Sentiment Time Series Clustering of Danmu Videos Based on BERT Fine-tuning and SBD-K-shape

Abstract

Purpose: This study aimed to construct a sentiment series generation method for danmu comment based on deep learning, and explored the features of sentiment series after clustering.

Methodology: This study consisted of two main parts: danmu comment sentiment series generation and clustering. In the first part, we proposed a sentiment classification model based on BERT fine-tuning to quantify danmu sentiment polarity. In order to smooth the sentiment series, we employed methods such as comprehensive weights. In the second part, the SBD-K-shape method was utilized to cluster the actual collected data.

Findings: The filtered sentiment series or curves of the micro films on Bilibili website could be divided into four major categories. For the first three types of sentiment curves, there is an apparently stable time interval, while the fourth type of sentiment curves show a clear trend of fluctuation in general. In addition, we found that 'disputed points' or 'highlights' are likely to appear at the beginning and the climax of the film, resulting in significant changes in the sentiment curves.

Originality: Our sentiment classification model based on BERT fine-tuning outperformed the traditional sentiment lexicon method, which provides a reference for utilizing deep learning as well as transfer learning for danmu sentiment analysis. BERTfinetuning-SBD-Kshape algorithm can weaken the effect of non-regular noise and temporal phase shift of danmu text.

Keywords: Time Series, Sentiment Curve, Danmu Sentiment Analysis, BERT Fine-tuning, Clustering, K-shape

1 Introduction

The danmu (or danmaku), an emerging type of user-generated comment, has mass appeal among young generations in recent years, especially in China and Japan. An increasing number of people are sending danmu comments online (e.g.,www.bilibili.com). Different from the traditional commentary system, danmu comments can be directly overlaid on the video screen after sending them, and will also be displayed when other viewers watch the same plot. In other words, viewers can express their own feelings and interact with other viewers when watching the same video clips, which creates a real-time sharing experience and adds a lot of engagement and interactivity. In addition to this, font size (large or small), font color and comment mode (position of danmu comments) can be chosen by users, which expresses a measure of individuality (Wu et al., 2021). According to Bilibili's report (Bilibili website, Southeast Asia's leading anime, comics, and games (ACG) community where people can create, watch and share engaging videos) released in 2021, a cumulative total of over 10 billion danmu comments has been sent. According to the data, in the first quarter of 2023, the average daily active users of Bilibili grew steadily, reaching 93.7 million, and the average daily usage time was 96 minutes.

Danmu is a special kind of real-time commentary which has a variety of values, such as content delivery and emotional interaction, and is of great significance to the study of customer group segmentation, video dissemination mechanisms among many other aspects. A study (Yang, 2021) dependent on danmu translation comments contributed to the development of a multimodal process analysis oriented to social semiotics, adding to more in-depth research of this practice and broadening the scope of lay translation research. For generation of user profiles, a survey (Yang and Yu, 2022) took the teaching videos of Bilibili as an example, calculated the content features and behavioral features of users sending danmu comments and clustered the users according to the features to get the user profile to explore the inner relationship of each feature. Another survey (Zhang et al., 2023) established the Bilibili danmu video interaction ritual model based on the interaction ritual chain theory and designed an early warning mechanism for user sentiment driving the formation of the interaction ritual chain, helping to better grasp public opinion dynamics.

Compared with traditional comments, danmu comments are more real-time and concise for the audience to express their true feelings to a particular moment or plot, which carries the properties of information, emotion and time. As a result, behind the sequential danmu comments, there is an invisible sentimental curve which reflects the viewing experiences changes over time (Nickerson, 1998). Therefore, information mining and sentiment analysis based on danmu content can be carried out, so as to grasp the underlying emotional behavior and pattern of users, which will contribute to the refined operation of the video market (Hong et al., 2018).

The essence of the sentiment curve is a sequence of time series, where precise quantification of sentiment value is the basis of follow-up work with regard to news and social platforms. Sentiment Informed Time-series Analyzing AI (SITALA) was proposed to study news-generated sentiment time series, thus predicting epidemic transmission (Prathamesh, 2021). Bhullar et al. (2022) collected social media information for time series sentiment analysis of rescue operations based on machine learning, and combined it with explainable artificial intelligence to improve the transparency and understandability of model decisions to optimize the management of rescue operations. However, few scholars pay attention to research on the sentiment curve of danmu comments and related applications.

To address such an issue, we conducted our own research. This paper mainly explored this topic using danmu comments of 150 micro films collected on the Bilibili website. All these contents will be inputed into the BERT fine-tuned model to calculate the sentiment polarity probability value which will later be used to acquire the sentiment curves. And then three time series clustering algorithms are applied to cluster all these films into several groups according to their sentimental trend.

More specifically, the contribution of this paper includes the following aspects:

A. Danmu comments consist of many Internet slang words and phrases which are informal and colloquial, so the traditional sentiment quantization method based on commonly used linguistic datasets can hardly be transferred to the field of danmu comments. Numerous studies have examined the emotional tendency based on constructing a sentiment lexicon including joy, compliment, anger, sorrow, fear, disgust, surprise, etc (Huang and Shi, 2022). However, the performance of sentiment lexicon in the task of sentiment analysis for danmu texts is still worth further discussion. This paper proposed calculating the sentiment polarity probability value based on BERT fine-tuned models, which achieved the classification effect on randomly selected danmu comment test data far beyond the previous sentiment dictionary method.

