A Hybrid Medical Text Classification Framework: Integrating Attentive Rule Construction and Neural Network

Xiang Lia, Menglin Cui, Jingpeng Li, Ruibin Bai, Zheng Lu, Uwe Aickelin

University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo,
315100, Zhejiang, China.

First published 2021

A Hybrid Medical Text Classification Framework: Integrating Attentive Rule Construction and Neural Network

Xiang Li[a], Menglin Cui[b,*], Jingpeng Li[c], Ruibin Bai[b], Zheng Lu[b], Uwe Aickelin[d]

[a] *Technology Department, Ping An Health Cloud, Shanghai 200232, China*
[b] *School of Computer Science, University of Nottingham, Ningbo, Zhejiang 315100, China*
[c] *School of Computer Science and Mathematics, University of Stirling, Stirling FK9 4LA, United Kingdom*
[d] *School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3010, Australia*

Abstract

The main objective of this work is to improve the quality and transparency of the medical text classification solutions. Conventional text classification methods provide users with only a restricted mechanism (based on frequency) for selecting features. In this paper, a three-stage hybrid method combining the threshold-gated attentive bi-directional Long Short-Term Memory (ABLSTM) and the regular expression based classifier is proposed for medical text classification tasks. The bi-directional Long Short-Term Memory (LSTM) architecture with an attention layer allows the network to weigh words according to their perceived importance and focus on crucial parts of a sentence. Feature words (or keywords) extracted by ABLSTM model are utilized to guide the regular expression rule construction. Our proposed approach leverages the advantages of both the interpretability of rule-based algorithms and the computational power of deep learning approaches for a production-ready scenario. Experimental results on real-world medical online query data clearly validate the superiority of our system in selecting domain-specific and topic-related features. Results show that the proposed approach achieves an accuracy of 0.89 and an F_1-score of 0.92 respectively. Furthermore, our experimentation also illustrates the versatility of regular expressions as a user-level tool for focusing on desired patterns and providing interpretable solutions for human modification.

Keywords: hybrid system; deep learning; attention mechanism; text classification

1. Introduction

Text classification is a well-established field related to Natural Language Processing (NLP). In the medical domain, accurate and precise decision making is often required. NLP for medical text is usually challenging because a great amount of domain knowledge is required to solve an even seemingly simple problem [1]. We are motivated by a real-world problem concerning online medical queries, which, as our main processing contexts, are in narrative formats that preserve the nature of ambiguity and informality. For example, given the medical category "female hypogastralgia", queries received from users include "*My underbelly aches during every menstrual period.*", "*I had a stomachache and menstruation didnt come on time.*", "*Feeling lower abdomen swells and backache after menstruation.*", *etc.* Forms of expression vary from one person to another, which makes it more difficult to discover the underlying patterns than those in ordinary written texts. Most current approaches rely on deep neural networks to perform text classification tasks, but such "black box" models hinder domain experts from easily verifying the evidence that supports decision making. To tackle such "black box" problem, we develop effective and interpretable algorithmic solutions by pushing regular expressions (REs) inside the classification process. Our choice of regular expressions is motivated by two factors. First, REs provide a simple and natural syntax for the succinct specification of sequential patterns [2]. Second, REs possessed sufficient expressive power for specifying a wide range of underlying patterns of texts. In our previous research, regex-based medical text classifier are constructed in an automated way to provide human experts with understandable and easy-to-modify classification solutions [3]. One of the main contributions of the work was the effective regular expression rules built upon a set of carefully selected feature words. However, like most conventional text classification methods, our previous approach provide users with frequency-based mechanism for feature selection. Feature words are selected based on relative word frequency in different categories. As a consequence, the feature selection process is typically characterized by lack of focus and causes inordinate computational costs just to deal with useless results. In this paper, we apply an attentive approach instead of conventional frequency-based methods for more effective feature selection and evaluate its efficacy in performing medical text classification tasks.

We propose to adopt the attention mechanism using a popular

*Corresponding author

Email addresses: xiangli.co@qq.com (Xiang Li), menglin.cui@nottingham.edu.cn (Menglin Cui), jingpeng.li@stir.ac.uk (Jingpeng Li), ruibin.bai@nottingham.edu.cn (Ruibin Bai), zheng.lu@nottingham.edu.cn (Zheng Lu), uwe.aickelin@unimelb.edu.au (Uwe Aickelin)

Recurrent Neural Network (RNN) setup (*i.e.* LSTM [4]) in order to derive attention-based variants for comparisons. LSTMs have been proven to be very effective to model word sequences and are powerful to learn long-range temporal dependencies. For the traditional LSTM network, it is not possible to derive the importance score for each word in the input document, but attention mechanism enables the model to focus on certain parts of the input document by assigning an attention weight to each term.

The proposed three-stage hybrid system that takes full advantage of both the expressive power of regular expressions and the non-linear complex modeling ability of deep neural networks to achieve high classification accuracy. The motivation of us developing such a hybrid system comes from the observation that the rule-based model in our previous work does not give prediction label to the instances that are not matched by any of the defined rules, which often leads to a result with high precision with low recall. The hybrid system first uses a threshold gate to filter deep learning results with high confidence scores; it then implements a heuristic method to construct regular expression rules by extracting and re-connecting feature words; finally, we run through a deep learning classifier to the rest of the data. Through our experiments, we find that the attentive bi-directional Long Short-Term Memory (ABLSTM) is able to identify keywords in a sentence relevant to its meaning and guide the regular expression construction process.

The main contributions presented in this paper can be summarized as follows:

1. We formulate the text classification problem with carefully selected features and develop novel and efficient algorithmic solutions for pushing regular expressions inside the deep learning approach.
2. The ABLSTM model is applied to lay stress on the importance of keywords or key phrases of a sentence. We take into account the feature word information by not only concatenating the attention weights into the sentence hidden representations, but also additionally applying the word weights to the construction of regular expression rules.
3. The proposed hybrid approach that combines gated neural networks and attention-guided regular expressions presents promising experimental results on real-world applications and shed light on the work towards developing interpretable ensemble systems that allow human interference.

The rest of our paper is structured as follows. Section 2 discusses the related work regarding medical text classification, deep learning approaches, and the attention mechanism. The problem scenario and preliminaries are described in Section 3. Section 4 gives a detailed description of our three-stage hybrid proposals with attention-guided regular expressions and ABLSTM method. Extensive experiments to justify the effectiveness of our proposals are presented in Section 5. Conclusions and future works are discussed in Section 6.

