Hindawi
Journal of Sensors
Volume 2021, Article ID 8013337, 7 pages
https://doi.org/10.1155/2021/8013337

Research Article

An Extended Text Combination Classification Model for Short Video Based on Albert

Yi Liu,[1] Yue Zhang,[2] Haidong Hu,[2] Xiaodong Liu,[1] Lun Zhang,[2] and Ruijun Liu[2]

[1]*Beijing Research Institute of Automation for Machinery Industry Co, Ltd, Beijing, China*
[2]*Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing, China*

Correspondence should be addressed to Ruijun Liu; liuruijun@btbu.edu.cn

Received 5 August 2021; Accepted 16 September 2021; Published 16 October 2021

Academic Editor: Haibin Lv

With the rise and rapid development of short video sharing websites, the number of short videos on the Internet has been growing explosively. The organization and classification of short videos have become the basis for the effective use of short videos, which is also a problem faced by major short video platforms. Aiming at the characteristics of complex short video content categories and rich extended text information, this paper uses methods in the text classification field to solve the short video classification problem. Compared with the traditional way of classifying and understanding short video key frames, this method has the characteristics of lower computational cost, more accurate classification results, and easier application. This paper proposes a text classification model based on the attention mechanism of multitext embedding short video extension. The experiment first uses the training language model Albert to extract sentence-level vectors and then uses the attention mechanism to study the text information in various short video extensions in a short video classification weight factor. And this research applied Google's unsupervised data augmentation (UDA) method based on unsupervised data, creatively combining it with the Chinese knowledge graph, and realized TF-IDF word replacement. During the training process, we introduced a large amount of unlabeled data, which significantly improved the accuracy of model classification. The final series of related experiments is aimed at comparing with the existing short video title classification methods, classification methods based on video key frames, and hybrid methods, and proving that the method proposed in this article is more accurate and robust on the test set.

1. Introduction

Short videos refer to short videos of 3 to 5 minutes in length, which are produced by users of short video APP using shooting clips. Major short video apps in China include Tik Tok, Bilibili, Weibo, and WeChat, which have recently launched short video functions. Foreign platforms include YouTube and Twitter. Its content covers all aspects of people's daily life, as a new means of social entertainment, and it is easy to arouse people's resonance. The length of the video is generally less than 5 minutes or 10 minutes. In the beginning, the creators of short video started to fight with KOL alone, and then the media circle, the talent circle, the brand circle, and the general public took an active part in the short video industry, showing a trend of "universal participation." As a way of Internet content transmission, the content and field of short video are constantly subdivided, presenting a trend of all flowers in content vertical categories. With the increase of competitive pressure, major video platforms are paying more and more attention to the diversity and pertinence of content, constantly expanding the emerging category market, and attracting attention to niche content and specific categories, so as to meet the diversified needs of users in this segmented era.

At present, watching short videos has become the main way of public entertainment, and the communication method of commenting on the bullet screen after watching short videos is becoming more and more popular. When people participate in this way of interaction, they also produce a huge amount of short video extended text data (video title, video introduction, video comment, and video barrage). How to mine the valuable information in the massive video extended text data is a valuable research direction. Video classification based on video extended text data can solve the problem of information overload, enable content service

providers to provide users with more personalized content recommendation services, analyze and mine valuable information in the data, and at the same time bring more benefits to the short video platform. The requirements for short video classification mainly include the following aspects: first, the videos uploaded by users shall be classified into the existing classification label organization system of the video platform, as shown in Figure 1. Second, we need to carry out multilabel text classification and multiclass text classification on the basis of the existing classification labels. Multiclass refers to the category of the short video platform to which the video belongs. Multilabel classification needs to select an appropriate label for the video from the existing label system of the short video platform. As shown in the figure, this video belongs to the financial management category on the channel. In addition, it also contains a series of labels such as finance, knowledge, learning, required production, finance, investment and financial management, fund, and A-share. Third, different from previous video classifications, different short video classifications focus not only on video background and character segmentation, but also on the video provided by HMDB, UCF101, and other data sets [1, 2]. For short videos, the classification is multiple and complex. The same person is chatting, it may be about investment, and it may be about feeling. To sum up, the previous data sets and training methods are not suitable for the current short video classifier that takes entertainment as the boundary.

