

Research on text summarization classification based on crowdfunding projects

Gang Zhou*

School of Business, Central South University, 410000, Changsha, Hunan, China

Abstract. In recent years, artificial intelligence technologies represented by deep learning and natural language processing have made huge breakthroughs and have begun to emerge in the field of crowdfunding project analysis. Natural language processing technology enables machines to understand and analyze the text of crowdfunding projects, and classify them based on the summary description of the project, which can help companies and individuals improve the project pass rate, so it has received widespread attention. However, most of the current researches are mostly applied to topic modeling of project texts. Few studies have proposed effective solutions for classification prediction based on abstracts of crowdfunding projects. Therefore, this paper proposes a sequence-enhanced capsule network model for this problem. Specifically, based on the work of the capsule network, we propose to connect BiGRU and CapsNet in order to achieve the effect of considering both the sequence semantic information and spatial location information of the text. We apply the proposed method to the kickstarter-NLP dataset, and the experimental results prove that our model has a good classification effect in this case.

1 Introduction

Crowdfunding is a new model of Internet finance. Its operating principle is that the promoters openly raise capital from the public through the Internet platform to support the development of the project, and provide investors with equity and in-kind returns after the project is finally successful. Crowdfunding provides new financing channels and marketing methods for individual entrepreneurs and start-ups. The crowdfunding market has huge potential. Kickstarter, the world's largest crowdfunding platform, has raised \$1.3 billion for more than 70,000 projects, with more than 7 million investors participating. Regrettably, more than 50% of projects still fail to raise funds[1]. The effect of crowdfunding financing determines the rise and fall of crowdfunding platforms. Crowdfunding behavior is largely determined by investors' subjective factors, and an important factor that affects subjective judgment is the persuasiveness of language. This is a typical user-generated content (UGC), and the project initiator can describe the project in any type of language style. Different language styles will change investors' perception of project prospects, and in turn affect their willingness to invest.

* Corresponding author: zgcsu353862@163.com

On the Internet, users often face information overload. In the crowdfunding community, there are a large number of projects going online every day, and investors cannot read the detailed text of each project carefully, and will only selectively browse the introduction of some projects. The text introduction of the crowdfunding project provides a function similar to the abstract of an academic paper. When reading academic papers, readers generally choose to read the abstract first, and detailed research papers only if they can arouse interest[2,3]. The project summary has attracted the attention of investors for its short and conciseness, and it has also been valued by the crowdfunding platform. In crowdfunding projects, investors will also read the introduction of the crowdfunding project first, and then continue to research the project and participate in investment if they are interested. In practice, project sponsors use different forms to describe and promote projects. In view of the important role of the project summary, it is worth studying according to its persuasiveness[4,5].

This article uses text mining methods to predict the success of crowdfunding project abstracts, hoping to help entrepreneurs increase their success rate in the near future[6,7]. This paper proposes a sequence-enhanced capsule network model for modeling. The proposed method is evaluated on a large public data set containing 215,513 projects on the Kickstarter platform. Our contributions can be summarized as follows:

1.A method of predicting success based on the summary of crowdfunding projects is proposed to assist entrepreneurs to improve the success rate of projects.

2.A sequence-enhanced capsule network model is proposed. We propose to connect BiGRU and CapsNet in order to achieve the effect of considering the sequence semantic information and spatial location information of texts at the same time, achieving better feature extraction results, and improving prediction performance.

2 Method

In this section, we first introduce our distributed feature extraction method in Section 2.1. Next, a sequence-enhanced capsule network model is introduced in Section 2.2. The model frame is shown in Fig. 1.