2. Related work

The problem of text classification has a long history. Classification for medical-related text is considered as a special case of text classification. Before deep learning research became popular, most text classification tasks used statistical machine learning methods. Researchers mainly focused on feature-based and kernel-based methods [5, 6, 7], which are limited by conditions such as manual feature engineering and dependence on existing NLP toolkits. A number of learning algorithms have been applied to text which had been vectorized using a TF-IDF weighting method, including Support Vector Machines (SVM) [8, 9], regression models [10], nearest neighbor classification [11], Bayesian models [12], and inductive learning [13]. These algorithms assume that independent keywords or key phrases are important to the text category and extract vector features representing those keywords or key phrases using statistical methods [14]. These methods have been successfully applied to medical text classification tasks on patient record notes [15] and other text documents in diseases like diabetes and cancers [16, 17], but the assumption is an oversimplification that brings some shortcomings. While independent keywords and phrases are important, there are other linking words which also give meaning to a text. The way words relates to each other can also provide context and disambiguation. Without this, we potentially lose some information.

Deep learning methods have seen substantial success for text classification because such methods can automatically and effectively learn underlying features and interrelationships in data. Machine learning models based on deep architectures, such as the RNN [18, 19, 20], Convolutional Neural Network (CNN) [21, 22, 23], Factor-based Compositional embedding Model (FCM) [24], and word embedding-based models [25], have achieved state-of-the-art performances in many natural language processing fields including text classification [26].

More recently, there have been research efforts to incorporate attention mechanisms into CNNs and RNNs that are typically used in NLP applications. Attentive neural networks, which pioneered for machine translation [27], have recently seen successes in the field of text classification [28, 29, 30, 31]. Bahdanau *et al.* [32] applied the attention-based model to machine translation, which allows the decoder to watch different parts of the source sentence at each step of the output generation rather than to encode the full source sentence into a fixed-length vector, and explicitly find a soft alignment between the current position and the input source. Since then, the attention mechanism has been adopted to text classification tasks when the weight of every component needs to be evaluated respectively. Er *et al.* [33] proposed attention pooling based CNN to represent sentences, which uses an intermediate sentence representation generated by the Bidirectional Long Short-Term Memory (BLSTM) as a reference for local representations produced by the convolutional layer to obtain attention weights. Yang *et al.* [34] contributed to designing the Hierarchical Attention Network (HAN) for document classification, which has two levels of attention mechanisms applied at the word and sentence level. Wang *et al.* [35] proposed an attention-based LSTM method with target embedding, which was proven to be an effective way to enforce the neural model to attend to the related part of a sentence. The attention mechanism is used to enforce the model to attend to the important part of a sentence, in response to a specific aspect. Likewise, Yang *et al.* [36] proposed two attention-based bidirectional LSTMs to improve classification performance. Liu and Zhang [37] extended the attention modeling by differentiating the attention obtained from the left context and the right context of a given target/aspect. Instead of using the attention-based neural network for direct text classification, in this work, we use the attention mechanism for identifying high-quality feature words for the automated construction of regular expression rules.

While deep learning models have seen widespread successes, they treat all the words as a block of input without explicitly giving any words or phrases special treatment. We would like to leverage the advantages of both the rule-based text categorization approaches, which focuses on keywords and their combinatorial patterns, and the modern deep learning approaches, which learn underlying relationships, for medical text classification. Hybrid systems that combine rule-based approaches with machine learning techniques have attracted much research interests recently. In general, two types of hybrid methods are presented to combine machine learning with explicit rules. One approach utilizes rules to verify the machine learning output [38, 39], while a more prevailing approach leverages on rule-based algorithms to identify the desired features to feed machine learning models. In the medical domain, Wang *et al.* [40] explored machine learning approaches trained from weak supervision labels derived from a rule-based engine, but the proposed paradigm is not effective for complex multiclass classification tasks. Yao *et al.* [41] employed CNNs to capture additional features generated by rule-based approaches such as Unified Medical Language System (UMLS) and Concept Unique Identifiers (CUIs).

While these works mainly focus on well-structured text data in formal languages, we previously employed a hybrid system that combines regular expression based classifiers with machine learning methods to deal with real-world clinic narratives collected from online medical consultation [3]. The innovation and contribution of our work is that the proposed three-step hybrid system is more proficient in narrative medical text classification. Our experimental results on real-life data showed the proposed method can be a better fit for the production-ready online medical guidance scenario compared with existing literature.

3. Preliminaries

3.1. Problem description

The problem defined in this study is described as follows: given a set of text queries Q and a set of predefined classes C, our task is to classify each query $q \in Q$ to a certain class $c \in C$ based on a set of previously labeled instances. This typical multi-classification problem is often tackled by supervised learning approaches such as SVM, CNNs, RNNs, and expert systems. We propose a hybrid system that combines rule-based classifiers with deep neural networks.

3.2. Feature word selection

Apart from syntactic structures, feature words (or keywords) are also important features to identify one category from another. Traditional NLP approaches for keyword extraction normally rely on the frequency information of the given text. Term Frequency (TF) [42], as one of the most typical methods, measures the occurrence of each term in the document and assigns it to the feature space. All terms in vocabulary are treated equally in TF. Inverse Document Frequency (IDF) [43] is a method that measures how much information the word provides by assigning higher weights to rarer words in the document. IDF is often used in conjunction with term frequency in order to lessen the effect of implicitly common words in the corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. TF-IDF is useful, most importantly in automated text analysis, for scoring words in machine learning algorithms for NLP.

Inspired by TF-IDF which works on a single document from a given corpus, we propose the relative frequency model over positive and negative sample sets for keyword extraction in our case. Specifically, for a given class c, the training instances labeled as c are treated as the positive set, while other instances compose the negative set. Feature words are selected based on the comparative word frequency of each word in the two sets. We define average word frequency f_w^c as the number of times a given word $w \in W$ presents in the query set Q^c divided by the total number of sentences of all queries in Q^c. That is,

$$f_w^c = \frac{\sum_{q \in Q^c} f_{w,q}^c}{\sum_{q \in Q^c} l_q}, \qquad (1)$$

where $f_{w,q}^c$ denotes the number of times that word w occurs in the query q in class c and l_q is the length (number of sentences) of Q^c. The average word frequency indicates how popular a given word is in each class. Since a word can be used in more than one class of queries, we introduce the term "relative word frequency" to measure the relative popularity of a word between different classes. The relative word frequency is defined as:

$$RF_w^c = \frac{f_w^c}{f_w^{\overline{c}}}, \qquad (2)$$

where $\overline{c}$ is the complementary set of c given the full set C. Therefore, a sorted list by RF_w^c in descending order gives the list of relatively most popular words in class c compared to all other classes. Regular expression rules R_P and R_N are generated based on the calculation of word similarities and co-occurrence by predefined filtering mechanisms. The detailed algorithm of regular expression generation is introduced in our previous work [3].