2. Research Status at Home and Abroad

Learning the features of short videos requires a large number of training samples, and there are few short video data sets with correct labels and short video descriptions. Therefore, model learning is faced with great difficulties. Currently, commonly used data sets containing short video descriptions include MPII and MSR-VTT [3], which contain videos of different behaviors and video descriptions. The content of this paper is the short video classification method of short video platform. Currently, there is no publicly labeled data set, so you need to build the data set required for the research. The former text-short video clustering method mostly uses unsupervised or semisupervised method, and the category information of short video is mostly learned from text-short video. However, at present, major platforms have corresponding short video classification systems, and how to classify according to the established classification system is the current problem. Supervised learning can satisfy this situation, but due to the lack of correctly labeled data, the recall rate of the model does not meet the value of use. There are three existing short video clustering methods. One is based on the visual similarity of short video key frames, the second is based on short video title text clustering, and the last method is based on the multimodal fusion of text and vision. Short video clustering method based on title text clustering is widely used in the industry due to its simplicity and high efficiency. However, due to the sparse semantic characteristics of short text, the final clustering effect of short video titles is not good.

3. Paper Structure

This paper focuses on the use of rich text information (bullet screen, comments, thumb up number, etc.) on short video websites to improve the existing short text method of high-dimensional semantic sparse problem, to achieve efficient short video topic clustering. In this paper, a clustering method of short video topics was proposed, which fused several kinds of short text information, such as short video titles, comments and bullet screen, and a large amount of unlabeled data was introduced by using the data enhancement method based on knowledge graph on the basis of semisupervised learning to improve the robustness of the model. Finally, the validity of the proposed method is verified by the real data of Bilibili short video website.

4. Related Work

4.1. Albert. The structure of Albert's whole model still follows the skeleton of Bert [4], using Transformer [5] and GELU [6] activation function. There are three specific innovations: one is the factorization of embedding parameters, the sharing of parameters across layers, and finally the abandonment of the original NSP task and the use of SOP task.

4.2. Attention Mechanism. Deep learning model simulates the learning mechanism of human brain [7]. The mechanism of attention mimics the part of the brain that people first notice when they hear a word, a picture, or an audio clip. This is where the famous attention mechanism model comes in. The attention mechanism is most commonly used in natural language processing (NMT) (machine translation) [8], among which the most classic model is the Seq2Seq model. The Seq2Seq model is a typical encoder-decoder framework. In this framework, the encoder encodes the input into the context variable C, and each output Y is decoded indiscriminately using this C. What the attention model needs to do is to encode the encoder into different C according to each time step of the sequence. When decoding, combined with each different C, the decoding output will be carried out, so that the result will be more accurate.

4.3. Semisupervised Learning. Current semisupervision methods can be summarized into three categories:

(1) Graph label propagation method based on graph convolution and graph embedding [9, 10]

(2) The target data is used as the latent variable for prediction [11]

(3) Force consistency/smoothness

Forced smoothing only makes the model less sensitive to smaller noises [12, 13]. A common approach is to add some noise to a sample and force the model's output to be as similar as possible with and without noise. Intuitively, a good model should be able to accommodate small perturbations that do not change the properties of the sample. And there is always a variety of different scenarios because of the perturbation function.

