Research on text classification of telecom user query

Zhenhua Liu*
Beijing Research Institute, China Telecom Co., Ltd., Beijing, China

ABSTRACT

Text classification plays a more and more important role in many practical applications. However, there are a large number of unevenly distributed but important data about the complaints of Chinese telecom users. How to efficiently and accurately implement these telephone consultation contents with only a small number of samples to specific departments and improve the user experience has become an urgent problem to be solved. In this paper, we propose a multi model-based machine learning method—telecom textCNN predictor which can effectively improve the accuracy and recall of the model on the long tail data, and automatically mark the text, so as to accurately locate the class of the problem. The results show that our model is helpful for business personnel to quickly identify the type of complaint business, conduct targeted business acceptance and user services, and improve user perception and produces better accuracy than the SVM model.

Keywords: Text classification, textCNN, SVM, deep learning, telecom

1. INTRODUCTION

Text categorization is becoming more and more crucial in numerous practical applications. Many businesses are considering creating apps that use text categorization techniques in the digitisation process, especially in light of recent developments in natural language processing (NLP) and text mining[1]. Some studies suggested enhancing SVM to significantly increase classification accuracy and speed up information search among diverse online Chinese language and literature resources with high matching degree[2]. While Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM) and Ant Colony Optimization (ACO) based Ranking Algorithm are used in some studies to create a classification-based model that is inspired by deep learning and generates extremely accurate and exact results[3]. To determine the topics that customers are most interested in when purchasing online, text mining was utilized to extract various inquiries from Q&A systems and online comments[4]. Although deep learning has been used to enhance text classification ability, there is still room for improvement in fields like law[5]. Important information processing systems are widely used, which can be said to be everywhere. However, in the complaint of Chinese telecom users, this problem can be defined as a short text multi classification problem[6]. There are a large number of unevenly distributed data. How to efficiently and accurately implement these calls to specific departments according to the consultation content and improve the user experience has become an urgent problem to be solved.

In this data set shown in Figure 1, some categories have 1500 data, while some categories have only 10 or even one data, which is the problem of data imbalance.

In this paper, we first propose a feature engineering scheme based on TFIDF (term frequency–inverse document frequency) and MLP (Multilayer Perceptron) model, then propose an optimization model based on textCNN neural network.

2. RELATED WORK

Text classification is becoming more and more crucial in numerous practical applications. Ghiassi[7] offers a comprehensive method for sentiment analysis of Twitter and spam filtering of YouTube comments that incorporates a new clustering technique, another clustering algorithm (YAC2), and a transferable domain feature engineering approach. A method of question-answering put out by Kim[8] automatically provides users with statistics on infrastructure degradation from textual input. Stitini[9] came to the conclusion that the relationship between contextual data and classification strengthens and improves the outcomes of recommendations.

* liuzhh11@chinatelecom.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125061P
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661973

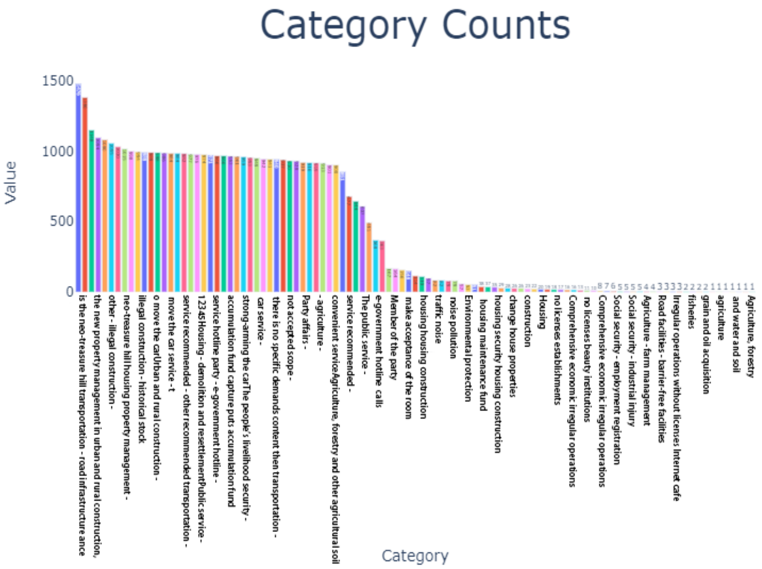

Figure 1. The distribution of the telecom user query dataset.

