Sentiment analysis based on Chinese BERT and fused deep neural networks for sentence-level Chinese e-commerce product reviews

Hong Fanga, Guangjie Jiangb and Desheng Lib

Abstract
Driven by the rapid development of Internet, more e-commerce product reviews are available on e-commerce platforms, which can help enterprises make business decisions. Currently, bidirectional encoder representations from transformers (BERT) applied in the embedding layer contributes to achieve promising results in English text sentiment analysis (SA). This paper proposes a novel model Chinese BERT with fused deep neural networks (CBERT-FDNN), extracting richer and more accurate semantic and grammatical information in Chinese text. First, Chinese BERT with whole word masking (Chinese-BERT-wwm) is used in the embedding layer to generate dynamic sentence representation vectors. It is a Chinese pre-training model based on the whole word masking (WWM) technology, which is more effective for Chinese text contextual embedding. Second, multi-channel and multi-scale convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) are designed to capture further crucial features in the feature extraction layer. To obtain more comprehensive sentence attributes, these features are concatenated together. Last, the model is evaluated on 100,000 sentence-level Chinese e-commerce product reviews for sentiment binary classification. The accuracy and F1 score can achieve 94.37% and 94.34%, respectively. Compared with the baseline models, the experiments show that our proposed model has higher accuracy and better prediction performance.

ARTICLE HISTORY
Received 24 March 2022
Accepted 5 September 2022

KEYWORDS
Product reviews; sentiment classification; Chinese BERT; CNN; BiLSTM

1. Introduction
A survey (Pang & Lee, 2008) showed that 81% of consumers shopping online have explored the products at least once before purchasing, and people tend to make decisions based on electronic word-of-mouth. Significantly, the rapid development of Internet and the ongoing COVID-19 pandemic have accelerated the online shopping needs of consumers. Many product reviews are stored on websites, which are essential for enterprises to explore consumer preferences and make service decisions (Jain et al., 2021; Wu & Chang, 2020; Zhang et al., 2021).

Sentiment analysis (SA), widely recognized by the academic community, is known as opinion mining (OM). It is a process in which the overall emotional colour analysis is calculated and classified, or the entity attributes involved in the text opinion are extracted. Then the emotional tendency of the text is identified (Liu, 2012; Zhang et al., 2018). Nasukawa & Yi (2003, October) mention that the fundamental problem of SA is to identify how the sentiment is expressed in a text and whether the expression indicates a positive or negative view of a topic. SA can generally be regarded as a text classification problem. Polarity classification is one of the essential tasks of SA (Ju & Li, 2012, July). Conventional text polarity classification is a binary classification to determine whether the emotional colour of the text belongs positive or negative. According to the fine-graininess, the polarity classification task of text is usually divided into three levels (Birjali et al., 2021; Hemmatian & Sohrabi, 2019), aspect-level (Alqaryouti et al., 2020; Mowlaei et al., 2020; Wang et al., 2017), sentence-level (Chen et al., 2017; Mai & Le, 2021), and document-level (Moraes et al., 2013).

SA has been one of the hottest tasks in natural language processing (NLP) since 2000 (Appel et al., 2018), and research on SA has also flourished. SA has been widely used in public opinion analysis such as news (Krishnamoorthy, 2018), Twitter (Georgiadou et al., 2020), precise marketing and intelligence mining of products such as movies (Sharma & Mishra, 2016, October), hotels (Lucini et al., 2020), airlines (Chang et al., 2020), e-commerce products (Yang et al., 2020), human-machine dialogue (Li et al., 2022), medical services (Rajput, 2020), and other fields.
The SA technology can be generally divided into three categories. The first category is dictionary-based. The classification results’ accuracy depends on the sentiment dictionary’s completeness and accuracy. However, some emerging fields are short of special sentiment dictionaries. The current research trend is constructing a new emotion dictionary through the existing emotion dictionaries, language rules, or corpus (Bernabé-Moreno et al., 2020). The second category is based on machine learning (ML) algorithms (Pang et al., 2002). Traditional ML classification algorithms include support vector machine (SVM) (Wang et al., 2012), random forest (Probst et al., 2019), Bayesian (Jensen & Nielsen, 2007), and clustering (Ahani et al., 2019), etc. Among them, SVM is an early industry-recognized model with outstanding performance in the polarity classification task of text (Alomari et al., 2017, June). Deep learning (DL) (LeCun et al., 2015) is a popular branch of ML. It was firstly proposed in 2006 (Hinton & Salakhutdinov, 2006). DL algorithms can process massive data efficiently, extract features automatically, and have strong generalization capability. The classic deep neural networks (DNN) of DL algorithms include convolutional neural networks (CNN) (Kim, 2014), long short-term memory (LSTM) (Gers & Schmidhuber, 2001), etc. Their essences are the more complex artificial neural networks. At the same time, the third category method is a mixture of dictionaries and ML algorithms.

