Expert Systems With Applications 187 (2022) 115977

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A deceptive reviews detection model: Separated training of multi-feature learning and classification

Ning Cao [a,1], Shujuan Ji [a,2,*], Dickson K.W. Chiu [b,3], Maoguo Gong [a,c,*,4]

[a] Key Laboratory for Wisdom Mine Information Technology of Shandong Province, Shandong University of Science and Technology, Qingdao, China
[b] Faculty of Education, The University of Hong Kong, China
[c] Key Laboratory of Intelligent Perception and Image Understanding, Xidian University, PO Box 224, Xi'an 710071, Shanxi, China

ARTICLE INFO

Keywords:
Deceptive reviews detection
Separated training
Convolutional neural network
Recurrent neural network
Self attention

ABSTRACT

The increasing online reviews play an essential role in the e-commerce platform, which profoundly affects the purchase decisions of consumers. However, rampant dishonest sellers manipulate other buyers or robots to post deceptive reviews for profit. Recently, the detection of deceptive reviews has attracted general research attention, which mainly comprises two directions, traditional methods based on statistics and intelligent methods based on neural networks. These methods use a single feature or multiple features for classifier design. To make full use of different features for better feature representation of detecting deceptive reviews, this paper proposes a new feature fusion strategy and verifies its performance by comparing it with other feature fusion strategies. First, we utilize three independent models for feature extraction: the TextCNN, the Bidirectional Gated Recurrent Unit (GRU), and the Self-Attention are used to learn local semantic features, temporal semantic features, and weighted semantic features of reviews, respectively. Secondly, after obtaining different feature representations from the fully connected layers of these three models, we concatenate them together to form the final documental representation. Finally, we use a full connection layer and the sigmoid function to further learn and complete deceptive review detection. Experiments on three balanced and unbalanced in-domain small datasets (hotel, restaurant, doctor) and mixed-domain datasets show that our model is superior to baselines. Experiments on large-scale data with various imbalanced proportions verify the effectiveness of our method. We also analyze the results of different datasets from the perspective of part of speech to improve the model's interpretability.

1. Introduction

The prosperity of the Internet has led more and more sellers to enter e-commerce platforms, and consumers are also aware of the convenience of online shopping, which produces a large number reviews of products and services that affect consumers' purchasing decisions. However, dishonest sellers take advantage of consumers' psychology and hire people or robots to post deceptive reviews to improve their reputations. Deceptive review is a kind of opinion-based false information that may affect consumers opinions or decisions (Kumar & Shah, 2018), and will undoubtedly harm the interests of potential consumers and honest sellers. Evidence has shown that around 14–20% of online reviews on Yelp are deceptive (Ott, Cardie, & Hancock, 2012; Fei, Mukherjee, Liu, Hsu, Castellanos, & Ghosh, 2013). However, Ott, Choi, Cardie, and Hancock (2011) show that the recognition accuracy of deceptive reviews is only 57.3%. Therefore, it is crucial to effectively detect deceptive reviews to protect the fairness of e-commerce platforms.

The detection of deceptive reviews is basically a classification task (Ott et al., 2011; Ren, Ji, & Zhang, 2014), and researchers used different features to accomplish this task. For example, Martinez-Torres and Toral (2019) used bags of words and sentiment features of reviews to detect

* Corresponding authors at: Key Laboratory for Wisdom Mine Information Technology of Shandong Province, Shandong University of Science and Technology, Qingdao, China.
E-mail addresses: 836300237@qq.com (N. Cao), jane_ji2003@aliyun.com (S. Ji), dicksonchiu@ieee.org (D.K.W. Chiu), Gong@ieee.org (M. Gong).
[1] ORCID: 0000–0002-5117–4516.
[2] ORCID: 0000–0003-2650–0161.
[3] ORCID: 0000–0002-7926–9568.
[4] ORCID: 0000–0002-0415–8556.

https://doi.org/10.1016/j.eswa.2021.115977
Received 4 February 2021; Received in revised form 9 July 2021; Accepted 23 September 2021
Available online 28 September 2021

deceptive reviews. Ott et al. (2011) used part-of-speech (POS), Linguistic Inquiry and Word Count (LIWC), and n-gram features of reviews for detecting deceptive reviews. Reviewer features were used to detect deceptive reviews (Jindal & Liu, 2008; Mukherjee, Kumar, Liu, Wang, Hsu, Castellanos, & Ghosh, 2013). Topic information was also used to identify deceptive reviews or to implement text categorization tasks. Zhu, Li, and Luo (2013) used external corpora and LDA topic models to implement classification tasks. Li, Cardie, and Li (2013) proposed the TopicSpam model, a generative LDA-based topic model, for deceptive review detection. Some researchers (Ren & Ji, 2017; Li, Qin, Ren, & Liu, 2017) used neural network-based representation learning algorithms such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to detect deceptive reviews, which can automatically extract features and achieve competitive results.

Recently, the attention mechanism has also been extensively studied. Its first application in natural language processing (NLP) tasks is neural machine translation (Bahdanau, Cho, & Bengio, 2014). The attention mechanism appears in more and more NLP tasks, such as aspect and opinion terms co-extraction (Wang, Pan, Dahlmeier, & Xiao, 2017), reading comprehension (Cui, Chen, Wei, Wang, Liu, & Hu, 2016), sentence representation (Wang, Zhang, & Zong, 2016), and sentiment classification (Nguyen, & Nguyen, 2020). Some researchers also used the attention mechanism in text classification tasks (Yang, Yang, Dyer, He, Smola, & Hovy, 2016; Zheng, Cai, Shao, & Chen, 2018).

In summary, traditional statistics-based learning methods have good interpretability, but they have two main problems: (1) feature engineering is needed, and the acquired features are discrete. (2) the sparsity problem prevents effective learning of semantic information. In comparison, the neural network-based method can automatically obtain richer semantic features, but it has poor interpretability and does not fully consider different semantic features.

To address the above problems, we study the differences of the classical neural network structure (e.g., CNN, RNN, attention mechanism) in feature representation for deceptive reviews detection, and propose a new method, namely separated training of multi-feature learning and classification (ST-MFLC), to improve the detection effect of deceptive reviews. The main contributions of this paper are as follows:

(1) To extract multi-level features and obtain sufficient text representation, we independently use TextCNN, Bidirectional GRU (Bi-GRU), and Self-Attention models to obtain local semantic features, temporal semantic features, and weighted semantic features in the text, respectively.
(2) To make full use of the learned features, we propose a new multi-feature fusion method. We first extract the output from the full connection layer of TextCNN, Bi-GRU, and Self-Attention models as independent feature representation. Then, we splice these features into the final text representation.
(3) Generally, there are two kinds of multi-feature fusion methods, serial connection method (Jain, Sharma, & Agarwal, 2019) and parallel connection method (Guo, Zhang, Wang, Wang, & Cui 2018; Madisetty & Desarkar, 2018). Feature learning and classification also have two ways, synchronous method (Ren & Ji 2017; Jain et al., 2019; Guo et al., 2018; Madisetty & Desarkar, 2018) and separate method (Fan, Lu, Li, & Liu, 2016; Cao, Ji, Chiu, He, & Sun, 2020), and most solutions are based on the former. For the multi-feature series-connected fusion method, feature weakening may occur when learning different features in sequence, while the synchronously parallel-connected method cannot achieve the learning of all the features, causing insufficient usable features. To solve this problem, this paper explores a new feature fusion method, i.e., we train three different models independently to sufficiently capture different features, and then fuse these features in parallel so as to classify the reviews into deceptive and true ones, which realizes the separated training of multi-feature learning and classification.
(4) To verify the model's performance, we design three sets of experiments based on the expanded gold standard small dataset (Li, Ott, Cardie, & Hovy, 2014) covering in-domain and mixed-domain applications and large-scale Yelp dataset. We implement two models according to different feature fusion strategies (i.e., multi-feature series/parallel-connected model) as the baselines. At the same time, we recover the model of Ren and Ji (2017) and Li et al. (2017) for comparison.
(5) The experimental results on the balanced/unbalanced in-domain datasets and mixed-domain datasets show that our ST-MFLC model is superior to the baselines, and our separated training of multi-feature learning and classification method has good stability and is suitable for deceptive reviews detection. Notably, the analysis of the results from the perspective of part-of-speech further improves the interpretability of our model.

In the remaining parts, Section 2 compares related work. Section 3 details our model. Section 4 introduces our experiment setup and analyzes the experimental results. Section 5 concludes the paper with our future work location.

2. Related work

Existing methods for detecting deceptive reviews can be divided into two categories: one is based on statistics, and the other is using neural networks for identifying deceptive reviews.