B. In the clustering stage, the social attribute of danmu leads to the phase shift between sequences. In addition, the comments will concentrate on several highlights within certain time windows, leading to the highly discontinuous of emotional changes, which further increases the difficulty of generating and mining sentiment curves. To solve problems mentioned above, we cluster these sentiment series by applying Shaped-Based Distance measurements after moving average, which greatly enhances clustering performance.

The rest of the paper is organized as follows. In the next section, we provide a systematic review of related work. In Section 3, we present the research route, methodology including sentiment classification based on BERT fine-tuning, calculation of comprehensively weighted sentiment polarity probability value and clustering of time series based on SBD-K-shape, Fuzzy c-Means and agglomerative hierarchical clustering. Section 4 introduces data collection, results and related analysis. In the final section, we summarize our findings, discuss the theoretical contributions and practical implications, and conclude with limitations and suggestions for future research.

2 Literature review

2.1 Danmu comment sentiment analysis

Sentiment analysis has been an important area of research in natural language processing all the time, as it enables automated extraction of subjective information like sentiment polarity from text data. Since Pang et al. (2002) studied the sentiment analysis of film reviews, sentiment analysis technology has been widely used in the business community. The value of danmu comments is that they contain many special attributes such as content delivery (Jiang, 2014),emotional interaction (Chen, 2015) and immediacy (Zhan, 2014). Emotional interaction (Zhang, 2015) has an important impact on the effect of video transmission. Therefore, it is of great significance to carry out sentiment analysis for danmu comments.

There are three main methods of danmu comment sentiment analysis: sentiment analysis based on sentiment dictionaries and rules, sentiment analysis based on traditional machine learning, and sentiment analysis based on deep learning. Murakami et al. (2011) proposed a video recommendation ranking method based on the danmu sentiment, using a sentiment dictionary algorithm to classify the sentiment of danmu comment texts, and weighting different sentiment categories to calculate sentiment values to determine the final sentiment value ranking of videos. Hong et al. built a dictionary of common words used in online danmu to perform sentiment analysis on danmu texts, and classified all users, who posted danmu comments, based on sentiment values by improving the traditional k-means clustering algorithm. Such results help to understand the sentimental similarities and differences of viewers who watch specific types of videos (Hong et al., 2018). Yamamoto et al. (2013) implemented the support vector machine to classify the sentiment of music clip videos using adjectives, word length and chorus text as features in the danmu text.

Over the years, researchers have developed several efficient deep learning algorithms such as RNN and LSTM to perform sentiment analysis, and the advent of pre-trained language models such as BERT has further advanced this field. Meanwhile, due to the irregular and arbitrary nature of danmu comment texts, traditional sentiment quantification methods based on canonical texts (sentiment dictionaries) cannot be directly and effectively transferred to the danmu domain. Therefore, some scholars have started to try to study the use of deep learning methods for danmu comment sentiment analysis as well. Su et al. (2023) proposed a sentiment analysis method based on a hierar-chical attention mechanism called HAN to analyze the sentiment polarity of bullet chatting texts or danmu comments, which could help teachers to improve teaching methods, adjust teaching contents and schedule in time so as to improve the quality of teaching. Wang et al. proposed a network structure model combining LSTM and RNN for danmu comment sentiment classification, and the classification results were used to annotate the key frames extracted by HC-FCM algorithm, which could effectively achieve accurate annotation of key frames. Although some scholars have launched deep learning-based sentiment analysis of danmu comments, the majority of studies still used sentiment dictionaries. The emerging deep network-based danmu sentiment analysis approach still needs further research.

2.2 BERT fine-tuning

Sentiment classification is a challenging task because it involves understanding the nuanced opinions and emotions expressed in natural language. In the past, traditional machine learning methods, such as naive Bayesian classification and support vector machine, did not have outstanding accuracy in sentiment classification tasks, and thus fail to meet the requirements of business analysis. However, recently, deep learning-based methods have shown great promise in achieving high accuracy in sentiment classification tasks. Among these methods, BERT has emerged as a state-of-the-art language model for many NLP tasks (Devlin et al., 2019).

The full name of BERT is Bidirectional Encoder Representations from Transformers. It pre-trains the deep two-way representation of unmarked text through the common condition of left and right contexts. BERT has pre-trained a large number of unmarked text corpora, including the entire Wikipedia (2.5 billion words) and book corpus (800 million words).

BERT is a large-scale neural network architecture with a large number of parameters, and its number of parameters can range from 100 million to more than 300 million. Therefore, training BERT model from zero on a small dataset will lead to over-fitting. Normally, the training of the BERT model needs to start with a large dataset, and then a relatively small dataset is used to retrain the model. This process is called model fine-tuning.