Fu et al. (2017) used the HowNet sentiment dictionary to help complete sentence-level text SA. Yekrangi and Abdolvand (2021) proposed a method based on the Twitter corpus to build a unique sentiment dictionary in the financial field. Chen et al. (2018) made a comparison model of LSTM and bidirectional LSTM (BiLSTM) based on word to vector (Word2vec) (Mikolov et al., 2013) for SA on mobile phone reviews. Behera et al. (2021) proposed an approach of hybrid CNN and LSTM for the sentiment classification of reviews in different domains. Li et al. (2020) proposed a multi-channel BiLSTM model based on self-attention for text sentiment classification. Basiri et al. (2021) proposed an attention-based bidirectional CNN-recurrent neural network (RNN) deep model (ABCDM) for five categories of reviews and tweets sentiment polarity analysis. Priyadarshini & Cotton (2021) worked on a novel LSTM-CNN-grid search-based DNN model for sentiment analysis. Li et al. (2021) studied the positive effect of the bidirectional encoder representations from transformers (BERT) (Arora et al., 2020; Devlin et al., 2018) model applied to the sentiment classification of investor comments in the Chinese stock field. Cai et al. (2020) collected about 100,000 comment samples from the Chinese Internet of energy market field as a dataset. They proposed a BERT-BiLSTM-based model to study the sentiment classification of comments.

It is difficult for existing general dictionaries to achieve satisfactory classification results in some professional fields (Sharma & Dutta, 2021). Therefore, constructing a new and professional emotion dictionary still takes some time. The models based on DNN models for text polarity classification are gaining more attention now, and the accuracy of classification results cannot be separated from the comprehensive extraction of text local features and contextual features. This paper proposes a novel model CBERT-FDNN (Chinese BERT with fused deep neural networks), efficiently processing many Chinese e-commerce comments. The model CBERT-FDNN could obtain more comprehensive text information for lots of sentence-level Chinese comments and improve the classification accuracy. Experimental results verify the feasibility of the proposed model, and the classification accuracy could reach 94.37%.

The rest of this paper is organized as follows. Section 2 presents the introduction of the proposed model CBERT-FDNN in detail. Section 3 shows the experimental process and results discussion. Finally, the whole paper is summarized in Section 4.

2. Proposed model

This section presents the classification model CBERT-FDNN we proposed. The contextual embedding model is described in Section 2.1, how the CNN and BiLSTM extract the features are presented in Sections 2.2 and 2.3, respectively. Finally, Section 2.4 describes the classifying principle of the fully connected layer.

The architecture of CBERT-FDNN is shown in Figure 1, includes the following three parts:

1. Contextual embedding layer. This layer is the input layer in this model. Chinese sentences are input into Chinese BERT with whole word masking (Chinese-BERT-wwm) to obtain more accurate pre-trained contextual embedding. Importantly, it is a 768-dimensional dynamic sentence vector \( v_s \) starting with [CLS] as the output representation for the whole sentence.

2. Feature extraction layer. We simultaneously extract further features from \( v_s \) through CNN of multi-channel and multi-scale and BiLSTM. 1-dimensional (1-D) convolution is used to extract vital local features for the text sentiment polarity classification task. The filters whose number is 64 are used to learn different features and obtain more information. We set up three channels with six different convolutional kernels in the convolutional layer, in the three channels the sizes of the kernels are 3, 5, and 7, respectively. Each channel has two local
Figure 1. The whole architecture of the CBERT-FDNN model.