2.1. Statistics-based methods

Deceptive review detection is fundamentally a text classification problem, in which selecting useful features to improve the performance of classification is crucial. Jindal and Liu (2008) were the first to raise the issue of deceptive review detection. They used features of reviews, reviewers, products, and duplicated reviews to establish a logistic regression model for detecting deceptive reviews. Mukherjee et al. (2013) used reviews and reviewer features to identify deceptive reviews based on the Bayesian method. Yoo and Gretzel (2009) collected a dataset containing 40 truthful hotel reviews and 42 deceptive reviews. They analyzed the difference between truthful reviews and deceptive reviews from the perspective of language structures. Ott et al. (2011) collected a gold standard dataset of deceptive reviews from Amazon Mechanical Turk. They used parts-of-speech (POS), Linguistic Inquiry and Word Count (LIWC), and n-gram features to detect deceptive reviews. Feng, Banerjee, and Choi (2012) used deep syntactic features extracted by context-free grammar parse trees and n-gram features to identify deceptive reviews. Feng and Hirst (2013) computed compatibility between reviews and product profile, and then combined the n-grams and the deep syntactic features (Feng et al., 2012) for detecting deceptive reviews.

Li et al. (2014) expanded a new gold standard dataset containing hotel, restaurant, and doctor domains, and added deceptive reviews from experts. They utilized Unigram, POS, and LIWC features of reviews for training their model and studied deceptive review detection on cross-domain. Chen and Chen (2015) used features such as bags-of-words, content characteristics, submission time, and sentiment on brands to detect deceptive reviews. Sadman, Gupta, Haque, Poudyal, and Sen (2020) used Vader Sentiment Analysis and Jaccard Similarity for detecting deceptive reviews. Savage, Zhang, Yu, Chou, and Wang (2015) used anomalous rating deviation to detect deceptive reviews. Li, Huang, Yang, and Zhu (2011) used review-related features and reviewer-related features to design a two-view method for detecting deceptive reviews. Martinez-Torres and Toral (2019) considered sentiment polarity (i.e., positive, negative) of reviews, and used Term Frequency-Inverse Document Frequency (TF-IDF) and the unique attributes word in

different review categories (i.e., truthful and deceptive reviews) to analyze and identify deceptive reviews. Ahmed, Traore, and Saad (2018) used the n-gram model, Term Frequency (TF), and TF-IDF to extract features and compared the performance of six classification algorithms in detecting opinion spam and fake news.

Some researchers also used the topic model to obtain the topic-related features for text classification. Blei et al. (2003) proposed the Latent Dirichlet Allocation (LDA) topic model. Many researchers made improvements based on the LDA topic model. Li, Ouyang, Zhou, Lu, and Liu (2015) proposed supervised labeled latent Dirichlet allocation (SL-LDA), which is an extension of L-LDA. They used SL-LDA for document categorization. Dong, Ji, Zhang, Zhang, Chiu, Qiu, and Li, (2018) extended the LDA topic model to the unsupervised topic-sentiment joint probabilistic model (UTSJ) and used it to detect deceptive reviews. Du, Zhu, Zhao, Zhao, Han, and Zhu (2020) proposed the Sentence Joint Topic Sentiment Model. They used sentence structure information and sentiment knowledge of reviews to improve the LDA topic model. Then they designed a voting system of multiple-classifier for deceptive review detection.

2.2. Neural network-based methods

Recently, neural networks have become popular and play an essential role in NLP tasks. Compared with feature engineering methods with the sparseness problem, neural networks can learn continuous representation for NLP tasks (Le & Mikolov, 2014). In the research of deceptive information detection, the most commonly used methods are CNN and RNN (Ren & Ji, 2019). In recent years, the attention mechanism has been widely studied and applied in NLP tasks (Hu, 2019).

Kim (2014) proposed a CNN model for sentence-level classification tasks, which has multiple filters of different sizes. Kalchbrenner, Grefenstette, and Blunsom (2014) used CNN and dynamic k-max pooling for the semantic modeling of sentences. Wang (2017) presented a hybrid Convolutional Neural Networks framework to integrate metadata with text and used this framework for detecting fake news. Li et al. (2017) used CNN to learn the sentence representation, and then used KL-divergence to implement a sentence-weighted neural network model to get the document representation for detecting deceptive reviews. Zhao, Xu, Liu, Guo, and Yun (2018) explored word order characteristics and designed a word order-preserving CNN network for detecting deceptive reviews, which used the word order persevering k-max pooling method.

Ma, Gao, Mitra, Kwon, Jansen, Wong, and Cha (2016) used different RNN structures (tanh-RNN, Single-layer LSTM/GRU, and Multi-layer GRU) to detect Rumors in microblogs. Ren and Ji (2017) utilized CNN to learn the sentence representation and gated recurrent neural network to get the document representation from sentence representation. Then, they used their neural network model to detect deceptive reviews. Based on knowledge bases such as WordNet and ConceptNet, Jain et al. (2019) used CNN and LSTM for spam detection. Lai, Xu, Liu, and Zhao (2015) proposed a Recurrent Convolutional Neural Networks (RCNN). They first used bidirectional recurrent neural networks to capture the left and right contexts of a word and combined them with the original word to form a new vector. Then they used the max-pooling layer to complete text classification. Zhang, Du, Yoshida, and Wang (2018) improved the RCNN model and used it to identify deceptive reviews. They considered that the words in deceptive reviews and truthful reviews correspond to different context information.

Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin (2017) proposed the Transformer architecture, which is completely based on the Self-Attention mechanism for translation tasks without recurrence and convolutions. Yang et al. (2016) used a hierarchical attention mechanism (i.e., word attention and sentence attention) and Gated Recurrent Units to achieve document classification. Wang, Liu, and Zhao (2017) proposed a neural network based on the attention mechanism for detecting deceptive reviews. They first used Multi-Layer Perceptron (MLP) and CNN to extract linguistic and behavioral features, respectively. Then they used a feature attention module to learn the attention scores between linguistic and behavioral features. Wang, Wang, Yang, and Lian (2021) used graph convolutional network and self-attention mechanism to learn the global sentence representations for fake news detection. Letarte, Paradis, Giguère, and Laviolette (2018) used Self-Attention for sentiment classification. Gao, Ramanathan, and Tourassi (2018) proposed a hierarchical model and used convolution and Self-Attention to implement text classification.

2.3. Combined methods and research gap

In addition to the above methods, some researchers used different feature combination methods for deceptive information detection via text classification. Li et al. (2017) combined their SWNN model with POS features and first-person pronoun features to detect deceptive reviews. Cao et al., (2020) used neural networks and the LDA model to obtain fine-grained features based on word vectors and coarse-grained topic features to detect deceptive reviews. Madisetty and Desarkar (2018) proposed an ensemble approach to detect twitter spam. They used five CNN-based methods with different word vector settings and one traditional feature-based method in the Ensemble. Guo et al. (2018) presented a hybrid CNN-RNN attention-based model for text classification. They first used the CNN and RNN layers to extract different features. Then they utilized an attention mechanism to combine the output of the CNN and RNN layers.

Statistics-based methods require manual design of features and do not perform well enough. Existing neural network-based methods or the combination method based on neural networks and statics mainly have the following problems. (1) The model that extracts a single semantic feature (local feature, temporal feature, weighted feature) does not comprehensively consider multiple features. (2) The serial consideration of multiple features, such as considering local features first and then temporal features, is not consistent with human writing habits. (3) Simply considering multiple features in parallel synchronously, the method that considers the word-sentence-document structure, and the method that is simple combination of neural networks and statistics cannot effectively learn every feature in the reviews.

To solve the above problems, we propose a new model for deceptive review detection, which focuses on the feature learning method based on neural networks. Compared with existing feature fusion methods, different features in the review text can be learned independently to obtain better feature representation in our model. Moreover, we use the commonly used neural network structure (e.g., CNN, RNN, attention mechanism) to extract features and fuse these multiple feature representations, which can effectively improve the detection effect of deceptive reviews.

3. Our approach

The habit of human expression has some basic logic and patterns. For example, consumers usually comment on some aspects of the product to express their views (Li, Huang, & Zhu, 2010), and the reviews are coherent by using some conjunctions, such as “however” and “although”. As human expressions typically have a focus, product comments normally focus on the satisfaction or disappointment of the product. The self-attention mechanism is often used to stress the importance of different parts of a document (Wang et al., 2021), which is why we use the self-attention mechanism to enhance text representations in our model.

This section introduces our deceptive review detection model based on separated training of multi-feature learning and classification in detail. Based on the previous analysis, we use TextCNN, Bi-GRU, and Self-Attention network to learn the different feature representations in the review text. TextCNN may use multiple filters of different sizes for extracting different n-gram features to better capture local features. GRU

proposed by Cho, Van Merriënboer, Bahdanau, and Bengio (2014) is similar to Long-Short-Term Memory (LSTM) networks but faster (Chung, Gulcehre, Cho, & Bengio, 2014) and it can capture temporal features. Self-attention has been widely used and achieved good results (Vaswani et al., 2017) as it can capture weighted features. As each word has a different contribution to the review's semantic information, the original semantic information is enhanced by weighting and fusing the importance of each word.

Fig. 1 illustrates the architecture of the deceptive review detection model based on separated training of multi-feature learning and classification proposed in this paper. First, we use three independent sub-models to learn different feature representations in the review text in parallel. Text classification tasks comprise mainly three types of commonly used models: convolutional neural networks, recurrent neural networks, and attention mechanism. Therefore, we select TextCNN, Bi-GRU, and Self-Attention models, respectively, as the feature extraction methods in this paper. The TextCNN model is used to learn local semantic features; the Bi-GRU model is used to learn temporal semantic features; and the Self-Attention model is used to learn weighted semantic features. Second, the output of the full connection layer is used as the feature representation of the review text, and the three different feature representations obtained through the three sub-models are spliced together. Finally, a fully connected layer and sigmoid function are used to realize the learning and classification of the final text representation.

