Cross-domain Fake Review Detection Based on Deep Learning Multi-Level Generic Features Extraction Fusion

Ge Ren, Huajun Wang, Yong Yang*
College of Computer Science and Technology, Xinjiang Normal University, Xinjiang, Urumqi, 830054, China.
E-mail: iphone202409@163.com
*Corresponding author

Keywords: fake review detection, deep learning, generic feature extraction, cross-domain detection, pre-trained language model

Received: September 4, 2024

The fake review detection aims to identify fake reviews that affect regular competition of online marketplaces. Existing research on fake review detection mainly focuses on deep learning and feature-based methods. Feature-based methods make it difficult to obtain potential semantic information, while deep learning methods are less likely to consider multi-granularity information based on text structure. Neither is very satisfactory when it comes to cross-domain detection. In this paper, we present a cross-domain fake review detection method based on multi-level generic feature extraction fusion. The important information prevalent in different domains is purposefully extracted from the review text based on its structure for cross-domain fake review detection. At the word level, in order to obtain potential semantic information, the paper combined GloVe and TF-IDF weights as well as CNN to extract multi-granularity semantic information from the word level. In order to obtain sentence-level generic features, sentence-level syntactic information of comments is extracted through lexical annotation techniques. In order to obtain finer-grained emotion features at the document level, the paper annotates the dataset comments with distilBERT, a pre-trained language model that has been fine-tuned on the emotion classification task. Moreover, the generic features extracted at each layer that are applicable across domains are fused using a multi-head attention mechanism. Finally, classification is performed in the classification layer. Experimental results on public datasets show that the model proposed in this paper has significantly improved performance in cross-domain detection, achieving 83% and 78.0% accuracy on the restaurant and doctor domain datasets, respectively. It outperforms the state-of-the-art method by 10.7% on the restaurant dataset.

Povzetek: Prispevek uvaja večnivojski model generičnega izločanja značilk za zaznavanje lažnih ocen med domenami, ki z združitvijo semantičnih, sintaktičnih in čustvenih značilk dosega vrhunsko točnost.

1 Introduction

With the increasing number of user-generated reviews, deceptive reviews are more of a concern than ever [1]. Because of the development of web technology and the popularity of e-commerce, more and more people are utilizing e-commerce to obtain products and services. However, with millions of products and services available on e-commerce websites and many companies adopting the online marketing model, it is difficult for people to find the most suitable products according to their needs. Product reviews comprise of detailed experience of the customer(s) with the product(s). They help the consumers in their purchase decision. 80% of customers value online reviews as much as personal recommendations. In other words, the customer's decision to buy the product depends on the reviews they see on the e-commerce platform or social media at the time. Online reviews are becoming increasingly important, while fake reviews are being sought after by more merchants.

Fake reviews are increasingly prevalent on various websites due to the existence of various social media platforms where anyone can freely criticize or comment on any company or product at any time without any obligation or restriction. For businessmen, positive evaluations can lead to great profits, while negative evaluations often result in adverse financial impacts [3]. As a result, certain companies may hire individuals or teams to publish false positive evaluations of their products. Similarly, to suppress competitors, certain companies may ask hired individuals to publish false negative evaluations of competitors' products. It is considered that reviews posted by individuals who have not physically interacted with the product under review are not authentic [4]. These fake reviews have caused great distress to consumers, damaged the legitimate rights and interests of other businesses, and undermined the order of fair competition in the online market.

A lot of work has been done on spam detection, such as email spam [5], SMS spam, web spam [6], social media spam [7], search engine spam [8], and video spam [9]. However, fake reviews are slightly different and are basically found on product review websites. Their purpose is to either give a positive review of a particular product

for profit and promotion or an unjustified negative review to disparage a competing brand or product. The problem of spammy comments was first raised in product reviews by Jindal and Liu in 2007 and is thus considered the first literature study in this area [10]. Over time, the number of spam comments increases, from individual spam comment senders to group spam comment senders. If not detected or removed in a timely manner, they will largely damage e-commerce businesses and social media, which is considered a trusted source of public opinion and may lose its luster. However, over the past few years, both the industry and the research community have made significant contributions to the fight against fake reviews. Researchers have obtained strong spoofing detection performance in some cases. However, these studies have mainly focused on single domains, and the datasets used are usually small. There is a lack of information on how small-scale, single-domain detection models can be generalized to new domains or real-world data. This work aims to fill this gap. The main contributions of this paper include:

(1) In this paper, we propose a novel multi-level generic features extraction fusion method to improve the cross-domain detection of fake reviews.

(2) We fine-tune the pre-trained language model distilBERT to extract the fine-grained emotion features of fake reviews.

2 Related work

Fake reviews are also commonly referred to as spam reviews, deceptive opinions, and spam opinions. The writer of the reviews is known as a spammer. Whether it is spam comment detection, deceptive comment detection, or spam comment detection, the primary challenge in detecting fake reviews is the classification of reviews as either "true" or "fake". Feature-based and deep-learning methods play an important role in the detection of fake reviews [11]. The research on cross-domain detection is also increasing. This section will review relevant studies from these three perspectives.

2.1 Feature-based methods

The detection of fake reviews represents a significant challenge in the field of natural language processing. In 2008, Jindal and Liu first proposed to train a model using features based on the reviewer, the content of the reviews, and the product itself [12]. Subsequently, researchers have studied fake review detection mainly in terms of behavioral and textual features. Behavioral features indicate the statistical significance of users' comments and behaviors based on their past and current reviews, such as positive rating, average review length, weighted rating bias, proportion of negative reviews, maximum content similarity, duplicate reviews, extreme rating behaviors, and first review rate [13]. Text features include semantic, syntactic, lexical, and metadata features of reviews that help to identify fake reviews. Between fake reviews and real reviews, these features have been demonstrated to differ in many studies. However, these methods seldom consider whether differences in features exist in other domains.

