Tibetan text classification based on RNN

Wei Yan¹, Hui Cao* and Zeyu Cui¹

¹Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

*Corresponding author’s e-mail: caohui321@sina.com

Abstract. In this paper, a deep learning RNN model is used to classify Tibetan texts. The core idea is to first preprocess the Tibetan news corpus, and then use Tibetan syllables to construct a Tibetan syllable table based on the lexical and grammatical structure of Tibetan, embed the syllables in the sentence, and represent each syllable as a fixed Numerical vector. Secondly, the RNN cyclic neural network model is constructed. First, the text of different lengths is filled or truncated into a sequence length of uniform length. For each input text, the vector representation of text syllables is input on each time step of RNN to train the RNN model. The test samples were then used to evaluate the accuracy of model classification by introducing recall rate, precision rate and F-test. Finally, compared with traditional machine learning Logistic algorithm, polynomial naive Bayes algorithm and KNN algorithm, the results show that RNN model has better classification effect.

1. Introduction

With the rapid development of network information technology, the communication between different countries and ethnic groups has become more and more frequent. With the help of the express train of high-speed informatization, the development of Tibetan informatization has penetrated into every corner of The Tibetan ethnic group in an unstoppable manner, and the data of various electronic texts related to The Tibetan language has increased sharply. General Secretary Xi Jinping has stressed the importance of "building a strong sense of community among the Chinese nation" and protecting and developing the languages of ethnic minorities while strengthening education in the country's common spoken and written languages. Tibetan language is 5 million Tibetan compat riots native language, Tibetan mainly presents the news text in a network, different websites for all kinds of press forward, cause of Tibetan data on the Internet too onerous, it is difficult to obtain high accurate data information, we use computer technology to classify Tibetan texts for quick retrieval, access, sorting and analysis.

The basic classification method of text is matching according to some feature words of the data to be classified. It is impossible to achieve complete matching. Therefore, it is necessary to filter the matching results to select the best match, so as to achieve classification. From text categorization matching method and knowledge engineering method to statistical learning method, classification technology is becoming more and more mature. However, the statistical learning method uses word frequency information to classify, which is easy to cause problems such as sparse data features and semantic sensitivity. With the continuous maturity of deep learning technology, RNN neural network model shines in the field of natural language processing. It not only has strong feature extraction ability, but also can obtain contextual text information, thus improving classification accuracy.
2. Related work
At present, the Tibetan text classification method mainly adopts the traditional machine learning algorithm. In 2013, Wang Yong et al.[1]. Made reference to the existing classification corpus of Chinese, combined with the existing Tibetan corpus, and made improvements to the polynomial model. In feature selection, after considering synonyms, the text classification effect has been further improved. 2016, Gu Hongyun[2] from the SVM (support vector machine) for complex tedious Chinese characters set based on rapid classification of mature application, put forward the application of SVM to the Tibetan text classification, in the same condition with other classifiers experiment contrast, support vector machine (SVM) model is verified in the Tibetan text classification has a good classification effect. In 2019, Su Huijing et al.[7]. Constructed feature vectors of text words and processed them with relevant dimension reduction. Then, Euclidian distance algorithm was used to obtain the similarity between the prediction samples and the training samples, and finally, the category of samples was predicted according to the nearest neighbor voting principle of K. The experiment shows that KNN model has good performance in Tibetan text classification.

3. Data preprocessing
The Tibetan language combines syllables into one word by means of syllable symbols. Just like the Chinese language, there is no separation symbol between words. The research on Tibetan word segmentation is a key technology in Tibetan information processing, which has been studied for many years, but no product has been produced. The serial processing feature of RNN deep learning network extends the selection of language processing units. In this paper, the Tibetan representation obtains the syllables through the syllables, and then stores the syllables into the content list to construct the syllables table. In this paper, 10 kinds of news corpus related to Tibetan were obtained through crawler technology, and the category names and article contents were extracted respectively and sorted into a TXT text, so as to be easy to view and use. After sorting out the corpus text, we will extract the category tag list and the content list. The list of categories is shown in Figure 1.

![Figure 1. List of categories.](image1)

The list of contents is shown in Figure 2.

![Figure 2. List of content.](image2)

Next, a dictionary is built to convert syllables into sequences of numbers. The syllable dictionary is shown in Figure 3.

![Figure 3. Syllable dictionary.](image3)
Next, begin to convert the tags and syllables in the sentence into a sequence of numbers. Where the value in the tag is converted to a one-hot form. The one-hot sample table for the category label is shown in Figure 4.

![Figure 4. The category label One-hot sample table.](image)

The text data set is then converted to fixed-length ID sequence representation using the PAD_sequences provided by Keras, and Tibetan syllables are mapped to word_index table according to the syllable table index, as shown in Figure 5.