The commonly used method for feature fusion is to obtain different feature representations through different network layers and splice them instead of obtaining features through separate models. Unlike the method based on other multi-feature fusion strategies, the method proposed in this paper has three advantages: (1) The different features in the review text can be learned independently and in parallel. (2) Different sub-models focus on the learning of different features and can obtain better feature representations. (3) Splicing different independent feature representations together and learning based on the final text representation can achieve the separated training of multi-feature learning and classification, so as to obtain better classification performance.

4. Experiments

To evaluate the performance of our model, we design three sets of experiments. The first set of experiments is to verify the performance of our model on the balanced/unbalanced in-domain small datasets. The second set of experiments is on the mixed-domain datasets. The third set of experiments is on the balanced/unbalanced in-domain large-scale datasets.

4.1. Data description and experimental settings

In the first and second sets of experiments, we use the small gold standard dataset of deceptive reviews detection published by Li et al. (2014), which contains three domains: doctor, hotel, and restaurant.

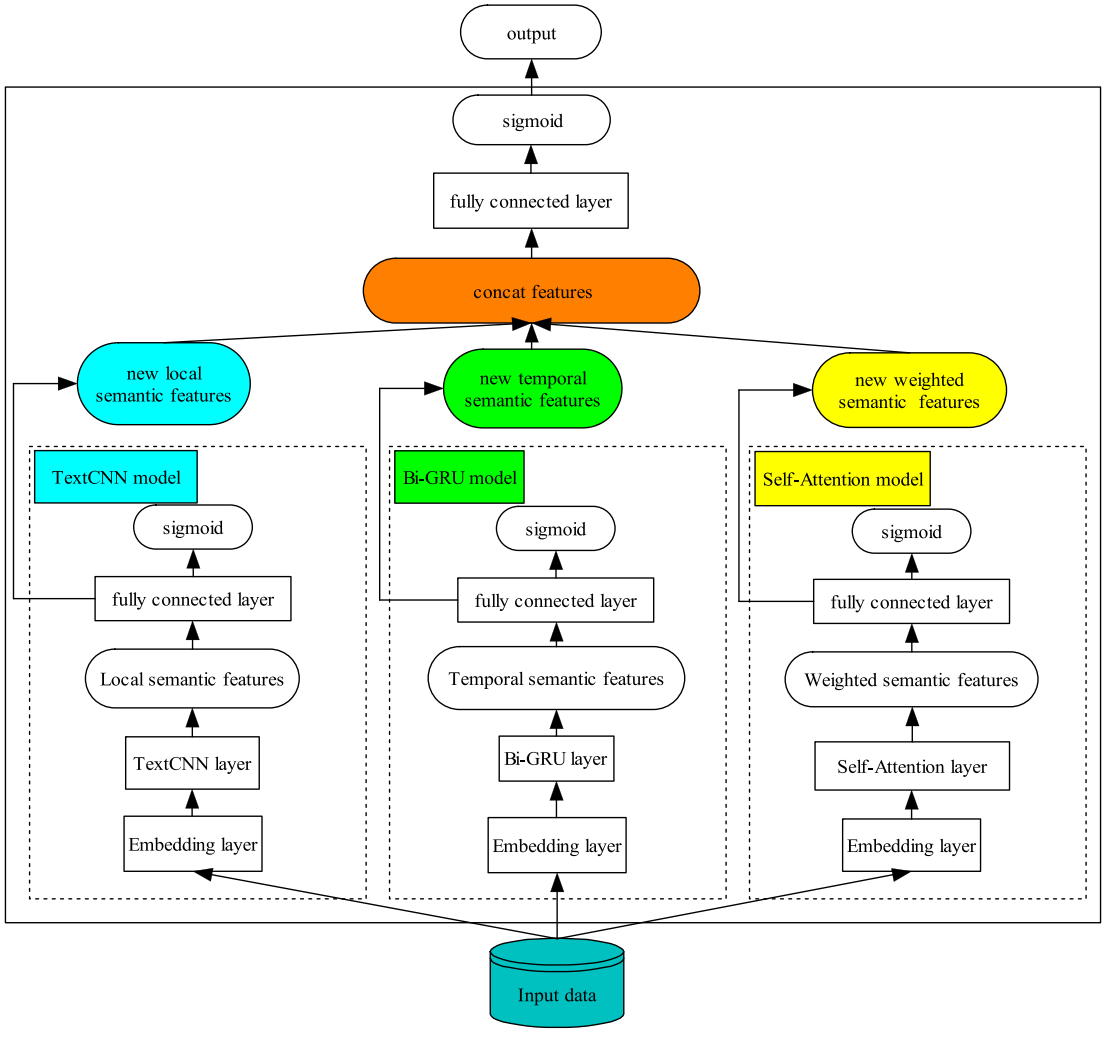

Fig. 1. The architecture of our ST-MFLC model.

Table 1
Statistics of the original golden standard dataset.

Domain	Turker	Employee	Customer	Deceptive %	Total reviews
Hotel	400/400	140/140	400/400	50	1880
Restaurant	200/0	0	200/0	50	400
Doctor	356/0	0	200/0	64	556

Table 1 illustrates the statistics of this dataset, where "Customer" represents the real review, "Turker" and "Employee" represent the deceptive review.

To comprehensively evaluate our model, we construct balanced and unbalanced, in-domain and mixed-domain datasets based on the original dataset. For in-domain datasets, we only use reviews from Turker and Customer. For mixed-domain datasets, we use reviews from all sources to construct the Mixed[1] dataset. At the same time, to make the proportion of deceptive reviews in the mixed-domain dataset equal to that in the unbalanced in-domain dataset, we randomly select some deceptive reviews to construct the Mixed[2] dataset. Table 2 lists the statistics of our constructed small datasets.

In the third set of experiments, we use the large-scale Yelp restaurant dataset (Mukherjee, Venkataraman, Liu, & Glance, 2013). Similar to the first set of experiments, we construct balanced and several groups of unbalanced datasets based on Yelp dataset. Table 3 illustrates the statistics of the Yelp-related dataset.

We use the same strategy to pre-process each dataset. Our process has the following steps: word segmentation, converting characters to lowercase, lemmatization, and removing stop words.

In these three sets of experiments, we use 5-fold cross-validation to evaluate models. We use the same data split to train and test each model. We use accuracy (A), precision (P), recall (R), and F1 score (F1) as the evaluation criteria in our experiments, where P, R, and F1 are calculated using macro-average. In particular, to objectively compare the effect of models on the unbalanced dataset, we also use Area Under Curve (AUC) as the evaluation criteria.

4.2. Baseline methods

To compare our model with other various models in detecting deceptive reviews, we select and recover several typical baselines such as TextCNN, Bi-GRU, Self-Attention, the neural model of Ren and Ji (2017), and the SWNN model of Li et al. (2017). Notably, to verify the performance of our method more completely, we construct two models based on different feature fusion strategies as the baseline. The main characteristics of baselines are as follows.

1. **Multi-feature parallel-connected model:** We construct a synchronously multi-feature parallel-connected fusion model as shown in Fig. 2, using the TextCNN layer, Bi-GRU layer, and Self-Attention layer to learn different features in reviews, and directly splicing features to complete the detection of deceptive reviews.

Table 2
Statistics of the experimental constructed small dataset.

Dataset	Deceptive	True	Deceptive%	Total reviews
Hotel^{y}	800	800	50	1600
Restaurant^{y}	200	200	50	400
Doctor^{y}	200	200	50	400
Hotel^{n}	550	800	40.7	1350
Restaurant^{n}	135	200	40.3	335
Doctor^{n}	135	200	40.3	335
Mixed^{1}	1636	1200	57.7	2836
Mixed^{2}	800	1200	40	2000

Table 3
Statistics of the Yelp-related dataset.

Dataset	Deceptive	True	Deceptive%	Total reviews
Yelp^{1}	8303	8303	50	16,606
Yelp^{2}	8303	12,455	40	20,758
Yelp^{3}	8303	19,373	30	27,676
Yelp^{4}	8303	33,212	20	41,515
Original Yelp (Yelp^{5})	8303	58,716	12.4	60,719

2. **Multi-feature series-connected model:** We compare different ways of multi-feature series-connected, and select a model with the best performance as the baseline. As shown in Fig. 3, the structure of the model is connected in sequence Bi-GRU-Self-Attention-TextCNN. The process of feature learning and classification is synchronous. We do not consider the separated multi-feature serial connection method because the effect of this method is similar to that of the synchronous one.
3. **TextCNN:** Multiple filters and pooling layers of different sizes.
4. **Bi-GRU:** Bi-directional Gated Recurrent Unit, a kind of recurrent neural network.
5. **Self-Attention:** A kind of attention mechanism.
6. **Li's SWNN** (Li et al., 2017)**:** Learn sentence representation from reviews, and then learn document representation. Both levels of learning are completed through CNN.
7. **Ren's Neural** (Ren & Ji, 2017)**:** Learn sentence representation through CNN, and then learn document representation using gated recurrent neural networks.

