Vietnamese Document Classification Using Hierarchical Attention Networks

Khanh Duy Tung Nguyen(✉), Anh Phan Viet, and Tuan Hao Hoang

Le Quy Don Technical University, Hanoi, Vietnam
tungkhanhmta@gmail.com, anhpv@mta.edu.vn, haoht@lqdtu.edu.vn

Abstract. Automatic document classification is considered to be an important part of managing and processing document in digital form, which is increasing. While there are a number of studies addressing the problem of English document classification, there are few studies that deal with the problem of Vietnamese document classification. In this paper, we propose to employ a hierarchical attention networks (HAN) for Vietnamese document classification. The HAN network has the two-level architecture with attention mechanisms applied to the word level and sentence level from which it reflects the hierarchical structure of the document. Experimental results are conducted on the Vietnamese news Database which is collected from the Vietnamese news Web sites. The results show that our proposed method is promising in the Vietnamese document classification problem.

Keywords: Document classification · Hierarchical attention Networks

1 Introduction

Automatic document classification is to assign text into an appropriate category for easier management. It is a fundamental problem in the field of natural language processing. For supervised learning, given a set of documents $D = \{d_1, d_2, \ldots, d_n\}$ and the corresponding label set $C = \{c_1, c_2, \ldots, c_m\}$, the task is to find a mapping function $f : D \mapsto C$ that is able to capture the text meaning to decide the label for each document.

Automatic document classification studies are becoming increasingly important recently because of the increasing number of digital documents from a variety of sources. Automatic document classification largely supports in document management and processing in enterprises, organizations, and government agencies.

The traditional methods extract surface features such as n-gram or bag of words (BoW) and then apply common learning algorithms to build predictive models [5,11]. Recently, deep neural networks to automatically learning text

S. C. Satapathy et al. (eds.), *Frontiers in Intelligent Computing: Theory and Applications*, Advances in Intelligent Systems and Computing 1014,
https://doi.org/10.1007/978-981-13-9920-6_13

features have been applied efficiently for document classification [9,13]. Tang et al. compressed a document into a single vector by modeling it from words to sentences using Gated Recurrent Unit (GRU) networks [9]. Yang et al. applied GRU networks with attention mechanism to visualize the words and sentences according to the degree of their contribution to the document meaning [12].

Although there are many studies dealing with automatic document classification in English, this number is quite limited with Vietnamese documents. Vietnamese language processing is different from English, especially word separation, which is covered in Sect. 2.1. Current existing Vietnamese document classification methods rely primarily on Naive Bayes [3], SVM [4,8], neural networks [10] on features extracted from documents such as n-grams and bag of word. These methods depend heavily on the parameter selection and feature extraction process, nor does it focus on meaningful words that distinguish between the document classes. In order to focus on important words that are meaningful in distinguishing between document classes, we propose to employ a hierarchical attention networks (HAN) [12] in this research, that is, effective thanks to the attention mechanism and has proved to be effective in the problem of English document classification [13].

Our experimental results are conducted on the Vietnamese news dataset published by Hoang et al., in [4] that contains more than 33,000 news in 10 classes. The results show the prospect of our proposed method in solving the problem of Vietnamese document classification.

The rest of this paper is organized as follows. Section 2 presents our proposed method. Section 3 describes more details about the dataset and experimental setting. Results and discussion are presented in Sect. 4. Finally, Sect. 5 is the conclusion and future work.

2 Approach

In this section, we will present our approach to the problem of Vietnamese document classification. The whole process of Vietnamese document classification is described in Fig. 1. The Vietnamese document classification process is divided into three steps: Document preprocessing, data presentation, and apply the hierarchical attention networks.

2.1 Document Preprocessing

This is the basic step in the field of natural language processing. First, we remove special characters from the document. Second, we use vnTokenizer [6]—the good word segmentation for Vietnamese to segment the documents into words. The last step is to remove stopwords by using Vietnamese stopwords list available on the github[1] which contains nearly 2000 stopwords.

[1] https://github.com/stopwords/vietnamese-stopwords.