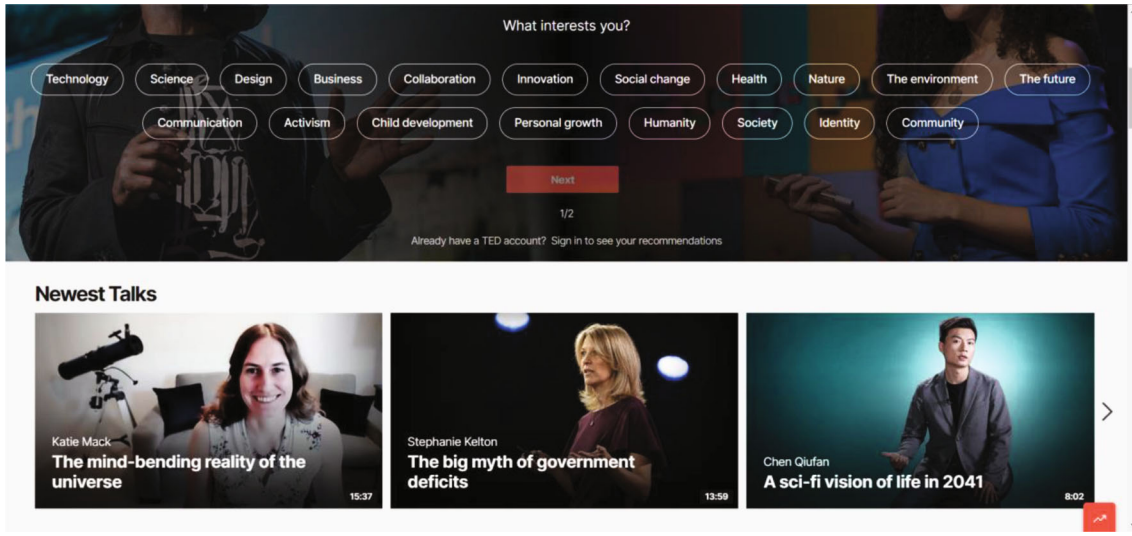

Figure 1: The source of the picture is TED (an example of a short video platform classification).

Learn the perturbation function from supervised data and get the optimal data enhancement method. A good data enhancement method can greatly improve the results of the model, and the data enhancement method can be applied in various fields. To enhance the unlabeled data, the main way is to do TF-IDF word replacement for the topic words in the unlabeled data through Chinese knowledge graph. The optimization method used is to minimize the KL divergence between the enhanced data and the real data. The training method is implemented based on the Bert model. Expand the training set with unlabeled master data. Compared with each model under supervised learning, it is found that its evaluation index on the test set, including precision rate, recall rate, and F1 value, is better than that of the model under supervised learning training.

5. System Framework

5.1. Albert-Based Extended Textual Ensemble Classification Model for Short Videos. The overall operation process of the model is as follows: first, input text data (title, description, comment, and bullet screen) and then use the Albert pretraining model to obtain the dynamic sentence vector containing the overall information of the text. The output of Albert model has two forms: one is character-level vector, that is, the vector representation corresponding to each character of the input short text. The other is the sentence-level vector, that is, the vector of the leftmost [CLS] special symbol output by Albert model. Albert model believes that this vector can represent the semantics of the whole sentence, as shown in Figure 2.

Assume that the input text T is sent to Albert pretraining model according to sentence granularity, and the vector of special symbol [CLS] at the leftmost of Albert model output is obtained. Then, the text is transformed into dimensional sentence vector matrix. Intuitively, the titles and introductions of short videos have a high correlation with the categories of short videos, followed by comments and bullet screens. In order to learn the correlation weight of these four types of texts from the sample data, the attention mechanism is introduced here. The first technique is a factorization of the embedding parameterization. The researcher decomposes the large lexical embedding matrix into two smaller matrices, thus separating the size of the hidden layer from the size of the lexical embedding. This separation makes it easier to add hidden layers without significantly increasing the number of lexical embedding parameters. The second technique is cross-layer parameter sharing. This technique avoids the increase in the number of parametrizations with the increase in network depth. The weights of the four types of texts are determined based on the number and frequency of their occurrence in the data set. After the vectorization of the text data is completed, the number of comments and barrage of different short videos is not consistent, and a short video only has one title and one introduction. For the task of this paper, the category of each group of text data is usually mainly determined by the title and the introduction sentence. The purpose of introducing the attention mechanism is to enhance the weight of these parts in the classification process, while reducing the weight of the barrage and comments, so as to achieve better classification effect.

The model architecture is shown in Figure 3: the model consists of three parts: the feature extraction layer, the encoder, and the decoder. The input to the model is text data, which is fed to the Albert pretraining model after adding tokens. The contextual information is put into the word embedding to obtain a dynamic sentence vector containing the overall information of the text. Then, the new sentence vector is input to the coding layer for feature extraction to capture the feature information of the text, and finally, the attention mechanism is introduced to obtain the final probabilistic representation of the input text for the purpose of text classification. The encoder is a two-layer LSTM [14, 15] (long-short-term memory, LSTM) network [14]. The decoder has the same architecture, and the attention layer is used to link between the encoder and the

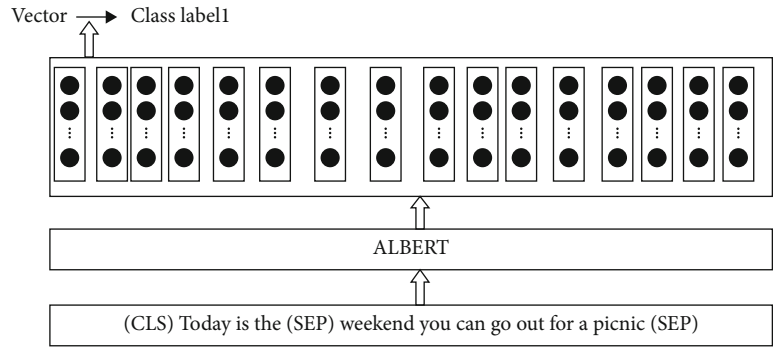

Figure 2: Albert model believes that [CLS] vector can represent the semantics of the whole sentence.