For each model, we use the same setting in some common parameters. We use Adam optimizer and use the default learning rate, which is 0.001. The size of all hidden layers is set as 64, such as CNN, GRU, Self-Attention, and fully connected layer. The widths of five convolutional filters are set as 1, 2, 3, 4, and 5. We set the text sequence size to 130, the embedding size is set as 128, but the size of the dictionary depends on different datasets. In the training process, we monitor the accuracy of the training set to achieve early stopping, which helps avoid overfitting.

4.3. In-domain results and analysis on the small dataset

In the in-domain experiment based on the small dataset, we compare our model with baselines in accuracy, precision, recall, F1 value, and AUC on the balanced/unbalanced hotel, restaurant, and doctor datasets.

Table 4 illustrates the comparison results on the hotel domain. On the balanced Hotel^{y} dataset, among the models based on the single feature (TextCNN, Bi-GRU, and Self-Attention), the TextCNN model performs best, followed by the Self-Attention model. Moreover, the Self-Attention model has the best AUC (0.95) in all baselines. Li's SWNN model performs best in accuracy, precision, recall, and F1 value (87.6, 87.6, 87.6, 87.6), which shows that the structure of learning sentence representation first and then learning document representation is helpful to represent review texts better. The performance of the multi-feature parallel-connected model is similar to that of the series-connected model, but they are not as good as models based on single feature. This shows that the multi-feature parallel-connected or series-connected method has poor performance, because it simply learns different features, which means it cannot learn each feature deeply. The ST-MFLC model in this paper performs best (88, 88.1, 88, 88, 0.952), which shows that fusing the features independently is beneficial to improve the model's performance.

For the unbalanced Hotel^{n} dataset, the TextCNN model performs best in the baselines (86.8, 86.6, 86.3, 86.3, 0.949). In the model based on word-sentence-document feature, Li's SWNN model is better than the Self-Attention, Bi-GRU, and series/parallel connection models. Our ST-MFLC model achieves the best performance (88.1, 88.1, 87.4, 87.6, 0.952), indicating that the method in this paper is also applicable to

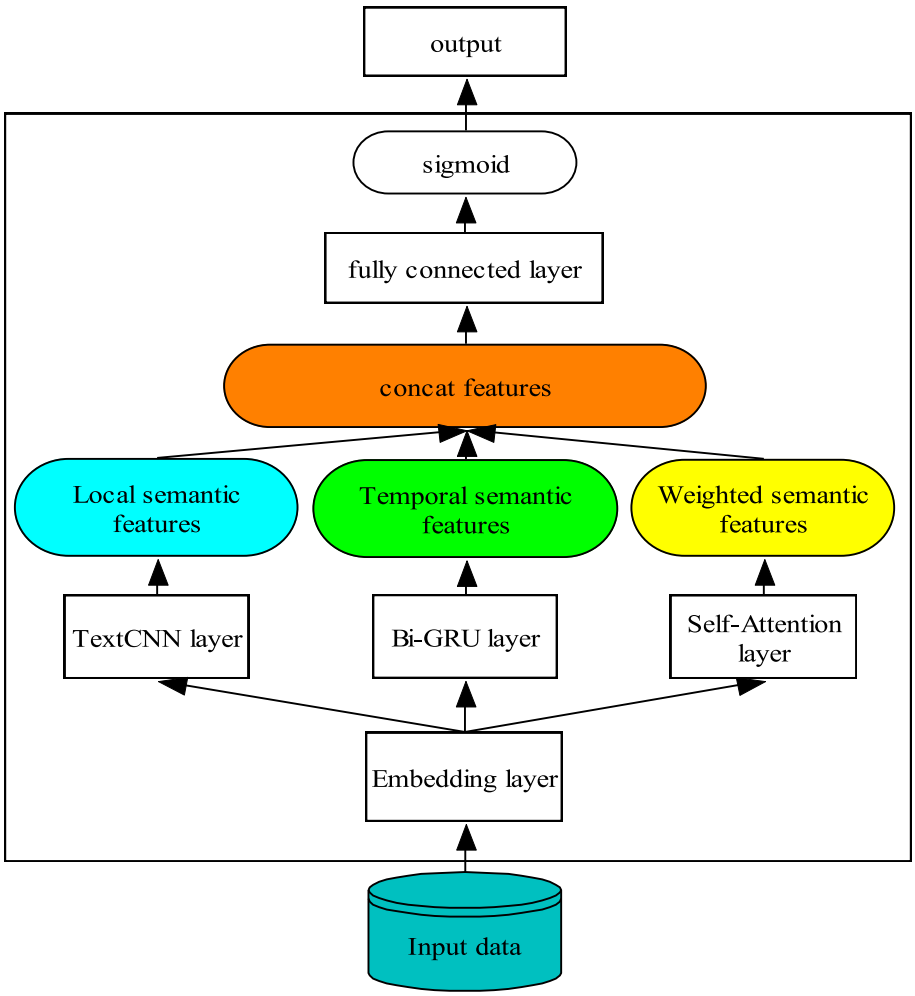

Fig. 2. The architecture of the multi-feature parallel-connected fusion model.

unbalanced data.

Table 5 illustrates the comparison results on the restaurant domain. On the balanced Restaurant^{y} dataset, the Self-Attention model is better than Bi-GRU and TextCNN models. It performs best in the baselines (83.8, 84.4, 83.8, 83.7, 0.908), which shows that the attention mechanism has a good effect on text representation. Compared with the series-connected model, the parallel-connected model has better performance. The ST-MFLC model proposed in this paper has the best overall performance (85, 85.3, 85, 85, 0.921), because our separated training of multi-feature learning and classification strategy helps obtain good feature representation.

On the unbalanced Restaurant^{n} dataset, Ren's Neural model achieves the best overall performance among baselines (82.1, 82.9, 81.8, 81.3, 0.902). The Bi-GRU model has the best AUC (0.91) in all baselines. The multi-feature series-connected model is better than parallel-connected in accuracy, precision, recall, and F1 value, but worse in AUC. Our ST-MFLC model has the best performance (85.7, 86.2, 84.3, 84.7, 0.912). The recall of most models is relatively low, but our model can still maintain a high level.

Table 6 illustrates the comparison results on the doctor domain. For the balanced Doctor^{y} dataset, the multi-feature parallel-connected model performs best in the baselines (88.3, 89.1, 88.8, 88.7, 0.956). The Self-Attention and Li's SWNN models are better than TextCNN and Bi-GRU models. The performance of the multi-feature series-connected model is better than the models based on single feature. Our ST-MFLC model has the best overall performance (90.3, 90.3, 90.3, 90.2, 0.961).

For the unbalanced Doctor^{n} dataset, the multi-feature parallel-connected model also has the best overall performance (86.3, 87, 85.6, 85.6, 0.931) among the baselines. Li's SWNN and Ren's Neural models have similar performances. Li's SWNN model achieves the best AUC (0.934) among the baselines. Our ST-MFLC model achieves the best performance in all evaluation metrics (87.5, 88, 86, 86.6, 0.944).

Based on the above in-domain experimental analysis, we can get the following conclusion.

Conclusion 1: Our separated training of multi-feature learning and classification method is better than baselines in in-domain experiments based on the small dataset, including models based on a single feature (TextCNN, Bi-GRU, and Self-Attention), multi-feature parallel/series-connected fusion model, and models based on word-sentence-document structure.

To better analyze the experimental results and improve the interpretability of our model, we study reviews from the perspective of text length and part-of-speech. Fig. 4 shows the kernel density estimation plot of the length of truth/deceptive review text in each domain, which visually shows the distribution characteristics of sample data. The horizontal axis is the length of the text, and the vertical axis is the density of the length of the text. In the balanced data, the average length of the text in the doctor domain is relatively small, most of which are less than 50, while the text length in the restaurant and hotel domains is relatively large. The distribution of truth and deceptive doctor datasets is quite different. The average length of the truth doctor dataset is smaller than that of the deceptive doctor dataset. However, for the restaurant and hotel datasets, there is no obvious difference between truth and deceptive data. Compared with the balanced dataset, the difference between