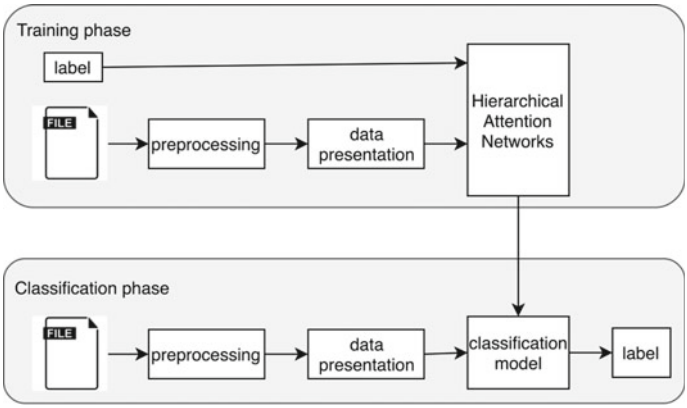

Fig. 1. Text classification process

2.2 Data Presentation

Data separation To apply the HAN network that applied two levels of attention mechanisms is word level and sentence level, and we also divide the input documents into two levels: sentence level and word level. The representation of the data as follows:

$$d = \{s_1, s_2, \ldots, s_n\}, s_i = \{w_{i1}, w_{i2}, \ldots, w_{im}\}, \quad i = \overline{1,n}$$

where d is input document, s_i is the ith sentence in d, n is the number of sentences in d, w_{ij} is jth word in s_i and m is the number of words in s_i.

Represent word into vector The input of the HAN network is the set of words in the sentences, which will be explained in Sect. 2.3. Thus, the primary input of the HAN network can be considered as words. So we need to convert these words into vectors and consider these vectors as the input of the network. There have been many studies concerning the expression of words into vectors [1,7]. To represent words into vectors, we use a pre-trained word vectors dataset which is available on github.[2] This dataset contains 294 languages (including Vietnamese), trained on Wikipedia using `fastText`. These vectors in dimension 300 were obtained using the skip-gram model described in [1] with default parameters.

2.3 Hierarchical Attention Networks

Figures 2 and 3 show the architecture of the hierarchical attention network for document classification [12]. The network generates the document vector representations by using two sequence-based encoders for words (Fig. 2) and sentences (Fig. 3), and then stacks a softmax layer for classification. The highlight

[2] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md.

of this model is the ability to visualize the important words and sentences due to the application of the hierarchical attention mechanism. The remainder of this section will describe the details of the network components.

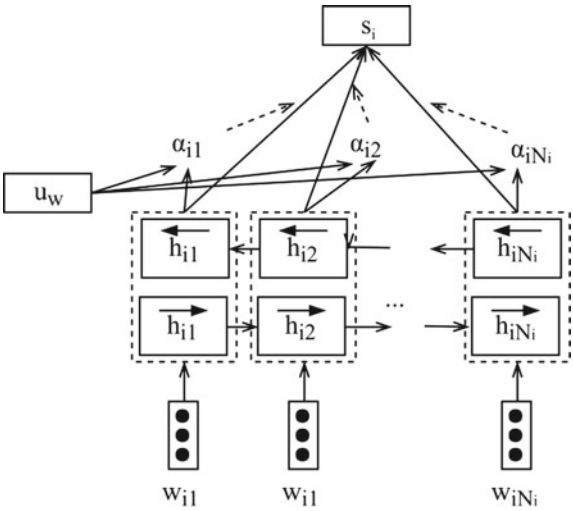

Fig. 2. Architecture of the hierarchical attention network: the sentence-level layers

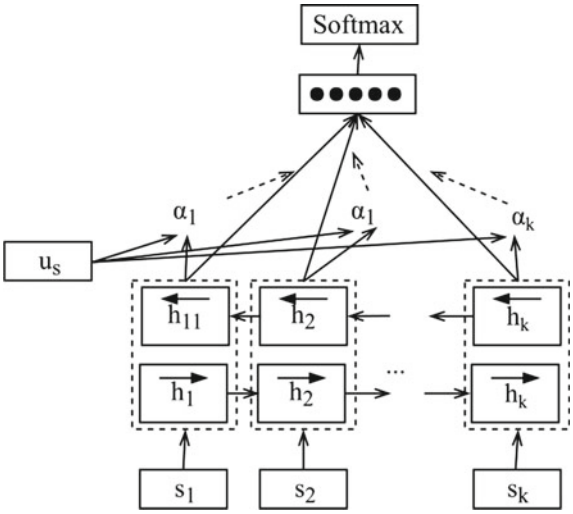

Fig. 3. Architecture of the hierarchical attention network: the document-level layers