3.3. Regular expression based classifiers

In this subsection, we revisit the regular expression based medical text classifier using the constructive heuristic approach proposed in our previous paper [3] and discuss the potentiality to refine the model. Regular expressions are often used in the situation where interpretable solutions are required. In current practice, rules are created by experienced annotators with clinical-related knowledge. However, the manual construction of regular expressions is often time-consuming and error-prone. To reduce human efforts, we proposed a heuristic method to automate the composition of regular expressions to obtain interpretable and explicit rules. The regular expression construction process can be summarized as follows:

1. Define positive and negative instances based on the medical category.
2. Filter positive feature words w_{P_i} and negative feature words w_{N_i}.
3. Construct R_P ($w_{P_1} * w_{P_2} * ... * w_{P_i}$) as positive regular expression rules and R_N ($w_{N_1} * w_{N_2} * ... * w_{N_i}$) as negative regular expression rules, where $*$ denotes a regular expression connector, which is either AND (denoted by ".*" in the regular expression), OR (denoted by "|"), or adjacency (denoted by ".{a, b}", where a and b are non-negative integers).
4. Generate the complete regular expression R by connecting R_P and R_N with logic function NOT (denoted by "#_#"). Note that the function NOT is not an embedded regex operator, but uniquely introduced in our scenario.

Note that rather than a boolean query with proximity operator, the regular expression identifies the hidden pattern of texts. Therefore, a co-occurrence matrix is constructed to provide information on the syntactic structure and word correlations of input texts. Co-occurrence here is referred to as the frequency of two words occurring together in a certain order in every text query of the input data. Let p be the size of vocabulary in the whole corpus, a matrix $\boldsymbol{M}$ with a size of $p \times p$ will be produced:

$$\boldsymbol{M}(i, j) = \sum_{q \in Q} \begin{cases} 1, & \text{if } pos_q(w_i) < pos_q(w_j) \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where i and j indicate the i-th and j-th word w_i and w_j respectively, $0 \leq i, j \leq p$, and $pos_q(w)$ represents the index (position) of a given word w in text query q.

In addition, special operators "?!" (zero-width negative look-ahead assertion) and "?<!" (zero-width negative look-behind assertion), which eliminate undesired words or short phrases, are also used in the construction of regular expressions. More details can be found in [3]. An example of a regular expression for the medical category "female hypogastralgia" provided below explicitly demonstrates that the regex generated by our previously proposed constructive heuristic method is fully interpretable to humans and highly flexible for modification.

```
.*(belly|stomach).{0,8}(?<!no|not)(ache|pain).*
#_#.*(pregnant|abortion|male).*
```

Recall our running examples illustrated in the *Introduction*, the sample regular expression captures the underlying pattern of the first two input texts "*My underbelly aches during every menstrual period*" and "*I had a stomachache and menstruation didnt come on time*" but fails to derive generalized patterns of the third example "*Feeling lower abdomen swells and backache after menstruation*". This is because its feature terms "*abdomen*" and "*swell*" have relatively low occurrence over the corpus.

Although experimental results on massive amounts of real-world medical data have verified the feasibility of regex rule-based method, we observe that the selection of feature words based on the word frequency has limitations when identifying terms with high topic relevance but low occurrence frequency. A heuristic process relying on deep neural networks to identify feature words would contribute to the overall performance enhancement. Furthermore, in previous settings, five hyper-parameters are set to define the threshold of relative word frequency selection criteria, thus the performance of the generated regular expressions is highly dependent on the training of these hyper-parameters; on the contrary, an attentive network-based approach effectively avoids the tuning of abundant parameters and narrows the weight of each word in a given sentence between 0 and 1.

4. Methodology

In this study, we aim to solve the multi-class medical text classification problem using a three-stage hybrid system which combines the threshold-gated neural network model and the attention-guided rule-based method. Specifically, we try to define word weights in narrative medical texts through the attention-based bidirectional LSTM network. Gated neural network models are trained to obtain high classification performances and regular expressions are constructed in a heuristic process guided by the word weights derived by ABLSTM architecture.

4.1. RNN and LSTM

Recurrent neural network [44] is a special kind of feed-forward neural network which is useful for modeling time-sensitive sequences. At each time t, the model receives input from the current example and also from the hidden layer of the network's previous state. The output is calculated given the hidden state at that time step. The recurrent connection makes the output at each time associated with all the previous inputs. The LSTM model [4] addresses the problem by re-parameterizing the RNN model. The core idea of LSTM is to introduce "gates" to control the data flow in the recurrent neural unit. The advantage of the LSTM nodes is that they can be set up in the process of synthesizing to control how much information should be received, forgotten, or passed back in the current synthesis step. Through these gate controls, the gradient of the long-term dependencies cannot vanish, which ensures the proficient learning ability of LSTM for long texts.

4.2. Attentive recurrent architecture

Although neural networks are effective in dealing with text classification tasks, obvious drawbacks in such "black box" approach cannot be ignored. In this paper, we describe a bi-directional LSTM architecture with an attention layer that allows the network to weigh words in a sentence according to their perceived importance. Attentive LSTM model can extract part of the subset of a given input, where it focuses on the word or phrase level importance of given queries. We utilize the intermediate result of neural network outputs for efficient feature selection (*i.e.* attentive word weight) to assist the rule-based approach to construct interpretable and explainable text classification solutions. Fig. 1 shows the overall architecture of the model.

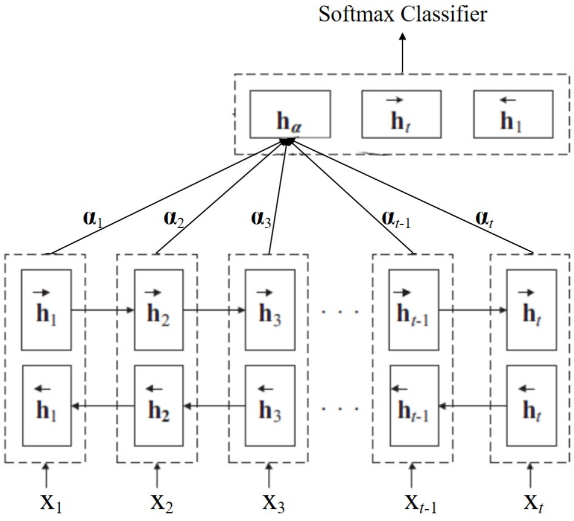

Softmax Classifier
$\mathbf{h}_a$
$\overrightarrow{\mathbf{h}}_t$
$\overleftarrow{\mathbf{h}}_1$
α_1
α_2
α_3
α_{t-1}
α_t
$\overrightarrow{\mathbf{h}}_1$
$\overrightarrow{\mathbf{h}}_2$
$\overrightarrow{\mathbf{h}}_3$
$\overrightarrow{\mathbf{h}}_{t-1}$
$\overrightarrow{\mathbf{h}}_t$
$\overleftarrow{\mathbf{h}}_1$
$\overleftarrow{\mathbf{h}}_2$
$\overleftarrow{\mathbf{h}}_3$
$\overleftarrow{\mathbf{h}}_{t-1}$
$\overleftarrow{\mathbf{h}}_t$
x_1
x_2
x_3
x_{t-1}
x_t