Several studies have used the BERT fine-tuning method for sentiment analysis in various domains, including restaurant reviews. For example, in a paper by Devlin et al. (2019), BERT was fine-tuned on a large corpus of text data to achieve state-of-the-art results on several NLP tasks, including sentiment analysis. Similarly, in a study by Zhang et al. (2020), BERT was fine-tuned on the SST dataset to classify them as positive or negative. In conclusion, the use of pre-trained language models such as BERT and the BERT fine-tuning method have greatly advanced the field of sentiment analysis, significant to the domain of danmu comments as well.

2.3 Time series clustering

A time series is essentially classified as dynamic data because its feature values change as a function of time, which means that the value(s) of each point of a time series is/are one or more observations that are made chronologically. Time series data is a type of temporal data which is naturally high dimensional and large in data size (Keogh and Kasett, 2002). Time series data are of vital importance in various domains ranging from science, business, finance, economics, healthcare to government.

On this basis, the clustering of time series also has great application and there are various related studies. For discovery of interesting and actionable business information such as previously unknown complementary products or substitutes, and hidden supply chain information, an improved k-means clustering method was proposed (Tan et al., 2015) to create small clusters of similar time series, and those clusters with very small intra-cluster variability were used to find similar time series.

A fuzzy approach with the Partitioning Around Medoids strategy (Pierpaolo et al.,2019) was suggested to cluster multivariate financial time series by considering the dynamic time warping distance, applied to stocks composing the FTSE MIB index to identify common time patterns and possible outliers. Time series could be utilized to represent dynamic customer behavior, and clustering on this basis helps to identify market segments and enables integration with the RFM model (Abbasimehr and Shabani, 2020). Although time series have been studied in depth for a long time, there is still great scope for research on clustering of sentiment time series.

3 Methodology

3.1 Research route

This research focuses primarily on the precise generation and clustering of sentiment series to explore the characteristics of sentiment curves. The overall technical route is shown in **Fig.1**.

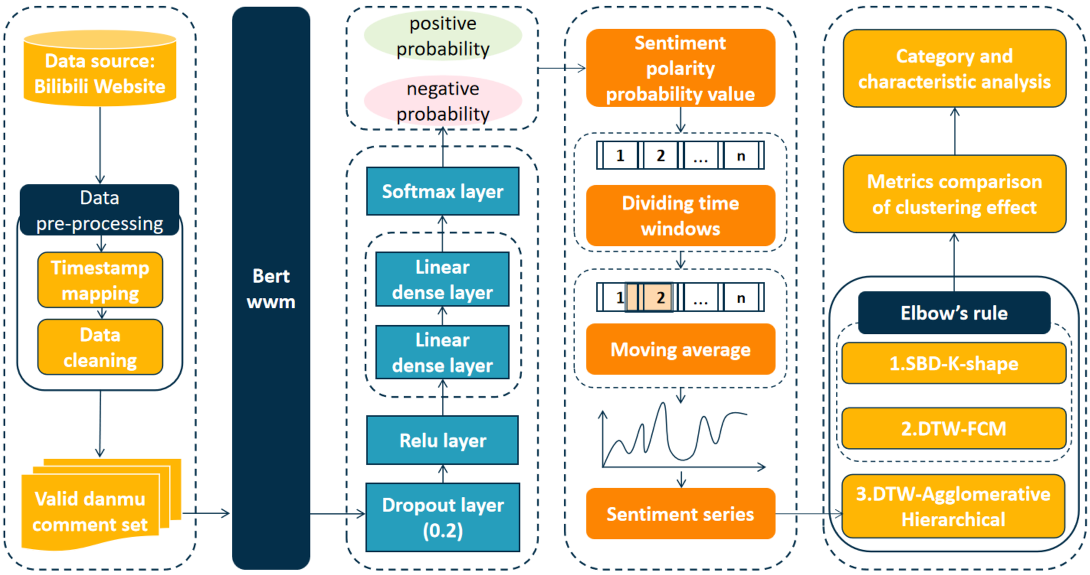

Fig.1 Overall Technical Route

In order to assure the specification of data sources, appropriate danmu video platform and danmu comment text are selected. Web crawlers and database technology are also employed to collect, save and preprocess danmu comments from relevant video platforms.

For the generation of sentiment series, the depiction of sentiment series is the basis of danmu comment sentiment series clustering. Accurately depicting the trend of the audience's sentiment polarity with time depends on the precise quantification of the danmu comment sentiment value contained in the unit text. In view of danmu comments with sentiment polarity continuity and the interactions between danmu comments to be considered within a certain time window, we should focus on the most prominent sentiment polarity. Under such circumstances, the precise quantification of the sentiment polarity probability value, as a good representative of the sentiment value, is of paramount significance. Based on the review of danmu comment texts and other sentiment analysis research, we utilized BERT fine-tuning to fulfill sentiment classification for quantification. To achieve a good transfer learning effect and adapt to the Bilibili danmu comment domain, we selected data from the same website and preprocessed it. For the base BERT model, we selected the Chinese BERT-wwm, a pre-trained BERT model that has outstanding performance in Chinese tasks based on Whole Word Masking technology, released by the Joint Laboratory of HIT and iFLYTEK Research in order to further promote the research and development of Chinese information processing. After that we added our self-defined classification section, namely a dropout layer, a Relu layer, two Linear dense layers and a Softmax layer. We also researched five fine-tuning methods and determined to use frozen parameter method, learning rate decay and bias correction to improve the fine-tuning efficiency. Through this part, the positive and negative probabilities of the danmu comment text are output. In the model validation phase, a comparison experiment of our method with the Hownet sentiment dictionary was set up. We randomly sampled a certain number of danmu comments and calculated the confusion matrix to verify the model classification effectiveness. During the actual generation stage of the sentiment curve, in order to improve the accuracy and smoothness of the sentiment time series, we proposed comprehensively weighted sentiment polarity probability and moving average processing, which to some extent ensures the balance between morphological distortion and time series smoothness.