2.1. Contextual embedding layer – Chinese-BERT-wwm

The embedding layer aims to generate feature vectors, representing the text data as mathematical expression vectors that the computer language can understand. Chinese-BERT-wwm is an outstanding deep and bidirectional contextual embedding model. It outputs a 768-dimensional dynamic vector $v_i$ starting with [CLS] as the whole sentence’s vectorized representation.

The original pre-trained model BERT was published by Google (Devlin et al., 2018) in 2018. In the tasks of SA, BERT achieves better performance than other previous 1-D or shallow bidirectional word vector models on more complex, unfamiliar sentences (Arora et al., 2020). And soon after, Google provided the BERT-Base-Chinese model for the Chinese language, which uses character-based tokenization. In 2021, Yiming Cui et al. created a series of models of Chinese BERT with the whole word masking (WWM) technology proposed by Google based on the BERT-Base-Chinese model (Cui et al., 2021), Chinese-BERT-wwm is one of them. In WWM, if the part of a term is replaced by Mask, the corresponding and rest of the term are replaced by Mask at once. Mask includes replacement with [MASK] label and keeps the original word, which is replaced with another word randomly. Therefore, the model’s ability to capture boundary relationships between words is better.

Besides, Chinese-BERT-wwm considers Chinese word segmentation (CWS), which is ignored in the previous BERT-Base-Chinese model, and the language technology platform (LTP) (http://ltp.ai) of Harbin Institute of Technology is used as word segmentation tool in Chinese-BERT-wwm. WWM means all the corresponding Chinese characters that form the whole term will be masked, and the training corpus includes both simplified and traditional Chinese Wikipedia. Therefore, the Chinese-BERT-wwm has a more flexible and robust Chinese text characterization ability. Suppose the rest of the term in a Chinese sentence is replaced with the [MASK] label in WWM, as shown in Table 1.

| Table 1. An example of Chinese sentence with word masking and segmentation. |
|---------------------------------------------------------------------------|
| Original sentence | 使用语言模型来预测下一个词的概率 |
| Original Sentence with Chinese Word Segmentation | 使用语言模型来预测下一个词的概率 |
| Original MASK input | 使用语言 [MASK] 来 [MASK] 下一个词的概率 |
| Whole word masking input | 使用语言 [MASK] [MASK] 来 [MASK] [MASK] 下一个词的概率 |
2.2. CNN feature extraction layer

CNN is a class of feedforward neural networks containing convolutional computing and deep architecture. CNN is characterized by local perception, weight sharing, multi-convolutional kernels, and down sampling. The connection between neurons in the upper and lower layers is local in the CNN. A specific neuron and all neurons in the latter layer share the same weight. Compared to the traditional neural networks, CNN has a small number of parameters, making the operation faster. CNN has a gradually growing popularity in processing text sentiment analysis tasks, whose unique components are the convolutional and pooling layers.

The convolutional layer is named because of the convolution operation. Generally, the more intuitive cross-correlation calculation replaces the convolution calculation in practical applications. The other name of the convolutional core is the filter or kernel. A convolutional window is also called the convolutional kernel window, consisting of the convolutional kernel array. Usually, in the 1-D cross-correlation operation, the sliding method is from top to bottom or left to right. The sliding distance is defined as the stride. For multiple channels, the output is obtained by adding the operation results of each corresponding channel, as Figure 2 is shown.

The pooling layer includes basic parameters such as filter size, stride, etc. For the input feature map, Pooling means choosing a pooling way to compress it to retain the main features while removing redundant information, improving the computation speed and the robustness of the extracted features. Standard pooling methods include Max Pooling, Average Pooling, etc. One of the most popular methods is Max Pooling because of its high efficiency. The practical effect of Max Pooling is to keep its maximum if a feature is extracted in the filter. Specifically, the largest number of the input matrix in the corresponding filter window is the corresponding value of the output matrix.