The majority of strategies now in use concentrate on creating intricate deep learning models to explore the relationship between context and target elements. Convolutional neural networks (CNN) and natural language processing (NLP) are the foundation of Du's novel automatic defective text categorization system[10] (AutoDefect), which uses a hierarchical two-level encoder. For learning text representations level-by-level, Ma[11] developed a level-by-level HMTC technique based on a two-way gated recursive unitary network model with a hybrid embedding.

The lack of training data and imbalanced data present a barrier in the Mechatronics text classification task. Bilal[12] proposed a churn prediction model based on a combination of clustering and classification algorithms employing an ensemble. Wang[13] created TTCP, which uses a unique loss function that takes both the quantity of predictions and the location of the ground truth label into account. Furthermore, the outcomes did not much improve as a result of these sophisticated models.

3. METHOD

3.1 TF-IDF

A weighting method that is frequently used in information retrieval and information exploration is term frequency inverse document frequency. A statistical technique called TF-IDF can be used to determine how important a word is to a document inside a corpus or group of documents. A word's significance rises in direct proportion to how frequently it appears in a text, but it falls in direct proportion to how frequently it appears in a corpus.

The word "term frequency" (TF) denotes how frequently an entry appears in the text. The TF formula is used to express this number, which is typically normalized (typically word frequency divided by the total number of words in the article) to avoid bias against long documents (the same word may occur more frequently in a long document than in a short one, regardless of whether the word is important or not):

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

whereas, $n_{i,j}$ represents the number of times the term t_i appears in the document d_j, and $TF_{i,j}$ represents the frequency of the term t_i in the document d_j.

3.2 SVM

A Support Vector Machine (SVM) is a binary classification model. Its basic model is a linear classifier with maximum intervals defined in the feature space, which makes it different from a perceptron; the SVM also includes kernel techniques, which makes it essentially a non-linear classifier.The learning strategy of the SVM is interval maximisation, which can be formalised as a problem for solving convex quadratic programming, which is also equivalent to the minimisation of the regularised hinge loss function.The learning algorithm of the SVM is an optimisation algorithm for solving convex quadratic programming.

$$L(w,b,a) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} a^i \left(1 - y_i(w^T x_i + b)\right) \tag{2}$$

3.3 TextCNN

Textcnn model is a model that uses convolutional neural network to deal with NLP problem, which is proposed in the article “Convolutional natural networks for sense classification” proposed by Yoon Kim Compared with the traditional RNN/LSTM models in NLP, CNN can extract important features more efficiently, which occupy an important position in classification. The architecture of textCNN is shown in Figure 2.

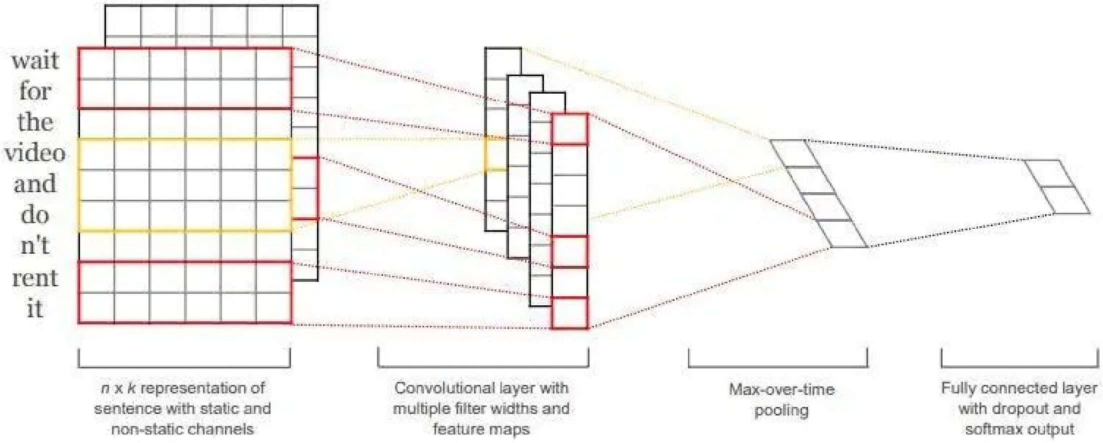
wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 2. The architecture of textCNN model.