Liu et al. presented a hierarchical attention network that purposefully uses different attention mechanisms at two layers (word-to-sentence layer and sentence-to-text layer) to capture multi-granular, important, and comprehensive semantic information [14]. Zhang presented a fake review detection model that fuses text features, commenter behavioral features, and temporal features [15]. The model utilizes BERT and Bi-LSTM, excellent performers in deep learning models. Features of different dimensions are extracted for deep fusion in the model-building stage. de Arriba-Pérez et al. presented an online solution for identifying and explaining spam reviews. The proposed method explores online reviews for stream-based spam classification with drift detection. In addition, it explores self-explainable Machine Learning models for transparency [16]. Liu and Pang presented an unsupervised unified framework to address the detection of fake reviews senders by computing review bias [17]. Wang et al. presented a fake reviews recognition algorithm F-Text GCN combining Gaussian Mixture Model and Text GCN, which performs multi-labeled node composition by combining lexical and non-textual features in the review text to extract the structural and terminological differences between normal and fake reviews [18]. Wang et al. presented an attention framework embedding each label in the same space as the word vector was introduced for measuring the compatibility of embedding between the labels and the text sequences [19]. The proposed method preserves the explanatory power of word embeddings while simultaneously capitalizing on alternative information sources, apart from input text sequences. Melleng et al. found that a combination of sentiment and emotion works better than either one alone by analyzing the effectiveness of sentiment and emotion representations based on different text capture rate estimates in a fake reviews classification task [20]. Zhang proposed an ensemble-based approach for spam detection in digital communication [21]. The approach combines the advantages of the ensemble approach with the semantic understanding provided by Word2Vec's word embeddings, aiming to enhance the representation of textual data in natural language processing efforts. Xue et al. present a method for detecting fake reviews based on opinion bias [22]. The overall bias is calculated by iterating user opinion deviation from the majority opinion impact through a three-level trust propagation framework. From this, researchers determine the reliability of users, reviews, and products.

Table 1: Feature-based methods.

References	Methods	Dataset Domain	Accuracy	F1-score	cross domains
Liu et al. [14]	hierarchical attention model	Hotel, Restaurant, Doctor	91.0	92.8	√
Zhang et al. [15]	Multi-Feature Fusion Model	Yelp Chicago Restaurant Dataset	-	0.953	×
de Arriba-Pérez et al. [16]	Online Detection Spam Reviews with Data Drift Adaptation	Yelp dataset (MediaWiki dataset)	78.75 (86.13)	78.44 (85.89)	×
Liu et al. [17]	a novel review deviation model	Amazon review datasets	78.62	79.13	×
Wang et al. [18]	Fake-review Text GCN	China Ecommerce Platform Mobile Review	-	82. 92	×
Wang et al. [19]	Label Embedding Attentive Model	The DBPedia and four other datasets	99.02	-	×
Melleng et al. [20]	Approach that combines sentiment and emotion	Ott dataset, Yelp Zip and Yelp NYC	-	65.3	×
Zhang et al. [21]	an ensemble-based approach	Collected from various online platforms,	95.1	-	×
Xue et al. [22]	Approach based on aspect-specific content-aware trust propagation	SemEval-2014 dataset, Yelp Yelp dataset	79.6	0.79	×

Table 1 aggregates the above studies of feature-based methods. The accuracy and F1 scores are the best performance obtained by each method tested on a single data. It is easy to see that the performance of many methods on a single dataset has reached a high level. At the same time, it can be easily seen that the number of methods with cross-domain detection capability in feature-based method research is very small. From the different datasets in the table, it is also not hard to think of the complexity of the application domain of fake review detection. Therefore, a strong domain-adaptive method for fake review detection is of great significance.

2.2 Deep learning methods

Deep learning methods have been increasingly utilized in the detection of fake reviews, with studies showing that neural networks can outperform traditional methods [23]. Deep learning methods are able to extract useful data features more quickly than traditional machine learning. Deep learning can also capture semantic information of text using word embedding methods. Fang et al. proposed a dynamic knowledge graph-based fake review detection method by considering the correlation between the semantics of the review text and the time, as well as the effect of multi-source information on detecting fake reviews [24]. Firstly, according to the features of product reviews online, use the model to extract four types of entities: products, reviews, reviewers, and stores. Then, features related to time series are incorporated to create a dynamic graph network during the construction of the knowledge graph.

You et al. presented an attribute-enhanced domain adaptive embedding model that captures domain relevance using the attributes of the reviewer, the item, and the review [25]. Bathla et al. optimized the detection of spam comments by introducing aspect extraction and replication [26]. They use only aspects extracted from the comments and their respective sentiment to detect spam comments. The findings of the experimental analysis indicate that the presented method demonstrates superior performance compared to recent methods. Cai et al. presented a co-attention model, which fuses multiple features to classify comments and uses domain adversarial to train the model for improving the robustness of the model [27]. Li et al. presented a hard attention neural network model by incorporating weights of sentences into the composition process of the document representation [28]. Cheng et al. designed a novel framework based on graph neural networks for detecting spammers by obtaining information from different combinations of social networks in different subgraphs [29]. While these methods have achieved strong deception detection performance in some cases, these studies have focused on a single domain and typically used small datasets. Current research lacks information on how small-scale, single-domain fake review detection models can be generalized to new domains and real-world data. This work aims to optimize the cross-domain detection capabilities of small-scale, single-domain fake review detection models.

2.3 Cross-domain detection

In recent years, cross-domain detection has become increasingly important in the direction of fake review detection. In 2014, Li et al. attempted to capture the general differences between fake and true reviews from language usage [30]. Three linguistic feature models were

used including LIWC, POS, and unigram. The final results were not satisfactory, with the highest accuracy of 78.5% for the cross-domain component, much lower than the prediction results within the same domain. Liu et al. presented a neural network method with bidirectional long short-term memory (BiLSTM) and feature combination to learn the representation of deceptive reviews [1]. The method improves the F1 value to 87.6% on the mixed-domain dataset and also performs more robustly than all baseline methods on the cross-domain dataset. Wei et al. proposed a cross-domain detection model based on Stimuli Organism Response (S-O-R) combining LIWC (Linguistic Inquiry and Word Count) with the addition of word2vec quantitative features to overcome the decrease in accuracy [31]. Table 2 summarises the fake review detection methods across domains, comparing the accuracy of single-domain detection, cross-domain fake detection methods should still be improved.