For the exploration of time series clustering, we utilized the Elbow criterion to detect the effect of different clustering categories on the overall clustering distortion degree, thus scientifically determining the optimal number of clusters. In this study, the K-shape algorithm was used for sentiment series clustering. The time series similarity of this algorithm adopts Shaped-Based Distance (SBD). Before the time series cluster division, the displacement time window between this time series and the cluster center is dynamically calculated in advance with a certain risk of edge shape loss to ensure that there is no lag or antecedence between the two sentiment time series. With regard to the principle of cluster centroid update, the K-shape algorithm equivalently transforms the Steiner tree problem when calculating the cluster centroid, which greatly improves the efficiency of cluster centroid iteration and avoids distortion in the iterative process of sentiment time series clusters. As a result, the extracted centroid series will maintain a consistent trend with the source clusters in terms of trend and change magnitude, thus accurately characterizing the high aggregation morphology of time curve clusters. In order to compare the effects of combinations of different time series similarity settings and algorithms, DTW

(Dynamic Time Warping) -FCM (Fuzzy C-means), and DTW-Agglomerative hierarchical clustering were set to compare with our method on the Davies-Bouldin Index (DBI) and other metrics.

3.2 Data collection

This study selected micro films as the research subject, which could reduce a measure of difficulty during the data cleaning process and enhance performance by limiting the video length and style. The Selenium WebDriver was used to collect danmu comments on the 150 most viewed (mostly over 1000 times) micro films on the Bilibili website (the largest online video commenting website in Asia) including movie serial numbers, contents and corresponding time stamps. Since we used the Chinese dataset, we removed the danmu comments containing other languages and those containing meaningless characters. Invalid or blank contents were deleted and only movies with danmu numbers over 200 could ultimately be retained. 24 samples were deleted, leaving the final dataset with 126 films and 134,157 danmu comments.

For the training set required for BERT fine-tuning, we selected the comment data of Bilibili's popular anime(https://aistudio.baidu.com/aistudio/datasetdetail/107440/0). We downloaded the second version, containing data collected on Bilibili as of June 6, 2021. These comment data also come from Bilibili, most of which are short comments, and can therefore be deemed to share similar characteristics with Bilibili 's danmu comments. After removing nulls, duplicates and other pre-processing operations, we labeled positive comments as 1 and negative comments as 0. Then 3000 positive comments and 3000 negative comments were randomly selected as the pre-training dataset for BERT fine-tuning.

Table 1 Preprocessed Danmu Comment Data (Part) Collected From Bilibili

Relative time of sending	Danmu comments	Translation
107.14500	好细节呀，画的好漂亮	[That's a cute drawing in great details]
111.09900	我敲这么美	[So pretty!]
150.55700	两分钟过去我还是没看懂	[Two minutes passed and I still didn't get it]
219.84500	好家伙，吓死人了	[Scary as hell]
230.63800	密闭恐惧症犯了	[I have trypophobia!!]

Table 2 Preprocessed Comment Data (Part) For Fine-tuning

Comments	Translation	Label
好可爱！！！也很有意义～很有趣	[So cute!! Also very meaningful ~ very interesting]	1
普及知识又有内容的番真的很棒啊	[It's so great to popularize knowledge for audience]	1
良心番良心番，推给初中当科普，推给高中当复习，推给学医的当点心	[Recommend for junior high school students as a science video, Recommend for high school students as a review,Recommend for medical students as a dessert!]	1

果然还是不喜欢歌舞片。	[I don't like musical movies.]	0
假 3D，渣特效，吃老本。	[Fake 3D, poor special effects]	0

3.3 Generation of sentiment series

3.3.1 Sentiment classification model based on BERT fine-tuning

According to the research findings of Zhang et al. (2020) and Xiong et al. (2020), there are many ways to fine-tune BERT, mainly including the following approaches:

A. Bias correction. It is implemented by adding bias correction to the optimizer.

B. Re-initialized weights. When fine-tuning, we can maintain the weights of the BERT bottom, and re-initialize the weights of the top layer randomly, so that these parameters can be re-learned on the task.

C. Pre-trained weight decay. Pre-trained weight decay is adapted for fine-tuning pre-trained models by subtracting the first term of pre-training parameters from the objective.