A phrase can be thought of as a series of words first. The sequence is n words long, each word is represented by a vector $\boldsymbol{x_i}$, and k is the dimension that each word carries. Therefore, the phrase is expressed as:

$$x_{i:n} = x_1 \oplus x_2 \oplus \cdots \oplus x_n \tag{3}$$

So $x_{i:i+j}$ is a left closed and right closed interval.

The width of the convolution kernel and the dimension of the word embedding are the same since textCNN uses a one-dimensional convolution. The number of words used in each window is represented by the convolution kernel’s height h. Convolution kernel, then $\mathrm{w} \in \mathrm{R}^{\mathrm{hk}}$.

The outcome of the convolution operation is, for each sliding window result (scalar) $\mathrm{c_i}$

$$\mathrm{c_i} = \mathrm{f}\left(\mathrm{wx_{i:j+h-1}}\right) + \mathrm{b} \tag{4}$$

where$b \in R$, and f is a nonlinear function, such as tan$h$ or ReLU.

4. EXPERIMENT

4.1 Baseline

In this experiment, we use real data to test our predictions in real-world applications. In 2021, we collected the content records of telecom calls in 13 cities in a month. They are recorded as 20 dimensions, namely id, order number, question, reply, handler, handling date, type of question, etc. From these features we extract effective data information. We trained our model with a total of 7W pieces of data, and then tested it on 49 categories of data.

We use ranking metrics such as precision recall fi-score to check whether our classifier can accurately identify these 49 categories. We are comparing the accuracy generated by linear, randomforest, and MLP as our baselines.

Table 1. Baseline comparison.

Model	Precision	Recall	F1-score
Svm	82%	83%	77%
MultinomialNB	74%	74%	74%
Logistic	75%	75%	72%
RandomForest	71%	71%	70%

As shown in Table 1, in our comparative experiment, we can see that the SVM model has achieved the best classification results. Our SVM model has achieved 82% precision, 83% recall and 77% F1 score, while the worst is the performance of randomforest model, which has only 70% precision, 70% recall and F1 score.

According to the results of baseline, we can know that our feature engineering is still good, but there is still room for further optimization.

4.2 TextCNN versus baseline

We set the max feature of the embedding layer to 20000, the word vector size of the convolution layer to 64, and the pool size of max poolingid_ to 2. We set three dense layers. After a non-linear variation of the dense layers, we extract the association between these features and finally map them to the output space for classification. The structure of the model is shown in Figure 3. We used 10,000 real data for testing, including 8,000 training sets, 1,000 validation sets, and 1,000 test sets. The precision and loss curves of the model are shown in Figure 4. The model achieved the best in the 8th round. The effect, the precision reached 91%.

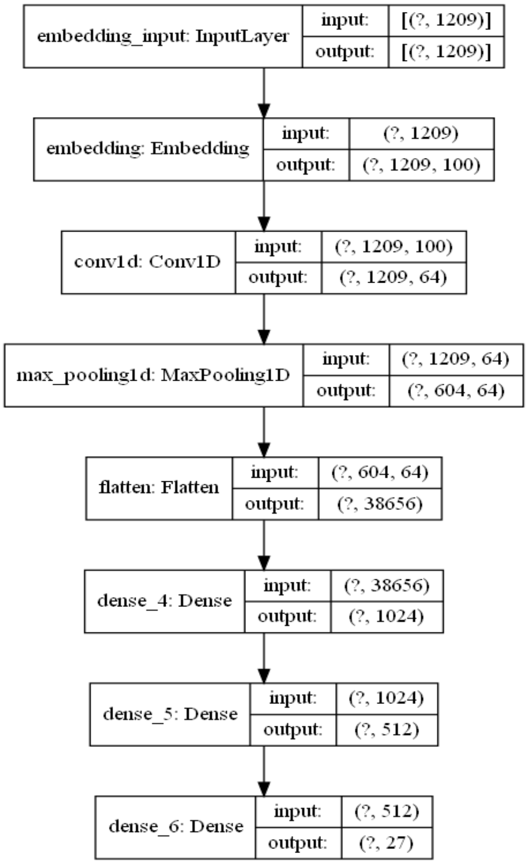

Figure 3. The architecture of our telecom textCNN predictor model.