In recent years, improving generalisation by extracting domain-invariant features has performed well in cross-domain text classification [32]. Ben-David et al. proposed PERL, a domain-adaptation model that fine-tunes a massively pre-trained deep contextualized embedding encoder (BERT) with a pivot-based Masked Language Modeling objective [33]. PERL outperforms strong baselines across 22 sentiment classification DA setups, improves in-domain model performance, increases its cross-configuration stability. Wu et al. proposed a novel Adversarial Soft Prompt Tuning method (AdSPT) to better model cross-domain sentiment analysis [34]. AdSPT uses a novel domain adversarial training strategy to learn domain-invariant representations between each source domain and the target domain. Experiments on a publicly available sentiment analysis dataset show that our model achieves the new state-of-the-art results for both single-source domain adaptation and multi-source domain adaptation. These methods take full advantage of the commonality between the source and target domains to enhance the generalisation of the model. By referring to these methods above, this paper proposes generic features applicable to fake review detection across domains. The generic features are the differences between genuine and fake reviews that are prevalent across domains. The generic features in this paper are semantic features, syntactic features and emotion features.

Table 2: Cross-domain fake review detection methods.

References	Methods	Source domain	Target domain (Accuracy)
Liu et al. [1]	A variant of Bi-LSTM	Hotel domain dataset	Restaurant domain dataset (81.3%); Doctor domain dataset (66.8%)
Liu et al. [14]	A hierarchical attention model	Hotel domain dataset	Restaurant domain dataset (77.5%); Doctor domain dataset (67.3%)
Li et al. [28]	A sentence-weighted neural network	Hotel domain dataset	Restaurant domain dataset (69.0%); Doctor domain dataset (61.0%)
Li et al. [30]	Sparse Additive Generative Model	Hotel domain dataset	Restaurant domain dataset (78.5%); Doctor domain dataset (64.7%)
Wei et al. [31]	SOR characteristic weight + Word vector	Hotel and Restaurant domain datasets	Doctor domain dataset (69.06%)

3 Method

Review texts have a hierarchical structure: words form sentences, sentences form documents, and the composition of both the words that form sentences and the sentences that form documents are similar [35]. In addition, some of the features have been shown to play important roles in many different fake reviews research tasks, and this paper argues that these features are generalizable in fake review detection tasks and play an important role in optimizing the cross-domain detection capability of the model. Based on these two points, this paper constructs a multilevel generic feature extraction fusion model (MFE) to detect fake reviews, as in Figure 1.

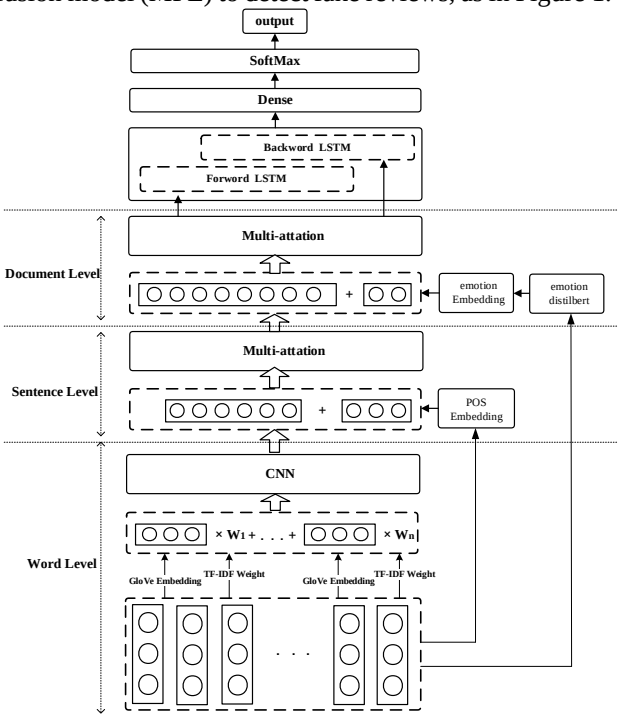

Figure 1: MFE model structure.

The model architecture of this paper consists of three levels, i.e., the word level (detailed in Section 2.1), which uses CNNs to extract multi-granularity (convolutional kernel sizes of 2, 3, and 4) information based on word embeddings and TF-IDF weights. The sentence level (detailed in Section 2.2) uses a combination of part-of-speech (POS) features and multi-head attention to fuse the extracted syntactic information with word-level features. The document-level (detailed in Section 2.3) uses a combination of pre-trained large language models and multi-head self-attention to fuse emotion features with sentence-level features to extract document-level features of reviews. Finally, the learned representation information is classified using Bi-LSTM and SoftMax classifiers to identify fake reviews.

3.1 Word-level

Global vectors for word representation (GloVe): Glove is a word continuous vector representation algorithm that converts words into meaningful vectors. It is worth noting that combining pre-trained word embeddings such as Word2Vec and GloVe in deep learning models may reduce accuracy [36]. Word embeddings may lead to problems such as high dimensionality, high sparsity, and ignoring textual sentiment information [37]. It is crucial to improve the accuracy of pre-trained word embeddings. Onan et al.'s experimental results show that the weighted word embedding scheme is an efficient text representation scheme that outperforms the traditional word embedding scheme [38]. It is argued that this scheme is equally useful for fake review detection. The GloVe used in this paper was pre-trained in Wikipedia 2014 and Gigaword5 to generate word vector matrices consisting of 300-dimensional vectors using 6 billion tokens and 400K vocab. If the input text review having n words is denoted as $S=\{t_1,t_2,...,t_n\}$, Each word is turned into a d (d=300 in this paper) dimensional word vector. Thus, the dimension space of each word is denoted as R^d, the $R^{n\times d}$ denotes the dimension space of the input text, and the word vector matrix generated after word embedding is denoted as

$$\boldsymbol{S}=\{\boldsymbol{t}_1,\boldsymbol{t}_2,...,\boldsymbol{t}_n\}\in R^{n\times d}.$$

Term frequency-inverse document frequency (TF-IDF): It has been demonstrated that the weighted average of word embeddings can enhance the performance of unsupervised natural language processing (NLP) tasks. In this paper, the TF-IDF is used to obtain the term weight information of the fake review texts in the training set, which is then used to reassign weights to the word embedding vectors in the sentences. TF-IDF is an unsupervised term weighting method that can be used for text mining and information retrieval. It evaluates the importance of a word by calculating the frequency of occurrence of the word in a document and the frequency of occurrence of the document containing the word in the corpus. The process is as follows Equation (1):

$$W_{d,t}=\mathrm{tf}\times\log\left(\frac{N}{\mathrm{df}+1}\right) \tag{1}$$

$W_{d,t}$ is the weight value of word t in document d, N is the total number of documents in the corpus, tf represents the frequency of occurrence of word t in document d, and represents df the number of documents in the corpus that contain word t.