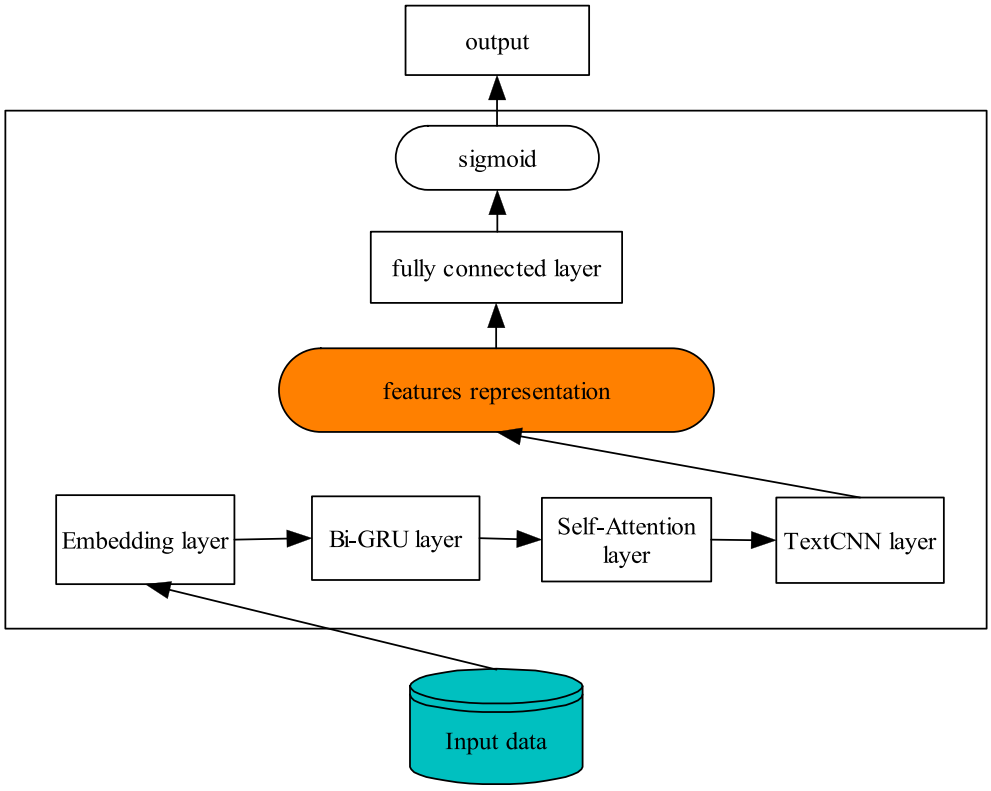

Fig. 3. The architecture of the multi-feature series-connected model.

Table 4
The experimental results on the Hotel dataset.

Model type	Model	Balanced Hotel^{y} dataset					Unbalanced Hotel^{n} dataset				
		A	P	R	F	AUC	A	P	R	F	AUC
single feature	TextCNN	87.1	87.1	87.1	87.1	0.948	86.8	86.6	86.3	86.3	0.949
	Bi-GRU	86	86.1	86	86	0.936	84.7	84.3	84.1	84.1	0.924
	Self-Attention	86.9	87	86.9	86.9	0.95	86.1	85.7	85.6	85.6	0.948
word-sentence-document feature	Li's SWNN	87.6	87.6	87.6	87.6	0.944	86.3	86.2	85.5	85.7	0.938
	Ren's Neural	83.6	84	83.6	83.6	0.925	84.7	84.5	83.6	83.9	0.928
multi-feature	parallel-connected	84.7	84.9	84.8	84.7	0.935	85.2	84.9	85.1	84.8	0.93
	series-connected	84.7	85	84.7	84.7	0.92	85.3	85.4	84.5	84.7	0.921
	ST-MFLC	**88**	**88.1**	**88**	**88**	**0.952**	**88.1**	**88.1**	**87.4**	**87.6**	**0.952**

Table 5
The experimental results on the Restaurant dataset.

Model type	Model	Balanced Restaurant^{y} dataset					Unbalanced Restaurant^{n} dataset				
		A	P	R	F	AUC	A	P	R	F	AUC
single feature	TextCNN	83.8	84.1	83.8	83.7	0.897	80.6	82.9	77.3	78.3	0.89
	Bi-GRU	83.5	83.9	83.5	83.5	0.899	81.2	82.4	78.6	79.4	0.91
	Self-Attention	83.8	84.4	83.8	83.7	0.908	82.1	75.9	79.7	77.2	0.873
word-sentence-document feature	Li's SWNN	80.7	81.1	80.7	80.7	0.872	81.2	82.8	78.6	79.3	0.881
	Ren's Neural	80.3	80.4	80.3	80.2	0.885	82.1	82.9	81.8	81.3	0.902
multi-feature	parallel-connected	82.3	83.3	82.3	82.1	0.896	80.6	80.6	79.2	79.4	0.895
	series-connected	80.3	81.1	80.3	80.1	0.884	82.4	83.6	81	81.1	0.841
	ST-MFLC	**85**	**85.3**	**85**	**85**	**0.921**	**85.7**	**86.2**	**84.3**	**84.7**	**0.912**

truth and deceptive restaurant/hotel datasets increases under the unbalanced dataset, and the difference in the hotel dataset is larger.

Figs. 5–7 show the kernel density estimates of words with different part-of-speech in the domain of doctors, restaurants, and hotels, respectively. For the doctor dataset, nouns and verbs have an approximate distribution, and the number of nouns and verbs accounts for the largest proportion in the review text. In balanced and unbalanced deceptive doctor datasets, the distribution of words with the same part-of-speech has little difference, but the difference between truth and deceptive doctor datasets is large. For the restaurant dataset, the difference between the three curves is very small. For the hotel dataset, nouns and adjectives, adverbs, and verbs have similar distributions. In the density plot of nouns and adjectives, there is an obvious distribution difference between truth data and balanced/unbalanced deceptive

Table 6
The experimental results on the Doctor dataset.

Model type	Model	Balanced Doctor^{y} dataset					Unbalanced Doctor^{n} dataset				
		A	P	R	F	AUC	A	P	R	F	AUC
single feature	TextCNN	84.7	84.9	84.7	84.7	0.941	84.8	85.4	82.8	83.6	0.921
	Bi-GRU	86.5	86.8	86.5	86.5	0.946	85.4	86.6	83.4	84.1	0.932
	Self-Attention	87	87.6	87	86.9	0.951	85.7	85.6	85	85.1	0.924
word-sentence-document feature	Li's SWNN	86.8	87.1	86.8	86.7	0.945	85.4	86.2	83.8	84.4	0.934
	Ren's Neural	86.3	86.5	86.3	86.2	0.951	85.1	85.5	83.6	84.1	0.924
multi-feature	parallel-connected	88.8	89.1	88.8	88.7	0.956	86.3	87	85.6	85.6	0.931
	series-connected	87.8	88.3	87.8	87.7	0.947	86	86.5	84.4	85	0.919
	ST-MFLC	**90.3**	**90.3**	**90.3**	**90.2**	**0.961**	**87.5**	**88**	**86**	**86.6**	**0.944**

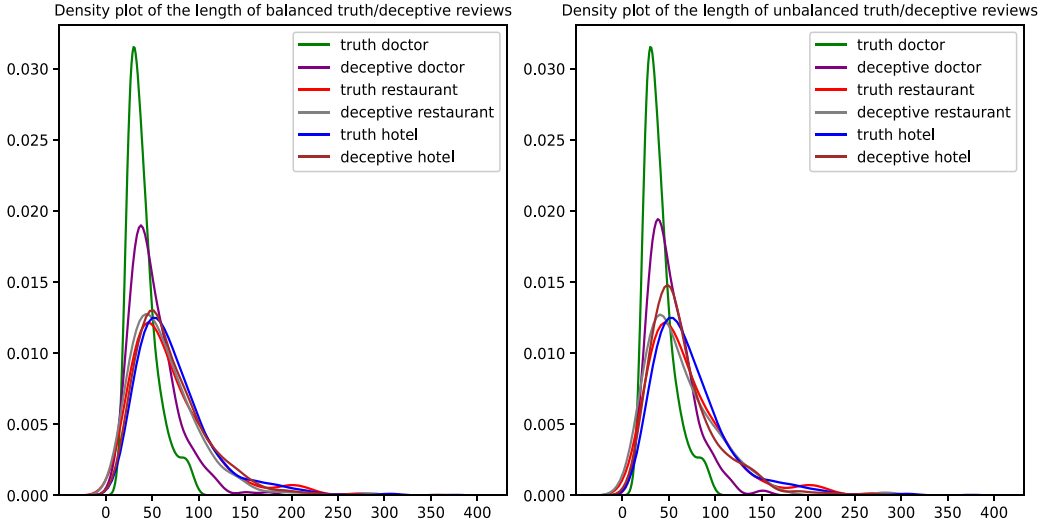

Fig. 4. The density plot of the length of truth/deceptive reviews.

datasets. However, in the density plot of adverbs and verbs, there is little difference between truth data and balanced deceptive data, but there is a big difference between the truth data and unbalanced deceptive data. Based on the above observation and analysis, we can draw the following conclusions.

Conclusion 2: From the perspective of text length, compared with restaurant and hotel domains, the difference between the truth and deceptive data in the doctor domain is larger, which may be one reason why the model has the best results on the Doctor domain datasets.

Conclusion 3: From the perspective of part-of-speech, comparing the part-of-speech distribution in the doctor, hotel, and restaurant domains, the difference between truth and deceptive datasets is the largest in the doctor domain, and the smallest in the restaurant domain. The differences in the part-of-speech of the dataset in different domains may make the model have the best learning effect in the doctor domain and the worst learning effect in the restaurant domain.

4.4. Mixed-domain results and analysis

To further verify the performance of our model, the second set of experiments compares our model with baselines on the mixed-domain datasets. The data in the real world is complex and diverse, so it is more valuable to explore the performance of models in a mixed domain.

Table 7 shows the experimental results of various models on the Mixed datasets. For the Mixed^{1} dataset, the Bi-GRU and Self-Attention models have similar performance, but both are worse than the TextCNN model. The multi-feature parallel/series-connected model is worse than TextCNN, Self-Attention, and Bi-GRU models based on the single feature. Li's SWNN is better than the multi-feature series-connected model. The performance of Ren's Neural model is not very good. Our ST-MFLC model performs best (84.2, 84.1, 83.3, 83.6, 0.924).