D. Frozen parameter. Frozen parameter method is often used in the training of some large models, mainly for some models with numerous parameters. The frozen parameter can reduce the iterative calculation of parameters and speed up the training speed without affecting the accuracy of the results too much.

E. Learning rate decay. It is a method that applies higher learning rates for top layers and lower learning rates for bottom layers.

For bias correction: its actual implementation is to improve the optimizer, which is relatively simple in operation and does not conflict with other methods. This method was utilized in this paper.

For re-initialized weights and frozen parameter: Considering that the dataset focused on in this paper is a relatively small dataset compared with other fine-tuning tasks as well, a large pre-training model needs to be adopted. On this basis, in order to avoid overfitting and reduce training time, the frozen parameter method is the most effective. And because the frozen parameter method freezes the weight parameters, it actually conflicts with the re-initialized weights method to some degree. If the two methods are combined, the accuracy may be improved, but the training time will be greatly prolonged. Therefore, we only used the frozen parameter method instead of the re-initial weight method in this paper.

For pre-trained weight decay: this method has little impact on the performance of text classification tasks and even causes performance degradation (Zhang et al., 2020). Therefore, this method was not used for sentiment classification in this paper.

For learning rate decay: AdamW optimizer is adopted, an adaptive learning rate optimizer (Xiong et al., 2020), to implement this method. The sentiment analysis in this paper was implemented on the basis of this optimizer.

There are two approaches to implementing the frozen parameter method:

A. Partial training: train some layers and freeze other layers at the same time. Keep the weight of the initial layer unchanged, and only retrain the higher level. This approach requires determining how many layers need to be frozen and how many layers need to be trained.

B. Freeze the whole architecture: This approach is to freeze the whole pre-training model, add some of its own neural network layers, and then train the new model. Only the weights of the additional layers will be updated during training.

The second method was utilized in this paper, which is relatively simple and requires low computational performance. This is extremely instructive for the rapid deployment of BERT fine-tuning based on other small datasets, because the number of layers that need to be frozen varies greatly according to the characteristics of datasets.

To sum up, the actual structure of our work was to freeze the entire BERT model during fine-tuning. Since the danmu comments of Bilibili are basically Chinese, we chose the Chinese pre-training model BERT-wwm. It's a Chinese pre-trained BERT model which has outstanding performance in Chinese tasks based on Whole Word Masking technology, released by Joint Laboratory of HIT and iFLYTEK Research in order to further promote the research and development of Chinese information processing. After that, a dropout layer, a Relu layer, two Linear dense layers and a Softmax layer were added. An improved AdamW optimizer was employed for bias correction and learning rate decay. For the entire sentiment classification network, the input is danmu comment sentences, and the output is the probability values of the text as positive and negative comments. By comparing the probability values, the polarity of the danmu comment can be determined.

3.3.2 Calculation of comprehensively weighted sentiment polarity probability value

Based on the results of sentiment classification, a specific time window should be determined as the sampling time window of a single ordinal of time series. Meanwhile, a proper method should be used to quantify the single ordinal sentiment polarity probability value and preprocess the sentiment time series to achieve the precondition of clustering.

When classifying sentiment, the time window needs to be selected under three conditions: the existence of a strong correlation between danmu comment sentiment, the continuity of danmu comment sentiments in general and the control of the series dimension. In light of danmu comments with sentiment polarity continuity, with the interactions between danmu comments to be considered within a certain time window, we focus on the most prominent sentiment polarity. Therefore, the single ordinal sentiment polarity probability value is calculated as shown in Equation (1).

$$S_t = \sum_{S_{ti}=P} S_{ti} \times O_P - \sum_{S_{ti}=N} S_{ti} \times O_N \tag{1}$$

Where $\sum_{S_{ti}=P} S_{ti}$ and O_P refer to the sum of sentiment probability values of positive danmu comments and the ratio of the number of positive danmu comments to the total number of danmu comments within a time window, respectively.

3.3.3 Moving average processing

The aforementioned method of sampling the danmu data won't result in omissions or overlap, but because the distribution of sentiment values is not uniform and independent, non-overlapping sampling is likely to cause significant fluctuations at the window's edge. According to the time series theory, the moving average process can filter the high frequency noise of sentiment values, making the time series morphologically smoother, as shown in Equation (2).

$$Seq'[n] = \frac{1}{M}\sum_{k=0}^{M-1} Seq[n-k] \quad (2)$$

Where *Seq[n]* refers to the original time series, *Seq'[n]* refers to the time series after moving average, *M* refers to the step size of moving average. The smoothness of the series increases with the value of *M*.

3.4 Clustering of time series based on SBD-K-shape algorithm

A. K-shape

Clustering of time series is a basic and significant unsupervised mining method that is widely used in emotion prediction, outlier detection of key performance indicators, trend mining, etc.

One of the most commonly used algorithms for partitioning clusters is k-Means (Hartigan and Wong, 1979), whose final goal is to minimize the total distance (typically Euclidian distance) between all objects in a cluster from their cluster center (the mean value within a cluster).