Gated Recurrent Unit (GRU) is introduced by Cho et al. [2] to solve the problem of vanishing gradient coming with the standard recurrent neural network. GRU uses update gate and reset gate to control the amount of information passing to the output. In which, the update gate z_t decides the amount of past and new information being updated to the new state h_t. Meanwhile, the reset gate r_t is used to determine the amount of the past information to forget. At time step, t, z_t, and r_t are computed as follows:

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \tag{1}$$

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \tag{2}$$

where x_t is the vector of the tth element. $W^{(z)}/W^{(r)}$ and $U^{(z)}/U^{(r)}$ are the weights for x_t and previous state h_{t-1} in update/reset gates, correspondingly.

The current memory content uses the reset gate to store relevant information from the past:

$$h'_t = \tanh(Wx_t + r_t \odot (Uh_{t-1})) \tag{3}$$

The final memory at the current time step t now is computed from the added new information and past information as follows:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \tag{4}$$

Sequence encoder using bidirectional GRU Given a sequence $S = \{s_1, s_2, \ldots, s_N\}$, where s_i, $i \in [1, N]$, is represented as a real-valued vector, we use a bidirectional GRU to encode the contextual information of the sequence. To do that, the bidirectional GRU contains the forward GRU $\overrightarrow{f}$ and backward GRU $\overleftarrow{f}$ to read the sequence from left to right and inverse direction:

$$\overrightarrow{h}_t = \overrightarrow{\text{GRU}}(x_t) \tag{5}$$

$$\overleftarrow{h}_t = \overleftarrow{\text{GRU}}(x_t) \tag{6}$$

The state at the time step t is determined by concatenating the forward and backward hidden states, $h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$. The h_t contains the information of the whole sequence centered around x_t.

Attention mechanism The GRU reads a sequence and compresses all information to a single vector. Several elements may lead to the loss of information since they have different contributions to the meaning of the sequence. Applying attention mechanism partially solves this problem. It allows to look over the original sequence and focus on informative elements. To do this, a context vector c_t is plugged between the GRU and the encoded vector to compute the probability distribution for each element. Math is shown below:

$$\alpha_{ts} = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'=1}^{N} \text{score}(h_t, \bar{h}'_s)} \tag{7}$$

$$c_t = \sum_{s=1} \alpha_{ts} \bar{h}'_s \tag{8}$$

$$a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t]) \tag{9}$$

GRU hierarchical attention To build the classifier, firstly, we apply bidirectional GRU with attention at both sentence level and document level to encode a document to a fixed-length vector. After that, a fully connected layer and a softmax are stacked on the top for classification.

3 Experiments

3.1 Dataset

In this section, we describe the dataset we used to obtain our experimental results. While there are many public datasets for the problems of categories classification in English, this number for Vietnamese is very limited. Fortunately, there is a dataset of Vietnamese news articles published by Vu Cong Duy Hoang and his colleagues [4]. This dataset is available on github.[3]

The dataset was collected from four largest Vietnamese online newspapers: VnExpress,[4] TuoiTre Online,[5] Thanh Nien Online, and[6] Nguoi Lao Dong Online.[7] According to the authors presented in [4], preprocessing data was conducted automatically by removing HTML tags, stopwords, spelling normalization via various heuristics techniques, and Teleport software. The process then had been manually reviewed and corrected by linguists. Finally, the relatively large and sufficient corpus is obtained which includes more than 80,000 documents. The dataset has two levels: Level 1 and Level 2. Level 1 contains documents listed in the top 10 categories from popular news Web sites mentioned above. In this research, we use the dataset which is the training dataset from Level 1 and it is depicted in Fig. 4.

Understanding the dataset is very important for selecting parameters in our proposed models. Therefore, we made some statistics on the training dataset (including 33,759 articles). We saw that there are 70,860 unique words in 33,759 articles, each article contains about 23.7 sentences and each sentence contains about 9.81 words.

The dataset is randomly split by the ratio 3:1:1 for training, validation, and testing process.