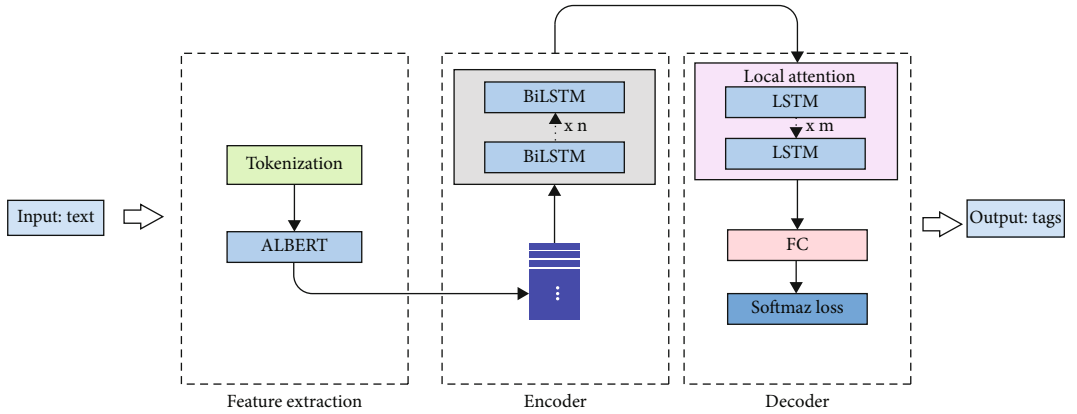

Figure 3: The structure of the model.

Table 1: Number of data sets.

Data set	Training set	Validation set	Test set	Unlabeled data set
Title data set	1800+	100+	100+	30000+
Description data set	1800+	100+	100+	30000+
Comment data set	180000+	10000+	10000 +	1000000+
Barrage data set	N	N	N	N

Table 2: Categories of data.

Category number comparison table Category	Number
1. Finance and economics	01
2. Real estate	02
3. Stock	03
4. Education	04
5. Science and technology	05
6. Social	06
7. Politics	07
8. Sports	08
9. Game	09
10. Entertainment	10

decoder [16]. It is the last hidden state of the encoder. The final output of the current decoder time step is obtained by cascading input to the feedforward neural network between the output of the current decoder time step and the context vector of the current time step [17]. Use the additive/Concat formula for scoring:

$$c_t = \sum_{j=1}^{T_x} a_{tj} h_j, \tag{1}$$

$$a_{tj} = \frac{\exp\left(e_{tj}\right)}{\sum_{k=1}^{T_x} \exp(e_{tk})}, \tag{2}$$

$$e_{tj} = V_\alpha^T \tanh\left(W_a \left[s_{t-1}; h_j\right]\right), \tag{3}$$

where c_t is the semantic vector at moment t and e_{ij} is the influence degree of the state of the hidden layer S_t of the encoder at moment j in the encoder on the state of the hidden layer h_t at moment t in the decoder, and then, the probability of e_{ij} is normalized into a_{ij} by the softmax function.

The encoder records the key sentence vector as it reads the text. Likewise, the decoder reads the same text while noting the key sentence vector. The encoder reports to the decoder as it reads each word. After completing feature selection, they

aid	author	mid	tag	title	class	description
202962238	Xu Da Prawn	354765	Indie Games, Sing	Prawn Xu teaches	9	Weibo @XuPrawnLol
203107467	The Ark of Tom	775300	Games, Mobile Ga	Tomorrow's Ark" r	9	Recasting the Future The Ark ofTomorrow
245575879	Oracle Handica	744700	Game Tips, Mobile	4K scenery, you kr	9	The 4K live recording vlog is here! The wo
245613767	Lao Dai is here	142762	4K, MSI, Games, S	4K "Cyberpunk 20	9	Updated the first episode, let everyone wa
245640078	Rohan Explanat	896064	Ancient Style, Mot	Old Rohan went t	9	The audience forgives me for being so ma
288172435	Cold snow with	140869	Original God, Mot	Battle Mage Barb	9	BGM: kimeru - Make You FreeNiko - Nig
330515456	Old Tomato	546198	Live Tips, Single pl	Medieval Overlorc	9	Game: Medieval Dynasty