K-Shape is based on a similar principle to K-means, however, it improves the distance measurement and cluster center calculation methodologies. For one thing, it supports amplitude scaling and shift invariance, and for another, it offers multiple application scenarios due to its high computational efficiency and automatic parameter setting. In this study, we propose the SBD-K-shape algorithm, which measures and updates the cluster center based on shape-based distance (SBD). This new algorithm fundamentally ensures that the trend and magnitude of the wave remain constant in the case of time shifting, which takes phase change into consideration.

B. FCM (Fuzzy c-Means)

Unlike k-Means and k-Shape, which cluster each sample with certainty, the FCM (Fuzzy c-Means) algorithm creates 'soft' clusters based on fuzzy set theory, so an object has a degree of membership in each cluster. Let $X = \{x_1, x_2, ..., x_k\}$be the dataset of n samples, the set of cluster be $C = \{c_1, c_2, ..., c_c\}$ and the set of c cluster centers be $V = \{v_1, v_2, ..., v_c\}$. The objective function of FCM is shown in Equation (3), which represents the square sum of error from samples to the cluster center (James et al., 1984):

$$J = \sum_{i=1}^{c}\sum_{j=1}^{k} (u_{ij})^m {dis_{k,k}^{DTW}}^2 \quad (3)$$

Set $P = \{p_1, p_2, ..., p_m, ..., p_k\}$ and $Q = \{q_1, q_2, ..., q_n, ..., p_k\}$ two sequences and $dis_{i,j}^{DTW}$ represents the distance between two time series measured by Dynamic Time Warping algorithm (DTW), shown in Equation (4):

$$dis_{k,k}^{DTW} = dis_{m,n}^{ED} + min\{dis_{m-1,n-1}^{DTW}, dis_{m-1,n}^{DTW}, dis_{n,m-1}^{DTW}\} \quad (4)$$

As shown in Equation (4), $dis_{k,k}^{DTW}$ is the sum of the minimum cumulative distance between $dis_{m,n}^{ED}$ and adjacent elements, whose process is (m,n) searching from (0,0) to (k,k), constantly trying to obtain the minimum cost of the warping path length.

The constraints that need to be satisfied are:

$$\begin{cases} \sum_{k=1}^{c} u_{ik} = 1, \quad i = 1,2,\ldots,n \\ 0 \leq u_{ik} \leq 1, \quad i = 1,2,\ldots,n, \quad k = 1,2,\ldots,c \\ 0 < \sum_{i=1}^{n} u_{ik} < n, \quad k = 1,2,\ldots,c \end{cases} \tag{5}$$

Based on the Lagrange multiplier method, U and V are calculated by the Equations (6) below, respectively.

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{dis_{i,k}^{DTW}}{dis_{j,k}^{DTW}}\right)^{\frac{2}{m-1}}} \quad and \quad v_k = \frac{\sum_{i=1}^{n}(u_{ik})^m x_i}{\sum_{i=1}^{n}(u_{ik})^m} \tag{6}$$

Where $U = (u_{ik})$ is an n × c matrix to measure degrees of membership, where u_{ik} is the possibility that the x_i sample belongs to the c_c cluster. m is a fuzziness index whose default value is two in most cases.

B. Agglomerative hierarchical clustering

Agglomerative hierarchical clustering is a bottom-up algorithm. It first regards each sample as a cluster, and then starts to merge clusters with high similarity according to certain rules, and finally the algorithm ends when all samples form a cluster or reach a certain condition. Determining the similarity between clusters is the key to this algorithm, and the similarity is determined by the inter-cluster distance, with high similarity for small inter-cluster distance (Bouguettaya et al., 2015). This algorithm is well suited for processing small amounts of data.

3.4.1 Distance measurement of time series

The distance measurement is an essential indicator of the similarity between two time series, which is the basis of clustering, classification and outlier detection of time series. Euclidean distance and dynamic time warping (DTW) are frequented in calculating distance measurement. Since this study mainly focus on clustering of sentimental trend, selecting shape-based distance measurements that can handle amplitude and phase distortions is crucial. When two series are out of phase, the Euclidean distance is highly sensitive, whereas DTW can effectively eliminate this defect, which is extensively used in template matching, hand gesture recognition and data mining. However, distance measures such as DTW that satisfy such requirements are computational overhead, so SBD is proposed as an alternative.

A. Cross-correlation measure

Cross-correlation measure is to determine the similarity of two sequences $P = \{p_1, p_2, \ldots, p_m, \ldots, p_k\}$ and $Q = \{q_1, q_2, \ldots, q_n, \ldots, p_k\}$, even if they are not properly aligned. Cross-correlation keeps q static and slides p over q to compute their inner product for each shifts of p. A Shift of a sequence are denoted as follows:

$$\overrightarrow{p_{(s)}} = \begin{cases} (\overbrace{0,\dots,0}^{|s|}, p_1, p_2, \dots, p_{k-s}), & s \geq 0 \\ (p_{1-s}, \dots, p_{k-1}, p_k, \underbrace{0,\dots,0}_{|s|}), & s < 0 \end{cases} \tag{7}$$

Where s is all possible translations in time series P, and $s \in$[-m, m]. If $s \geq 0$, the time series of P shifts s units to the right; If $s < 0$, the time series of P is shifted s units to the left.