For the Mixed^{2} dataset, the Self-Attention model has the best overall performance in the baselines. Li's SWNN and Ren's Neural models are better than the Bi-GRU model. The multi-feature series-connected model performs poorly, while the performance of the multi-feature parallel-connected model is also limited, which shows that the series/parallel -connected strategy is not suitable for mixed data. The ST-MFLC model in this paper performs best (83.2, 82.8, 82.2, 82.3, 0.906), which indicates that our strategy can make better use of different features than other methods. Based on the above mixed-domain experimental analysis, we can get the following conclusion.

Conclusion 4: The multi-feature parallel/series-connected fusion model is not quite suitable for processing mixed-domain datasets. However, the separated training of multi-feature learning and classification method performs best on both the Mixed^{1} dataset and the Mixed^{2} dataset, which shows that the method proposed in this paper has more advantages and is more suitable than other models to deal with mixed-domain datasets.

4.5. In-domain results and analysis on the large-scale dataset

In the third set of experiments, we explore the performance of each model under large-scale data, and use five datasets with different unbalanced proportions (i.e., 50%, 40%, 30%, 20%, 12.4%) for comprehensive analysis. Tables 8–10 show the experimental results on the large-scale Yelp datasets.

For the balanced Yelp^{1} dataset, the TextCNN model has the best results in the baselines (83.3, 83.3, 83.3, 83.2, 0.911), while the Self-

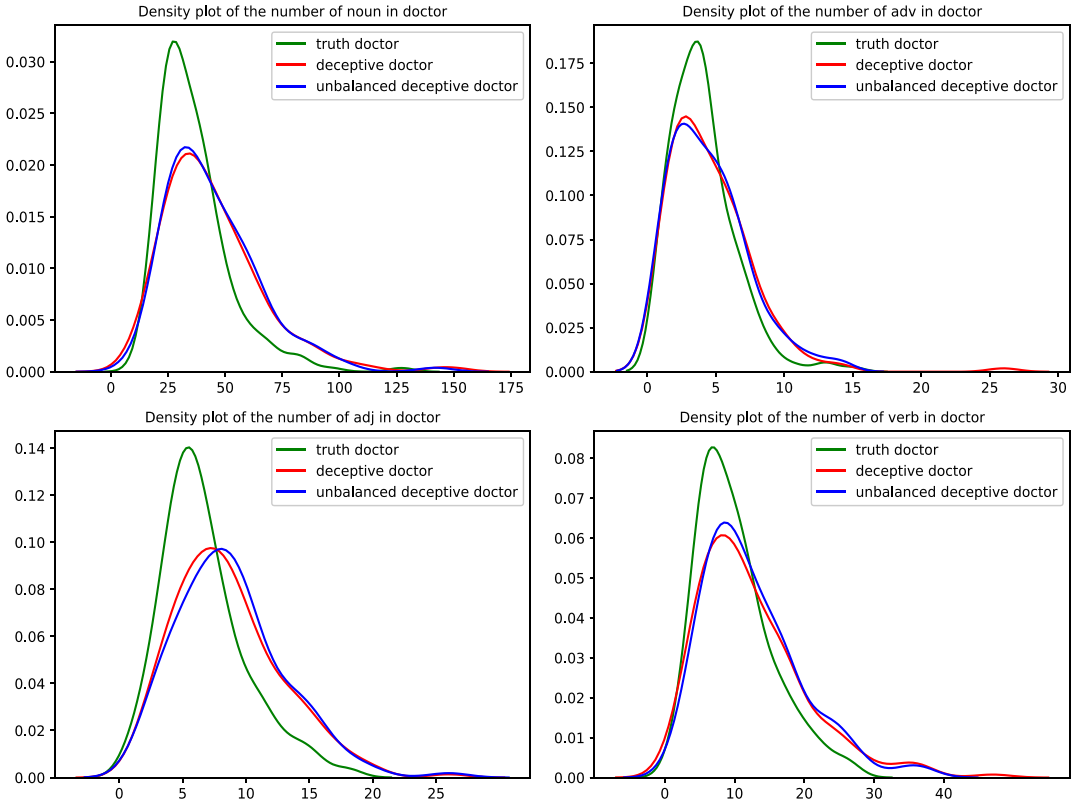

Fig. 5. The density plot of the number of words with different parts of speech in doctor reviews.

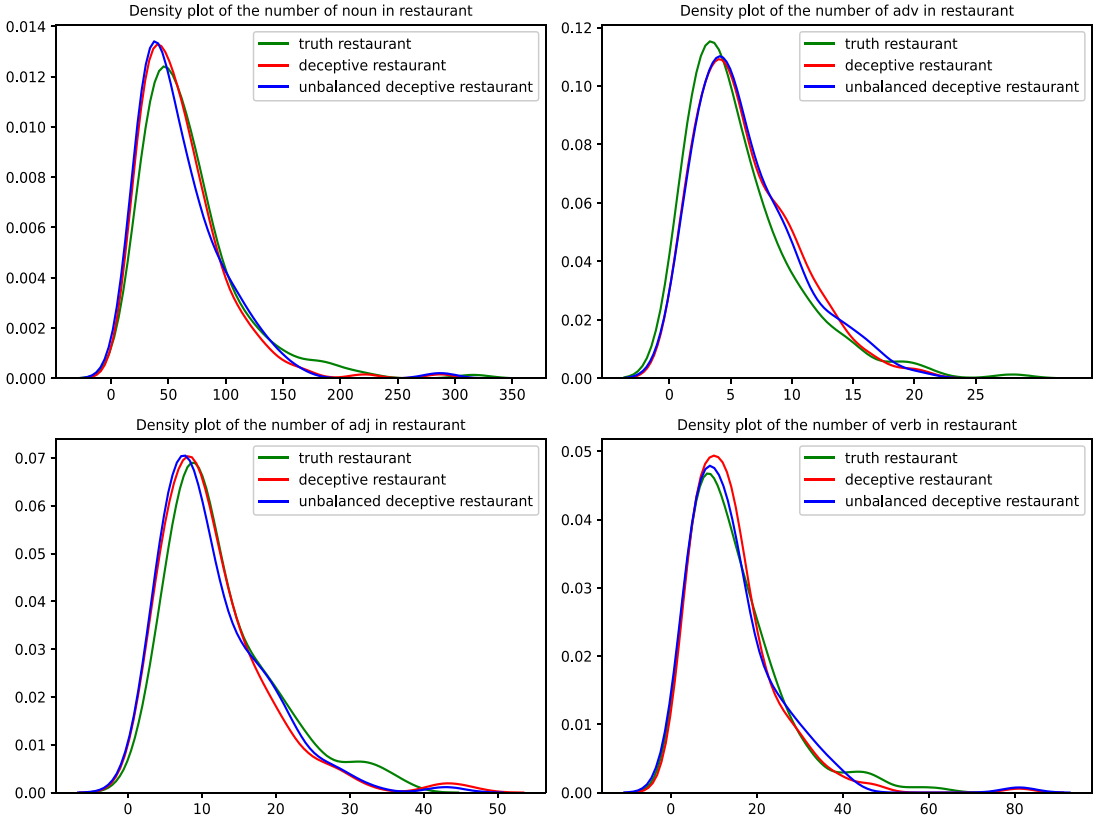

Fig. 6. The density plot of the number of words with different parts of speech in restaurant reviews.