Figure 1: Attentive bidirectional LSTM structure

Assume a sentence S is segmented into t words, *i.e.*, $S = [I_1, ..., I_t]$, where I_i represents the i-th word. Let $w_i \in R^d$ denote the vector representation of the word I_i. We apply the bi-directional LSTM to get summarizing annotations of word embedding from the sentence. The bi-LSTM contains the forward LSTM which reads the words from I_1 to I_t and a backward LSTM which reads from I_t to I_1.

$$\begin{aligned}\overrightarrow{\mathbf{h}_i} &= \overrightarrow{\text{LSTM}}(w_i);\ i \in [1, t],\\ \overleftarrow{\mathbf{h}_i} &= \overleftarrow{\text{LSTM}}(w_i);\ i \in [t, 1],\end{aligned} \tag{4}$$

A word-level neural representation for a given word I_i is obtained by concatenating the forward hidden stage $\overrightarrow{\mathbf{h}_i}$ and backward hidden state $\overleftarrow{\mathbf{h}_i}$:

$$\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}], \tag{5}$$

Bi-LSTM neural unit summarizes the information of the whole sentence S. In the traditional LSTM model, the vector of $\overrightarrow{\mathbf{h}_t}$ and $\overleftarrow{\mathbf{h}_1}$ is usually concatenated as a text representation, which hardly captures the information about the importance of each word to the whole sentence. Since the words that reflect the subject in a text are primarily a few keywords, the importance of each word is different. Therefore, we introduce the attention mechanism, which can drive the model to address the significance of "hot" words to the meaning of a sentence. The following formulas are applied to compute the attention weight α_i between the i-th word and the sentence S.

$$u_i = tanh(W_w \mathbf{h}_i + b_w) \tag{6}$$

$$\alpha_i = softmax(u_i) = \frac{exp(u_i u_w)}{\sum_i exp(u_i u_w)} \tag{7}$$

$$\mathbf{h}_a = \sum_{i=1}^{t} \alpha_i \mathbf{h}_i \tag{8}$$

where W_w and u_w are projection parameters, b_w is the bias parameter, and $\mathbf{h}_a$ is the resulting weighted feature vector that summarizes all the word-level information in a sentence. Then, $\mathbf{h}_a$ together with LSTM forward and backward results forms the vector representation of the text, which can be demonstrated as:

$$\mathbf{s} = [\mathbf{h}_a; \overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_1}], \tag{9}$$

where $\mathbf{s}$ is a high-level representation of the sentence that can be used as final features to predict the label y for sentence classification by a softmax layer.

$$y = softmax(W_s \mathbf{s} + b_s) \tag{10}$$

Let $\hat{y}$ be the ground truth of the category label, the goal of training is to minimize the cross-entropy error between y and ground truth $\hat{y}$ for all training data.

$$loss = -\sum_k \sum_j y^j log \hat{y}^j + \lambda ||\theta||^2 \tag{11}$$

where k is the index of the sentence, j is the index of the category, λ is the L_2-regularization term, and θ is the parameter set.

4.3. Word weight calculation

For a given category $C = [S_1, S_2, ..., S_j, ...]$, two steps are put forward to calculate the weight (or importance) of each word in a given category based on the word attention weights derived from attentive LSTM method. Let α_w^k denote the attention weight of the word w in the k-th sentence and α_w^C denote the weight of the word w in the category C.

Weighted average category word weight calculation based on the weighted average method is represented by the following equation:

$$\alpha_w^C = \frac{\sum_n \alpha_w^k}{n} \tag{12}$$

where n is the number of sentences in C that contain the word w. The resulted weight is normalized to obtain the importance of each word in the category C.

Occurrence filter in each sentence S, the scarce terms are often name, location, or other noise words that should be abandoned in our extraction process. All extracted words for every sentence in the category C are filtered by the minimal occurrence, for example, 10 times in a certain class. Then, the results are sorted in descending order based on the weighted average α_w^C.

$$\alpha_w^C = \begin{cases} \alpha_w^C, & \text{if } \beta_w^C \geq occur_{min} \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where β_w^C denotes the occurrence counts of the word w in all sentences in C and $occur_{min}$ is a pre-defined threshold.

These two steps help address the true relevant terms in one category. Some disturbing words with high weighted average and low occurrences would be filtered out, but terms with high weighted average and modest occurrences are kept. These core words are neither the most frequent words nor words with high weighted average by chance. We expect this method to extract semantically relevant keywords with deeper connections to the given category. The performance of the system based on the above-mentioned approaches is demonstrated in Section 5.

4.4. Hybrid method for text classification

In this subsection, we present the overall framework of the three-stage hybrid method to deal with the medical text classification task. The diagram of the proposed method is presented in Fig. 2. Compared with traditional machine learning methods such as decision trees (*e.g.* random forest, gradient boosted trees), naïve Bayes and SVM, neural network approaches demonstrate better performance when dealing with large data sets by learning underlying patterns and features in data. Specifically, we find that the attentive bidirectional LSTM model is able to effectively capture the interdependencies of words in a sentence, and in the meantime provide information on the importance (or weight) of each individual word by the attention mechanism. Word weights, as a side output of ABLSTM, are important features to guide the construction of regular expressions. A hybrid method leveraging both the computational power of neural networks and the interpretability of regular expressions is proposed to improve the text classification performance. The overall system contains three parts:

1. an ABLSTM model to extract key features, where the output confidence score is gated by a predefined threshold to ensure high precision results;
2. a regular expression based classification model whose feature words are provided by ABLSTM model in stage 1;
3. a non-gated ABLSTM model to give prediction labels to the remaining unlabeled instances.