In order to enhance the text representation, we assign TF-IDF weights to Glove word embeddings (as shown in Equation (2)):

$$\boldsymbol{V}_t=W_{d,t}\times\boldsymbol{t} \tag{2}$$

where $\boldsymbol{t}$ is the word vector matrix of word t obtained from GloVe. To obtain enriched information in the weighted word embedding layer, we use Text-CNN to obtain multi-granular features, which have three convolutional filters with widths of 2, 3, and 4, respectively. Aggregate them so that features of different widths can be extracted simultaneously in the next stage. Defined as the weighted word vector matrix is computed through a convolutional layer in order to capture both local and intrinsic features as follows Equation (3):

$$\boldsymbol{h}_i=f\left(\boldsymbol{V}_{i:i+t-1}\times\boldsymbol{W}_i+b_i\right) \tag{3}$$

where f is a nonlinear activation function, i represents the i-th word vector in V, t is the size of the convolution kernel, W is the weight matrix, and b is the bias. $\boldsymbol{h}\in R_{n-t+1}$ is the feature maps generated by the convolutional layer. Once the convolutional layer produces the feature maps, the maximum pooling layer minimizes the data dimensions and abstracts the important features. This is shown in the following Equation (4):

$$\boldsymbol{p}_i=Max[\boldsymbol{h}_i] \tag{4}$$

Where the $\boldsymbol{p}_i\in R_{(n-t+1)/2}$ is the feature map obtained after the maximum pooling layer. The feature mappings $\boldsymbol{p}_1$, $\boldsymbol{p}_2$ and $\boldsymbol{p}_3$ obtained from each of the three convolutional filters are spliced to finally obtain word-level feature information.

3.2 Sentence-level

Among the results of the fake reviews studies, fake reviews contain more adverbs (RB), verbs (V), pre-determiners (PDT), and pronouns (PRP); truth reviews contain more prepositions (In), determiners (DT), nouns (N), and adjectives (JJ) [30]. The lexical annotation feature is very helpful in most of the experiments for recognizing fake reviews; therefore, it has also been selected as one of the generic features in this paper. The lexical labeling task is to assign lexical labels from a given set of tags to each word in a given sentence. It is the process of classifying and labeling the words in a sentence, which is actually a multi-categorization task. Each word is

generally assigned a corresponding lexical property by lexical categorization based on its components in the syntactic structure or linguistic form. In this section, SpaCy's English model is used to calculate the number of Adjectives (ADJ), Nouns (NOUN), Pronouns (PRON), Verbs (VERB), and Adverbs (ADV) in each review. After fusing the lexical information of the review with the word-level features, the sentence-level features of the review are extracted by the multi-head attention mechanism.

Making associations by using a multi-attention mechanism can help the model to rationally allocate attention to each feature of the input and allow the model to adaptively learn the mapping relationship between the inputs and outputs, thus improving the model's performance ability. Its calculation equation is as follows Equations (5)-(7):

$$\boldsymbol{h}_i = Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \tag{5}$$

$$\boldsymbol{H} = concat(\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3)\boldsymbol{W}^O \tag{6}$$

$$\boldsymbol{S}^R = RELU(\boldsymbol{W} \cdot \boldsymbol{H} + b) \tag{7}$$

Where, $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are the attention input matrices, $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$ and $\boldsymbol{W}_i^V$ represent the weight matrix of the i-th header, $\boldsymbol{W}^O$ is the learnable weight matrix, *concat* represents the splice function, $\boldsymbol{W}$ is the parameter matrix, *RELU* is the activation function, b is the bias term, and $\boldsymbol{S}^R$ is the sentence-level feature representation of the review.

3.3 Document-level

Sentiment features are one of the features that all review texts possess. The experimental results of Melleng et al. demonstrate the effectiveness of emotion and sentiment-based representations for the fake review detection [20]. Therefore, we target seven emotions- fear, anger, joy, disgust, sadness, surprise, and neutrality- for review emotion detection [39]. Because of the lack of datasets of fake reviews with fine-grained emotion annotations, we use the "distilbert-base" model from the HuggingFace library. This model has been pre-trained and can be fine-tuned for emotion classification tasks. The model was then fine-tuned on a combined balanced dataset, with the output being one of seven emotions representing the predominant emotion of the review. The main part of the dataset is the HuggingFace Emotion Dataset [40], which consists of English Twitter messages. The rest of the dataset comes from the ISEAR (International Survey on Emotion Antecedents and Reactions) [41], Daily Dialogue [42], and Emotion Stimulus [43] datasets as a complement to the six basic and neutral emotions to form a balanced dataset. Each emotion in the final combined dataset was represented with approximately 2000 review samples. After fine-tuning the training, the "distilbert-base" model achieved 91.55% correctness on the test dataset of the combined dataset, with an F1 score of 90.67%. Subsequently, the model was used to classify the emotion of the fake reviews dataset to be used subsequently, and the classification results were used as the emotion labels of each review. The fine-tuned pre-trained model flow is shown in Figure 2.

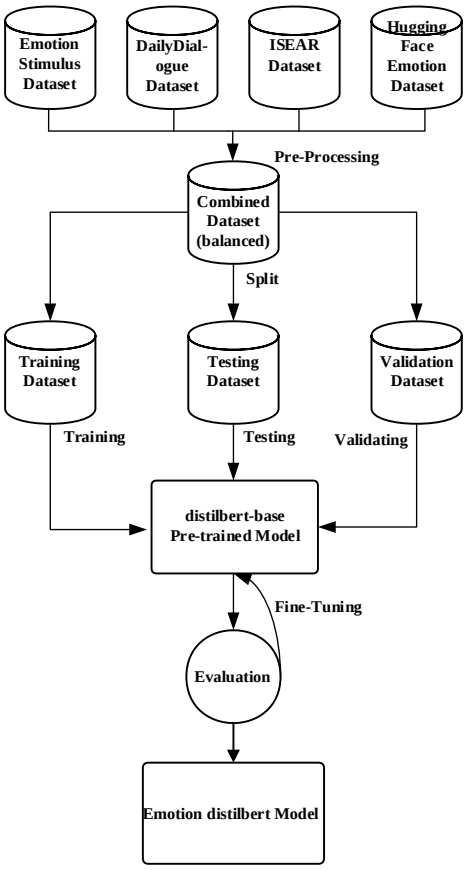

Figure 2: Fine-tuning the pre-trained model.