Figure 4: Video information table.

rpid	member	comment	aid	mid	like	sex	class	time
3800459177	Pigeons do not co	Waited for more than 50 years to	755610233	168687092	1896	Confidential	9	2020/12/10 12:19:21
3811455931	North wind	Ma you choose the female role is	755610233	441674406	990	Confidential	9	2020/12/13 10:01:30
3805965283	to be a cat in the r	I awakened strange attributes. Be	755610233	41711951	1395	Female	9	2020/12/11 21:40:48
3808785459	Beeping dry cup	Who remembers things is Pepi w	755610233	398097933	966	Confidential	9	2020/12/12 16:53:25
3805780508	Jinjin diary	What makes you think no one lik	755610233	20423532	2596	Male	9	2020/12/11 22:26:00
3800415324	green mountain	Good guys, waiting for more tha	755610233	299973946	1246	Female	9	2020/12/10 12:03:15
3809147522	have a summer an	V said it was too hard to say whe	755610233	501321155	848	Male	9	2020/12/12 18:27:41
3800433758	white Feather	Live contentist matters to avoid r	755610233	12022519	1208	Confidential	9	2020/12/10 12:10:43
3802544458	Adrianna's little he	Don't worry, today is the day of p	755610233	6019912	1526	Confidential	9	2020/12/13 16:37:13

Figure 5: Comment information table.

classify the text documents sentence by sentence based on the jointly selected combined key sentence vector.

5.2. Semisupervised Training Method Based on Knowledge Graph Data Enhancement. In order to enhance the robustness of the model, we used an unsupervised data enhancement method based on unsupervised data in the training process to solve the difficulty of obtaining data without publicly labeled data sets and labeled data. UDA [18] is a new technology proposed by Google, which has proved its effectiveness in the NLP field. Furthermore, due to the particularity of short video extension text data, such as colloquial bullet comments, extensive use of memes, and the creation of short video titles in order to attract attention, and the use of the latest hot events to increase popularity, we can try instead commonly used unsupervised data enhancement reverse translation, TF-IDF word replacement to enhance the data. The TF-IDF [19] (Frequency-Inverse Document Frequency) algorithm is a weighted statistical method commonly used in information retrieval and data mining. It is used to evaluate the importance of words to a text or corpus. The importance of a word is proportional to its frequency of occurrence in the text, but inversely proportional to its frequency of occurrence in the corpus. The main idea is that if the word or phrase occurs more frequently in the text (high TF value) and rarely in other words or phrases (low DF value and high IDF value), the word or phrase is considered to represent the text well and can be used for classification. In order to ensure access to the latest structured entity data, CN-DBpedia provides an API interface for academic research. When inquiring, we directly access CN-DBpedia API to query relevant entity information. Consistency training is adopted throughout the training process. Firstly, the unlabeled data is enhanced by TF-IDF word substitution for the subject words in the unlabeled data through the Chinese knowledge graph. This method is different from the previous reverse translation method in the NLP field. Then, both the preenhanced data and the postenhanced data are sent into the network to produce a prediction result. The two results are calculated as a KL divergence as the unsupervised cross entropy loss, and the supervised consistency loss is added together to make back propagation (BP). Augment data and KL divergence of unlabeled data using minimized unlabeled data. Equation (4) is the final consistency loss function, as follows:

$$\min_\theta \Gamma(\theta) = E_{x,y^* \in L}[-\log p_\theta(y^* \mid x)] + \lambda E_{x \in U} E_{\hat{x} \sim q\left(\hat{x}, x\right)} \left[D_{\mathrm{KL}} \left(p_\vartheta(y \mid x) \middle\| p_\vartheta\left(y \mid \hat{x}\right) \right) \right]. \tag{4}$$

By minimizing the loss of consistency, data enhancement enables labeled information to be transferred from labeled data to unlabeled data. For most experiments, λ is set to 1 and different batch sizes are used for labeled and unlabeled data. Use a larger batch for unlabeled data results in better training results.

Due to the emergence of related texts in spoken short videos, the use of a large number of popular words, and the stigma of daily use, it is difficult to effectively implement unlabeled enhanced data when traditional back translation methods are used for data enhancement. In order to overcome this situation, the TF-IDF Chinese word knowledge map replacement method based on CN-DBpedia is used here.