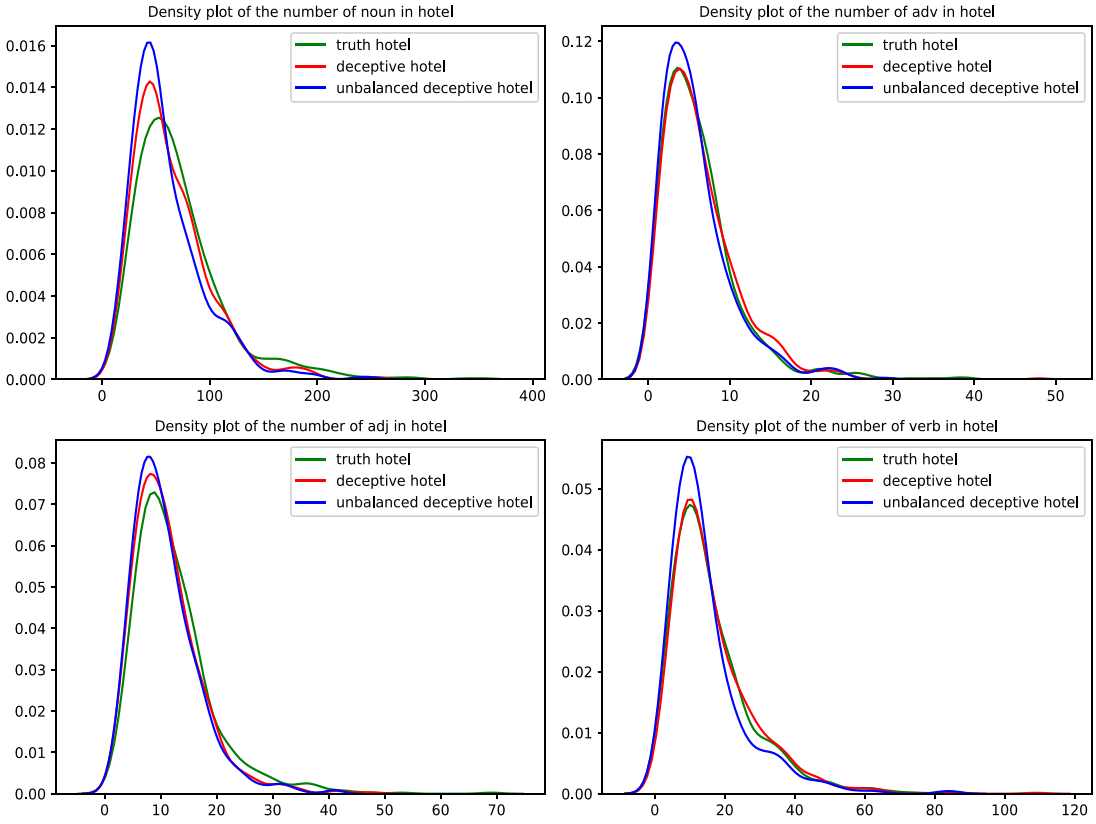

Fig. 7. The density plot of the number of words with different parts of speech in hotel reviews.

Table 7
The experimental results on the Mixed dataset.

Model type	Model	Mixed[1] dataset					Mixed[2] dataset				
		A	P	R	F	AUC	A	P	R	F	AUC
single feature	TextCNN	83.7	83.6	82.9	83.2	0.914	81.4	81	80	80.3	0.881
	Bi-GRU	82.9	82.6	82.2	82.4	0.907	78	77.4	76.8	76.8	0.856
	Self-Attention	83	82.8	82.3	82.5	0.911	81.7	81.2	80.7	80.8	0.895
word-sentence-document feature	Li's SWNN	82.3	81.9	82	81.9	0.903	79	78.6	77.5	77.7	0.877
	Ren's Neural	80.4	80.1	80	80	0.881	78.4	78.1	76.6	77	0.866
multi-feature	parallel-connected	82.6	82.5	81.8	82	0.902	80.7	80.9	78.7	79.2	0.879
	series-connected	81.9	81.6	81.2	81.4	0.887	78.7	78.4	77	77.3	0.852
	ST-MFLC	**84.2**	**84.1**	**83.3**	**83.6**	**0.924**	**83.2**	**82.8**	**82.2**	**82.3**	**0.906**

Table 8
The experimental results on the Yelp[1] and Yelp[2] datasets.

Model type	Model	Yelp[1] dataset					Yelp[2] dataset				
		A	P	R	F	AUC	A	P	R	F	AUC
single feature	TextCNN	83.3	83.3	83.3	83.2	0.911	79.4	78.7	**78.2**	**78.4**	0.863
	Bi-GRU	81.1	81.1	81.1	81.1	0.892	76.4	75.6	75.2	75.3	0.835
	Self-Attention	79.1	79.2	79.1	79	0.87	76	75.1	74.6	74.8	0.816
word-sentence-document feature	Li's SWNN	82.1	82.4	82.1	82	0.905	77.8	77.3	76.8	76.8	0.853
	Ren's Neural	80.1	80.1	80.1	80.1	0.881	76.2	75.3	74.6	74.9	0.829
multi-feature	parallel-connected	82.7	82.8	82.7	82.7	0.908	77.9	77	76.7	76.8	0.854
	series-connected	79.9	80	79.9	79.9	0.876	76.3	75.4	75.5	75.4	0.824
	ST-MFLC	**83.8**	**83.8**	**83.8**	**83.8**	**0.914**	**79.6**	**79.1**	78	**78.4**	**0.864**

Attention model has the worst performance. Li's SWNN and Ren's Neural models are better than the multi-feature series-connected model, but worse than parallel-connected model. Our ST-MFLC model achieves the best results, which has demonstrated the merits of our model on large-scale data.

For the Yelp[2] dataset with an unbalanced proportion of 40%, the TextCNN model obtains the best value on recall and F1 (78.2, 78.4), it also has a high AUC (0.863). This result proves the importance of local semantic features. Li's SWNN and parallel-connected models have similar performance, and they have a good AUC. Compared with the

Table 9
The experimental results on the Yelp[3] and Yelp[4] datasets.

Model type	Model	Yelp[3] dataset					Yelp[4] dataset				
		A	P	R	F	AUC	A	P	R	F	AUC
single feature	TextCNN	79.2	75.4	**73.5**	74.3	0.823	**80.7**	**68.8**	**63.6**	**65.3**	**0.746**
	Bi-GRU	75.7	70.9	70.4	70.6	0.789	77.3	63.4	61.7	62.3	0.72
	Self-Attention	76.8	72.4	70.5	71.2	0.794	78.1	64.4	62.1	62.9	0.727
word-sentence-document feature	Li's SWNN	77.3	73.5	72	72.3	0.81	79.2	67	62.8	63.6	0.729
	Ren's Neural	76	71.5	70.4	70.8	0.788	77.6	63.7	61.6	62.3	0.715
multi-feature	parallel-connected	77.6	73.4	72.3	72.8	0.817	80	67.5	62.1	63.5	0.742
	series-connected	75.9	71.4	71.6	71.4	0.787	77.2	63.3	61.2	61.7	0.709
	ST-MFLC	**79.6**	**76.2**	73.4	**74.5**	**0.827**	80.3	68.2	63.5	65	**0.746**

Table 10
The experimental results on the Yelp[5] dataset.

Model type	Model	Yelp[5] dataset				
		A	P	R	F	AUC
single feature	TextCNN	84.8	58.8	54.5	55	0.665
	Bi-GRU	82.9	55.8	54.2	54.7	0.656
	Self-Attention	84.1	55.3	53	53.3	0.65
word-sentence-document feature	Li's SWNN	84.2	57.1	54.3	54.8	0.649
	Ren's Neural	82.6	55.2	53.8	54.2	0.644
multi-feature	parallel-connected	82.9	57.1	**55.4**	**55.8**	0.666
	series-connected	83.7	55.6	53.3	53.6	0.651
	ST-MFLC	**85.4**	**58.9**	54.1	54.8	**0.681**

TextCNN and our ST-MFLC models, the overall performance of other methods is poor. And our ST-MFLC model has the best overall performance (79.6, 79.1, 78, 78.4, 0.864). This shows that our model can effectively learn different features to improve the effect of deceptive review detection.

For the Yelp[3] dataset, its imbalance proportion is 30%. The TextCNN model has better performance than other baselines, and it has the best recall (73.5) in all models. Li's SWNN and Ren's Neural model perform better than the Bi-GRU model, and Li's SWNN is also better than the Self-Attention model. The effect of Bi-GRU model is not very good, probably because the simple temporal semantic features cannot help identify deceptive reviews. Our ST-MFLC achieves the best overall performance (79.6, 76.2, 73.4, 74.5, 0.827).

For the Yelp[4] dataset with an unbalanced proportion of 20%, the TextCNN has the best results among all methods (80.7, 68.8, 63.6, 65.3, 0.746). The series-connected model has the worst performance, due to the weakening of features as they are passed from layer to layer, whereas he parallel-connected model has better results than the series-connected model. Our ST-MFLC model achieves only slighter weaker performance as the TextCNN model on AUC (0.746). However, our ST-MFLC model still outperforms other baselines except for TextCNN.

For the most unbalanced Yelp[5] dataset, the unbalanced proportion is 12.4%. The multi-feature parallel-connected model obtains the best performance on recall and F1 (55.4, 55.8), and it has the best AUC (0.666) in baselines. The TextCNN model also has a good performance. Our ST-MFLC achieves the best results in accuracy, precision, and AUC (85.4, 58.9, 0.681). This shows that our strategy of using different semantic features can help improve the detection of deceptive reviews on large-scale unbalanced datasets. Based on the above mixed-domain experimental analysis, we can get the following conclusion.

Conclusion 5: Compared with baselines on large-scale data, our separated training of multi-feature learning and classification method has better performance in most cases of different unbalanced proportions. This shows that our method can effectively use different features to deal with large-scale data and unbalanced data.

Based on the experimental results in Sections 4.3, 4.4, and 4.5, we can get the following conclusion.

Conclusion 6: Our model has good stability in different situations (e.g., different domains, different scales, different unbalanced proportions) than other models.

5. Conclusion

This paper proposes a new deceptive reviews detection model based on separated training of multi-feature learning and classification, which makes full use of local semantic features, temporal semantic features, and weighted semantic features. To verify the performance of our method, we select TextCNN, Bi-GRU, Self-Attention, Li's SWNN, and Ren's Neural models as baselines. Besides, to further verify the performance of our method, we construct the multi-feature parallel/series-connected fusion model as a baseline. We construct three sets of experiments on the small public gold standard dataset and the large-scale Yelp dataset. In in-domain experiments based on the small dataset, the results show that our ST-MFLC model is superior to the baselines on both balanced and unbalanced datasets. In mixed-domain experiments, we construct two mixed datasets, in which the proportion of deceptive reviews is 57.7 and 40, respectively. The results show that the method proposed in this paper has a higher performance than baselines. In in-domain experiments based on the large-scale Yelp dataset, we construct five datasets, and the deceptive reviews account for 50%, 40%, 30%, 20%, and 12.4%, respectively. Comprehensive experiments show that our method can better detect deceptive reviews in different situations than baselines.

