Research on Chinese Social Media Metaphor Text Detection Model based on Multi-Feature Fusion

Jingxiang Zhang
Computer Science and Technology, Beijing University of Technology, Beijing, China

Abstract: In this paper, a lightweight metaphor text detection model based on multi-feature fusion is proposed to address the challenge of identifying non-literal emotional expressions, such as metaphor and irony, in Chinese social media[1], which remain difficult to detect due to their non-literal semantics, dynamic linguistic patterns, and the limited generalization ability of existing methods.

The proposed model incorporates two parallel dynamic feature extractors. CPDI (Contextual PMI-based Dynamic Incongruity) is designed to capture dynamic changes in emotional polarity within specific contexts, while templateGAN[2] is employed to discover and learn emerging online linguistic patterns through adversarial generation. These features are encoded using a dual-channel BiLSTM network and further enhanced via an attention mechanism. Subsequently, the representations are concatenated and nonlinearly transformed through an adaptive fusion layer before being fed into a classifier for prediction.

Experimental results demonstrate that the proposed model achieves significant improvements over baseline models on Chinese metaphorical text detection tasks. The overall accuracy reaches 0.9033, with macro-averaged Precision, Recall, and F1 scores of 0.9036, 0.9032, and 0.9033, respectively, indicating its effectiveness in identifying metaphorical expressions.

Keywords: Metaphor Detection; Multi-Feature Fusion ; Attention Mechanism ; biLSTM ; templateGAN ; cPDI

1. Introduction

With the advent of the Web 2.0 era, the expression of language on social media platforms is becoming more and more diversified and complicated[2], especially for the extensive use of non-literal and implicit emotional expressions such as Metaphor, Irony and 'Yin and Yang Qi'. Traditional text sentiment analysis techniques mostly rely on explicit commendatory and derogatory words. For this kind of metaphorical expression that is inconsistent with the literal meaning, it often makes a wrong judgment, resulting in a misunderstanding of the user 's true intention.

Therefore, this paper proposes a Chinese social media metaphor text detection model based on multi-feature fusion, aiming to achieve efficient and accurate identification of dynamic network terms. The main work and innovations of this paper are as follows: A dual-channel model architecture combining dynamic emotional features and text template features is proposed. The architecture uses two parallel feature extractors to capture metaphorical signals from the two dimensions of dynamic semantics of words and macro-patterns of sentences, and effectively integrates them through subsequent fusion modules. CPDI (Contextual PMI-based Dynamic Incongruity) feature extractor is designed and implemented. TemplateGAN, a generative adversarial network for mining and generalizing emerging network speech templates, is designed and implemented. A lightweight dual-channel BiLSTM network is constructed, which combines attention mechanism and adaptive fusion layer. This design effectively controls the complexity of the model while ensuring the performance of the model, making it easier to deploy in practical applications. The attention mechanism is used to highlight key features, and the adaptive fusion layer enables the model to learn independently how to optimally combine heterogeneous information from two channels.

2. Related Work

There are three main research areas related to this study: computational research on metaphor and irony, deep learning–based text classification, and feature fusion techniques.

2.1 Computational Study of Metaphor and Irony

Metaphor and irony are typical forms of non-literal language, and their computational study has evolved from rule-based approaches to data-driven methods.

Rule-based and dictionary-based approaches represent the earliest attempts in this field. In traditional machine learning frameworks, metaphor detection is typically formulated as a binary classification problem. Researchers extract various textual features, such as Bag-of-Words (BoW), N-grams, part-of-speech (POS) tags, and sentiment polarity scores, and then train classifiers such as support vector machines (SVM), Naive Bayes, or decision trees. These methods can automatically learn patterns from data and exhibit improved adaptability. However, their performance is highly dependent on the quality of feature engineering, which increases both the complexity and workload of the research process.

In recent years, methods based on emotional incongruity have emerged as an important research direction. The CPDI feature extractor proposed in this paper is inspired by this idea, while placing greater emphasis on dynamically quantifying emotional conflict at the lexical level.

2.2 Text Classification based on Deep Learning

With the rapid development of deep learning, the performance of text classification tasks has been significantly improved. Deep learning models can automatically learn hierarchical feature representations from raw text, thereby avoiding the need for labor-intensive feature engineering.