The full predicting procedure of the proposed hybrid system is presented in algorithm 1. Now we explain each step in detail. An ABLSTM classifier is firstly trained based on the training data. The confidence score of each prediction is computed and recorded along with the output. Then a threshold gate is posed to the confidence score to screen the more confident (more likely to be true) predictions. Only the predictions with confidence scores above a certain threshold are kept. The rest of the queries are processed in the second stage classifier. The second stage is a rule-based classifier composed of a set of regular expressions. The construction of regular expressions is guided by the attention weights computed by ABLSTM model. Words with higher weights are regarded as stronger features. The detailed construction process can be referred to in [3]. Each query will go through the regex rules matching scheme and certain category label will be assigned to the query if it is matched with one of the regular expressions in the given category. The final stage is a non-gated ABLSTM classifier, which processes the queries which are not labeled by the preceding two stages.

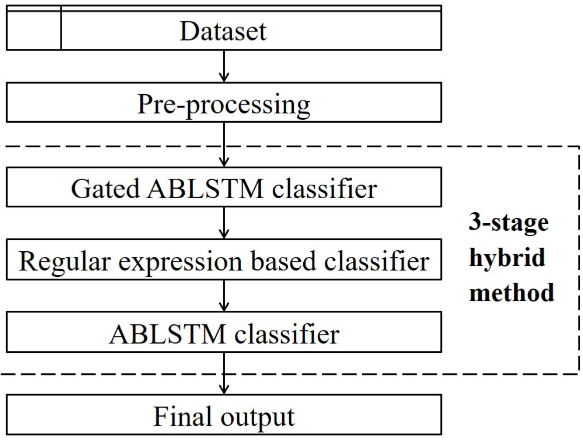


Figure 2: The three-stage hybrid system

Algorithm 1 Procedure of the three-stage hybrid method

```
Require: a text corpus with $k$ sentences and a confidence threshold $\lambda_c$.
Ensure: class labels for each of the $m$ sentences.
An ABLSTM model is trained. Confidence scores $\hat{y}_j$ of all instances are computed. A sorted list of relevant word $W_e$ is extracted using (13).
A constructive heuristic using $W_e$ as seed words is performed to generate a set of regular expression rules RE.
if max($\hat{y}_j$) ≥ confidence threshold $\lambda_c$ then
    return argmax($\hat{y}_j$) as predicting label.
else
    RE based matching scheme is performed
    if RE scheme is performed and matched then
        return predicting label
    else
        return argmax($\hat{y}_j$) as predicting label
    end if
end if
```

5. Experiments

In this section, we carry out a comprehensive experimental evaluation on the performance of the proposed hybrid method. Specifically, we try to address both the interpretability and the effectiveness of our approach in comparison with state-of-the-art algorithms. We report the accuracy, precision, recall, and F_1-score as our evaluation metrics.

5.1. Data and pre-processing

We use online consultation data provided by our collaborator, a major online healthcare provider in the Chinese market, for both training and testing of our system. A collection of patient queries labeled by medical categories is treated as our input to perform the text classification task. The categories are manually labeled by a team of medical experts in our collaborating company. Table 1 gives an illustration of what the data set looks like. Note that the original corpus that we use in our experiment is in Chinese, but we translate the corresponding text to English for the demonstration purpose. In fact, the system is not specifically designed for a particular language, since the core idea of word weight extraction using the output of the recurrent neural network attention layer proposed by this work can be generalized to deal with narrative medical texts in different languages. In our setting, a total of $150,000$ effective records within 100 medical categories are collected from our collaborating institution's online operational streams. The ratio of training, validation and test set is 80%, 10% and 10%.

During the pre-processing step, Chinese word segmentation method *jieba* is applied to the input text. Unlike English words that are naturally separated from each other by whitespace, the operation of word segmentation is necessary for Chinese text. After word segmentation, stop words, symbols, punctuation, *etc.* are removed from the tokenized data. We pre-train a *word2vec* [45] model over a large-scale unannotated corpus and encode meaningful linguistic relationships between words into learned word embeddings. The *word2vec* model is trained on medical text records with the dimension of 100 to produce effective word embedding. A demonstration of word similarity based on *word2vec* model is shown in Table 2 The similarity between words is not measured by their spelling or shared characters, but instead the embedded semantic information.

For the setting of hyper-parameters, we apply the RMSprop optimizer in the setting of the learning rate for 0.001, the gradient moving average decay factor for 0.9, and no learning rate decay.

5.2. Attention-guided rule-based models

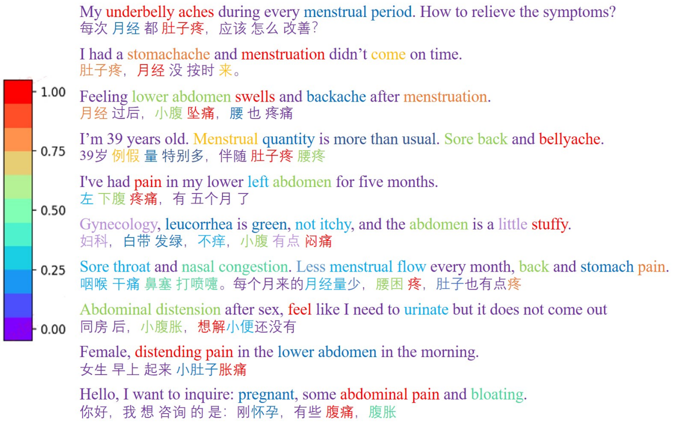
As introduced in Section 4.2, ABLSTM features, *i.e.*, word weights as the output of the attention layer, are helpful in guiding the rule construction. The heat map in Fig. 3 visualizes the weights placed on different words by ABLSTM model for the category"female hypogastralgia". Words in red color are relatively important, while words in purple are less important. With regards to the query "*Feeling lower abdomen swells and backache after menstruation*" which fails to be matched by the regular expression rule due to low frequency features, ABLSTM assigns the term"*abdomen*" with a weight of 0.6 and "*swell*" with a weight of 1.0. The attention mechanism makes the model to lay stress on scarce but crucial terms by assigning them with high attention weight.

Table 2: Demonstration of word similarity by word2vec model

word1	**word2**	**similarity**
baby	children	0.93
baby	fever	0.06
cold	fever	0.43
chill	fever	0.36
cold	chill	0.82
cough	sneeze	0.69
flu	influenza	0.95
pain	hurt	0.81
bellyache	stomachache	0.78

We also compare the keywords extracted by different approaches, namely, term frequency, relative word frequency [3], and ABLSTM. Table 3 illustrates that keywords derived from the relative word frequency approach and the attention-based approach are much topic-related and domain-specific compared to the term frequency approach. Specifically, we can observe from the result that terms like "vaginitis", "pelvic inflammation", and "abdominal pain" have relatively low occurrence (low term frequency) over the corpus but they rank high (top 3) by the ABLSTM approach. The above-mentioned three phrases are all common syndrome under the medical category "female hypogastralgia" and can be regarded as crucial features for the classification task. Our corpus, mainly in narrative formats, contains a fair number of feature words with relatively low occurrences, in which case the conventional term frequency method often fails to give them high rankings. This suggests the efficiency of our method to capture the important terms with semantic consideration.