![Figure 5. Word_index sample table.](image)

Finally, the scrambled sample data is input to the RNN model and randomly sampled. The processed data format is as follows. The data format is shown in Table 1.

| Data    | Shape       | Data    | Shape       |
|---------|-------------|---------|-------------|
| x_train | [50000, 600] | y_train | [50000, 10] |
| x_val   | [5000, 600]  | y_val   | [5000, 10]  |
| x_test  | [10000, 600] | y_test  | [10000, 10] |

4. **RNN network model structure**

A total of 30 Tibetan letters consonants, 4 vowels symbols, Tibetan glyph structure with a letter as the core, the rest of the letters are based on additional and fold up and down before and after writing, combined into a complete word table structure, usually Tibetan text in between adjacent syllables as the basic unit, the content of the syllable point between the content may be a word or word, so you can use the word embedded get each syllable fixed dimension vectors. Tibetan text can be regarded as a sequence of data. In the traditional neural network model, from the input layer to the hidden layer and then to the output layer, the layers are fully connected, and the nodes between each layer are unconnected. But ordinary neural networks are powerless for many problems. For example, if you want to predict what the next word in a sentence will be, you usually need to use the first word, because the words before and after a sentence are not independent. However, RNN cyclic neural network is very good at dealing with sequence problems, and there is a strong correlation between the sequence data before and after. RNN is reflected by the sharing of weight and bias of each unit and cyclic calculation.
RNN is called a cyclic neural network, in which the current output of a sequence is also related to the previous output. The figure below is a typical RNN neural network structure diagram, as shown in Figure 6.

![Figure 6. RNN structure.](image)

As can be seen from the figure above, the RNN network contains an input $x$, an output $h$ and a neural network unit $A$. Different from the common neural network, the neural network unit $A$ of RNN network is not only related to the input and output, but also has a loop with itself. This network structure reveals the essence of RNN: the network state information of the last moment will act on the network state of the next moment. The RNN network can also be expanded as a time series. The RNN time series diagram is shown in Figure 7.

![Figure 7. RNN time series diagram.](image)

As you can see from the figure above, after the network receives the input $x_t$ at time $t$, the value of the hidden layer is $S_t$ and the output value is $O_t$. The thing to notice here is that the value of $S_t$ depends not only on the value of $X_t$, but also on the value of $S_{t-1}$. The following formula can be used to represent the calculation method of RNN network.

$$O_t = g(V \cdot S_t)$$  \hspace{1cm} (1)

$$S_t = f(U \cdot X_t + W \cdot S_{t-1})$$  \hspace{1cm} (2)

5. Experimental results and analysis

The experimental programming language python3.6, the deep learning framework Tensorflow1.14. The experimental data set is crawled from Tibetan news websites and divided into 10 categories, including sports, business, property, home, education, technology, fashion, politics, game, entertainment, with 6500 pieces of data in each category. From each of the 10 categories, 5000 pieces of data are selected as the training set, 500 pieces of data are selected as the verification set, and 1000 pieces of data are selected as the test set.

5.1. Network model parameters

The experimental training is batch training, and the parameters of RNN network model are shown in Table 2.
Table 2. RNN network model parameters.

| Parameter                  | Value  | Parameter               | Value  |
|----------------------------|--------|-------------------------|--------|
| Word vector dimension      | 64     | Number of hidden layers | 2      |
| Sequence length            | 600    | Hidden layer neurons    | 128    |
| Vocabulary size            | 5000   | Drop_out                | 0.8    |
| Learning rate              | 0.001  | Training size of each batch | 128 |
| loss function              | Cross entropy | Total iteration order | 10     |

5.2. Evaluation criteria

In order to evaluate the performance of the algorithm, the accuracy rate, recall rate, $F1$ value and confusion matrix are used as the evaluation criteria of text classification. In text classification, precision rate $P$ measures the precision of the category, which refers to the ratio of the number of samples correctly predicted to belong to the marked category $TP$ and the total number of samples predicted to belong to the category ($TP+FP$), where $FP$ is defined as the number of samples wrongly predicted to belong to a category by the classifier[8]. Recall rate $R$ measures the recall rate of a category, which refers to the ratio between the number of samples correctly predicted to belong to the marked category $TP$ and the number of samples actually belonging to the marked category ($TP+FN$), where $FN$ represents the number of samples actually belonging to the marked category but wrongly predicted to belong to other categories by the classifier. $F1$ value is a harmonic average that takes into account both precision and recall rates, with no preference for precision or recall rates. The mixed matrix includes all the categories in the experiment, and intuitively presents the process state of classification performance based on the correct and wrong number relation predicted by classification of each category[11]. The classification performance evaluation scale definition is shown in the following formula.

$$\text{Accurate rate: } P = \frac{TP}{TP + FP}$$  \hspace{1cm} (3)

$$\text{Recall rate: } R = \frac{TP}{TP + FN}$$  \hspace{1cm} (4)

$$\text{F1 value: } F1 = \frac{2P + R}{P + R}$$  \hspace{1cm} (5)

5.3. Results and evaluation of classification performance

The classification performance results include the classification performance mixed matrix and the overall classification performance. The classification process and results are used to evaluate the performance. The classification performance results are shown in Table 3 and Table 4.