Then the cross-correlation sequence $CC_\omega(P,Q) = [c_1, c_2, \dots, c_\omega]$ is obtained whose length is $2k-1$, defined as Equation (8):

$$CC_\omega(P,Q) = R_{\omega-m}(P,Q),\ \omega \in \{1,2,\dots,2k-1\} \tag{8}$$

Where $R_k(\mathrm{P},\mathrm{Q})$ is defined as Equation (9):

$$R_k(P,Q) = \begin{cases} \sum_{l=1}^{m-k} x_l + ky_l & k \geq 0 \\ R_{-k}(Q,P) & k < 0 \end{cases} \tag{9}$$

Calculate the value of ω that maximizes $CC_\omega(\mathrm{P},\mathrm{Q})$, and acquire the optimal displacement of P relative to Q (where s=$\omega - m$). Then normalization for $CC_\omega(P,Q)$ might be required and this paper chose the coefficient normalization which is defined as follows:

$$NCC_{n,\omega}(P,Q) = \frac{CC_\omega(P,Q)}{\sqrt{R_0(P,P)R_0(Q,Q)}} \tag{10}$$

Where $NCC_\omega(P,Q) \in$[-1,1] with $R_0(P,P)$ and $R_0(Q,Q)$ represent the value of two identical sequences without relative displacement. The larger the normalized value of $NCC_\omega(P,Q)$, the higher the positive correlation between the two sequences.

B. Shape-based distance (SBD)

After normalization of the series, the position ω is chosen where $NCC_\omega(P,Q)$ is maximized and we derive the following distance measure:

$$SBD(\vec{x},\vec{y}) = 1 - \max_\omega \frac{CC_\omega(\vec{p},\vec{1})}{\sqrt{R_0(\vec{p},\vec{p}) \cdot R_0(\vec{q},\vec{q})}} \tag{11}$$

Where $SBD(\vec{x},\vec{y})$ takes values between 0 to 2, with 0 indicating perfect similarity for time series.

3.4.2 Cluster center of time series

The easiest way to extract centroids within a cluster is to calculate the arithmetic mean of all series. k-Means, the most prevalent clustering method, uses this strategy, but it fails to accurately capture the class characteristics of each cluster. Then K-shape is proposed, to cast the centroid computation as an optimization problem (steiner tree optimization) with the goal of determining the minimum of the sum of squared distances to all other time series.

To avoid such problems, we cast centroid computation as an optimization problem where the objective is to find the minimum of the sum of squared distances for all other time series, as shown in Equation (12).

$$\overrightarrow{c_c}^* = \underset{\overrightarrow{c_c}}{argmin} \sum_{\overrightarrow{p_i} \in P} NCC(\overrightarrow{p_i}, \overrightarrow{c_c})^2 \quad c_k \in R \tag{12}$$

Where, $\overrightarrow{p_i}$ is the i-th class clustering series set, and $\overrightarrow{c_c}^*$ is the cluster center extracted by this method.

As the cross-correlation measure extracts the similarity rather than the difference between the two time series, we can assimilate Equation (12) as a maximization problem, as shown in Equation (13). During each iteration, we use the previously computed centroids as reference and align all sequences towards this reference sequence.

$$\begin{aligned} \overrightarrow{c_c}^* &= \underset{\overrightarrow{c_c}}{argmax} \sum_{\overrightarrow{p_i} \in P} (\overrightarrow{p_i}^T \cdot \overrightarrow{c_c})^2 \\ &= \underset{\overrightarrow{c_c}}{argmax}\, \overrightarrow{c_c}^T \sum_{\overrightarrow{p_i} \in P} (\overrightarrow{p_i} \cdot \overrightarrow{p_i}^T)^2 \cdot \overrightarrow{c_c} \end{aligned} \tag{13}$$

Equation (13) is normalized as shown in Equation (14):

$$\overrightarrow{c_c}^* = \underset{\overrightarrow{c_c}}{argmax} \frac{\overrightarrow{c_c}^T \cdot Q^T \cdot S \cdot Q \cdot \overrightarrow{c_c}}{\overrightarrow{c_c}^T \cdot \overrightarrow{c_c}} = \underset{\overrightarrow{c_c}}{argmax} \frac{\overrightarrow{c_c}^T \cdot M \cdot \overrightarrow{c_c}}{\overrightarrow{c_c}^T \cdot \overrightarrow{c_c}} \tag{14}$$

We set $\overrightarrow{c_c} = \overrightarrow{c_c} \cdot k$, where $Q = I - \frac{1}{m} O$, I is the identity matrix, O is a matrix with all ones and $M = Q^T \cdot S \cdot Q$.

The maximum value of $\overrightarrow{c_c}^*$ is the eigenvector corresponding to the maximum eigenvalue of matrix M, namely the extracted cluster center curve.