Since the obtained emotion labels are consecutive numbers from 0-6, in this paper, they are one-hot coded and then spliced with sentence-level feature representations. Finally, the spliced vectors are fused and extracted through the application of a multi-head attention mechanism, thereby facilitating the acquisition of a document-level feature representation.

3.4 Classifier

Bi-LSTM: Bi-LSTM is composed of two independent LSTMs that are responsible for processing the input sequence in two directions (forward and reverse). The LSTM cell consists of an input gate ($\boldsymbol{i}_t$), a forget gate ($\boldsymbol{f}_t$), an output gate ($\boldsymbol{o}_t$), and a cell state ($\boldsymbol{c}_t$). Given a review $\boldsymbol{M} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_n\}$, the LSTM processes it word by word. For each position, given the previous hidden state $\boldsymbol{h}_{t-1}$ and cell state $\boldsymbol{c}_{t-1}$, an LSTM cell generates the next hidden state $\boldsymbol{h}_t$ and cell state $\boldsymbol{c}_t$ using $\boldsymbol{i}_t$, $\boldsymbol{f}_t$ and $\boldsymbol{o}_t$. As shown in the following Equations (8)-(12):

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{mi}\boldsymbol{m}_t + \boldsymbol{P}_{mi}\boldsymbol{h}_{t-1} + b_i) \tag{8}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{mf}\boldsymbol{m}_t + \boldsymbol{P}_{mf}\boldsymbol{h}_{t-1} + b_f) \tag{9}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{mo}\boldsymbol{m}_t + \boldsymbol{P}_{mo}\boldsymbol{h}_{t-1} + b_o) \quad (10)$$

$$\boldsymbol{c_t} = \boldsymbol{f_t} \odot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \odot tanh(\boldsymbol{W}_{mc}\boldsymbol{m}_t + \boldsymbol{P}_{mc}\boldsymbol{h}_{t-1} + b_c) \quad (11)$$

$$\boldsymbol{h_t} = \boldsymbol{o_t} \odot tanh(c_t) \quad (12)$$

where σ represents the sigmoid function, ⊙ is represents the dot product operation, b is the bias vector, and $\boldsymbol{W}$ and $\boldsymbol{P}$ are both weight matrices. The output of the LSTM network is a series of hidden vectors $[\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n]$. Each $\boldsymbol{h}_t$ contains all the information of the entered review, focusing on the part of the input review around the t-th word. In contrast to the unidirectional LSTM network, the bidirectional LSTM network introduces a second LSTM layer in which the hidden states connect to the hidden states flowing in the opposite direction. The equation is as follows Equation (13):

$$\boldsymbol{h}_t = \left[\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}\right] \quad (13)$$

$\boldsymbol{h}_t$ denotes the output of the t-th word of the input. It extracts the combined output of forward passes and backward passes to obtain long dependency features with complete sequence information of all words before and after them, which helps to some extent in document-level long text information extraction.

As shown in Figure 1, the document-level feature representation goes through Bi-LSTM and fully connected layers to obtain the final feature representation, after which the classification results of the reviews are obtained through the activation function SoftMax.

4 Experiment

4.1 Datasets and evaluation metrics

Datasets: This paper uses the publicly available datasets from Li et al. [38]. These datasets are produced via domain experts and crowdsourcing. The datasets include the three domains "restaurant" "hotel", and "doctor". The statistical data for each domain's dataset is shown in Table 3. Each domain consists of three data sources: Customers, Experts, and Turkeys (Anonymous online workers). The customers with actual consumption experiences provided true reviews. The fake reviews were written by hired experts and online workers (turkeys) on the Amazon Mechanical Turk website. The experts are staff members with a high level of expertise in their domain. Li et al. asked selected hotel employees, restaurant employees (waiters or cooks), and real doctors to write positive fake reviews of their corresponding domains [38]. Reviews from all sources were used to categorize the "hotel" domain. In the "doctor" and "restaurant" domains, only "turkey" and "customer" reviews were used for categorization due to the limited availability of expert reviews. The datasets were split in a ratio of 8:1:1 for training, validation and testing. The random seed was set to 42 and shuffling strategy was used on the split data.

Table 3: Statistics for the three domain datasets.

Domain	Turkey	Expert	Customer
Hotel	800	280	800
Restaurant	200	0	200
Doctor	356	0	200

Evaluation metrics: In this paper, we use Accuracy (A), Precision (P), Recall (R), and F1 score as measures of method performance. Accuracy quantifies the model's ability to predict both fake reviews and real reviews. The precision rate represents the proportion of all predicted fake reviews that are correct. The recall is defined as the proportion of correctly predicted fake reviews that are identified in relation to the total number of genuine fake reviews. The F1 score is defined as the harmonic mean of precision and recall. The F1 score reflects the model's ability to predict deceptive reviews.

4.2 Experimental environment and parameter

Experimental environment: Since the deep learning parameters are large and computationally intensive and require more resources, the hardware and software environments used in this paper are shown in Table 4.

Table 4: Experimental environments

environments	configurations
RAM	60G
CPU	AMD EPYC 9654
GPU	NVIDIA GTX 4090，24G
operating system	Ubuntu Server 22.04.2
environment	Python, PyTorch

Parameter settings: During the model training process, the proposed model was initialized for embedding by combining TF-IDF weighting and 300-dimensional GloVe along with other parameters. This section compares the effect of the CNN in the model on the overall performance under different window sizes and dropout rates. The datasets used for the experiments are three domain mixed datasets. From Table 5 and Table 6, we can easily find that the model achieves the best results when the sentence window size is set to [2, 3, 4] and the dropout rate is set to 0.5. In this paper, we have fine-tuned the pre-trained language model distilbert, the optimisation function is Adam, the dropout is 0.1, the learning rate is 2e-5, and the training takes 43 seconds across 5 epochs. The learning rate of the complete model is 5e-5, epoch set to 40, and the training was stopped when the development accuracy did not improve for 5 consecutive epochs. The hotel domain took the longest time to train, spending 1921 seconds.