Convolutional neural networks (CNNs) capture key N-gram patterns (i.e., local features) in text by applying convolutional filters of varying sizes. Recurrent neural networks (RNNs) and their variants, such as long short-term memory networks (LSTMs) and gated recurrent units (GRUs), are capable of modeling contextual dependencies by processing sequential information through recurrent structures.

The introduction of the attention mechanism further enhances model performance. It enables the model to dynamically focus on the most relevant parts of the input when processing long sequences. As a result, the model can automatically identify key words or phrases that are critical for determining the emotional tendency of the text and assign them higher weights, thereby improving both classification accuracy and interpretability.

2.3 Feature Fusion Technology

In many complex tasks, features derived from a single source are often insufficient for accurate prediction. Feature fusion techniques aim to effectively integrate heterogeneous features extracted from different sources or through different methods, thereby achieving stronger discriminative power than any individual feature.

Early fusion and late fusion are two classical fusion strategies. Intermediate-level fusion methods based on simple operations, such as concatenation, element-wise summation, or element-wise multiplication, are currently the most widely used approaches.

More advanced strategies involve dynamic weighted fusion based on attention mechanisms. These methods not only combine features but also dynamically adjust their relative importance according to the input. Through an additional attention network, the model learns to assign greater importance to more informative features under different conditions, enabling more flexible and adaptive fusion. In this paper, the attention mechanism is applied to each channel of the dual-channel BiLSTM. This design allows the information within each channel to be refined and weighted prior to fusion.

3. A metaphor Detection Model based on Multi-Feature Fusion

The whole model is mainly composed of the following parts:

(1) Input & Embedding Layer: Receive the text sequence and use the pre-trained word vector model (Word2Vec [1]) to convert it into a low-dimensional, dense distributed vector representation.

(2) Dual-Channel Feature Extraction Module: the core of the model, including two parallel processing flows:

Semantic Channel: A Bidirectional Long Short-Term Memory (BiLSTM [2]) network is used to encode word vector sequences to capture contextual semantic dependencies of text. The channel is equipped with a cross-modal polarity difference enhancer (CPDI) for dynamically capturing and quantifying emotional conflict features at the lexical level.

Template channel: First, a template generation

adversarial network (TemplateGAN) is used to learn from the input text and generate a dynamic grammar-emotion template with stronger generalization ability. Then, another BiLSTM network encodes the template sequence. The channel is equipped with a special TemplateGAN feature extractor.

(3) Attention Layer: After the BiLSTM layer of each channel, attention mechanisms[3] are introduced to identify and focus on the semantic and template features that are most critical to metaphor judgment.

(4) Adaptive Feature Fusion Layer: The feature vectors obtained by the two channels after attention weighting are concatenated, and the adaptive weights are designed for fusion.

(5) Output Layer: The fused feature vector is passed through a nonlinear transformation (fully connected layer), and finally sent to a Softmax classifier to output the probability that the text is metaphorical or non-metaphorical.

4. Experimental Results and Analysis

The experiment is carried out in a general PC environment. The operating system is Windows 11 home Chinese version (64-bit), the processor is 13th Gen Intel (R) Core (TM) i5-13500H (12 cores, 16 threads, base frequency of about 2.6GHz), the physical memory is 16 GB, and the storage is NVMe solid state drive (capacity of about 953.86 GB). The graphics card is Intel Iris Xe Graphics, and CUDA acceleration is not enabled. The software environment is Python 3.8.20 and PyTorch 2.8.0 + cpu, supporting the use of numpy, tqdm, scikit-learn, jieba, hanlp, gensim and other dependencies. Training and evaluation are completed on the CPU.

4.1 Dataset

The experiment in this paper uses the Chinese review data set 'Hotel Review' in the project. The data files are 'Hotel Review / train.tsv', 'Hotel Review / dev.tsv' and 'Hotel Review / test.tsv ', which are separated by tabs and contain two columns: 'label' (binary label) and 'text _ a' (review text). The data size is about : 3280 training sets, 397 validation sets, and 417 test sets. In the training phase, ' train.tsv ' and ' dev.tsv ' are first merged to construct the vocabulary and word vector, and then the training and verification sets are divided by 80 % / 20 % using the leave-out method. Preprocessing includes Chinese word segmentation (jieba), vocabulary construction (minimum word frequency threshold 5), and training Word2Vec vector based on training corpus (dimension 50).

