International Journal of Wireless and Mobile Computing

ISSN online: 1741-1092 - ISSN print: 1741-1084
https://www.inderscience.com/ijwmc

Sentiment analysis method for e-commerce review on weak-label data and deep learning model

Zihao Zhou, Jie Chen, Junhui Wu

DOI: 10.1504/IJWMC.2023.10061392

Article History:

Received:	30 September 2022
Last revised:	20 November 2022
Accepted:	15 December 2022
Published online:	07 February 2024

Sentiment analysis method for e-commerce review on weak-label data and deep learning model

Zihao Zhou, Jie Chen and Junhui Wu*

College of Electronic and Information Engineering,
Tongji University,
Shanghai 201804, China
Email: zzhlov@163.com
Email: chenjie1968@tongji.edu.cn
Email: junhui_wu@163.com
*Corresponding author

Abstract: For inaccurate weak-label data of e-commerce reviews, the traditional manual labelling method is time-consuming, and it is necessary to solve the problem of polysemy and imbalance in Chinese reviews to improve the performance of sentiment analysis model. This paper collects agricultural product reviews on Jingdong platform. Firstly, the improved SO-PMI method is used to construct a domain sentiment dictionary, by combining the review sentiment tendency calculated by the dictionary with the weak-label data of user ratings, an unsupervised generation of high-quality training sets is realised. Secondly, two basic learners, Bidirectional Long Short Term Memory (BiLSTM) and Convolutional Neural Network (CNN), are combined in the sentiment analysis model, and the character, word, part-of-speech vector features are extracted in parallel. In addition, an attention mechanism is embedded in the channel, and using Focal Loss during model training process. The experimental results show that the accuracy of the method proposed in this paper reaches 97.34%, which is 4.64% higher than that of directly using weak-label data for training. Compared with single-channel CNN and BiLSTM model, the accuracy is improved by 1.55% and 0.99% respectively. Therefore, this method improves the accuracy of sentiment analysis of e-commerce reviews.

Keywords: sentiment analysis; deep learning; weak-label data; sentiment dictionary; multi-channel network.

Reference to this paper should be made as follows: Zhou, Z., Chen, J. and Wu, J. (2024) 'Sentiment analysis method for e-commerce review on weak-label data and deep learning model', *Int. J. Wireless and Mobile Computing*, Vol. 26, No. 1, pp.9–18.

Biographical notes: Zihao Zhou is a postgraduate student at the Tongji University, China. His research interests include natural language processing and deep learning.

Jie Chen is now an Associate Professor of the College of Electronic and Information Engineering, Tongji University, China. His research direction is machine learning and data mining.

Junhui Wu is now an Associate Professor of the College of Electronic and Information Engineering, Tongji University, China. His research direction is natural language processing and data mining.

This paper is a revised and expanded version of a paper entitled 'Sentiment Analysis Method for Agricultural Product Review Based on Corpus Characteristics and Deep Learning Model' presented at '2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)', Guangzhou, China, 5–7 August 2022.

1 Introduction

With the promotion of e-commerce, more users purchase products through e-commerce platforms, such as Jingdong Fresh, Tmall Supermarket, Hema Fresh, etc., generating a large amount of reviews. Review sentiment analysis is to use natural language processing (NLP) technology to identify the semantic feature expression in the text, and output the sentiment tendency of the review according to two categories (positive, negative sentiment), three categories (positive, neutral and negative sentiment) or multi-category (happiness, sadness, anger, fear, etc.) (Lin and Wu, 2022). The result can not only improve the browsing experience of customers, but also provide information value for merchants to improve products or services (Mullen and Collier, 2004).

On the one hand, most of e-commerce reviews are accompanied by rating labels that express users' subjective sentiment attitudes. Although it can save the trouble of

manually labelling the training set, it lacks uniform constraints and has greater randomness, Qu et al. (2012) defined that as weak-label data. The specific performance is that some high-rating negative reviews and low-rating positive reviews are mixed, and these noisy samples will seriously interfere with the performance of the sentiment analysis model (Zhu et al., 2022). On the other hand, e-commerce reviews have different text lengths, unbalanced sentiment categories, and more buzzword expressions (Kishore et al., 2022). For short reviews, the surface semantics of individual words are sometimes difficult to fully express text information, but adding the part-of-speech (POS) features of words to judge the sentiment tendency is very helpful (Yan et al., 2021a). For long reviews, there are more polysemy and out of vocabulary (OOV) problems. Although deep learning models have achieved impressive results in competition datasets compared with traditional sentiment dictionaries or machine learning methods, it is still a challenging task to perform sentiment analysis on e-commerce reviews with vague and ambiguous semantics.

This paper proposes an e-commerce review sentiment analysis method with weak-label data characteristics. Firstly, a framework for unsupervised correction of label of training sets is proposed, and then a deep learning model is built to output the review sentiment label prediction result. The contributions of this paper are summarised as follows: The improved SO-PMI method is used to build a domain sentiment dictionary, and the dictionary-based sentiment tendency labelling results and user rating data are combined to unsupervised generate high-quality domain training sets. In the deep learning model, the character, word and part of speech (POS) vector features in the review are simultaneously extracted as the input of different channels, and the bidirectional long short term memory (BiLSTM), convolutional neural network (CNN), attention mechanism and Focal Loss are embedded in the model, so that the model can extract richer sentiment information, and focus on the training of minority class samples, to improve the performance of sentiment analysis.

2 Related work

2.1 Weak-label data process

Weak-label data broadly refers to ratings or emojis with emotional meanings, which are used as sentiment labels for review datasets, but these labels are often not accurate enough and needs further processing. Previously, when scholars conducted sentiment analysis on comments with weak-label data, they mostly relied on supervised training for correction of manually labelled features, which still required a lot of labour costs. For example, Täckström and McDonald (2011) mixed the comment data with ratings and the manually labelled data to construct a conditional random field (CRF) model to achieve the task of sentiment analysis. Lu et al. (2022) designed a fine-tuned supervised contrastive learning framework to learn sentiment semantics from weak-label data, and validated it on the Amazon review dataset. In contrast, some scholars used an unsupervised mixed sentiment dictionary and weak label data method to automatically label reviews (Fan et al., 2018), but when using the SO-PMI method, the frequent items in the review were ignored, resulting in the selection of seed sentiment words is not representative, and the SO-PMI method still cannot accurately classify weak sentiment polar words and neutral words.