2.3. BiLSTM feature extraction layer

The main architecture of BiLSTM is composed of forward LSTM and backward LSTM. LSTM is very classic and popular in DL algorithms. Compared with the conventional RNN, the unique network structure of LSTM is based on retaining the previous iterative mode. It determines the output information’s effectiveness by introducing a specially designed gate mechanism and cell state. Figure 3 is the structural diagram of a single LSTM cell model.

In Figure 3, an LSTM unit contains an input gate, a forget gate, an output gate, and a linear cell state throughout the entire time step (Table 2).

Firstly, $x_t$ is input the forget gate to get the output $f_t$, and then gets its output $i_t$ and $\tilde{C}_t$ through the input gate, then we can calculate to get $C_t$, and $h_t$ are obtained through the product operation in the output gate, the operator $\odot$ signifies the element-wise multiplication. The
Table 2. Defined symbols and meanings.

| Symbol  | Meaning                                                                 |
|---------|-------------------------------------------------------------------------|
| \( h_t \) | The LSTM cell hidden layer output at the current time step              |
| \( h_{t-1} \) | The LSTM cell hidden layer output from the last time step               |
| \( o_t \) | The output of the output gate                                           |
| \( C_t \) | The cell state of the current time step                                 |
| \( C_{t-1} \) | The cell state of the last time step                                    |
| \( i_t \) | The output of the input door                                            |
| \( f_t \) | The output of the forget door                                           |
| \( x_t \) | The input                                                               |
| \( W \) | The continuously updated weights used in the calculation process        |
| \( b \) | The continuously updated bias terms used in the calculation process     |
| \( \sigma \) | The Sigmoid function                                                    |

The calculating mechanism of \( h_t \) is expressed as the following:

\[
\begin{align*}
    h_t &= o_t \odot \tanh(C_t), \\
    o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \\
    C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \\
    \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \\
    i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\
    f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f),
\end{align*}
\]

In the BiLSTM, the backward LSTM can reverse the data along the time, insert it into the LSTM, and then reverse the result and splice it with the information encoded by the forward LSTM. Therefore, BiLSTM can better capture the bidirectional semantic features, which are genuinely context-based and more representative. Its structure is shown in Figure 4, in which \( x_t \) represents the input at time \( t \), and \( h_t \) means the output stitching of the hidden state of the bidirectional LSTM unit at time \( t \). One BiLSTM neuron outputs the concatenation vector of 2-dimensional (2-D) hidden layer state vectors as

\[
h_t = \begin{bmatrix} h_{f_t}^r \ h_{f_t}^l \end{bmatrix}.
\]

where \( h_{f_t}^r \) is the forward hidden layer vector encoded by the forward LSTM at time \( t \), which includes the sequence information characteristics from left to right. \( h_{f_t}^l \) is the backward hidden layer vector encoded by the backward LSTM at time \( t \), which includes the sequence information characteristics from right to left, and \( h_t \) is the concatenation vector of \( h_{f_t}^r \) and \( h_{f_t}^l \), which includes bidirectional sequence features.

2.4. Fully connected layer

The fully connected layer maps the samples from the feature space to the label. Firstly, \( y_i \) is obtained from the matrix transformation of the feature vector \( V_i \) as

\[
y_i = w_z V_i + b_z.
\]

Then the probability of the classification is output by using the Softmax activation function as

\[
P_i = \frac{\exp(y_i)}{\sum_{j=1}^{n} \exp(y_j)}.
\]

\( y_i \) is the classification category of the output. In this paper, the output sentiment polarity has two classes, so \( n \) is set to 2.

3. Experiment and discussion

In this section, the used dataset and the evaluation criteria are introduced. We study the hyper-parameters impact on model performance and select suitable hyper-parameters for our model. Epoch and Dropout are taken as examples to show. Some popular models are used as the baseline comparison models. The experimental results show that compared with the other six groups of baseline models, the model that we have proposed, CBERT-FDNN has the highest scores for the evaluation criteria, which proves that our model is effective.