Table 5: Performance comparison of different sentence filter windows, when the dropout is set to 0.5

Filter window	Accuracy	F1-score
1, 2, 3	86.2	85.5
2, 3, 4	**87.1**	**86.7**
3, 4, 5	86.7	86.1
4, 5, 6	86.1	85.2
5, 6, 7	86.0	85.3

Table 6: Performance comparison of different dropouts, when the sentence filter window is set to [2, 3, 4]

dropout	Accuracy	F1-score
0.1	85.8	85.6
0.2	86.7	86.3
0.3	86.3	86.1
0.4	86.5	86.2
0.5	**87.1**	**86.7**

4.3 Baseline methods for comparison

To validate the efficiency of the MFE model proposed in this paper for cross-domain fake review detection, this paper will compare and analyze MFE with some state-of-the-art baselines.

Bi-LSTMWF-POS-I[1]: A variant of Bi-LSTM where word representations are combined from Glove, POS, and First-Person Pronouns embeddings.

DSRHA[17]: A two-layer hierarchical attention model that extracts multi-granular information from reviews.

SWNN[28]: A sentence-weighted neural network model.

SWNN-POS-I[28]: A variant of SWNN that adds the features POS and first-person pronouns to SWNN.

HAN[35]: A hierarchical attention neural network model.

MFE: The proposed multilevel generalized feature extraction fusion model.

4.4 Cross-domain tests

This section compares the cross-domain capabilities of the MFE method with different neural network methods. MFE extracts generic features for fusing fake reviews from three different levels (word level, sentence level, and document level). The experimental results prove that the method proposed in this paper is effective. The model is trained on the hotel domain dataset and tested on the restaurant and doctor domain datasets.

The experimental results of the methods in this paper and the baseline method are shown in Table 7. In the experiments, MFE obtained good results in both the doctor and restaurant domains. In the restaurant domain, MFE has the highest accuracy and precision, far outperforming the other baseline methods, and the DSRHA method performs the best on F1 scores and recall. The good performance of DSRHA on recall may be attributed to the local representation extracted by the unique two-layer convolutional structure. It is lower than the MFE method in terms of accuracy and precision. It suggests that using only multi-granular CNNs to extract word embedding information may not be sufficient, and the use of TF-IDF in conjunction with multi-granular CNNs can better identify important feature information in word embedding that can be used to distinguish between true and false reviews. In the domain of doctors, MFE performs well in terms of accuracy, precision and F1 score, with only a slight lack of recall. The overall performance is significantly better than other baseline methods. In terms of accuracy, our method outperforms the second-ranked DSRHA method by 10.7%. This is closely related to the inclusion of more fine-grained emotion features. A comprehensive comparison of the experimental results for the restaurant and doctor domains is presented in Table 7, which shows that the results of the test in the restaurant domain are overall better than the results of the test in the doctor domain. This is due to the fact that the restaurant domain and the hotel domain have similar real-world and linguistic environments, whereas the doctor domain is very different from the hotel domain. Unsupervised domain adaptation techniques have been a major challenge in cross-domain text categorization. The accuracy (F1) of the Bi-LSTMWF-POS-I method is 81.3% (82.4%) in the restaurant domain, but only 66.8% (70.8%) in the doctor domain, a decrease of 14.5 (11.6). This suggests that unsupervised domain adaptation techniques have an equally important role in cross-domain fake review detection. The accuracy (F1) of this paper's method in the domain of doctors decreases by only 5.0 (7.9) compared to the domain of restaurants, which is much better than the other methods. This proves that this paper is feasible to strengthen the domain adaptation ability of the model by incorporating emotion features and lexical features that are common to different domains. The MFE method with strong domain adaptation ability also maintains high performance in the domain of doctors, where the language environments differ greatly. Overall, MFE has a strong and effective cross-domain detection capability. The effectiveness of the generic feature approach proposed in this paper in terms of domain adaptation greatly improves the cross-domain detection capability of the model.

Table 7: Cross-domain performance comparison.

Domain	Methods	Accuracy	F1-score	Precision	Recall
Restaurant	Bi-LSTMWF-POS-I*	81.3	82.4	77.8	87.5
	DSRHA*	77.5	**84.7**	75.8	**96.2**
	SWNN*	69.0	73.3	64.4	85.0
	SWNN-POS-I*	66.8	73.3	61.2	91.5
	HAN*	75.5	77.3	72.0	83.5
	MFE	**83.0**	83.3	**84.2**	82.5
Doctor	Bi-LSTMWF-POS-I*	66.8	70.8	63.1	80.5
	DSRHA*	67.3	73.5	78.1	69.4
	SWNN*	61.0	68.8	57.3	86.0
	SWNN-POS-I*	61.5	69.3	57.6	87.0
	HAN*	52.8	67.2	51.5	**97.0**
	MFE	**78.0**	**75.4**	**79.2**	72.0

4.5 Ablation tests

To verify the effect of different levels of features on the model performance, ablation experiments are conducted in this paper. In this section, MFE is used to denote the complete model, and MFE-EMO denotes the removal of the document layer structure on top of the complete model, i.e., the emotional features extracted from the document layer and the multi-head attention mechanism used for feature fusion. Similarly, MFE-EMO-POS denotes the removal of the sentence layer on top of the removal of the document layer, i.e., the output of the CNN network at the word layer is fed directly into the Bi-LSTM network.

In this paper, they were tested respectively for single-domain performance and cross-domain performance, and the experimental results are shown in Table 8. It can be clearly seen through Figures 3 and 4. The structure of the document and sentence layers and the generic features they incorporate have a significant effect on the overall performance of the model.

Compared with the complete MFE model, the MFE-EMO model has an average performance degradation of 8.9 percentage points on the three single-domain datasets and an average degradation of 4.35 percentage points in the cross-domain detection experiments. The MFE-EMO to MFE-EMO-POS has a degradation of 2.67 and 4 percentage points, respectively. The results show that emotion features in review texts are more variable than lexical features, which is more obvious in a single dataset, and our approach can better capture such differences, thus improving model performance.

Table 8: Impact of different hierarchical features on model performance.