2.2 Sentiment analysis model

The sentiment analysis model can be roughly divided into three types: based on sentiment dictionaries, machine learning and deep learning. Since deep learning can automatically extract richer text features, it performs well in NLP fields such as machine translation and entity recognition, and is also more widely applicated in sentiment analysis tasks. According to the structure of deep learning model, it can be divided into single neural network and combined neural network. In a single neural network study, Zhao et al. (2017) used a word embedding model to extract semantic features, and used a deep CNN to achieve sentiment classification of Twitter reviews. Because CNN has the characteristics of strong ability to capture local features and is not good at long-distance modelling, in contrast, using temporal neural networks such as gated recurrent unit (GRU) or long short-term memory neural network (LSTM) can better capture temporal factors of contextual semantic features, and demonstrated that it achieved better performance than traditional recurrent neural networks (RNNs) (Fan and Li, 2021). Another paper combined the advantages of multiple neural network neural network structures to extract deeper text sentiment semantic features through serial or parallelised multi-channel structures. For example, Hassan and Mahmood (2018) combined CNN and LSTM to build a serial model, and the experiment proved that the accuracy of the combined model was higher under the two-category and five-category sentiment analysis tasks. Yan et al. (2021b) combined CNN and BiLSTM to form a dual-channel structure, extracted word embedding features and POS features in the text in parallel, and proved that the performance of it was greater than that of the single-channel structure.

In the training process of the deep learning model, attention mechanism was first applied in the field of image recognition, and then it became mainstream to embed the mechanism into models in the field of sentiment analysis. For example, Cheng et al. (2020b) proposed to embed the attention mechanism in a multi-channel neural network model fused with CNN and BiGRU, aiming to pay closer attention to the word with important sentiment semantic features in the text. On the other hand, there is often a problem of sample imbalance in e-commerce reviews, which is manifested in the fact that users are more inclined to express positive sentiment opinions, resulting in the number of positive reviews in the dataset is far greater than the number of negative reviews. Direct training on the samples will result in the model being underpowered to identify the minority class of negative samples (Obiedat et al., 2022). The traditional over-sampling and under-sampling techniques to solve the above problems

are easy to cause the problem of model overfitting or loss of a large amount of sample. In contrast, Focal Loss improves the loss function from the perspective of the difficulty of category identification, so that the sentiment analysis model pays more attention to the learning of hard-to-identify samples (Cheng et al., 2020a).

3 Unsupervised annotation method for review

The process of unsupervised annotation of review sentiment labels can be divided into three stages: data collection, sentiment dictionary expansion based on improved SO-PMI, and sentiment label generation. The flow chart is shown in Figure 1.

3.1 Data collection and weak-label data

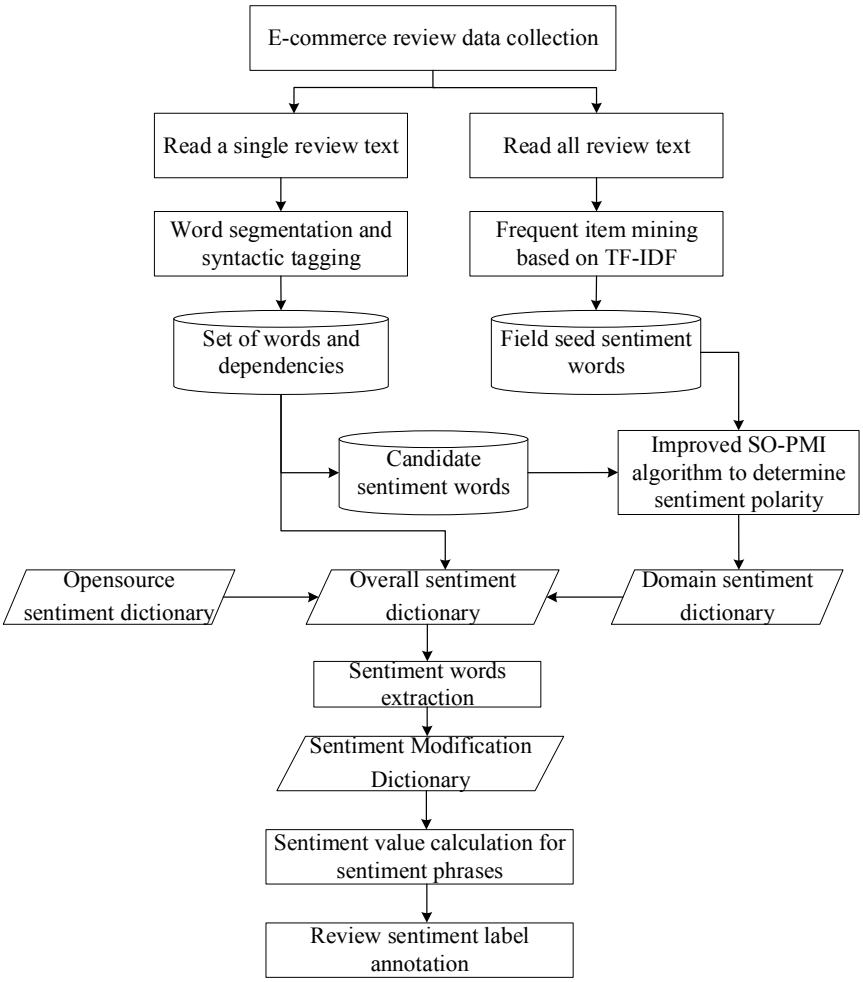
Firstly, the crawler technology of the selenium framework was used to obtain user reviews of 7 agricultural product categories including peanuts, corn, tea, pears, kiwi, sesame, and soybeans on the Jingdong Platform. After deduplication of the collected data, regular expressions are used to remove irrelevant information such as numbers and emojis in the review text, and the final review dataset obtained consists of text and user rating labels. A total of 30,000 reviews with a rating of 4 to 5 and 5,000 reviews with a rating of 1 to 2 are used as a dataset for sentiment analysis model training.