6. Experimental Analysis

6.1. Data Set. In this chapter, two kinds of data sets are used in the experiment design: annotated data set (labeled data) and unlabeled data set. There are currently no publicly labeled data sets in the field of short video classification and short video text research. Therefore, the annotated data

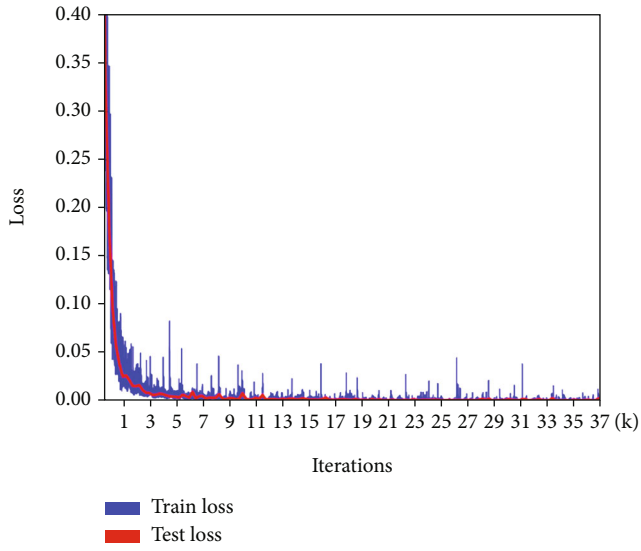

Figure 6: Convergence process of the model.

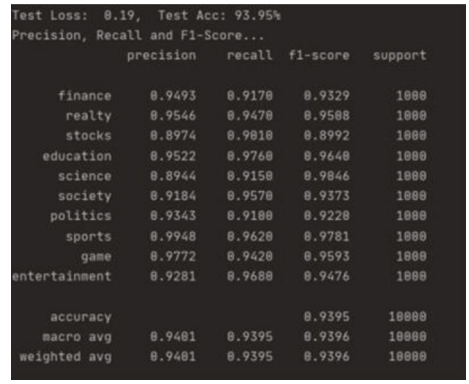

Figure 7: Accuracy, recall, and F1.

used in this paper comes from the data of Bilibili website obtained by the crawler. However, unlabeled data is easier to obtain than labeled data, and there is no need to consider labeling it when crawling. Therefore, we can directly crawl unlabeled comment data and bullet screen data on the Bilibili portal website. Then, mix and stack the current public NLP Chinese classification data set and crawled data set and scramble its order. An unannotated data set for verifying the semisupervised learning training method is based on CN-DBpedia knowledge graph data enhancement.

Tag Data has so far prepared 10 categories of data, 200 in each category, and captured more than 2,000 short videos. For each short video, get all comments and the number of likes up of the corresponding comments. Since there is no specific number of barrage data, it is determined according to the short video. The unannotated data is superimposed and shuffled using several publicly labeled data sets currently on the web. The data set is detailed in Table 1, and categories are shown in Table 2.

The data set consists of two sources: crawler acquisition and public data set. There are already public data sets that do not need to be processed. There is a lot of noise in the text data obtained by crawler. In order to improve the convergence speed of the model and reduce the calculation cost, the model has a lot of noise. In this paper, the original data are preprocessed, including processing special characters, filtering repeated or meaningless bullet screen text, and Chinese word segmentation. The data samples are shown in Figures 4 and 5.

6.2. Model Parameter Initialization and Convergence Process. The initial training parameters are set as follows: Epoch = 10, batch_size of supervised learning = 128, batch_size = 512 when there is no supervision, and Iter. When it goes to 2500, Val loss = 0.2 and does not decrease, and the model converges. Figure 6 is the convergence process of the model in the comment data, and Figure 7 is the accuracy rate, recall rate, and F1 value of the model in the ten categories of the validation set.

Table 3: Comparison on the test set between the multitext fusion model based on attention mechanism and other Bert-based composite models trained by supervised learning and semisupervised learning methods.