	Domain	Methods	Accuracy	F1-score	Precision	Recall
Uni-domain	Hotel	MFE-EMO-POS	71.0	71.1	79.4	64.3
		MFE-EMO	73.4	73.9	74.8	73.1
		MFE	**82.3**	**82.2**	**82.3**	**82.3**
	Restaurant	MFE-EMO-POS	71.3	71.1	72.6	69.7
		MFE-EMO	75.0	73.3	74.7	72.0
		MFE	**84.3**	**84.4**	**84.5**	**84.3**
	Doctor	MFE-EMO-POS	70.5	67.5	74.1	61.9
		MFE-EMO	72.4	57.7	60.7	55.0
		MFE	**80.9**	**81.3**	**81.8**	**80.9**
Cross-domain	Restaurant	MFE-EMO-POS	72.3	69.7	78.5	62.6
		MFE-EMO	78.6	78.6	81.6	75.9
		MFE	**83.0**	**83.3**	**84.2**	**82.5**
	Doctor	MFE-EMO-POS	72.0	68.2	74.0	63.3
		MFE-EMO	73.7	71.0	78.3	64.9
		MFE	**78.0**	**75.4**	**79.2**	**72.0**

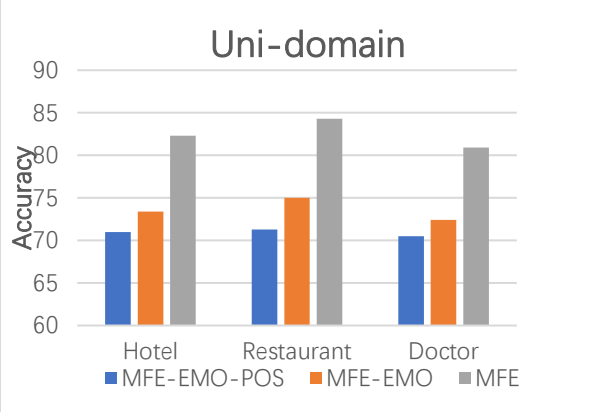

Figure 3: Uni-domain ablation tests.

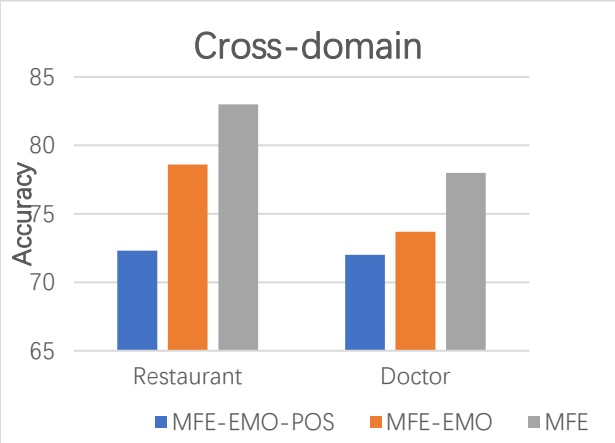

Figure 4: Cross-domain ablation tests.

Figure 4 compares the cross-domain detection results of the three models and shows that the performance of the models improves considerably in both the restaurant and doctor domains. This suggests that there are emotion and lexical differences between fake reviews and real reviews in different domains. Adding these generic features to the model can improve the generalization and cross-domain detection ability of the model.

5 Conclusion and future work

In this paper, we present a multi-level generic features extraction fusion method for the cross-domain detection of fake reviews. The method extracts multi-granular, generic, and significant information of reviews from word-level, sentence-level, and document-level, respectively. In addition, this paper conducts extensive experiments on a publicly available dataset of fake reviews. In the cross-domain detection experiments, the performance of fake review detection is significantly improved, with the best results achieved on the datasets in the restaurant and doctor domains. The experimental results clearly validate the state-of-the-art performance of the model in this paper in detecting fake reviews across domains and demonstrate the effectiveness of the model. In future work, on the one hand, more generic features can be introduced to further improve the performance of cross-domain fake review detection, and then the relationship between the number of generic features and the performance of cross-domain detection can be investigated; on the other hand, experiments can be carried out on datasets from more domains to investigate the impact of domain variability on the performance of the algorithm. Finally, a greater range of cross-domain detection research can also be attempted by combining the cross-domain approach of this paper with the cross-language approach.

Funding

This work was supported by the National Natural Science Foundation of China (62167008).

References

[1] Liu W, Jing W, Li Y. Incorporating feature representation into BiLSTM for deceptive review detection. Computing, 102(3): 701-715, 2020. https://doi.org/10.1007/s00607-019-00763-y.

[2] Chehal, Dimple, et al. Predicting the usefulness of e-commerce products' reviews using machine learning techniques. Informatica, 47(2): 275–284, 2023. https://doi.org/10.31449/inf.v47i2.4155.

[3] Ho Dac N N, Carson S J, Moore W L. The effects of positive and negative online customer reviews: Do brand strength and category maturity matter? Journal of Marketing, 77(6): 37-53, 2013. https://doi.org/10.1509/jm.11.0011.

[4] Heydari A, Ali Tavakoli M, Salim N, et al. Detection of review spam: A survey. Expert Systems with Applications, 42(7): 3634-3642, 2015. https://doi.org/10.1016/j.eswa.2014.12.029.

[5] Fette I, Sadeh N, Tomasic A. Learning to detect phishing emails. The 16th International Conference on World Wide Web. 649-656, 2007. https://doi.org/10.1145/1242572.1242660.

[6] Karimpour J, Noroozi A A, Alizadeh S. Web spam detection by learning from small labeled samples. International Journal of Computer Applications, 50(21): 1-5, 2013. https://doi.org/10.5120/7924-0993

[7] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks. The 26th Annual Computer Security Applications Conference. 1-9, 2010. https://doi.org/10.1145/1920261.1920263.

[8] Becchetti L, Castillo C, Donato D, et al. Link analysis for web spam detection. ACM Transactions on the Web (TWEB), 2: 1-42, 2008. https://doi.org/10.1145/1326561.1326563.

[9] Hess A, Klaue J. A video-spam detection approach for unprotected multimedia flows based on active networks. 30th Euromicro Conference. 461-465, 2004. https://doi.org/10.1109/EURMIC.2004.1333405.

[10] Jindal N, Liu B. Analyzing and detecting review spam. Seventh IEEE International Conference on Data Mining (ICDM 2007). 547-552, 2007. https://doi.org/10.1109/ICDM.2007.68.

[11] Crawford M, Khoshgoftaar T M, Prusa J D, et al. Survey of review spam detection using machine learning techniques. Journal of Big Data, 2: 1-24, 2015. https://doi.org/10.1186/s40537-015-0029-9.

[12] Jindal N, Liu B. Opinion spam and analysis. The 2008 International Conference on Web Search and Data Mining. 219-230, 2008. https://doi.org/10.1145/1341531.1341560.