If the user rating label is directly used as the sentiment label, referring to the division method of the mainstream e-commerce platform interface, the rating label with a score of 4 to 5 is converted into a positive sentiment label, and the rating label with a score of 1 to 2 is converted into a negative sentiment label. However, since the sentiment score is subjectively defined by the user and the e-commerce platform lacks unified rules and constraints, there are some user ratings that do not match the real sentiment expression. Examples are shown in Table 1.

Table 1 Examples of incorrect rating data

User review text	*User rating label*	*True sentiment label*
The raw materials of peanuts are not good, they are hard and bitter, and I will not buy them again.	5	negative
The soup colour and taste of the tea are quite satisfactory, and the price and performance ratio is still very good.	2	positive

Figure 1 Unsupervised annotation method flow for review



3.2 Expansion of the sentiment dictionary

At present, the most used sentiment dictionaries in the field of Chinese text analysis include the HowNet dictionary, the NTUSD dictionary from National Taiwan University, and the Tsinghua dictionary from Tsinghua University. The above three-part dictionary sets are combined and deduplicated to form a basic sentiment dictionary, which contains 5098 positive sentiment words and 8197 negative sentiment words, in which the polarity of positive sentiment words is marked as 1, and the polarity of negative sentiment words is marked as -1. But in the e-commerce review, there are often some common words with sentiment meaning in the field, such as the review in the field of agricultural products "These peanuts are moldy!", "moldy" is a word with negative sentiment but does not exist in the basic sentiment dictionary. Therefore, it is necessary to build a domain sentiment dictionary for expansion. Generally, the effect of dictionary expansion depends on the size of the dataset and the selection quality and quantity of seed sentiment words. The richer the dataset and the more comprehensive the seed sentiment words, the better the effect of domain sentiment dictionary expansion.

The sentiment tendency of seed sentiment words must be very clear, such as positive words "good", "scented" and negative words "poor", "rotten" in the field of agricultural products. Since the high-frequency sentiment words commonly expressed by users in the review often appear in the form of adjectives or nouns, the sentiment polarity of these words is relatively obvious, it is more appropriate to select them as seed sentiment words. According to the vocabulary set of all review texts, the frequent item mining algorithm TF-IDF is used to identify high-frequency adjectives or nouns in the dataset and count the word frequency. The calculation formula of the TF-IDF algorithm is described as

$$\begin{aligned} TF-IDF_{W,D_i} &= TF_{W,D_i} \times IDF_W \\ &= \frac{\text{count}(W)}{|D_i|} \times \log \frac{N}{\sum_{i=1}^{N} I(W, D_i)} \end{aligned} \tag{1}$$

where TF_{W,D_i} represents the frequency of keyword W in document D_i, IDF_W represents its inverse document frequency. $\text{count}(W)$ represents the frequency of the occurrence of the keyword W, $|D_i|$ is the frequency of all words in the document D_i, N is the total number of all documents, $I(W, D_i)$ represents the 0/1 variable of whether the document D_i contains the keyword. According to the Top-N sorting rule, the words that appear in the top 30 most frequently in the review text with a rating of 4 to 5, and appear in the basic sentiment dictionary are regarded as positive seed sentiment words. Similarly, from the review text with a rating of 1 to 2, negative seed sentiment words are screened out.

After the seed sentiment words are determined, the sentiment polarity of each candidate domain sentiment word is marked by SO-PMI algorithm. First, the calculation formula of the basic PMI algorithm is described as

$$PMI(word1, word2) = \log\left(\frac{P(word1 \,\&\, word2)}{P(word1) \times P(word2)}\right) \tag{2}$$

where $word1$ represents the candidate domain sentiment word that needs to be judged sentiment polarity, $word2$ represents the seed sentiment word, $P(word1)$ and $P(word2)$ represent the probability of occurrence of the candidate word and the seed sentiment word respectively, and $P(word1 \,\&\, word2)$ represents the difference between co-occurrence probability of the two words. Since the use of a single seed sentiment word is easy to cause discrimination bias in PMI calculation, the multi-word joint SO-PMI calculation method is used to discriminate, the calculation method is described as

$$\begin{aligned} SO-PMI(word1) = &\sum_{p \in pword2} PMI(word1, p) \\ &- \sum_{n \in nword2} PMI(word1, n) \end{aligned} \tag{3}$$

where p and n represent the positive and negative seed sentiment words set respectively. After calculating the PMI value corresponding to the candidate word and all the seed sentiment words, the SO-PMI values of candidate words are obtained by summation and subtraction. It is generally believed that the larger the absolute value of SO-PMI, the more significant the sentiment tendency of the word. If the value is greater than 0, the polarity of the labelled word is 1; otherwise, the polarity of the labelled word is –1. However, there is a large gap between the maximum PMI value and the minimum PMI value in the same sentiment words in practice, and there are a large number of neutral sentiment words with values close to 0, which are also classified as positive and negative sentiment words by general methods, it will affect the accuracy of the labelling results. Referring to previous scholars' research, Min-Max normalisation is performed on the PMI value of each word in this paper (Bao et al., 2021), and it is mapped to the $[0,1]$ interval, the calculation method is described as

$$senti_i = \frac{SO-PMI_i - \min_{1 \le j \le n}\{SO-PMI_j\}}{\max_{1 \le j \le n}\{SO-PMI_j\} - \min_{1 \le j \le n}\{SO-PMI_j\}} \tag{4}$$

where $SO-PMI_i$ represents the initial value of the word i, $\max_{1 \le j \le n}\{SO-PMI_j\}$ and $\min_{1 \le j \le n}\{SO-PMI_j\}$ represents the smallest and largest value of all words respectively, and $senti_i$ represents the lexical sentiment tendency value obtained by normalisation. According to the $senti_i$, the word with the most obvious sentiment tendency is selected as the extended domain sentiment word. A threshold value T needs to be determined, so that the sentiment words whose $senti_i$ is in the $[-T, T]$ interval are neutral words to be removed. After repeated experiments, this paper sets $T = 0.2$. In order to reduce the computational cost, on the one hand, the co-occurrence window of the SO-PMI algorithm is set to 5; on the other hand, after reading the review text and segmenting the words to form a word list, three conditions are set to determine whether to mark the word as a domain sentiment word:

1 The word does not exist in the existing basic sentiment dictionary and seed sentiment dictionary.

2 The POS of the word is adjective, noun and verb.

3 The absolute value of the sentiment tendency value $senti_i$ is greater than 0.2.

On this basis, after manually removing a small number of noise words, a sentiment dictionary in the field of agricultural products is constructed, which is combined with the basic sentiment dictionary and the domain sentiment dictionary to form an overall sentiment dictionary. Examples are shown in Table 2.