Experiments are also conducted to evaluate the performance of rule-based method with RF features and ABLSTM features calculated by two approaches introduced in Section 4.3. Results in Table 4 illustrate that the ABLSTM feature captured by the occurrence filter mechanism outperforms the other two approaches. Taking into account both the word weight and frequency of occurrence, the occurrence filter is able to determine keywords that are semantically relevant to the specific medical scenario, as it filters out the words with either low weight



My underbelly aches during every menstrual period. How to relieve the symptoms?
每次 月经 都 肚子疼，应该 怎么 改善？

I had a stomachache and menstruation didn't come on time.
肚子疼，月经 没 按时 来。

Feeling lower abdomen swells and backache after menstruation.
月经 过后，小腹 坠痛，腰 也 疼痛

I'm 39 years old. Menstrual quantity is more than usual. Sore back and bellyache.
39岁 例假 量 特别多，伴随 肚子疼 腰疼

I've had pain in my lower left abdomen for five months.
左 下腹 疼痛，有 五个月 了

Gynecology, leucorrhea is green, not itchy, and the abdomen is a little stuffy.
妇科，白带 发绿，不痒，小腹 有点 闷痛

Sore throat and nasal congestion. Less menstrual flow every month, back and stomach pain.
咽喉 干痛 鼻塞 打喷嚏。每个月来的月经量少，腰困 疼，肚子也有点疼

Abdominal distension after sex, feel like I need to urinate but it does not come out
同房 后，小腹胀，想解小便还没有

Female, distending pain in the lower abdomen in the morning.
女生 早上 起来 小肚子胀痛

Hello, I want to inquire: pregnant, some abdominal pain and bloating.
你好，我 想 咨询 的 是：刚怀孕，有些 腹痛，腹胀

Figure 3: Heat map of word weights calculated by the ABLSTM model

Table 1: Data set and labels

Text queries	Medical Category
I feel difficult to fall asleep every day and I always dream during the night.	Insomnia
Always sleep talking. Feel stressed.	Insomnia
Will I get pneumonia if I sometimes choke while eating?	Pneumonia
Baby has a fever and coughs. I'm worried if it's pneumonia.	Children cough
My son gets a cold and often sneezes.	Acute upper respiratory infection
5-year-old child burps out loud in the morning and it gets worse before sleep.	Children indigestion
Hiccup, uncomfortable throat.	Adult indigestion
I feel full by eating just a little and could not digest properly.	Adult indigestion
My right knee feels painful when I go upstairs. It doesn't hurt when I walk.	Knee pain
Can I smoke after abortion surgery?	Induced abortion
My husband smokes heavily. Does it matter if we want to have a child?	Pregnancy preparation
I had sex last night and saw blood on my underwear this morning.	Postcoital vaginal bleeding
My period hasn't come this month, but I see a little blood on my underwear.	Abnormal vaginal bleeding
My period hasn't come this month. It sometimes comes late and the bleeding is scanty.	Irregular menstruation
...	...

Table 3: Top 10 most important words extracted by different methods for category "female hypogastralgia"

Importance	Term frequency		Relative word frequency		ABLSTM	
	Word	Term frequency	Word	Relative frequency	Word	ABLSTM weight
1	inquire	3787	distending pain	10.47	vaginitis	93.86
2	ache	2736	belly	9.00	pelvic inflammation	89.71
3	belly	2604	twitch	9.00	abdominal pain	88.26
4	menstruation	2130	ache	6.50	ovarian cyst	85.67
5	stomach	1689	stomachache	5.97	stomachache	79.22
6	stomachache	1444	abdomen	5.42	pain	77.36
7	pregnant	1430	abdominal pain	5.27	dysmenorrhea	65.99
8	not	1131	backache	5.00	cervicitis	63.48
9	matter	1117	pain	4.50	bleed	61.81
10	pain	950	intermittence	4.23	cervical erosion	61.53

Table 4: Performance of rule-based models with different features

Model	Macro Precision	Macro Recall	Macro F_1	Accuracy
Relative word frequency	0.94	0.67	0.75	0.5833
ABLSTM weighted average	0.95	0.75	0.84	0.6762
ABLSTM occurrence filter	**0.95**	**0.80**	**0.87**	**0.7035**

(the word is not important in the corpus) or low occurrence (the word is a scarce term). In the following section, the experimental results of rule-based model with occurrence filtered ABLSTM features are presented in the hybrid system.

5.3. *Three-stage hybrid system*

Note that different from machine learning methods that make predictions on every test instance, the regular expression-based model assigns a prediction label to a certain instance only if the instance is matched with the rules defined by REs. In other words, there exist certain test samples that are not assigned to any category when we apply the rule-based method. Intuitively, this kind of method guarantees high classification accuracy with the sacrifice of recall. Experimental results in Table 5 verifies this speculation. Though the accuracy of rule-based model for all instances is relatively low, we observe that among the $9,370$ out of $15,000$ test data that are matched and labeled by the rule-based model, an accuracy of 0.9330 is achieved, which is higher than ABLSTM method. From another perspective, the rule-based model misses about 38% queries (*i.e.* no classification outputs) despite of high performance for those matched cases. Besides, it is also demonstrated in Table 5 that ABLSTM achieves higher recall and lower precision compared with the rule-based model (for all instances).

As previously introduced, the medical domain requires precise and accurate decision making. Therefore, we try to ensure classification precision and accuracy by trading off the amount of support (data with prediction labels). This is achieved by posing a threshold restriction on the output confidence of the ABLSTM model to screen the highly "confident" result. A set of thresholds from 0.1 to 0.9 are tested for the output result of the deep learning method. With the increment of the threshold, we observe the improvement of classification performance as expected, whereas the number of successfully classified (or supported) instances decreases, as shown in Table 6. For example, at the confidence threshold of 0.1, $14,914$ (out of $15,000$) instances are classified with an F_1-score of 0.88, which is not satisfactory for practical use. When the threshold is raised to 0.8, F_1-score increases to 0.96 but at the expense of a smaller number of successful classifications ($10,680$ out of $15,000$).