[13] Mohawesh R, Xu S, Tran S N, et al. Fake reviews detection: A survey. IEEE Access, 9: 65771-65802, 2021. https://doi.org/10.1109/ACCESS.2021.3075573.

[14] Liu Y, Wang L, Shi T, et al. Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. Information Systems, 103: 101865, 2022. https://doi.org/10.1016/j.is.2021.101865.

[15] Zhang D. Fake review recognition based on multi-feature-fusion. Hainan University, 2023.

[16] de Arriba-Pérez, Francisco, et al. "Online Detection and Infographic Explanation of Spam Reviews with Data Drift Adaptation." Informatica, 35(3): 483–507, 2024. https://doi.org/10.15388/24-INFOR562.

[17] Liu Y, Pang B. A unified framework for detecting author spamicity by modeling review deviation. Expert Systems with Applications, 112: 148-155, 2018. https://doi.org/10.1016/j.eswa.2018.06.028.

[18] Wang X, Liu G J, Chen Z H. A fake review recognition algorithm combining Gaussian mixture model and textual graph convolutional network. Computer Applications, 2024, 44(02): 360-368. https://www.nstl.gov.cn/paper_detail.html?id=cde8da5122f1960ffa980e7f84180486.

[19] Wang G, Li C, Wang W, et al. Joint embedding of words and labels for text classification. The 56th Annual Meeting of the Association for Computational Linguistics. 2018. https://doi.org/10.18653/v1/P18-1216.

[20] Melleng A, Jurek Loughrey A, Deepak P. Sentiment and emotion-based representations for fake reviews detection. The International Conference on Recent Advances in Natural Language Processing (RANLP 2019). 750-757, 2019. https://doi.org/10.26615/978-954-452-056-4_087.

[21] Zhang M. Ensemble-based text classification for spam detection. Informatica, 48(6): 71–80, 2024. https://doi.org/10.31449/inf.v48i6.5246.

[22] Xue H, Wang Q, Luo B, et al. Content-aware trust propagation toward online review spam detection. Journal of Data and Information Quality (JDIQ), 11(3): 1-31, 2019. https://doi.org/10.1145/3305258.

[23] Archchitha K, Charles E. Y. A. Opinion spam detection in online reviews using neural networks. 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer). 1-6, 2019. https://doi.org/10.1109/ICTer48817.2019.9023695.

[24] Fang Y, Wang H, Zhao L, et al. Dynamic knowledge graph based fake-review detection. Applied Intelligence, 50: 4281-4295, 2020. https://doi.org/10.1007/s10489-020-01761-w.

[25] You Z, Qian T, Liu B. An attribute enhanced domain adaptive model for cold-start spam review detection. The 27th International Conference on Computational Linguistics. 1884-1895, 2018. https://aclanthology.org/C18-1160/.

[26] Bathla G, Singh P, Singh R K, et al. Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. Neural Computing and Applications, 2022, 34(22): 20213-20229.

[27] Cai H Y, Yu K, Wu X F. Co-attention based deep model with domain-adversarial training for spam review detection. International Conference on Network, Communication and Computing. 14-18, 2021. https://doi.org/10.1145/3510513.3510516.

[28] Li L, Qin B, Ren W, et al. Document representation and feature combination for deceptive spam review detection. Neurocomputing, 254: 33-41, 2017. https://doi.org/10.1016/j.neucom.2016.10.080.

[29] Cheng L C, Wu Y T, Chao C T, et al. Detecting fake reviewers from the social context with a graph neural network method. Decision Support Systems, 179: 114150, 2024. https://doi.org/10.1016/j.dss.2023.114150.

[30] Li J, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam. The 52nd Annual Meeting of the Association for Computational Linguistics. 1566-1576, 2014. https://doi.org/10.3115/v1/P14-1147.

[31] Wei C S, Hsu P Y, Huang C W, et al. Devising a Cross-Domain Model to Detect Fake Review Comments. In Advances in Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12: 714-725, 2020. https://doi.org/10.1007/978-3-030-63119-2_58.

[32] Song R, Fausto G, Li Y et al. TACIT: A target-agnostic feature disentanglement framework for cross-domain text classification. arXiv:2312.17263, arXiv, 24 Dec. 2023. arXiv.org, http://arxiv.org/abs/2312.17263.

[33] Ben-David E, Rabinovitz C, Reichart R. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. Transactions of the Association for Computational Linguistics 8: 504-521, 2020. https://doi.org/10.1162/tacl_a_00328.

[34] Wu H, Shi X. Adversarial soft prompt tuning for cross-domain sentiment analysis. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1: 2438-2447, 2022. https://doi.org/10.18653/v1/2022.acl-long.174.

[35] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification. The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 1480-1489, 2016. https://doi.org/10.18653/v1/N16-1174.

[36] Kamkarhaghighi M, Masoud M. Content tree word embedding for document representation. Expert

Systems with Applications, 90: 241-249, 2017. https://doi.org/10.1016/j.eswa.2017.08.021.

[37] Araque O, Corcuera-Platas I, Sánchez-Rada J F, et al. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 77: 236-246, 2017. https://doi.org/10.1016/j.eswa.2017.02.002.

[38] Onan A, Toçoğlu M A. A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. IEEE Access, 9: 7701-7722, 2021. https://doi.org/10.1109/ACCESS.2021.3049734.

[39] Kolev V, Weiss G, Spanakis G. FOREAL: RoBERTa model for fake news detection based on emotions. International Conference on Agents and Artificial Intelligence. 429-440, 2022. https://doi.org/10.5220/0010873900003116.

[40] Saravia E, Liu H C T, Huang Y H, et al. CARER: Contextualized affect representations for emotion recognition. The 2018 Conference on Empirical Methods in Natural Language Processing, 3687-3697, 2018. https://doi.org/10.18653/v1/D18-1404.

[41] Scherer K R, Wallbott H G. Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology, 66(2): 310-328, 1994. https://doi.org/10.1037/0022-3514.66.2.310.

[42] Li Y, Su H, Shen X, et al. Dailydialog: A manually labelled multi-turn dialogue dataset. The Eighth International Joint Conference on Natural Language Processing. 986-995, 2017. https://doi.org/10.48550/arXiv.1710.03957.

[43] Ghazi D, Inkpen D, Szpakowicz S. Detecting emotion stimuli in emotion-bearing sentences. Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015. 9042, 2015. https://doi.org/10.1007/978-3-319-18117-2_12.