Table 2 Examples of overall sentiment dictionary

Dictionary type	*Sentiment polarity*	*Sentiment word example*
Basic Sentiment Dictionary	1	5098 words such as praise, joy, surprise, like
	–1	8197 words such as anger, doubt, sadness, negativity
Seed Sentiment Dictionary	1	30 words such as nice, big, sweet, even, cheap
	–1	30 words such as moldly, worst, slow, shabby
Domain Sentiment Dictionary	1	194 words such as delicious, clean, sweet, delicate, translucent
	–1	191 words such as unpalatable, bitter, spoiled, moldly, worthless, bitter

There are often sentiment modifier words related to sentiment words in the e-commerce review text, including degree adverbs and negative words, which also significantly affect the sentiment expression of the overall review (Liang et al., 2019). Based on sentiment modification dictionary of HowNet, this paper selects 220 degree adverbs and 58 negative words to form a sentiment modification dictionary. According to the effect of each sentiment modifier word on the strengthening, weakening or reversal of sentiment expression, the dictionary is divided into seven levels and assigned sentiment weights ranging from –1 to 2. The example of the constructed sentiment modification dictionary is shown in Table 3.

Table 3 Examples of sentiment modifier dictionary

Sentiment expression	*Sentiment modifier word example*	*Weight factor*
strengthen	69 words such as very, extremely	2
strengthen	42 words such as too, especially	1.5
strengthen	37 words such as still, more, more and more	1.25
weaken	29 words such as slightly, quite, somewhat	0.6
weaken	12 words such as mild, relative, weak	0.4
weaken	30 words such as over, too	0.2
reverse	58 words such as no, not	–1

3.3 Generation of sentiment labels

After word segmentation and LTP library-based dependency syntax analysis for each e-commerce review, firstly extract all sentiment words in the review text that appear in the overall sentiment dictionary to form a corresponding sentiment word set. Then find one or more sentiment modifier words that satisfy the dependent syntactic relationship with each sentiment word as an adverbial structure, and match them to form the corresponding sentiment phrase. For five different types of the sentiment phrase, this paper proposes a method to calculate the sentiment score of the review, the calculation method is described as

$$q_i = \begin{cases} V_{ne+dg+se} = sen(se) \times (-1)^{|ne|} \times \left(1 + 0.2 \times \left(-\dfrac{1}{\sum_{j=1}^{n} W_{dg}^{j}}\right)\right) \\ V_{dg+ne+se} = sen(se) \times (-1)^{|ne|} \times \left(1 + 0.2 \times \left(\sum_{j=1}^{n} W_{dg}^{j}\right)\right) \\ V_{dg+se} = sen(se) \times \left(1 + 0.2 \times \left(\sum_{j=1}^{n} W_{dg}^{j}\right)\right) \\ V_{ne+se} = sen(se) \times (-1)^{|ne|} \\ V_{se} = sen(se) \end{cases} \tag{5}$$

where $V_{ne+dg+se}$ is the sentiment phrase type of negative adverb + degree adverb + sentiment word, $V_{dg+ne+se}$ is the type of degree adverb + negative adverb + sentiment word, V_{dg+se} is the type of degree adverb + sentiment word, V_{ne+se} is the type of negative adverb + sentiment word, V_{se} is the type of only sentiment word. q_i represents the calculated sentiment value expressed by the sentiment phrase i, and $sen(se)$ represents the polarity of the sentiment word. The sentiment value calculation process also considers the semantic changes of multiple sentiment words. W_{dg}^{j} represents the weight of the degree adverb j in the sentiment phrase, n represents the number of degree adverbs, and $|ne|$ represents the number of negative words. By summarising the sentiment value calculation results of all sentiment phrases in the review text, the sentiment label of the overall review can be marked, and the formula is described as

$$flag = \begin{cases} 1 & \sum_{i=1}^{N} q_i > 0 \\ 0 & \sum_{i=1}^{N} q_i = 0 \\ -1 & \sum_{i=1}^{N} q_i < 0 \end{cases} \tag{6}$$

where $flag$ is the sentiment label of the marked review. When the value of $flag$ is 1, it means positive sentiment, when it is –1, it means negative sentiment, and when it is 0, it means neutral sentiment. Only keep reviews whose $flag$ matches the rating label to generate a high-quality training set. When $flag$ does not match the rating label, the review is removed as noise data. Examples of unsupervised review annotation results are shown in Table 4.