3.2 Experimental Setups

This section describes some of the issues related to the implementation for our research.

[3] https://github.com/duyvuleo/VNTC.
[4] www.vnexpress.net.
[5] www.tuoitre.com.vn.
[6] www.thanhnien.com.vn.
[7] www.nld.com.vn.

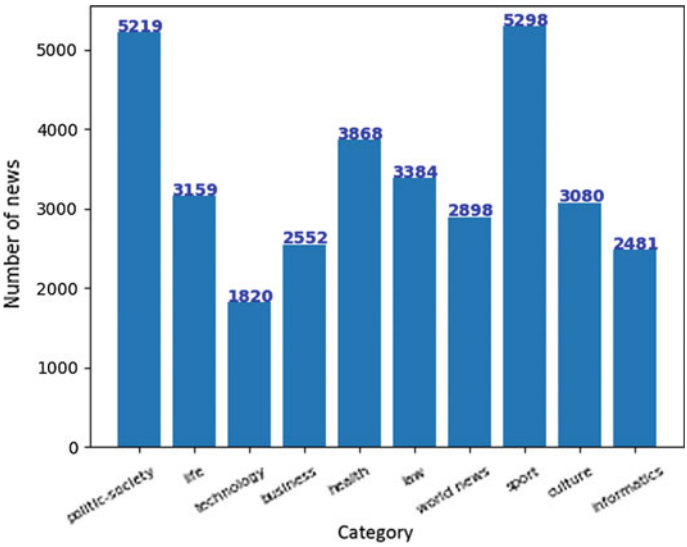

Fig. 4. Number of articles by category in training set

Algorithm 1 is used to represent each document into an $m \times n$ matrix. Then set of documents $\mathcal{D}$ becomes the set of matrices as shown in Fig. 5 where k is the number of documents, n is the maximum number of sentences (denoted by $MAX - SENTENCE$ in the algorithm) which is considered in a document, and m is the maximum number of words (denoted by $MAX - WORD$ in the algorithm) which is considered in a sentence and each word is a 300-dimensional vector. This means that documents longer than $MAX-SENTENCE$ sentences and sentences longer than $MAX - WORD$ words will be truncated and if the corresponding length is smaller, we make zero padding. This helps to limit the amount of computing. In line 3, we make a transformation of a word into a vector. We use `fastText` pre-trained word vectors provided by the Facebook AI Research group. In `fastText` file, followed by a word is a 300-dimensional vector as its representation. Based on the statistics in Sect. 3.1, we chose $MAX-SENTENCE = 15$ and $MAX-WORD = 100$, note that we only select 20,000 words with the highest frequency to create the dictionary.

We train and evaluate the HAN network through 20 epochs with the batch size of 50.

To evaluate our proposed approach, we also implemented other conventional classification algorithms including Naive Bayes, random forest, and SVM. In addition, the bag-of-words (BoW) feature is used in all conventional classification algorithms. We built the dictionary using 20,000 words with the highest frequency.

Algorithm 1 Document representation algorithm

INPUT: Raw of documents $\mathcal{D}$
OUTPUT: Matrix-formatted data of documents

```
docList ← {}
for d ∈ D : do
    docMatrix[n, m] ← Zero matrix
    for s_i ∈ d and i < MAX − SENTENCE do
        for w_ij ∈ s_i and j < MAX − WORD do
            if w_ij ∈ dictionary then
                docMatrix[i, j] = word2vec(w_ij)
    docList.add (docMatrix)
ruturn docList
```