2.1 Distributed feature extraction method

Text feature representation is very important for text classification, and it directly affects the classification results. Traditional text representation still stays at the surface processing of vocabulary. For example, the Vector Space Model (VSM) expresses text data through a vector composed of feature words and weights, and the Bag-of-Words model (Bag-of-Words), BoW) takes text as a collection of several words and the words are independent of each other. Traditional artificial feature extraction methods ignore the contextual information of the text and cannot express its relevance, which has a great impact on the semantic expression of the text. Some researchers map the original high-dimensional feature word space to a low-dimensional latent semantic space or topic space to explore the potential semantic structure of the text. For example, Latent Semantic Analysis (LSA) represents text as a semantic vector in a low-dimensional latent semantic space [8], which improves its sparsity while reducing dimensionality and denoising. However, the process of dimensionality reduction may cause classification problems. Loss problem, and the semantics of each dimension in the semantic space is not clear. The Latent Dirichlet Allocation (LDA) model expresses the text as a probability distribution of implicit topics [9], but its effect on the sparsity of short texts is limited. Literature [10] proposed the Word2Vec model, which can train the word vectors in the corpus quickly and efficiently. Compared with the LSA and LDA models, the Word2Vec model can effectively use contextual semantic

information. While solving the problem of dimensionality, it will not ignore The relative position of phrases in the corpus retains the semantic relevance of phrases.

Word2Vec includes Skip-Gram and Continuous Bag-of-Words Model (CBOW) two models, which can quickly train word vectors according to the given corpus. This paper maps the data set text to a 256-dimensional word vector matrix. The word vector matrix is used as the input of the model.

2.2 Distributed Feature Extraction Method

In this section, the capsule network (capsule) [11]and the bidirectional gated recurrent unit network (BiGRU)[12] are combined, the capsule network is used to extract the local features of the text, and the BiGRU network is used to extract the global features of the text. The two levels of the extracted Feature fusion obtains multi-level features of the text.

The bidirectional gated recurrent unit network (BiGRU) is a bidirectional gated-based recurrent neural network composed of a combination of forward GRU and backward GRU. The text is traversed through the network in two directions to obtain information containing the text context, which solves the problem that the GRU model can only contain the above information. The GRU model is a variant of the long short-term memory network (LSTM). The gated loop unit GRU is the 3 gates of LSTM[13] reduced to 2, which are respectively the update gates z_t , reset gate r_t .

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

The update gate is used to control the degree to which the state information at the previous moment is substituted into the current state. The larger the update gate value, the more state information is substituted at the previous moment; the reset gate is used to control the degree of discarding the information of the previous state, and the reset gate the smaller the value, the more information is discarded. The final hidden layer state h_t and \tilde{h}_t sum be calculated as:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{3}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \tag{4}$$

Expression x_t is the output at time t equals, h_{t-1} is the output of the hidden layer at the previous time. W_t and U_h is the weight matrix of the forget gate, b_t is offset term the forget gate, σ is Activation function called sigmoid.

Compared with LSTM, the GRU model is simple, has fewer parameters, and has high training efficiency. Therefore, this article chooses GRU as a tool for sequence information learning. The Bi-GRU model used in this article consists of a forward-propagating GRU and a backward-propagating GRU. The BiGRU module can fully input the relevant information before and after the sequence, extract the text information from two directions, and obtain the global features of the text vector.

In the capsule network, the convolution operation is first performed, and the capsule convolution operation is performed through the main capsule layer. After the squeeze

function operation, it is used as the output of the main capsule layer. After the dynamic routing protocol mechanism is calculated, it is connected to the classification capsule layer to classify the capsule. capsules share their weights with each other. The low-level feature information obtained through the initial capsule layer is aggregated into the high-level capsule layer (TextCaps) with an 8-dimensional vector through dynamic routing. Dynamic routing performs 3 iterations of important spatial hierarchical relationships between PrimaryCaps and TextCaps. The corresponding dynamic routing algorithm is given in Table 1, and the method used to calculate the coupling coefficient is described.