Table 4 Examples of unsupervised review annotation results

Review text	$\sum_{i=1}^{N} q_i$	*flag*	*User rating label*	*Operate*
The raw materials of peanuts are not good, they are hard and bitter, and I will not buy them again.	–2	–1	5	Remove
The soup colour and taste of the tea leaves are quite satisfactory, and the price and performance ratio is still very good.	2.54	1	2	Remove
Corn is nutritious and tastes great	2	1	5	Reserve

4 Design of sentiment analysis model based on deep learning

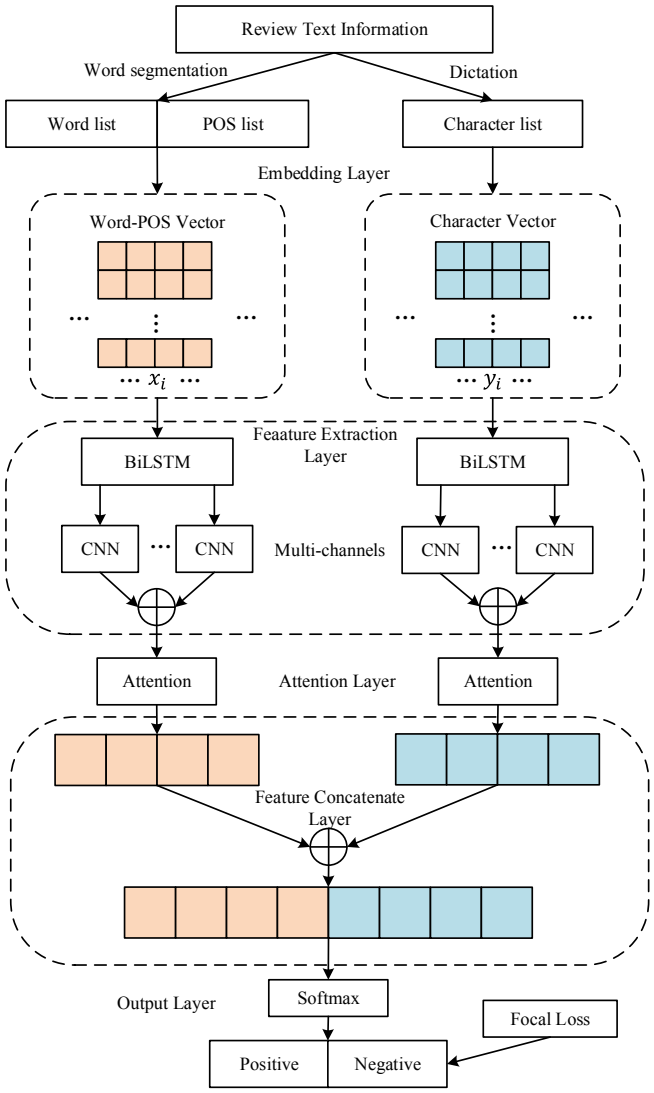
The e-commerce review sentiment analysis model designed in this paper is mainly composed of embedding layer, feature extraction layer, attention layer, feature aggregation layer and output layer, and has the characteristics of multi-channel structure. The embedding layer outputs word-POS vector and character vector respectively into different channels, and the feature extraction layer includes two network structures, BiLSTM and CNN. After each channel is embedded in the attention layer for feature weight redistribution, the features extracted by each channel are fused in the feature concatenate layer, and the sentiment prediction result is output. The model training process introduces the loss function of Focal Loss. The overall structure of M-AttBiLSTMCNN-FL model in this paper is shown in Figure 2.

4.1 Feature extraction layer

1) *BiLSTM structure:* The word-POS vector and character vector are respectively input into the BiLSTM structure of different channels for multi-dimensional text feature extraction. The structure of BiLSTM is composed of a forward LSTM and a backward LSTM network. The forward LSTM extracts the information h_{t-1} output by the previous node, and the backward LSTM extracts the information h_{t+1} output by the next node. The output of each cell state is jointly influenced by the forward cell state and the backward cell state, which can be used to capture bidirectional contextual semantic dependencies. A single LSTM is a special-structured RNN that alleviates the RNN's vanishing gradient problem by adding a "gating device". In LSTM, the area where information is stored at each moment as the Cell State, and the three gate structures that interact with it in the internal unit: Input Gate, Forget Gate and Output Gate. The calculation process of the LSTM internal unit at time t is described as

$$i_t = \delta(W_i[h_{t-1}, x_t] + b_t) \tag{7}$$

Figure 2 The structure of M-AttBiLSTMCNN-FL model



$$c_t^{'} = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{8}$$

$$f_t = \delta(W_f[h_{t-1}, x_t] + b_f) \tag{9}$$

$$c_t = f_t \bullet c_{t-1} + i_t \bullet c_t^{'} \tag{10}$$

$$o_t = \delta(W_o[h_{t-1}, x_t] + b_o) \tag{11}$$

$$h_t = o_t \bullet \tanh(c_t) \tag{12}$$

where c_t, i_t, f_t, o_t represent the Cell State, Input Gate, Forget Gate and Output Gate at time t respectively. W_c, W_i, W_f, W_o represent the respective weights, b_c, b_i, b_f, b_o represent the respective biases. δ represents the sigmoid activation function, c_{t-1} is the cell state at the previous moment, and $c_t^{'} \in R^D$ is the candidate state obtained by the tanh function, The timing information is transmitted through the cell state $c_t \in R^D$ at time t, and information $h_t \in R^D$ is output to the external state of the hidden layer.

2) *CNN structure:* The features extracted by BiLSTM are used as the input of multiple CNN structures with

different convolution kernel lengths to better capture the local semantic features of the review. For a convolution kernel with a set length of h, the feature is divided into $\{x_{0:h-1}, x_{1:h}, ..., x_{i:i+h-1}, ..., x_{n-h+1,n}\}$, and the feature map vector extracted after performing the convolution operation on each component $x_{i:i+h-1}$, the calculation formula is described as

$$c_i = relu(W \bullet x_{i:i+h-1} + b) \tag{13}$$

where $W \in R^{h\times(n+d)}$ represents the convolution kernel weight, $b \in R$ represents the bias. The feature vector is filtered by the pooling layer of the max-pooling strategy, and the most important feature $c_i^{"} = \max\{c_i\}$ is retained. Finally, the feature information obtained by sampling e convolution kernel is obtained, and aggregate the feature information extracted by m different CNN channels, then output all the extracted feature information C, the calculation formula is described as:

$$\begin{aligned} &C = [C_1, C_2, ..., C_m] = \\ &[(c_{11}^{"}, c_{12}^{"}, ..., c_{1e}^{"}), (c_{21}^{"}, c_{22}^{"}, ..., c_{2e}^{"})..., (c_{m1}^{"}, c_{m2}^{"}, ..., c_{me}^{"})] \end{aligned} \tag{14}$$