3.1. Dataset

The dataset used in this paper comes from Yang et al. (2020), which is a high-quality Chinese sentence-level book review with 100,000 sufficient data sorted out from the website http://www.dangdang.com/. The categories are balanced because it includes 50,000 positive comments and 50,000 negative comments, respectively. The labels of these comments are divided into 0 or 1, 0 represents negative comments, and 1 represents positive comments, as shown in Table 3.

We call the shuffle function to disrupt the data of different labels and then use k-fold cross-validation, where k is 10. 10-fold cross-validation refers to dividing the data set into 10 non-overlapping sub-datasets and selecting 9 of
3.2. Evaluation criteria

There are four evaluation indicators used in this paper, namely accuracy score, recall score, precision score, and F1 score. The closer their values are to 1, the better model performance is, which are defined as

\[
\text{Accuracy score} = \frac{TP + TN}{TP + FP + TN + FN}, \quad (10)
\]

\[
\text{Recall score} = \frac{TP}{TP + FN}, \quad (11)
\]

\[
\text{Precision Score} = \frac{TP}{TP + FP}, \quad (12)
\]

\[
\text{F1 Score} = \frac{2 \cdot \text{Recall score} \cdot \text{Precision Score}}{\text{Recall score} + \text{Precision Score}}, \quad (13)
\]

Specifically, TP means that the sentiment polarity of the predicted comment is consistent with the actual comment sentiment polarity, and both are positive. TN means that the sentiment polarity of the predicted comment is consistent with the actual comment sentiment polarity, and both are negative. FP means the sentiment polarity of the predicted comment is positive but negative, and FN means the sentiment polarity of the predicted comment is negative but positive. Recall score and precision score are contradictory measures from Equations (11) and (12). Usually, when one is high, the other will be low. F1 score is the weighted harmonic average of them.

3.3. Parameters setting

The batch size is set to 128, the learning rate is 0.001, the optimizer is Adam, which has fast convergence speed and high computational efficiency. The dropout rate is 0.3, the number of training rounds epoch is 25, and the loss function adopts the binary cross-entropy loss. Accuracy score, recall score, precision score, and F1 score are used as the model evaluation criterion, especially accuracy score and F1 score.

3.4. Experiments and results discussion

On the premise that other parameters are fixed, the optimization process of the proposed model is shown with Epoch and Dropout. Epoch is the time of the model trained on the training set. Choosing the appropriate Epoch and Dropout can help overcome the problems of under-fitting or over-fitting. We firstly observed the changes of the evaluation indicator with Epoch from 10 to 30 under the premise that Dropout is set to 0.1. The line graph is shown in Figure 5, and the appropriate Epoch value is chosen 25 for the model, which has the highest accuracy score and F1 score simultaneously.

Then, under the premise that Epoch is fixed at 25, we simulate the impact of Dropout changes on model performance. As shown in Figure 6, six different values 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 are chosen. The suitable Dropout value for the model is 0.3, which has the highest accuracy score and F1 score simultaneously.

We set up two sets of comparative experiments to evaluate the model performance. The first set of comparative experiments focuses on the input layer. We choose Word2vec in the input layer to obtain the word vectors and set them as the training set and the remaining one as the test set. This operation can verify the generalization ability of the model.
Table 4. The model performance comparison of different word embedding.

| Wordvector representations | Accuracy score | Recall score | Precision score | F1 score |
|----------------------------|----------------|--------------|----------------|---------|
| Word2vec                   | 0.9113         | 0.9170       | 0.9059         | 0.9112  |
| Proposed Model             | 0.9437         | 0.9449       | 0.9419         | 0.9434  |

and Chinese-BERT-wwm to generate the sentence representation vectors. The feature extraction layer is the same. The word vectors obtained by Word2vec include continuous bag-of-word (CBOW) and Skip-gram two ways, and the skip-gram is selected to generate word vectors by predicting nearby words from the central words in this paper. The final comparison results are shown in Table 4.