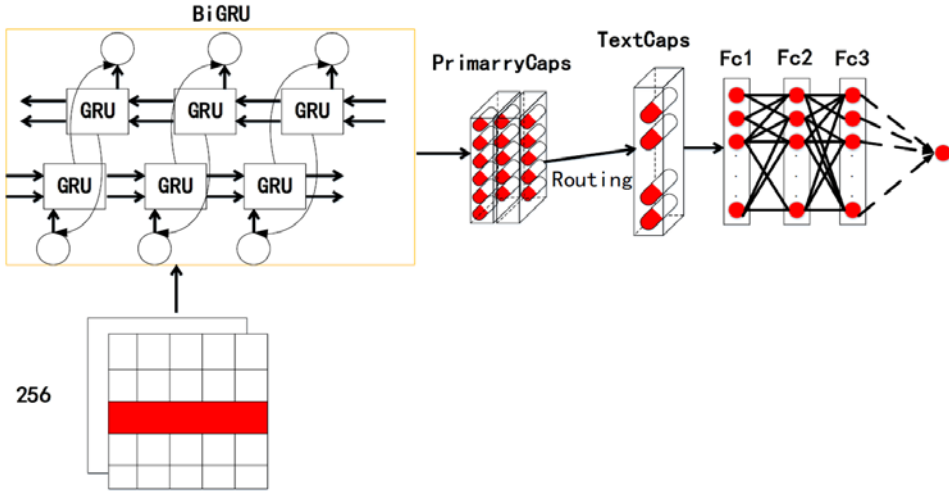


Fig. 1. Our model architecture.

The output result of the layer is expanded as a local feature vector of the text. As the input of the fully connected network, the last layer of the fully connected network uses the sigmoid function as the classifier for judgment.

Table 1. Dynamic routing algorithm.

Algorithm 1: Routing algorithm

Require: low-level capsules u , iterative number z

Ensure: high-level capsules r

1: Initialization: $b_{ij}=0$

2: for z iterations do

3: $c_{ij} = \text{soft max}(b_{ij}) \quad \forall u_i \in u$

4: $u_{j|i} = w_{ij}u_i \quad \forall u_i \in u$

5: $s_j = \sum_{ij} c_{ij} \hat{u}_{j|i} \quad \forall r_j \in r$

6: $v_j = \text{squash}(s_j) \quad \forall r_j \in r$ **cf.(5)**

7: $b_{ij} = b_{ij} + u_{j|i} v_j \quad \forall r_j \in r$ **cf.(6)**

8: end for

3 Experimental results and analysis

3.1 Dataset and Experimental Setup

In order to evaluate the performance of the model on the task of text similarity, this article uses the kickstarter-NLP dataset. Kickstarter was established in New York, USA in April 2009. It is a crowdfunding website platform designed to raise funds for companies with creative solutions. The kickstarter-NLP data set contains 215,513 crowdfunding projects, of which 172411 were successful, with the label set to 1, and 43102 failed, set to 0. In the experiment of this article, the data set is divided, 80% as the training set, 20% as the test set. The experiment carried out in this paper is implemented based on the keras framework.. It uses the Adam optimizer. The number of model iterations is 15. And the Dropout rate is set to 0.5.

3.2 Evaluation Index

This article uses accuracy, precision, recall and F1 value (F-measure) as the evaluation indicators of the model. Let TP be the number of correct classes predicted as correct; TN be the number of wrong classes predicted as wrong; FP be the number of correct classes predicted as correct; FN be the number of correct classes predicted Number. The calculation formula of the evaluation index is as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

3.4 Results analysis

Compared Methods To demonstrate the superiority of our TransCap for ASC tasks, we compare it with followings baselines:FastText (Joulin et al.,2016)[14], TextCNN (Kim et al.,2014)[15], Transformer (Vaswani et al.,2017)[16], RCNN(Girshick et al.,2014)[17], BLSTM-Attention (Zhou et al.,2016)[18], Most of them are frequently-used classical models.

The comparative experimental results of the model are shown in Table 2. It is clear that our BiGRU-Capsule model consistently performs better than all baselines on our kickstarter-NLP dases.The structure can more fully extract the information of the text context,and extract spatial location information.This method uses dynamic routing methods to store and cluster features, and encapsulate the features into a set of sentence-level semantic capsules. The similarity between the input of the lower-level capsule and the output of the higher-level capsule is obtained by the dot product algorithm, so the lower-level capsule will send its output to the similar high-level capsule, and the capsule in the classified capsule layer contains the highest The characteristics of the order, each capsule represents the category that the text may belong to, so as to achieve a higher accuracy. As shown in Fig.2. we also

compared the performance under different batches, it is clear that our method performs better when its batch is 128. Simultaneously BiGRU-Capsule performs better than its branch structure.