4.2 Evaluation Indicators

Metaphor detection is essentially a binary classification problem. In order to comprehensively evaluate the performance of the model, we use four widely used classification task evaluation metrics:

(1) Accuracy: The proportion of the number of correctly classified samples to the total number of samples. Its calculation formula is:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

(2) Precision: represents the proportion of positive examples in the samples predicted as positive examples (metaphors). It measures the ' precision ' of model predictions.

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

(3) Recall rate (Recall): represents the proportion of all samples that are actually positive and are successfully predicted as positive by the model. It measures the ' recall rate ' of the model for positive examples.

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

F1 value (F1-Score): The harmonic mean of precision and recall is an important index to evaluate the performance of the model.

4.3 Experimental Settings

The model is implemented based on PyTorch. The embedding dimension is 256, the BiLSTM hidden dimension is 128, the CPDI feature dimension is 1, the template feature dimension is 64, the syntactic feature dimension is 34, and the dropout is set to 0.3. The optimizer uses Adam, the initial learning rate is 1×10^{-3}, and the weight of the weight is attenuated by 1×10^{-5}. The loss function is cross entropy. The batch size is 32; the training rounds were 3; taking the weighted F1 of the verification set as the main index, the early stop strategy with a patience value of 10 is used to save the best weight ' best _ model.pth ' when verifying the F1 improvement.

4.4 Result Analysis

On the independent test set (a total of 1200 samples, with category support of 592 and 608, respectively), the overall accuracy of the model was 0.9033. The macro average Precision / Recall / F1 is 0.9036 / 0.9032 / 0.9033 ; the weighted average Precision / Recall / F1 was 0.9035 / 0.9033 / 0.9033, respectively. Classification : Precision = 0.9103, Recall =

0.8919, F1 = 0.9010 (support = 592) ; precision = 0.8968, Recall = 0.9145, F1=0.9055 (support = 608) for category 1.

Comparing the accuracy, precision, recall rate and F1 value of different methods, the results are shown in the following figure.

Table 1. Comparison of Sentiment Analysis Performance of Different Methods

Model	Accuracy	Precision	Recall rate	F1 score
CNN	0.7906	0.7847	0.7615	0.7729
Bi-LSTM	0.8364	0.8462	0.8354	0.8408
Bi-LSTMCNN	0.8426	0.8317	0.8552	0.8433
SAM+Bi-LSTM	0.8754	0.8961	0.8721	0.8839
The model	0.9033	0.9036	0.9033	0.9033

The experimental results show that the overall performance of Bi-LSTM is better than that of CNN because it can effectively capture context information. The Bi-LSTM-CNN[4] model has higher metrics than other models, which proves that the hybrid model does combine the advantages of CNN and LSTM [ ]. Compared with the traditional CNN, Bi-LSTM and more innovative fusion models, the multi-feature fusion model proposed in this paper has more than 90 % of the evaluation indicators, showing a significant improvement.

4.5 Summary and Improvement

The dual-channel BiLSTM with attention effectively encodes both semantic and template-level information. The Transformer encoder further models higher-order interactions across channels, enabling the model to achieve macro and weighted F1 scores of approximately 0.903.

The distribution of the two classes is approximately balanced. The recall exhibits a complementary pattern between the two classes: the recall for class 0 is slightly lower (0.8919), while that for class 1 is higher (0.9145), whereas the precision shows the opposite trend. This suggests that the current decision boundary is relatively more tolerant toward class 1. To further improve the recall of class 0, it may be beneficial to increase the loss weight assigned to class 0 or perform threshold calibration during the inference stage.

Additionally, considering the quality of template generation and the robustness of CPDI estimation for low-frequency words, it is advisable to incorporate class imbalance handling techniques (e.g., weighted loss or focal loss) and adopt more robust CPDI smoothing strategies. Furthermore, metric-based learning rate adaptation during the validation stage may help improve both overall performance and class-wise metrics.