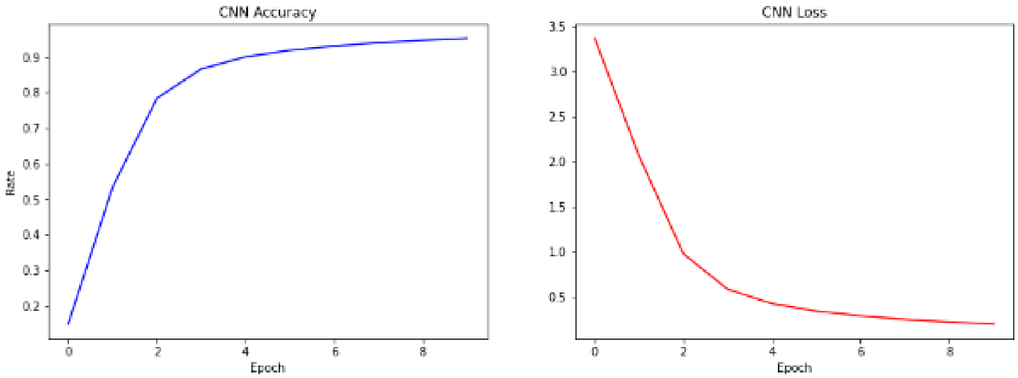

Figure 4. The accuracy and loss curves of our telecom textCNN predictor model.

5. CONCLUSION

To sum up, first of all, we propose a TF-IDF+SVM-based telecom text classification model, which is helpful for business personnel to quickly identify the type of complaint business, conduct targeted business acceptance and user services, and improve user perception. Then we use the extracted features to train the textCNN model and optimize our TTCP model. The results show that our model produces better accuracy than the SVM model.

REFERENCES

[1] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., et al., “Text classification algorithms: A survey,” Information, 10(4), 150(2019).

[2] Li, X., “Chinese language and literature online resource classification algorithm based on improved SVM,” Scientific Programming, (2022).

[3] Vaish, K., Deepak, G. and Santhanavijayan, A., “DSEORA: Integration of deep learning and metaheuristics for web page recommendation based on search engine optimization ranking,” Emerging Research in Computing, Information, Communication and Applications, Springer, Singapore, 873-883(2022).

[4] Chen, Y., Liu, D., Liu, Y., et al., “Research on user generated content in Q&A system and online comments based on text mining,” Alexandria Engineering Journal, (2022).

[5] Chen, H., Wu, L., Chen, J., et al., “A comparative study of automated legal text classification using random forests and deep learning,” Information Processing & Management, 59, (2022).

[6] Song, G., Ye, Y., Du, X., et al., “Short text classification: A survey,” Journal of multimedia, 9(5), 635(2014).

[7] Ghiassi, M., Lee, S. and Gaikwad, S. R., “Sentiment analysis and spam filtering using the YAC2 clustering algorithm with transferability,” Computers & Industrial Engineering, 107959(2022).

[8] Kim, Y., Bang, S., Sohn, J., et al., “Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers,” Automation in Construction, 134, 104061(2022).

[9] Stitini, O., Kaloun, S. and Bencharef, O., “Integrating contextual information into multi-class classification to improve the context-aware recommendation,” Procedia Computer Science, 198, 311-316(2022).

[10] Yang, D. U., Kim, B., Lee, S. H., et al., “AutoDefect: Defect text classification in residential buildings using a multi-task channel attention network,” Sustainable Cities and Society, 103803(2022).

[11] Ma, Y., Liu, X., Zhao, L., et al., “Hybrid embedding-based text representation for hierarchical multi-label text classification,” Expert Systems with Applications, 187, 115905(2022).

[12] Bilal, S. F., Almazroi, A. A., Bashir, S., et al., “An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry,” PeerJ Computer Science, 8, e854(2022).

[13] Wang, K., Liu, Y., Cao, B., et al., “TkTC: A framework for top-k text classification of multimedia computing in wireless networks,” Wireless Networks, 1-12(2022).