Instructed by the idea of leveraging the advantages of both ABLSTM and rule-based methods for better overall performance, the hybrid method puts between the two networks a attention-guided regex classifier that is invoked whenever the confidence of the network classification is low. Compared with the experimental results shown in Table 5, ABLSTM models with a certain level of threshold restriction (threshold above 0.8 in this experiment) outperform the rule-based model for matched instances in terms of all evaluation metrics. The second stage of the hybrid system is the attention-guided rule-based classifier. The regular expression construction process is similar to the approach introduced in [3], but the main innovation of this work is that we rely on the attention weight calculated by ABLSTM model instead of frequency count to extract feature words. At the final stage, a non-gated ABLSTM model is applied to handle the remaining unlabeled data.

Experimental results in Table 7 certify the effectiveness of our proposed system. The best result is obtained when the threshold of ABLSTM confidence score for output label is set to 0.8, which outperforms any other single methods in terms of precision, recall, F_1-score and accuracy. Experiments on tuning this parameter on the validation set also reveal that when the confidence threshold is set between 0.7 and 0.9, the model performance shows trivial differences. The hybrid system makes complementation and compatibility of the rule-based method and deep neural networks. In real-world applications, such a system is more suitable for the production-ready scenario than existing text classification models based solely on deep learning approaches. When the scenario changes or any problem incurs, modifying the explicit rules presented by interpretable regular expressions is more time-saving and practically achievable than retraining the deep learning models.

5.4. *Model robustness*

In this subsection, we evaluate the performance of the proposed method using different sizes of training data, so as to test the robustness of the hybrid system under data scarcity. We reduce the sizes of training and validation sets to 10% and 50% of the original ones and keep the same test set. Table 8 demonstrates the performances of ABLSTM and hybrid methods on different training sizes. From the experimental results, it can be observed that the performance of both approaches improves with the increased number of training data. Under data scarcity, when the training size is limited, while ABLSTM model shows deficient performance (78% precision and 60% recall with the training size of $12,000$), the hybrid method keeps a relatively high precision of 88% and recall of 70%. We argue that the decreased performance of ABLSTM model may due to the inherent complexity of deep learning methods. As one of main strengths of deep learning lies in being able to handle complex data and relationships, algorithms used in deep learning is usually resource-intensive. To extract from the data the complex patterns that is general enough, a large quantity of data is typically required by deep learning algorithms than traditional machine learning or rule-based models. The results further certify the role of the rule-based model as an effective complement to deep learning models.

6. Conclusions and future work

In this paper, we describe a three-stage hybrid system for medical text classification. While most text classification models from the literature treat all words equally and focus on the semantic relationships between words to get the overall meaning, our proposed approach takes one step further by capitalizing on important feature words. The attention-based bi-directional LSTM extracts the attentive weights from attention layers, which allows us to inspect discriminative terms in a particular sentence and guide the construction of regular expression rules for a particular class. The hybrid method combining attentive bi-directional LSTM and rule-based approach takes the advantages of both the computational power of neural networks to select effective features and the high interpretability

Table 5: Performances of ABLSTM and baseline methods

Model	**Macro Precision**	**Macro Recall**	**Macro F_1**	**Accuracy**	**Support**
Rule-based model for matched instances	0.94	0.94	0.94	0.9330	9,370
Rule-based model for all instances	0.94	0.67	0.75	0.5833	15,000
ABLSTM	0.93	0.83	0.87	0.8631	15,000

Table 6: The performance of ABLSTM model with threshold restrictions

Threshold	**Macro Precision**	**Macro Recall**	**Macro F_1**	**Accuracy**	**Support**
0.1	0.94	0.84	0.88	0.8665	14,914
0.2	0.94	0.85	0.89	0.8731	14,774
0.3	0.95	0.86	0.90	0.8926	14,210
0.4	0.95	0.86	0.90	0.8926	14,210
0.5	0.95	0.88	0.91	0.9056	13.693
0.6	0.95	0.89	0.91	0.9290	12,818
0.7	0.96	0.91	0.93	0.9489	11,586
0.8	**0.98**	**0.95**	**0.96**	0.9686	10,680
0.9	0.97	0.95	0.95	**0.9834**	8,914

Table 7: The performance of hybrid methods

Threshold	**Macro Precision**	**Macro Recall**	**Macro F_1**	**Accuracy**	**Support**
0.1	0.94	0.83	0.88	0.8638	15,000
0.2	0.95	0.84	0.88	0.8655	15,000
0.3	0.95	0.84	0.89	0.8684	15,000
0.4	0.95	0.84	0.89	0.8716	15,000
0.5	**0.96**	0.85	0.90	0.8826	15,000
0.6	0.95	0.86	0.91	0.8869	15,000
0.7	**0.96**	0.88	0.91	0.8868	15,000
0.8	**0.96**	**0.89**	**0.92**	**0.8901**	15,000
0.9	0.95	0.88	0.91	0.8893	15,000

Table 8: Performances of ABLSTM and hybrid methods on different training sizes

Model	**Training Size**	**Macro Precision**	**Macro Recall**	**Macro F_1**	**Accuracy**
ABLSTM	12,000	0.78	0.60	0.66	0.7558
	60,000	0.87	0.70	0.76	0.8105
	120,000	0.93	0.83	0.87	0.8631
Hybrid method (threshold = 0.8)	12,000	0.88	0.70	0.76	0.7875
	60,000	0.93	0.82	0.86	0.8643
	120,000	0.96	0.89	0.92	0.8901

of regular expressions to avoid black boxes. Experimental results clearly demonstrate that our hybrid method outperforms ABLSTM and regular expression based models in medical text classification tasks because the proposed multi-stage system restricts the output accuracy from high to low at each stage. The proposed hybrid classification method also opens a window for domain experts to interfere with the classification process and fine-tune the solutions with the help of automated heuristic approach and interpretable regex rules.

We believe that using deep learning guided regular expressions to tap into the sequential relationships among salient words is an important area of research to help improve text classification performance and support accurate decision-making. Meanwhile, introducing human-in-the-loop to fine-tune the regex solutions with specialized domain knowledge would promote human-machine collaborative intelligence. The attention mechanism introduced in this paper could be utilized to compose high quality initial solution for heuristic regex evolution by linking word selection probability with its attention weight. With the extraction of high quality information and medical concepts as a basis, the sequential pattern mining can be extended to automated knowledge graph construction to learn high quality knowledge bases that link diseases and symptoms directly from electronic medical records. The application of the proposed system is not limited to text classification. Future investigation will be concentrated on using ABLSTM and regular expression rules to perform other NLP tasks such as entity recognition and relation classification in other domains by inducing general patterns of named entities.