4.2 Attention layer

The essence of the attention mechanism is to focus on the key sentiment word in the review to help improve the model performance. The calculation steps are divided into three stages: similarity calculation, normalisation and weighted summation. First, calculate the similarity between the target sequence q with input length N and each key K_i, and obtain the value corresponding to K_i as the weight coefficient $F(Q, K_i)$, and then import the normalisation layer to calculate the attention distribution, the formula is described as:

$$\begin{aligned} a_i &= p(z = i \mid Q, K) = \text{softmax}(F(Q, K_i)) \\ &= \frac{\exp(F(Q, K_i))}{\sum_{z=1}^{N} \exp(F(Q, K_i))} \end{aligned} \tag{15}$$

where $z \in [1, N]$ represents the index position, $\text{softmax}(F(Q, K_i))$ represents the normalised processing result of the weight coefficient. After calculating the attention weight a_i of each keyword, weighted summation is performed with the value corresponding to each K to obtain the final output result. The formula is described as:

$$att(Q, K, V) = \sum_{i=1}^{N} a_i * Value_i \tag{16}$$

where $att(Q, K, V)$ represents the features after the key information is enhanced by the attention layer.

4.3 Feature concatenate and output layer

In the feature concatenate layer, the word-POS feature z^w and character feature z^c output by the two channels are spliced, and the final text feature information representation t formula is as follows:

$$t = [z^w, z^c] \tag{17}$$

Then input t into the fully connected layer, use the Softmax activation function to calculate the sentiment probability p of the review and determine the final sentiment polarity, the calculation method is as follows:

$$p = softmax(Wt + b) \tag{18}$$

4.4 Focal loss

In the model training of the binary classification task, $y = 0$ and $y = 1$ are used to represent the real category of the sample, and $y^{'}$ represents the predicted category of the sample. The traditional calculation method of the cross entropy loss function L_{CE} is described as:

$$\begin{aligned} L_{CE} &= -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \\ &= \begin{cases} -\log \hat{y}, & y = 1 \\ -\log(1 - \hat{y}), & y = 0 \end{cases} \end{aligned} \tag{19}$$

In the sentiment analysis task, the positive samples of the majority class in the corpus can be defined as easy-to-identify samples, and the negative samples of the minority class can be defined as difficult-to-identify samples. Focal Loss improves model performance by increasing the loss weight of negative sentiment samples in training. First, the problem of uneven proportion of positive and negative samples is balanced by adding a weight parameter α with a value range of $[0,1]$ in the loss function, then use the value range of $[0, +\infty]$ as the modulation factor β to reduce the loss weight of easy-to-distinguish samples, so that the model can focus more on the harder-to-distinguish samples. Finally, the calculation formula of Focal loss is obtained as

$$FocalLoss = \begin{cases} -\alpha(1 - p)^{\gamma} \log(p), & y = 1 \\ -(1 - \alpha) p^{\gamma} \log(1 - p), & y = 0 \end{cases} \tag{20}$$

where p represents the sample class probability distribution. When the hard samples with small p are misclassified, γ is close to 1 without causing too much loss, but when the easy-to-classify samples are misclassified, p is close to 1 and γ tends to 0, the loss weight is significantly reduced. In practical applications, it is necessary to adjust the parameters of α and γ at the same time to obtain the optimal model performance.

5 Experiment

Firstly, the experimental comparison of different models under the same annotated dataset is designed. Secondly, the experiments on four different annotated datasets of the proposed sentiment analysis model are designed, to verify the effectiveness of the proposed method in this paper.

5.1 Dataset preparation

In order to verify the effectiveness of the method in this paper, on the one hand, the dataset named "Weak_vert" without unsupervised annotation processing is reserved, that simply label reviews with a rating of 4 to 5 as positive, and reviews with a rating of 1 to 2 as negative. According to the ratio of 4:1, 28,000 reviews are divided as the training set and 7,000 reviews are divided as the test set. On the other hand, using the dataset named "Pmi_weak_vert" processed by unsupervised annotation, although some noise data is removed and the sample is reduced, 22,118 reviews are still divided as the training set and 5,547 reviews are divided as the test set. In addition, in order to verify that it is necessary to use SO-PMI method and weak label data in automatic annotation processing, in the third dataset named "Dic_weak_vert", the dictionary extension part of SO-PMI is removed, and only the basic dictionary is used for annotation. In the fourth dataset named "Pmi_vert", the weak-label data is removed, and only the dictionary based on SO-PMI expansion is used for annotation. A total of 4 different types of datasets were generated. In order to ensure the accuracy and consistency of the test set, manual correction annotations by a 3-person team were used in the test set of each dataset. Table 5 shows the descriptive statistics of the four datasets used in this paper.

Table 5 Descriptive statistics of the four datasets

Dataset name	*Number of datasets*	*Number of training set*	*Number of test set*
Weak_vert	35,000	28,000	7000
Pmi_vert	35,000	28,000	7000
Dic_weak_vert	28,033	22,426	5607
Pmi_weak_vert	27,665	22,118	5547

5.2 Model hyperparameter settings and performance evaluation metrics

After repeated comparison experiments, set the word vector dimension size of Word2vec model to 300. The fixed length of the word-POS vector is set to 50, and the fixed length of the character vector is set to 100. When the text sequence is less than this length, zero-padding is performed, and when it is greater than this length, truncation is performed. The deep learning model is based on the Keras framework, selecting the optimisation function Adam and setting the learning rate to 0.001. In each channel, three CNN structures with convolution kernel window sizes of 3, 4, and 5 are used, and the number of each convolution kernel is 128. The number of internal units in BiLSTM is also set to 128. To prevent overfitting of the model, weight-based L2 regularisation and dropout mechanisms are used. In the loss function of Focal Loss, after repeated trial and error, the weight parameter α is set to 0.25, and the modulation factor β is set to 1.8. All the specific hyperparameter settings of the model in this paper are shown in Table 6.