Table 2. Comparison of different methods.

model	Accuracy	precision	Recall	F1
FastText	0.77	0.76	0.769	0.772
TextCNN	0.78	0.776	0.786	0.781
Transformer	0.81	0.816	0.801	0.82
RCNN	0.79	0.79	0.78	0.795
BLSTM-Attention	0.82	0.826	0.819	0.826
BiGRU	0.8	0.79	0.784	0.796
Capsule	0.84	0.839	0.856	0.825
BiGRU-Capsule	0.87	0.864	0.854	0.877

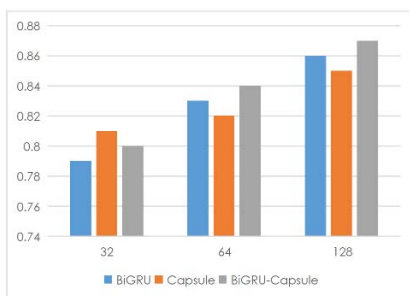


Fig. 2. Testing effects under different batches.

4. Conclusion

This paper proposes a sequence-enhanced capsule network model for text classification tasks based on crowdfunding project summaries. Based on the work of the capsule network, considering the influence of the sequence semantic information and spatial location information of the text, it is recommended to connect BiGRU and CapsNet in order. This strategy achieves good results and is superior to other compared models. This work will increase the pass rate of crowdfunding projects for entrepreneurs in the near future.

References

1. Cumming D , Hornuf L . The Economics of Crowdfunding || Crowdfunding Creative Ideas: The Dynamics of Project Backers[J]. 2018, 10.1007/978-3-319-66119-3(Chapter 8):151-182.
2. Ohtani, Bunsho. Preparing Articles on Photocatalysis — Beyond the Illusions, Misconceptions, and Speculation[J]. Chemistry Letters, 2008, 37(3):216-229.
3. Luke S G , Henderson J M , Ferreira F . Children's eye-movements during reading reflect the quality of lexical representations: An individual differences approach.[J]. Journal of Experimental Psychology Learning Memory & Cognition, 2015, 41(6):1675-1683.
4. Engineering; Findings from Beijing Wuzi University Provide New Insights into Engineering (An Integration Model Based On Graph Convolutional Network for Text Classification)[J]. Journal of Engineering,2020.

5. Zhao W , Ye J , Yang M , et al. Investigating Capsule Networks with Dynamic Routing for Text Classification[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.
6. Greenberg M D , Pardo B , Hariharan K , et al. Crowdfunding support tools: predicting success & failure[C]// CHI '13 Extended Abstracts on Human Factors in Computing Systems. 2013.
7. Yuan H , Lau R Y K , Xu W . The determinants of crowdfunding success: A semantic text analytics approach[J]. Decision Support Systems, 2016, 91(nov.):67-76.
8. Landauer, Thomas K.Dumais, Susan T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.[J]. Psychological Review, 1997.
9. Putri I , Kusumaningrum R . Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia[J]. Journal of Physics Conference, 2017, 801:012073.
10. Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer ence, 2013.
11. Sabour S , Frosst N , Hinton G E . Dynamic Routing Between Capsules[J]. 2017.
12. She D , Jia M . A BiGRU method for remaining useful life prediction of machinery[J]. Measurement, 2020, 167:108277.
13. Graves A , Mohamed A R , Hinton G . Speech Recognition with Deep Recurrent Neural Networks[J]. Acoustics Speech & Signal Processing. icassp.international Conference on, 2013.
14. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. 2016. "Bag of Tricks for Efficient Text Classification," arXiv: Computation and Language. 900
15. Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
16. Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.
17. Girshick R , Donahue J , Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// CVPR. IEEE, 2014.
18. Zhou P , Shi W , Tian J , et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016.