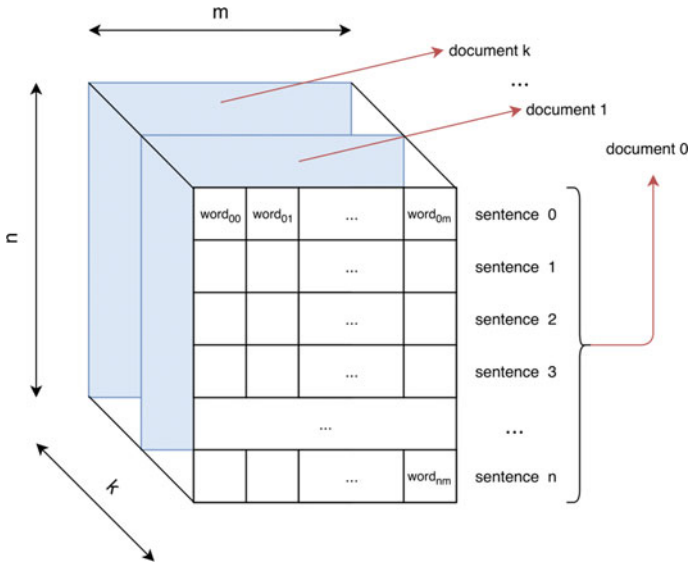

Fig. 5. Data presentation

4 Results and Discussion

Table 1 compares our method with some conventional machine learning approaches (random forest–BoW, Naive Bayes–BoW, SVM–BoW) according to accuracy, precision, recall, and F1-score. As can be seen, our approach outperforms the handcrafted-feature-based methods. According to our observation, BoW can obtain high performance because the words related to each topic are more frequently occurred in the articles belonging to such topic than in those of other topics. It is interesting that HAN network automatically learns the meaning of the text without any information of topic words, and can obtain higher performance.

Table 1. Comparison of approaches according to accuracy, precision, recall, and F1

Method	Accuracy	Precision	Recall	F1
Random Forest BoW	46.3	69.94	34.61	34.61
Naive Bayes-BoW	85.25	87.52	85.01	85.68
SVM-BoW	86.54	87.89	85.67	85.77
HAN	87.73	88.10	86.36	86.37

We also analyze the training process to check the convergence of the model. Figure 6 shows the accuracy on the training, validation, and testing sets through epochs. It can be seen, the model converges rapidly around 10 epochs and gets stable after that. The varying trend of validation and testing curves is similar. This shows the model stability to predict unseen instances.

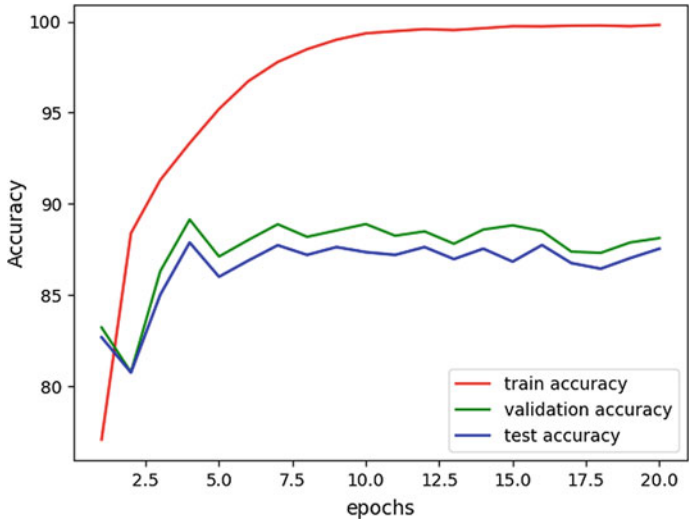

Fig. 6. Accuracy

Another improvement of our proposed model compared to conventional learning methods like Naive Bayes and SVM are the ability to extract important words in the classification process. To confirm that our model has the ability to select meaningful words for classification, we visualize the word attention layer. Figures 7 and 8 show results from some test data. We use different levels of yellow color to indicate the weight of words. The word highlighted in yellow has the greater weight comparing to others. The word marked with white color has almost zero weight. The results show that our model has the ability to select words containing meaningful information for the classification process.

For example, for the document in Fig. 7 with label 7, which denotes sports, our model accurately localizes the words huấn-luyện-viên (coach in English), đội-hình (team), and cầu-thủ (football player). For the document in Fig. 8 with label 8, which denotes culture, our model focuses on âm-nhạc (music), ca-sĩ (singer), hát (sing), and diva.