Due to resource constraints, the improvement strategies proposed in this paper follow the principles of being lightweight, robust, and reproducible, prioritizing methods with low computational cost, clear performance gains, and ease of practical deployment. Without such constraints, more advanced improvements could be explored.

First, a Chinese pre-trained language model could be fine-tuned, with CPDI and template vectors integrated into the same encoder as additional features to enable deeper cross-feature fusion. Second, multi-task joint learning across metaphor, sarcasm, and sentiment analysis tasks could be performed by sharing encoders and applying appropriate regularization to enhance robustness. Third, the training scale could be expanded (e.g., distributed training, mixed precision, and gradient accumulation), combined with techniques such as learning rate warm-up, cosine or one-cycle annealing, stochastic weight averaging, adaptive weight optimization, and automated hyperparameter search, with stable configurations selected via K-fold cross-validation.

Fourth, inference-stage enhancements could be introduced through model ensembling, along with temperature scaling, probability calibration, and classification threshold optimization. Fifth, data and model augmentation strategies could be applied, including continued pre-training, back-translation, synonym substitution, contrastive learning, and the incorporation of dependency graphs or graph-based encoding structures; multimodal extensions may also be considered if necessary. Finally, dataset improvements could be pursued by leveraging larger-scale corpora or constructing dedicated datasets for metaphorical text detection.

5. Conclusion

In the context of social networks, the identification of metaphor and sarcasm is of considerable practical importance. To address the limitations of existing methods in capturing emotional information and achieving high classification accuracy, this paper proposes a lightweight and reproducible model for Chinese metaphor detection. The model adopts a

dual-channel architecture to simultaneously characterize the semantic content and structural patterns of text, achieving stable performance without relying on extensive computational resources.

Specifically, the approach unifies Chinese word segmentation and vocabulary representation, and utilizes general-purpose pre-trained word embeddings. The semantic channel encodes contextual information using a bidirectional long short-term memory (BiLSTM) network, while the template channel employs a template-based generation mechanism to learn and generalize structure–emotion patterns in online discourse. The fusion layer integrates features through an attention mechanism and adaptive weighting, and the final predictions are produced by a probabilistic classifier. This design ensures controllable computational cost and better adaptability to emerging expressions.

Under consistent preprocessing and data partitioning settings, the proposed model outperforms several baseline methods-including support vector machines, convolutional neural networks, bidirectional long short-term memory networks, attention-based models, and Transformer encoders-in terms of test accuracy and weighted F1 score.

References

[1] Deng Dailing, Gao Wenxuan. Research hotspots and development trends of network buzzwords-visual analysis based on cnki database [J].Chinese character culture, 2024, (14): 29-32.DOI: 10.14014 / j.cnki.cn11-2597/g2.2024.14.052.

[2] Huang Guizhen, Liang Ting, Zheng Meng, et al. Review of Generative Adversarial Networks [J].Intelligent Internet of Things Technology, 2025,57(04):17-20.DOI: 10.26921/j.cnki.2096-6059.2025.04.003.

[3] Чирвоний, Олександр Сергійович (2024) The Evolution of Social Media Language: A Sociolinguistic Analysis of Recent Neologisms Закарпатські філологічні студії (35). pp. 127-132. ISSN ISSN 2663-4899

[4] HUANG X X,LIU G F,LIU X Y,et al. Sentiment classification depth model based on word2vec and bi-directional LSTM[J]. Application Research of Computers,2019,36(12):3583-3587,3596.

[5] LI Y S,WANG L M,CHAI Y M,et al Research on the construction method of dynamic emotion dictionary based on Bi-LSTM[J]. Journal of Chinese Computer Systems,2019,40(3):503-509.

[6] YIN W,SCHÜTZE H,XIANG B,et al. ABCNN ： attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computational Linguistics,2016,4：259-272.

[7] Joel Philip Thekkekara,Sira Yongchareon & Veronica Liesaputra.(2024).An attention-based CNN-BiLSTM model for depression detection on social media text.Expert Systems With Applications,249(PC),123834-.https://doi.org/10.1016/J.ESWA.2024.123834.