In Table 4, the four evaluation indicators of our proposed model are higher. When using Word2vec, accuracy score, recall score, precision score, and F1 score are reduced by 3.43%, 6.45%, 3.82%, and 3.41%, respectively compared with Chinese-BERT-wwm used. We analysed in principle and found that although Word2vec has undergone Jieba word segmentation, removal of stop words, and special symbols before the data set is trained, it generates static word vectors that do not consider the context. Resolving polysemy in Chinese requires consideration of context information seriously, both in generating sentence vectors and feature extraction. Chinese-BERT-wwm has precisely this advantage. The sentence representation vector it generates entirely takes care of the context, including more profound, accurate, and vivid semantic information. Hence, it can achieve better results.

The second set of comparative experiments focuses on the network models in the feature extraction layer. On the premise that the input layer is set to the Chinese-BERT-wwm model, the network model used in the feature extraction layer is changed. We set up six groups of classic baseline models to extract deep features for classification and compare them with our proposed model. The first group is SVM, the second group is RNN, the third group is LSTM, the fourth group is CNN, the fifth group is BiLSTM, and the sixth group is CNN_BiLSTM, which means that the feature vector extracted from the output of CNN is input into BiLSTM for feature extraction. Each set of comparative models uses 10-fold cross-validation, and we continuously optimize the hyper-parameters of each model. The final scores are displayed as shown in Table 5.

According to the data in Table 5, we rank the four model evaluation metric scores for seven sets of models, respectively. For accuracy score, Proposed Model > BiLSTM > LSTM > RNN > CNN > SVM > CNN_BiLSTM. For recall score, Proposed Model > RNN > BiLSTM > LSTM > CNN > SVM > CNN_BiLSTM. For precision score, Proposed Model > SVM > LSTM > CNN > BiLSTM > CNN_BiLSTM > RNN. For F1 score, Proposed

Table 5. Performance comparison for the proposed model and the baseline models.

| Model         | Accuracy score | Recall score | Precision score | F1 score |
|---------------|----------------|--------------|----------------|---------|
| SVM           | 0.9362         | 0.9311       | 0.9399         | 0.9355  |
| RNN           | 0.9394         | 0.9418       | 0.9367         | 0.9392  |
| LSTM          | 0.9396         | 0.9387       | 0.9398         | 0.9391  |
| CNN           | 0.9378         | 0.9352       | 0.9394         | 0.9373  |
| BiLSTM        | 0.9398         | 0.9406       | 0.9386         | 0.9395  |
| CNN_BiLSTM    | 0.9344         | 0.9307       | 0.9372         | 0.9338  |
| Proposed Model| 0.9437         | 0.9449       | 0.9419         | 0.9434  |

Figure 7. Performance comparison between the proposed model and the baseline models.

Model > BiLSTM > RNN > LSTM > CNN > SVM > CNN_BiLSTM. Our proposed model has the highest scores on the four evaluation indicators of the sentence-level Chinese e-commerce reviews dataset. The dynamic changes of accuracy score and F1 score are drawn in Figure 7.

In Figure 7, we can observe that the accuracy score and F1 score of our proposed model are higher than that of the other six groups of comparison models intuitively, which indicates that the thought of our model chooses CNN and BiLSTM in the feature extraction layer to extract features simultaneously and then performs feature fusion is effective and practical.

4. Conclusion

This paper mainly studies the sentiment classification of Chinese e-commerce product reviews at the sentence level. In this paper, the Chinese-BERT-wwm model is used in the input layer to get more representative sentence representation vectors for Chinese text. Combined with multi-channel multi-scale CNN and BiLSTM, the feature extraction layer extracts the sentence vectors’ crucial local and contextual features. The experimental results show that different models used in the input and feature extraction layers will impact the model’s classification performance, and the input layer model contributes more. Besides, our proposed CBERT-FDNN model can improve the sentiment classification accuracy of Chinese e-commerce product reviews compared with the baseline models.
Nevertheless, the potential of this work can still be further explored because it is currently a relatively efficient classification of two kinds of single emotional polarities for Chinese e-commerce products’ positive and negative reviews. In other cases, the emotions contained in the reviews may also be neutral, ambiguous, or a combination of multiple emotions. Many online comments data are derived from online sales of products on the Chinese e-commerce platforms. As a marketing method that conforms to the development trend of the new era, it provides a strong research basis for Chinese SA. Therefore, in the future work, we may also pay more attention to constructing our review dataset from the Chinese e-commerce platforms and studying how to accurately classify the reviews with compound emotions.