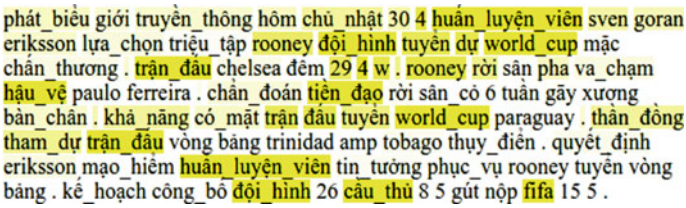
phát_biểu giới truyền_thông hôm chủ_nhật 30 4 huấn_luyện_viên sven goran eriksson lựa_chọn triệu_tập rooney đội_hình tuyển dự world_cup mặc chấn_thương . trận_đấu chelsea đêm 29 4 w . rooney rời sân pha va_chạm hậu_vệ paulo ferreira . chẩn_đoán tiền_đạo rời sân_cỏ 6 tuần gãy xương bàn_chân . khả_năng có_mặt trận đấu tuyển world_cup paraguay . thần_đồng tham_dự trận_đấu vòng bảng trinidad amp tobago thụy_điển . quyết_định eriksson mạo_hiểm huấn_luyện_viên tin_tưởng phục_vụ rooney tuyển vòng bảng . kế_hoạch công_bố đội_hình 26 cầu_thủ 8 5 gút nộp fifa 15 5 .

Fig. 7. A document with label 7 means sports

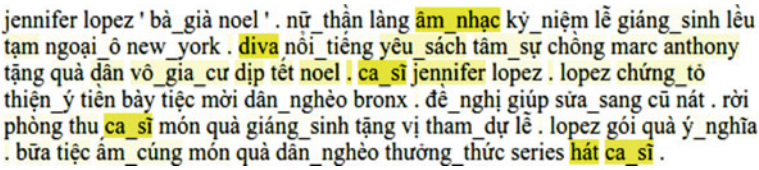
jennifer lopez ' bà_già noel ' . nữ_thần làng âm_nhạc kỷ_niệm lễ giáng_sinh lều tạm ngoại_ô new_york . diva nổi_tiếng yêu_sách tâm_sự chồng marc anthony tặng quà dân vô_gia_cư dịp tết noel . ca_sĩ jennifer lopez . lopez chứng_tỏ thiện_ý tiền bảy tiệc mời dân_nghèo bronx . đề_nghị giúp sửa_sang cũ nát . rời phòng thu ca_sĩ món quà giáng_sinh tặng vị tham_dự lễ . lopez gói quà ý_nghĩa . bữa tiệc âm_cúng món quà dân_nghèo thưởng_thức series hát ca_sĩ .

Fig. 8. A document with label 8 means culture

5 Conclusions and Future Works

Our study proposes a new and effective approach to the problem of Vietnamese document classification. We propose a new model with a two-level attention mechanism that is word level and sentence level, which illustrates the hierarchical structure of the document. We obtained better visualization using the highly meaningful words of a document. The results have shown the effectiveness of our proposed method. Visualization of attention layer demonstrates that our model is capable of selecting meaningful words for document classification.

Although our results show efficiency, it does not really harness the power of the hierarchical attention network (HAN), which is the ability to identify words and sentences that are focused and meaningful for the class. So, in the future, we will focus on analyzing and harnessing the power of the HAN network.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Ha, P.T., Chi, N.Q.: Automatic classification for vietnamese news. Adv. Comput. Sci. Int. J. **4**(4), 126–132 (2015)
4. Hoang, V.C.D., Dinh, D., Le Nguyen, N., Ngo, H.Q.: A comparative study on vietnamese text classification methods. In: 2007 IEEE International Conference on Research, Innovation and Vision for the Future, pp. 267–273, March 2007
5. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: European Conference on Machine Learning, pp. 137–142. Springer (1998)
6. Le, N.M., Do, B.N., Nguyen, B.D., Nguyen, T.D.: Vnlp: an open source framework for vietnamese natural language processing. In: Proceedings of the Fourth Symposium on Information and Communication Technology, pp. 88–93. ACM (2013)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
8. Nguyen, G.-S., Gao, X., Andreae, P.: Vietnamese document representation and classification. In: Australasian Joint Conference on Artificial Intelligence, pp. 577–586. Springer (2009)
9. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422–1432 (2015)
10. Van Toan, P., Thanh, T.M.: Vietnamese news classification based on bow with keywords extraction and neural network. In: 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), pp. 43–48. IEEE (2017)
11. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pp. 90–94. Association for Computational Linguistics (2012)
12. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: 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, pp. 1480–1489 (2016)
13. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems, pp. 649–657 (2015)