Table 6 Hyperparameter settings for the model

Hyperparameter name	*Value*
Size of Word Vector	300
Window Size	3,4,5
Number of convolution kernel channels	128
Hidden Size of BiLSTM	128
Mini-batch-size	32
Epochs	20
Learning Rate of Adam	0.001
Size of L2 Regularisation	0.001
Dropout Rate	0.5
Weight parameter α	0.25
modulation factor β	1.8

In the confusion matrix of the sentiment binary classification task, TP and TN represent the number of correct predictions for positive and negative samples, respectively, and FP and FN represent the number of false predictions for positive and negative samples respectively. On this basis, four commonly used machine learning verification indicators, accuracy, precision, recall and f1 value in the macro-average state are selected to measure the performance of the model (Schütze et al., 2008), and their formulas are as follows

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \times 100\% \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{23}$$

$$F1 = \frac{1}{n}\sum_{i=1}^{n} \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{24}$$

5.3 Benchmark model

First compare model M-BiLSTMCNNAtt-FL in this paper with other classic sentiment analysis models, then design the ablation experiment of the internal structure of the model.

1 *NaiveBayer:* A classic machine learning method that uses probability and statistical knowledge to classify algorithms. The model structure is simple and the operation speed is fast (Yang and Xia, 2016).

2 *SVM:* After the weighted average of the word vector of the text trained by Word2vec, it is used as the input of the SVM for classification.

3 *CNN:* An ordinary single-layer CNN classifier, which takes the word-POS vector as the CNN input, sets the convolution kernel size to 3 and the filter to 64.

4 *RCNN:* A model that combines RNN and CNN to capture broader contextual information, which takes the word-POS vector as the RCNN input.

5 *BiLSTM:* Use word-POS vector representations as input to this model.

6 *BiGRU:* On the basis of model 5), only the reset gate and update gate are kept, which has more concise network parameters (Fan and Li, 2021).

7 *W-BiLSTMCNN:* Ablation experiment, a single-channel network structure that uses only word-POS vector as input.

8 *C-BiLSTMCNN:* Ablation experiment, replacing the word-POS vector of Model 7) with character vector.

9 *M-BiLSTMCNN:* Ablation experiment, a two-channel structure that simultaneously extracts word-POS vector and character vector features.

10 *M-CNNBiLSTM:* Swap the positions of the CNN and BiLSTM in the model 9) structure.

11 *M-AttBiLSTMCNN:* Ablation experiment, adding attention layer based on model 9), but only use the traditional cross-entropy loss function.

5.4 Experimental results

First, the model performance is compared under the dataset named "Pmi_weak_vert", the results are shown in Table 7.

Table 7 Performance comparison of different sentiment analysis models

Model	*Accuracy*	*Precision*	*Recall*	*f1*
NaiveBayer	82.39	82.32	76.49	78.84
SVM	85.67	87.20	74.70	80.50
CNN	95.79	94.00	91.98	92.95
RCNN	95.81	93.49	92.60	93.04
BiLSTM	96.35	92.76	95.91	94.23
BiGRU	96.25	92.72	95.54	94.05
W-BiLSTMCNN	96.43	93.03	95.87	94.36
C-BiLSTMCNN	96.09	94.04	92.85	93.43
M-BiLSTMCNN	96.95	94.41	95.76	95.07
M-CNNBiLSTM	96.81	94.83	94.64	94.74
M-AttBiLSTMCNN	97.10	**95.00**	95.55	95.27
M-AttBiLSTMCNN-FL	**97.34**	94.83	**96.73**	**95.74**

It can be seen from the result that the performance of the CNN learner in sentiment analysis task is significantly better than that of the two classic machine learning models, NaiveBayers and SVM. Among the temporal structure neural network learners represented by BiLSTM and BiGRU, the former has better performance than the latter. Therefore, this paper considers the fusion of CNN and BiLSTM to build a model. The model M-BiLSTMCNN with backward embedded CNN structure performs slightly better than the model with forward embedded CNN structure, indicating that extracting the global features of the text first, and then focusing on the splicing order of local features can enable the learner to extract richer semantic features. The application of the attention mechanism can effectively make the model pay more attention to the important sentiment semantic features, and the accuracy and F1 value in the dataset are further improved by 0.15% and 0.20%. The Focal Loss is used in the model training process to further improve the model's recognition of a few negative reviews. The accuracy and F1 value of the model M-AttBiLSTMCNN-FL in this paper reach 97.34% and 95.74% respectively, achieved the best performance.

Then verity the performance of the sentiment analysis model M-AttBiLSTMCNN-FL on the constructed datasets generated by four different annotation methods, the results are shown in Table 8.

Table 8 M-AttBiLSTMCNN-FL model performance on different datasets

Dataset name	*Accuracy*	*Precision*	*Recall*	*f1*
Weak_vert	92.70	86.98	90.48	88.57
Pmi_vert	90.55	84.57	88.14	86.80
Dic_weak_vert	96.25	93.11	94.83	93.89
Pmi_weak_vert	**97.34**	**94.83**	**96.73**	**95.74**

From Table 8, it can be concluded that if only the weak-label data of e-commerce review ratings are directly converted into sentiment labels, more noise samples will significantly affect the training results of the model. However, if weak-label data is not used, and only the sentiment dictionary based on the SO-PMI method is used to annotate the review, the accuracy of the dictionary-based sentiment analysis method is often not as good as that of the deep learning, and the marked training set will have a large number of noise data, which reduces the accuracy of the model. If SO-PMI method is removed in the unsupervised annotation process, the joint annotation method that only uses the common sentiment dictionary and weak-label data can improve the training effect of the model to a certain extent, the accuracy and f1 value on this generated dataset reached 96.25% and 93.89% respectively. The unsupervised annotation method named "Dic_weak_vert" in this paper can not only make full use of the weak-label data of user ratings, but also expand the domain sentiment dictionary through the SO-PMI method. By identifying words with sentiment polarity in the field of agricultural products, the results of dictionary-based sentiment annotation are more reliable. The accuracy and f1 value of the model on the dataset generated by this method are further improved by 1.09% and 1.85%, reaching the optimal performance.