**Data availability statement**

Publicly available dataset was analysed in this study. It can be found here: https://github.com/ly2014/sentimen-analysis-based-on-sentiment-lexicon-and-deep-learning.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was supported by the Collaborative Innovation Platform of Electronic Information Master under grant no. A10GY21F015 of Shanghai Polytechnic University.

**References**

Ahani, A., Nilashi, M., Yadegaridehkordi, E., Sanzogni, L., Tarik, A. R., Knox, K., Samad, S., & Ibrahim, O. (2019). Revealing customers’ satisfaction and preferences through online review analysis: The case of Canary Islands hotels. *Journal of Retailing and Consumer Services, 51*, 331–343. https://doi.org/10.1016/j.jretconser.2019.06.014

Alomari, K. M., ElSherif, H. M., & Shaalan, K. (2017, June). Arabic tweets sentimental analysis using machine learning. In Salem Benferhat, Karim Tabia, & Moonis Ali (Eds.), *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 602–610). Springer, Cham.

Alqaryouti, O., Siyam, N., Monem, A. A., & Shaalan, K. (2020). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics, ahead-of-print* (ahead-of-print), 1–20. https://doi.org/10.1016/j.aci.2019.11.003.

Appel, O., Chichana, F., Carter, J., & Fujita, H. (2018). Successes and challenges in developing a hybrid approach to sentiment analysis. *Applied Intelligence*, 48(5), 1176–1188. https://doi.org/10.1007/s10489-017-0966-4

Arora, S., May, A., Zhang, J., & Ré, C. (2020). Contextual embeddings: When are they worth it?. arXiv preprint arXiv:2005.09117.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems, 115*, 279–294. https://doi.org/10.1016/j.future.2020.08.005

Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management, 58*(1), 102435. https://doi.org/10.1016/j.ipm.2020.102435

Bernabé-Moreno, J., Tejeda-Lorente, A., Herce-Zelaya, J., Porcel, C., & Herrera-Viedma, E. (2020). A context-aware embeddings supported method to extract a fuzzy sentiment polarity dictionary. *Knowledge-Based Systems, 190*, 105236. https://doi.org/10.1016/j.knosys.2019.105236

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing, 68*, 227–236. https://doi.org/10.1007/978-3-642-76153-9_28

Cai, R., Qin, B., Chen, Y., Zhang, L., Yang, R., Chen, S., & Wang, W. (2020). Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM. *IEEE Access, 8*, 171408–171415. https://doi.org/10.1109/ACCESS.2020.3024750

Chang, V., Liu, L., Xu, Q., Li, T., & Hsu, C. H. (2020). An improved model for sentiment analysis on luxury hotel review. *Expert Systems*. https://doi.org/10.1111/exsy.12580

Chen, H., Li, S., Wu, P., Yi, N., Li, S., & Huang, X. (2018). Fine-grained sentiment analysis of Chinese reviews using LSTM network. *Journal of Engineering Science & Technology Review, 11*(1), 174–179. https://doi.org/10.25103/jestr.111.21

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications, 72*, 221–230. https://doi.org/10.1016/j.eswa.2016.10.065

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for Chinese bert. IEEE/ACM transactions on audio, *Speech, and Language Processing, 29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fu, X., Liu, W., Xu, Y., & Cui, L. (2017). Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis. *Neurocomputing, 241*, 18–27. https://doi.org/10.1016/j.neucom.2017.01.079

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. *International Journal of Information Management, 51*, 102048. https://doi.org/10.1016/j.ijinfomgt.2019.102048

Gers, F. A., & Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks, 12*(6), 1333–1340. https://doi.org/10.1109/72.963769

Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review, 52*(3), 1495–1545. https://doi.org/10.1007/s10462-017-9599-6