Though our model has a good performance and outperforms baselines, it can be further improved. First, the emotional information in the reviews helps identify deceptive reviews, and deceptive negative reviews are more difficult to detect than positive reviews (Fusilier, Montes-y-Gómez, Rosso, & Cabrera, 2015). Secondly, we use three submodels to capture different features, but we have not considered the relationship and weights of different features. Thirdly, we plan to explore the influence of different parameters on our model. Finally, deceptive reviews may contain image information, and some researchers have studied the detection of multi-modal fake information (Yang, Liu, Zhou, & Luo, 2019; Wang, Ma, Jin, Yuan, Xun, Jha, Su, & Gao, 2018; Yang, Zheng, Zhang, Cui, Li, & Yu, 2018). We are planning to use multi-modal information to detect deceptive reviews.

CRediT authorship contribution statement

Ning Cao: Software, Writing – original draft. **Shujuan Ji:** Conceptualization, Methodology. **Dickson K.W. Chiu:** Writing – review & editing. **Maoguo Gong:** Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is supported in part by the Natural Science Foundation of China (No. 71772107), Qingdao social science planning project (No. QDSKL1801138), the training program of the major research plan of the National natural science foundation of China (No. 91746104), the National Key R&D Plan of China (Nos. 2017YFC0804406, 2018YFC1406203, 2018YFC0831002), Humanity and Social Science Fund of the Ministry of Education (No. 18YJAZH136), the Key R&D Plan of Shandong Province (No.2018GGX101045), the Natural Science Foundation of Shandong Province (Nos. ZR2018BF013), Shandong Education Quality Improvement Plan for Postgraduate, the SDUST Mountain and sea Talents Project and the SDUST Research Fund.

References

Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy, 1*(1), e9. https://doi.org/10.1002/spy2.2018.1.issue-110.1002/spy2.9

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Cao, N., Ji, S., Chiu, D. K. W., He, M., & Sun, X. (2020). A Deceptive Review Detection Framework: Combination of Coarse and Fine-grained Features. *Expert Systems with Applications, 156*, 113465. https://doi.org/10.1016/j.eswa.2020.113465

Chen, Y. R., & Chen, H. H. (2015). May). Opinion spam detection in web forum: A real case study. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 173–183).

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2016). Attention-over-attention neural networks for reading comprehension. arXiv preprint arXiv:1607.04423.

Dong, L. Y., Ji, S. J., Zhang, C. J., Zhang, Q., Chiu, D. W., Qiu, L. Q., & Li, D. (2018). An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews. *Expert Systems with Applications, 114*, 210–223.

Du, X., Zhu, R., Zhao, F., Zhao, F., Han, P., & Zhu, Z. (2020). A deceptive detection model based on topic, sentiment, and sentence structure information. *Applied Intelligence, 50*(11), 3868–3881.

Fan, Y., Lu, X., Li, D., & Liu, Y. (2016). October). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 445–450).

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013, June). Exploiting burstiness in reviews for review spammer detection. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1).

Feng, S., Banerjee, R., & Choi, Y. (2012). July). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 171–175).

Feng, V. W., & Hirst, G. (2013). October). Detecting deceptive opinions with profile compatibility. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 338–346).

Hernández Fusilier, D., Montes-y-Gómez, M., Rosso, P., & Guzmán Cabrera, R. (2015). Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management, 51*(4), 433–443.

Gao, S., Ramanathan, A., & Tourassi, G. (2018). July). Hierarchical convolutional attention networks for text classification. In *Proceedings of The Third Workshop on Representation Learning for NLP* (pp. 11–23).

Guo, L., Zhang, D., Wang, L., Wang, H., & Cui, B. (2018, October). CRAN: a hybrid CNN-RNN attention-based model for text classification. In International Conference on Conceptual Modeling (pp. 571-585). Springer, Cham.

Hu, D. (2019, September). An introductory survey on attention mechanisms in NLP problems. In Proceedings of SAI Intelligent Systems Conference (pp. 432-448). Springer, Cham.

Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short term memory neural network. *Annals of Mathematics and Artificial Intelligence, 85*(1), 21–44.

Jindal, N., & Liu, B. (2008). February). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 219–230).

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. arXiv preprint arXiv:1804.08559.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, February). Recurrent convolutional neural networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

Le, Q., & Mikolov, T. (2014). January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196).

Letarte, G., Paradis, F., Giguère, P., & Laviolette, F. (2018). November). Importance of self-attention for sentiment analysis. In *In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 267–275).

Li, F. H., Huang, M., Yang, Y., & Zhu, X. (2011, June). Learning to identify review spam. In Twenty-second international joint conference on artificial intelligence.

Li, F., Huang, M., & Zhu, X. (2010, July). Sentiment analysis with global topics and local dependency. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 24, No. 1).

Li, J., Cardie, C., & Li, S. (2013, August). Topicspam: a topic-model based approach for spam detection. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 217-221).

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014, June). Towards a general rule for identifying deceptive opinion spam. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1566-1576).

Li, L., Qin, B., Ren, W., & Liu, T. (2017). Document representation and feature combination for deceptive spam review detection. *Neurocomputing, 254*, 33–41.

Li, X., Ouyang, J., Zhou, X., Lu, Y., & Liu, Y. (2015). Supervised labeled latent Dirichlet allocation for document categorization. *Applied Intelligence, 42*(3), 581–593.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks.

Madisetty, S., & Desarkar, M. S. (2018). A neural network-based ensemble approach for spam detection in Twitter. *IEEE Transactions on Computational Social Systems, 5*(4), 973–984.

Martinez-Torres, M. R., & Toral, S. L. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Management, 75*, 393–403.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., & Ghosh, R. (2013). August). Spotting opinion spammers using behavioral footprints. In *In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 632–640).

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013, June). What yelp fake review filter might be doing?. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1).

Nguyen, H.-T., & Nguyen, L.-M. (2020). ILWAANet: An interactive lexicon-aware word-aspect attention network for aspect-level sentiment classification on social networking. *Expert Systems with Applications, 146*, 113065. https://doi.org/10.1016/j.eswa.2019.113065

Ott, M., Cardie, C., & Hancock, J. (2012). April). Estimating the prevalence of deception in online review communities. In *In Proceedings of the 21st international conference on World Wide Web* (pp. 201–210).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557.

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences, 385-386*, 213–224.

Ren, Y., & Ji, D. (2019). Learning to detect deceptive opinion spam: A survey. *IEEE Access, 7*, 42934–42945.

Ren, Y., Ji, D., & Zhang, H. (2014). October). Positive unlabeled learning for deceptive reviews detection. In *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 488–498).

Sadman, N., Gupta, K. D., Haque, A., Poudyal, S., & Sen, S. (2020). February). Detect Review Manipulation by Leveraging Reviewer Historical Stylometrics in Amazon, Yelp, Facebook and Google Reviews. In *In Proceedings of the 2020 The 6th International Conference on E-Business and Applications* (pp. 42–47).

Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2015). Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications, 42*(22), 8650–8657.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Wang, S., Zhang, J., & Zong, C. (2016). Learning sentence representation with guidance of human attention. arXiv preprint arXiv:1609.09189.

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.

Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2017, February). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

Wang, X., Liu, K., & Zhao, J. (2017). In *November). Detecting deceptive review spam via attention-based neural networks* (pp. 866–876). Cham: Springer.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... Gao, J. (2018). July). Eann: Event adversarial neural networks for multi-modal fake news detection. In *In Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857).

Wang, Y., Wang, L., Yang, Y., & Lian, T. (2021). SemSeq4FD: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection. *Expert Systems with Applications, 166*, Article 114090.

Yang, H., Liu, Q., Zhou, S., & Luo, Y. (2019). A spam filtering method based on multi-modal fusion. *Applied Sciences, 9*(6), 1152.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P. S. (2018). TI-CNN: Convolutional neural networks for fake news detection. arXiv preprint arXiv:1806.00749.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). June). Hierarchical attention networks for document classification. In *In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1480–1489).

Yoo, K. H., & Gretzel, U. (2009, January). Comparison of deceptive and truthful travel reviews. In ENTER (pp. 37-47).

Zhang, W., Du, Y., Yoshida, T., & Wang, Q. (2018). DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management, 54*(4), 576–592.

Zhao, S., Xu, Z., Liu, L., Guo, M., & Yun, J. (2018). Towards accurate deceptive opinions detection based on word order-preserving CNN. Mathematical Problems in Engineering, 2018.

Zheng, J., Cai, F., Shao, T., & Chen, H. (2018). Self-interaction attention mechanism-based text representation for document classification. *Applied Sciences, 8*(4), 613.

Zhu, Y., Li, L., & Luo, L. (2013, August). Learning to classify short text with topic model and external knowledge. In International Conference on Knowledge Science, Engineering and Management (pp. 493-503). Springer, Berlin, Heidelberg.