceleration parameter α was set to 0.3.

4.2 Comparison with Baseline

In order to verify the advantages of our proposed AGTG, some representative models in text sentiment analysis studies were selected as baseline models, and comparative experiments on model performance were conducted in the same experimental environment, mainly including the following models:

- BERT[0]: Devlin et al. proposed a bi-directional Transformer architecture for pre-training models, and the officially released Chinese model was pre-trained on a large Chinese corpus to make it highly capable of language comprehension.
- BERT-wwm[0]: Cui et al. propose a variant based on the BERT model to improve the pre-training effect of BERT by replacing the traditional split-word masking with whole-word masking, which makes full use of the characteristics of Chinese language and enables the model to better understand the relationships and semantic meanings between words, thus achieving better performance in NLP tasks.
- ERNIE[0]: Zhang et al. propose a Transformer-based pre-trained language representation model that aims to integrate entity information and relational knowledge to enhance the representation capability of pre-trained language models, employing a term masking mechanism to mask entities in the input text and requiring the model to predict the masked entities based on their context and external knowledge, thus integrating knowledge and linguistic semantic information together, allowing the model to learn entity representations and their relationships more efficiently, thus improving the performance of NLP tasks.
- RoBERTa[0]: Liu et al. propose a reinforced optimised pre-trained self-supervised model based on BERT that aims to improve the performance of BERT by using more pre-trained data, larger batch sizes and dynamic masks used to replace static masks.
- DeBERTa[0]: He et al. propose a self-supervised pre-training model based on BERT, introducing a decoding enhancement technique that allows the model to better capture relations in sentences by adding the output of the decoder to a self-attentive mechanism, in addition to using a decomposed attention mechanism to better distinguish between different features.

We conducted model comparison experiments under the baseline and AGTG models using three different sentiment analysis datasets. Below is a brief description of the three datasets we used. The main experimental results are shown in Table 1.

- QZS: The raw text data of this dataset was crawled from our public social networks, and after pre-processing, polarity annotation and corpus enhancement, it contains more than 70,000 Chinese data classified into three categories: positive, negative and neutral.
- SST-2: known as Stanford Sentiment Treebank 2, is a dataset for sentiment analysis. The dataset contains various types of sentences such as movie reviews, TV reviews and product reviews, each of which is labelled with the sentiment polarity they express.
- weibo_senti_100k: It is a Chinese sentiment analysis dataset containing 100,000 Chinese sentences from Sina Weibo. The sentences cover a variety of topics, such as entertainment, sports, politics, etc. Each sentence was labelled as positive, negative or neutral sentiment.

Table 1. Comparison results on the three datasets.

Model	QZS		SST-2		weibo_senti_100k	
	Acc	F1	Acc	F1	Acc	F1
BERT	84.2	85.1	93.2	90.9	87.9	89.1
BERT-wwm	84.7	85.5	94.8	92.4	88.2	90.3
ERNIE	85.4	86.5	95.1	92.5	88.4	89.9
RoBERTa	86.2	87.1	96.5	94.6	88.9	90.7
DeBERTa	86.3	87.4	**97.2**	96.1	89.1	91.9
AGTG	**86.9**	**88.0**	97.0	**96.3**	**89.4**	**92.0**

Based on the reported accuracy and F1 scores on the QZS, SST-2, and weibo_senti_100k datasets, DeBERTa and RoBERTa consistently outperform other models in terms of both accuracy and F1 score. However, it is worth noting that AGTG also achieves competitive results across all three datasets.

Specifically, on the QZS dataset, AGTG achieves an accuracy of 86.9% and an F1 score of 88.0%, which is comparable to DeBERTa and RoBERTa. On the SST-2 dataset, AGTG achieves an accuracy of 97.0% and an F1 score of 96.3%, which is only slightly lower than DeBERTa, but outperforms all other models including RoBERTa. Similarly, on the weibo_senti_100k dataset, AGTG achieves an accuracy of 89.4% and an F1 score of 92.0%, which is also competitive with DeBERTa and RoBERTa. We can conclude that AGTG shows a competitive performance in sentiment analysis tasks. In particular, AGTG is effective for the application area of social network text sentiment analysis.

4.3 Ablation Study

We investigate the effectiveness of these improvements for all the proposed improvement components of our proposed AGTG model: within-task pre-training, Dynamic-Static Word Embedding, and GRAND.AGTG represents our proposed model. We then denote the model without pre-training as -p, the model with dynamic pre-training removed as -d, the model with static pre-training removed as -s, and the model without GRAND algorithm as -G.

Table 2. Ablation studies for sentiment analysis tasks on two datasets.

Model	QZS		weibo_senti_100k	
	Acc	F1	Acc	F1
AGTG	**86.9**	**88.0**	**89.4**	**92.0**
$AGTG_{-p}$	83.7	85.1	85.6	87.2
$AGTG_{-d}$	82.4	83.9	86.1	88.0
$AGTG_{-s}$	86.3	87.6	88.9	90.9.
$AGTG_{-G}$	85.3	86.7	87.3	89.1

It can be seen from this that AGTG showed superior results to the other models in both the QZS and weibo_senti_100k social text datasets for both Acc and F1 metrics. For example, on QZS, removing within-task pre-training leads to a 3.2% decrease in accuracy of the model, while removing GRAND leads to a 1.6% decrease in accuracy. This suggests that AGTG is able to better capture the semantic information of the text when processing social text, thus improving the classification accuracy.

5 CONCLUSION

We propose a text sentiment analysis model, AGTG, for the social network domain. The model uses text data from the social network task domain for within-task pre-training in the pre-training phase, and uses a kinesthetic word embedding aggregation method to

enrich the semantic representation information of the text, and adds the GRAND method to adjust the loss function in the fine-tuning phase. Our experimental results show that AGTG can effectively improve the classification performance of sentiment analysis in the social network domain, providing directional ideas for related research in the social network domain.

REFERENCES

Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv, 2018, preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations [J]. arXiv, 2019, preprint arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach [J]. arXiv, 2019, preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Hampel F R. The influence curve and its role in robust estimation [J]. Journal of the american statistical association, 1974, 69(346): 383-393. https://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482962

Eckstein J. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming [J]. Mathematics of Operations Research, 1993, 18(1): 202-226. https://pubsonline.informs.org/doi/abs/10.1287/moor.18.1.202

Sun C, Qiu X, Xu Y, et al. How to fine-tune bert for text classification? [C]. Chinese Computational Linguistics: 18th China National Conference, Kunming, China, 2019: 194-206. https://link.springer.com/chapter/10.1007/978-3-030-32381-3_16

Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results [J]. Advances in neural information processing systems, 2017, 30. https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html

Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv, 2018, preprint arXiv:181004805. https://doi.org/10.48550/arXiv.1810.04805

Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514. https://ieeexplore.ieee.org/abstract/document/9599397/

Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced language representation with informative entities [J]. arXiv, 2019, preprint arXiv:190507129. https://doi.org/10.48550/arXiv.1905.07129

Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach [J]. arXiv, 2019, preprint arXiv:190711692. https://doi.org/10.48550/arXiv.1907.11692

He P, Liu X, Gao J, et al. Deberta: Decoding-enhanced bert with disentangled attention [J]. arXiv, 2020, preprint arXiv:200603654. https://doi.org/10.48550/arXiv.2006.03654