Model	Val Acc	Val loss	Test Acc	Test loss	Supervised/ semisupervised learning
Ours	94.08%	0.2	93.95%	0.19	Semisupervised
Albert	93.66%	0.29	91.63%	0.21	Semisupervised
Albert	92.97%	0.29	82.88%	0.44	Supervision
Albert _CNN	90.93%	0.3	80.31%	0.51	Supervision
Albert _DPCNN	91.93%	0.26	82.54%	0.47	Supervision
Albert _RNN	91.41%	0.3	82.22%	0.47	Supervision
ERNIE	90.16%	0.26	80.66%	0.51	Supervision

We compared the effects of a variety of composite models based on Albert proposed in recent years when the supervised learning method was used for training in the test set and found that the Albert model alone had the best effect, and details are in Table 3. Then, the Albert model and the multitext fusion model based on attention mechanism were trained by the semisupervised learning method. Finally, the submodel trained by semisupervised learning had the best performance, and its performance on the verification set was only slightly decreased, indicating that its generalization ability was also very good. Among them, the accuracy of the multitext fusion model based on the attention mechanism is 93.95% in the test set, which proves that our model is superior to other models.

7. Conclusions

We are in the era of information rapid growth, and for short video classification task, accuracy and efficiency are extremely important indicators. Traditional methods based on short video content understanding have high accuracy, but their efficiency is difficult to cope with the current growth rate of short video. The classification method based on the title of short video is faced with the dilemma of sparse meaning, and its accuracy is generally difficult to guarantee.

The deep learning method improves the accuracy by improving the model, while the hybrid method further improves the accuracy at the cost of time consumption. Therefore, this paper proposes a multitext embedded short video-related text classification model based on attention mechanism, which is aimed at reducing the time consumption of short video text classification as much as possible on the premise of guaranteeing high classification accuracy. Then, in order to ensure the quality and speed of text enhancement model training, a semi-supervised training method based on knowledge graph data enhancement was used, and a large number of unlabeled data were introduced to fully improve the robustness of the model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Beijing Natural Science Foundation (grant number 4202016), National Natural Science Foundation of China (grant number 62076012), National key research and development plan (2019YFC1605306), jointly funded by Beijing Municipal Education Commission and Municipal Natural Fund Committee (KZ202110011017), and National Numerical Wind Tunnel Project (NNW2019-JT01-004).

References

[1] D. S. Wishart, Y. D. Feunang, A. Marcu et al., "HMDB 4.0: the human metabolome database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D608–D617, 2018.

[2] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," 2012, https://arxiv.org/abs/1212.0402.

[3] N. Xu, A. A. Liu, W. Nie, and Y. Su, "Multi-guiding long short-term memory for video captioning," *Multimedia Systems*, vol. 25, no. 6, pp. 663–672, 2019.

[4] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite Bert for self-supervised learning of language representations," 2019, https://arxiv.org/abs/1909.11942.

[5] M. Jaderberg and K. Simonyan, *Spatial Transformer Networks*, MIT Press, 2015.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," 2018, https://arxiv.org/abs/1810.04805.

[7] H. Lu, M. Zhang, X. Xu, Y. Li, and H. T. Shen, "Deep fuzzy hashing network for efficient image retrieval," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 166–176, 2021.

[8] H. Lu, Y. Li, S. Mu, D. Wang, H. Kim, and S. Serikawa, "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2315–2322, 2018.

[9] H. Lu, Y. Zhang, Y. Li, C. Jiang, and H. Abbas, "User-oriented virtual mobile network resource management for vehicle communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3521–3532, 2021.

[10] H. Lu, Y. Tang, and Y. Sun, "DRRS-BC: decentralized routing registration system based on blockchain," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 12, pp. 1868–1876, 2021.

[11] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain intelligence: go beyond artificial intelligence," *Mobile Networks and Applications*, vol. 23, no. 2, pp. 368–375, 2018.

[12] P. Wang, D. Wang, X. Zhang et al., "Numerical and experimental study on the maneuverability of an active propeller control based wave glider," *Applied Ocean Research*, vol. 104, article 102369, 2020.

[13] Z. Chen, H. Lu, S. Tian et al., "Construction of a hierarchical feature enhancement network and its application in fault recognition," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4827–4836, 2021.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, https://arxiv.org/abs/1409.0473.

[15] H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao, and R. Lan, "Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 17, no. 1s, pp. 1–18, 2021.

[16] H. Paulheim, "Knowledge graph refinement: a survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2016.

[17] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, 2009.

[18] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," 2019, https://arxiv.org/abs/1904.12848.

[19] L. Shi and R. L. Xu, "Research on deep learning model based on Word 2vec and improved TF-IDF algorithm," *Computer and Digital Engineering*, vol. 49, no. 5, pp. 966–970, 2021.