6 Conclusion

This paper proposes a sentiment analysis method for e-commerce user reviews based on the existence of weak- label data. The sentiment analysis method with high accuracy can increase the browsing experience of e-commerce users and provide information value for merchants to improve their product or service. Firstly, in the unsupervised annotation of review sentiment label method, a method based on improved SO-PMI is proposed to expand the domain vocabulary of the basic sentiment dictionary. By combining the sentiment dictionary and the weak-label data of user ratings, the sentiment value calculation results are converted into sentiment labels of user reviews to construct a high-precision dataset. In the sentiment analysis model, the embedding model is first used to convert the review text into word

vectors, POS vectors and character vectors, and import them into two different channels to extract text features in parallel, so as to alleviate the polysemy and OOV problem existing in e-commerce reviews. The advantages of BiLSTM and CNN are fused inside each channel, and an attention mechanism is embedded, so that the model can better extract the local semantic and global semantic features of the review text. In the process of model training, Focal Loss is used instead of cross entropy loss function to alleviate the problem of dataset imbalance, so that the model pays more attention to the learning of minority negative samples.

The experimental comparison results of different models and datasets show that the construction of multi-channel neural network integrating word-POS vector and character vector feature extraction, the application of attention mechanism and Focal Loss can improve the performance of the model. The proposed unsupervised annotation method based on improved SO-PMI can eliminate the noise caused by weak-label data and improve the performance of model training. The method proposed in this paper can be extended to sentiment analysis tasks in other scenarios. Nevertheless, the research in this paper still has some shortcomings that need to be improved. Since the aspect word extraction and processing module is not introduced in this paper, the deep learning model cannot be used to complete the more fine-grained sentiment analysis task. The next research work will improve the model in this paper and conduct research on fine-grained sentiment analysis.

Acknowledgement

This research was supported by the national key research and development program of Ministry of Science and Technology of the People's Republic of China (Grant Number: 2020YFD1100603): Development and demonstration of online service trading platform for agricultural materials and agricultural products.

References

Bao, Q.H., Li, J.L, Shi, S.Z., Dai, Y. and Liu, X. (2021) 'Sentiment analysis of egg consumption online review based on DSLML', *Transactions of the Chinese Society for Agricultural Machinery*, Vol. 52, No. S1, pp.496–503.

Cheng, K., Yue, Y. and Song, Z. (2020a) 'Sentiment classification based on part-of-speech and self-attention mechanism', *IEEE Access*, pp.16387–16396.

Cheng, Y., Yao, L.B., Zhang, G.H., Tang, T.W., Xiang, G.X., Chen, H.M., Feng, Y. and Cai, Z. (2020b) 'Multi-channel CNN and BiGRU-based text sentiment analysis based on attention mechanism', *Computer Research and Development*, Vol. 57, No. 12, pp.2583–2595.

Fan, H. and Li, P.F. (2021) 'Sentiment analysis of short text based on FastText word vector and bidirectional GRU recurrent neural network – taking Weibo comment text as an example', *Information Science*, Vol. 39, No. 4, pp.15–22. Doi: 10.13833/j.issn.1007-7634.2021.04.003.

Fan, Z., Guo, Y., Zhang, Z.H. and Han, M.Q. (2018) 'Sentiment analysis of movie reviews based on dictionary and weakly annotated information', *Computer Applications*, Vol. 38, No. 11, pp.3084–3088.

Hassan, A. and Mahmood, A. (2018) 'Convolutional recurrent deep learning model for sentence classification', *IEEE Access*, pp.13949–13957.

Kishore, P.K., Prathima, K., and Eswari, D.S. (2022) 'Bidirectional LSTM-based sentiment analysis of context-sensitive lexicon for imbalanced text', *Intelligent System Design: Proceedings of India 2022*, pp.494–283.

Liang, X., Liu, P. and Wang, Z. (2019) 'Hotel selection utilizing online reviews: a novel decision support model based on sentiment analysis and DL-VIKOR method', *Technological and Economic Development of Economy*, Vol. 25, No. 6, pp.1139–1161.

Lin, X.Y. and Wu, S. (2022) 'Typhoon disaster network emotion analysis method based on semantic rules and word vector', *Journal of Geo-information Science*, Vol. 24, No. 4, pp.114–126. Doi:10.12082/dqxxkx.2022.210575.

Lu, S.S., Lu, G.Y., Gu, Z.Y. and Xu, F. (2022) 'Small sample sentiment classification based on weakly supervised contrastive learning', *Computer Research and Development*, pp.1–13.

Mullen, T. and Collier, N. (2004) 'Sentiment analysis using supportvector machines with diverse information sources', *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp.412–418.

Obiedat, R., Qaddoura, R. and Ala'M, A.Z. (2022) 'Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution', *IEEE Access*, pp.22260–22273.

Qu, L., Rainer, G. and Gerhard, W. (2012) 'A weakly supervised model for sentence-level semantic orientation analysis with multiple experts', *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.149–159.

Schütze, H., Manning, C.D. and Raghavan. P. (2008) *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, Vol. 39, pp.234–265.

Täckström, O. and McDonald, R. (2011) 'Semi-supervised latent variable models for sentence-level sentiment analysis', *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Yan, L.L., Zhu, X.D. and Chen, X. (2021a) 'Review text sentiment classification algorithm based on dual-channel fusion and BiLSTM-attention', *.Journal of University of Shanghai for Science and Technology*, Vol. 43, No. 6, pp.597–605.

Yan, L.R., Zhu, X.D. and Chen, X. (2021b) 'Sentiment classification algorithm of review text based on dual-channel fusion and BiLSTM-attention', *Journal of University of Shanghai for Science and Technology*, Vol. 43, No. 6, pp.597–605. Doi:10.13255/j.cnki.jusst. 20210102001.

Yang, S. and Xia, Z. (2016) 'A convolutional neural network method for Chinese document sentiment analyzing', *2016 2nd IEEE International Conference.*

Zhao, W., Guan, Z. and Chen, L., (2017) 'Weakly-supervised deep embedding for product review sentiment analysis', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30, No. 1, pp.185–197.

Zhu, D., Shen, X. and Hedderich, M.A. (2022) 'Meta self-refinement for robust learning with weak supervision', *arXiv preprint arXiv:2205.07290*.