Table 3. Classification performance mixed matrix.

|       | Sports | Business | Property | Home | Education | Technology | Fashion | Politics | Game | Entertainment |
|-------|--------|----------|----------|------|-----------|------------|---------|----------|------|---------------|
| Sports| 990    | 0        | 0        | 0    | 1         | 1          | 4       | 0        | 0    | 3             |
| Business| 3      | 980      | 1        | 1    | 2         | 3          | 2       | 8        | 0    | 0             |
| Property| 0      | 0        | 995      | 1    | 0         | 2          | 2       | 0        | 0    | 0             |
| Home  | 0      | 3        | 0        | 800  | 21        | 16         | 114     | 12       | 12   | 17            |
As can be seen from table 3 and table 4, the classification effect is very good, and the accuracy rate, recall rate and F1 value of each category, except for Home and Politics, are more than 0.9.

5.4. Performance comparison of classification models
In order to verify the effect of this model, Tibetan news corpus is segmented by Tibetan word segmentation software of Professor Qi Kunyu of Northwest University for nationalities. TF-IDF method is used for feature selection. Three traditional machine learning algorithms, logistic algorithm, polynomial naive Bayes algorithm and KNN algorithm, are selected to compare the classification effect. The comparison of model classification effect is shown in Table 5.
It can be seen from table 5 that on the same data set, the RNN model has better classification effect. Three traditional machine learning methods, logistic regression, polynomial naive Bayes and KNN, are based on statistical methods. The classification is based on word frequency and inverse document frequency characteristics, without semantic information. In this paper, the RNN model extracts the context features of sentences and contains the potential semantic information of sentences. Therefore, the classification effect is better. Experimental results show that the classification accuracy of this model is 1 to 3 percentage points higher than that of traditional machine learning methods.

6. Conclusion
For Tibetan text, the language structure is complex. Traditional machine learning only mechanically uses statistical methods to deal with Tibetan text classification problems, so its performance is relatively weak. The deep learning RNN model automatically solves the problem of feature extraction, not only can include the context information of the sentence, but also can capture the long-distance dependency in the sequence by constructing a deeper convolution layer. But at the same time, since the output of the last time step of the RNN model depends on the output of the previous time step, the speed of model training is slow. Secondly, although the training of sequence data can be solved theoretically, the problem of gradient disappearance exists, which is especially serious when the sequence is very long. These questions need further study.

References
[1] Wang Yong. Research on Tibetan Text Classification based on Naive Bayes [D]. Northwest University for Nationalities, 2013.
[2] Zhou Deng. Research on Tibetan text classification technology based on N-gram model [D]. Northwest University for nationalities, 2010.
[3] Jia Hongyun. Research and implementation of Tibetan text classification based on AdaBoost model [D]. Tibet University, 2019.
[4] Li Liang. Tibetan Pre-training Model based on ALBERT and its Application [D]. Lanzhou University, 2020.
[5] Wan, Q.B., Dong, F.M., Sun, S.F. (2020) Text classification method based on BiLSTM-Attention-CNN hybrid neural network. J. Computer applications and software., 37(09):94-98+201.
[6] Su, H.J., Qun, N., Jia, H.Y. (2019) Research and implementation of Tibetan text classification based on Gaussian NB model. J. Journal of Qinghai Normal University (NATURAL SCIENCE EDITION), 35 (04): 1-4 + 54.
[7] Su, H.J., Qun, N., Jia, H.Y. (2019) Research and implementation of Tibetan text classification based on KNN model. J. Plateau science research., 3(02):88-92.
[8] Qun, N., Jia, H.Y. (2018) Research and Implementation of Tibetan Text Classification based on Logistic Regression Model. J. Information and Computers (Theoretical), (05):70-73.
[9] Bao, H., Xi re Dan, Z., Guo, L.Y., Shang, H.J. (2019) Tibetan text classification based on random forest. J. Computer knowledge and technology., 15 (34): 178-180.
[10] Jia, H.Y., Qun, N., Su, H.J. (2018) Tsering Luozeng, Basang Zhuoma. Research and Implementation of Tibetan Text Classification based on SVM. J. Electronic Technology and Software Engineering., (09):144-146.
[11] Amer, A.A., Abdalla, H.I. (2020) A set theory based similarity measure for text clustering and classification. J. Journal of Big Data., 7(1):52-1590.
[12] Asim, M.N., Ghani, M.U., Ibrahim, M.A., et al. (2020) Benchmark Performance of Machine And Deep Learning-Based Methodologies for Urdu Text Document Classification. J. Neural Computing and Applications., :1-33.
[13] Ni, P., Li, Y.M., Chang, V. (2020) Research on Text Classification Based on Automatically Extracted Keywords. J. International Journal of Enterprise Information Systems (IJEIS), 16(4):1-16.
[14] Shanavas, N., Wang, H., Lin, Z.W., Hawe, G. (2020) Knowledge-driven graph similarity for text classification. J. International Journal of Machine Learning and Cybernetics., :1-15.

[15] Li, Q., Li, P.F., Mao, K.Z., Lo, E.Y.M. (2020) Improving convolutional neural network for text classification by recursive data pruning. J. Neurocomputing., 414:143-152.