Acknowledgments

The authors gratefully acknowledge Ping An Health Cloud for data and technical support. This work is supported by the National Natural Science Foundation of China (Grant No. 71471092), Natural Science Foundation of Zhejiang Province (Grant No. LR17G010001), and Ningbo Municipal Bureau of Science and Technology (Grant No. 2017D10034, 2019B10026, 2019F1028).

References

[1] A. Hall, G. Walton, Information overload within the health care system: a literature review, Health Information & Libraries Journal 21 (2) (2004) 102–108.

[2] M. Garofalakis, R. Rastogi, K. Shim, Mining Sequential Patterns with Regular expression constraints, IEEE Transactions on Knowledge and Data Engineering 14 (3) (2002) 530–552. doi:10.1109/TKDE.2002.1000341.

[3] M. Cui, R. Bai, Z. Lu, X. Li, U. Aickelin, P. GE, Regular Expression Based Medical Text Classification Using Constructive Heuristic Approach, IEEE Access 7 (2019) 147892–147904.

[4] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[5] J. D'Souza, V. Ng, Ensemble-based medical relation classification, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 1682–1693.

[6] B. Rink, S. Harabagiu, K. Roberts, Automatic extraction of relations between medical concepts in clinical texts, Journal of the American Medical Informatics Association 18 (5) (2011) 594–600.

[7] J. Kim, Y. Choe, K. Mueller, Extracting clinical relations in electronic health records using enriched parse trees, Procedia Computer Science 53 (2015) 274–283.

[8] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European conference on machine learning, Springer, 1998, pp. 137–142.

[9] J. Thorsten, T. Joachims, Learning to classify text using support vector machines, Comput Linguist 29 (4) (2002) 655–661.

[10] Y. Yang, C. G. Chute, An example-based mapping method for text categorization and retrieval, ACM Transactions on Information Systems (TOIS) 12 (3) (1994) 252–277.

[11] Y. Yang, Expert network: Effective and efficient learning from human decisions in text categorization and retrieval, in: SIGIR'94, Springer, 1994, pp. 13–22.

[12] K. Tzeras, S. Hartmann, Automatic indexing based on Bayesian inference networks, in: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1993, pp. 22–35.

[13] D. D. Lewis, M. Ringuette, A comparison of two learning algorithms for text categorization, in: Third annual symposium on document analysis and information retrieval, Vol. 33, 1994, pp. 81–93.

[14] F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys (CSUR) 34 (1) (2002) 1–47.

[15] R. Cohen, I. Aviram, M. Elhadad, N. Elhadad, Redundancy-aware topic modeling for patient record notes, PloS one 9 (2) (2014) e87555.

[16] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, R. A. Dudley, N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit, Journal of the American Medical Informatics Association 21 (5) (2014) 871–875.

[17] L. Wang, F. Chu, W. Xie, Accurate cancer classification using expressions of very few genes, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 4 (1) (2007) 40–53.

[18] Z. C. Lipton, D. C. Kale, C. Elkan, R. Wetzel, Learning to diagnose with LSTM recurrent neural networks, arXiv preprint arXiv:1511.03677.

[19] A. N. Jagannatha, H. Yu, Structured prediction models for RNN based sequence labeling in clinical text, in: Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing, Vol. 2016, NIH Public Access, 2016, p. 856. arXiv:1608.00612, doi:10.18653/v1/d16-1082.

[20] A. N. Jagannatha, H. Yu, Bidirectional RNN for medical event detection in electronic health records, in: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, Vol. 2016, NIH Public Access, 2016, p. 473. arXiv:1606.07953, doi:10.18653/v1/n16-1056.

[21] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in: COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers, 2014, pp. 2335–2344.

[22] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, arXiv preprint arXiv:1404.2188.

[23] M. Hughes, I. Li, S. Kotoulas, T. Suzumura, Medical text classification using convolutional neural networks, Stud Health Technol Inform 235 (2017) 246–250.

[24] M. Yu, M. Gormley, M. Dredze, Factor-based compositional embedding models, in: NIPS Workshop on Learning Semantics, 2014, pp. 95–101.

[25] K. Hashimoto, P. Stenetorp, M. Miwa, Y. Tsuruoka, Task-oriented learning of word embeddings for semantic relation classification, arXiv preprint arXiv:1503.00095.

[26] R. Johnson, T. Zhang, Semi-supervised convolutional neural networks for text categorization via region embedding, in: Advances in neural information processing systems, 2015, pp. 919–927.

[27] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Advances in neural information processing systems, 2015, pp. 1693–1701.

[28] T. Liu, S. Yu, B. Xu, H. Yin, Recurrent networks with attention and convolutional networks for sentence representation and classification, Applied Intelligence 48 (10) (2018) 3797–3806.

[29] G. Liu, J. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification, Neurocomputing 337 (2019) 325–

338.

[30] Y. Yu, M. Li, L. Liu, Z. Fei, F.-X. Wu, J. Wang, Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN, Journal of biomedical informatics 91 (2019) 103114.

[31] Y. Wang, H. Wang, X. Zhang, T. Chaspari, Y. Choe, M. Lu, An Attention-aware Bidirectional Multi-residual Recurrent Neural Network (Abmrnn): A Study about Better Short-term Text Classification, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3582–3586.

[32] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.

[33] M. J. Er, Y. Zhang, N. Wang, M. Pratama, Attention pooling-based convolutional neural network for sentence modelling, Information Sciences 373 (2016) 388–403.

[34] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.

[35] Y. Wang, M. Huang, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 606–615.

[36] M. Yang, W. Tu, J. Wang, F. Xu, X. Chen, Attention based LSTM for target dependent sentiment classification, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 5013–5014.

[37] J. Liu, Y. Zhang, Attention modeling for targeted sentiment, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 572–577.

[38] J. Y. Lee, F. Dernoncourt, P. Szolovits, Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks, arXiv preprint arXiv:1704.01523.

[39] C. Li, Z. Rao, Q. Zheng, X. Zhang, A set of domain rules and a deep network for protein coreference resolution, Database 2018.

[40] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, H. Liu, A clinical text classification paradigm using weak supervision and deep representation, BMC medical informatics and decision making 19 (1) (2019) 1.

[41] L. Yao, C. Mao, Y. Luo, Clinical text classification with rule-based features and knowledge-guided convolutional neural networks, BMC medical informatics and decision making 19 (3) (2019) 71.

[42] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Managementdoi:10.1016/0306-4573(88)90021-0.

[43] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval (1972). doi:10.1108/eb026526.

[44] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, nature 323 (6088) (1986) 533–536.

[45] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013. arXiv:1301.3781.