4 Results

4.1 Generation of sentiment series

4.1.1 Establishment of sentiment classification model based on BERT fine-tuning

Freeze the Chinese BERT-wmm as the base model, and then add our own custom classification layer. The output of the model is the positive and negative probability values of the text, which can be compared to determine the sentiment polarity. Then we employed improved AdamW optimizer for bias correction and learning rate decay. To fine-tune the pretrained model, a dataset of preprocessed Bilibili film reviews containing 3,000 positive reviews and 3,000 negative reviews was collected. The training set, verification set, and test set were divided according to the proportions 70:15:15 and then input into the network for training. We determined the batch size to be 64, the padding length to be 50, and the initial learning rate to be 1e-5 after several experiments.

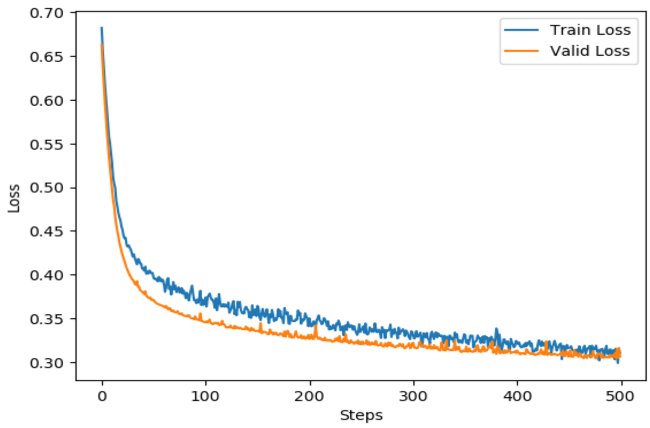

Fig.2 Training and validation loss curves for the BERT fine-tuning process

Table 3 Sentiment Classification Report On the Bilibili Comment Test Data

Label	Performance			Support (Samples)
	Precision	Recall	F1-score	
0(Negative)	0.87	0.87	0.87	450
1(Positive)	0.86	0.86	0.86	450

After 500 epochs, the training loss and validation loss converge basically converge, as shown in the **Fig.2**. Both training and validation losses were in a downward trend, and there wasn't any overfitting. The saved model parameters were used to predict the test set, and the final accuracy rate reached 86.65%. The final generated classification report is shown in the **Table 3**. The precision rate, recall rate and F1-score of negative and positive comments were 0.87 and 0.86, respectively, demonstrating high performance.

Next, we labeled 134,157 bullet screen comments from 126 micro films using the trained model and calculated the sentiment polarity probability value. Part of the results are shown in the **Table 4**.

Table 4 Results of danmu comment sentiment analysis (part)

Relative time of sending	Danmu comments	Predicted label	Sentiment polarity probability value
6.822	看我发现了啥 [Look what I found !]	1	0.945843339
13.391	真的好喜欢卓别林的作品 [I really like Chaplin's work]	1	0.960744917
29.639	什么是大师，大师就是过了一百年，大家也会承认你很优秀。 [What is a master? A master the one who still stands out from the crowd even after 100 years.]	1	0.960018456
31.2	童年阴影 [childhood trauma]	0	0.922767937

32.719	人类从历史里学到的最大的教训就是人类无法从历史中学到教训 [The greatest lesson mankind has learned from history is that mankind never learns]	0	0.991773129

4.1.2 Analysis and validation of sentiment classification model based on BERT fine-tuning

To varify the effectiveness of our proposed sentiment classification model based on BERT fine-tuning, we compared this model with the Hownet sentiment lexicon. Hownet sentiment lexicon method calculates the sentiment value of clauses according to predetermined criteria, and then summatively calculates the sentiment value of the entire danmu comments. If the emotion value is greater than 0, the text is judged as positive; if the emotion value is less than 0, it is judged as negative.

We randomly selected 200 danmu comments which had already been labeled by sentiment classification model based on BERT fine-tuning, and marked these values again manually. The report of model classification and evaluation results is shown in the **Table 5**.

Table 5 Sentiment Classification Comparison Report On The Labeld Danmu Comments

	Label	Performance			Support (Samples)
		Precision	Recall	F1-score	
1(Hownet)	0(Negative)	0.74	0.76	0.75	100
	1(Positive)	0.73	0.71	0.72	100
2(Our method)	0(Negative)	0.81	0.83	0.82	100
	1(Positive)	0.82	0.80	0.81	100

The accuracy of sentiment classification results obtained by Hownet sentiment lexicon method was 73.41%, whereas the accuracy of our method was greatly improved, reaching 81.64%. Our sentiment classification method demonstrated a significant improvement in precision, recall, and F1-score compared with the direct implementation of the Hownet sentiment lexicon method. The F1-score for negative and positive comments reached 0.82 and 0.81, respectively. It can be considered that the sentiment classification model based on BERT fine-tuning has excellent overall performance in the sentiment classification of danmu comments.

4.1.3 Sentiment time series and sentiment curves

In this study, the time window is selected according to the existence of strong correlation between danmu comment sentiment, general danmu comment sentiment continuity and series dimension control. By observing the relative time between the video content and the corresponding response, it is found that there is a relative dislocation of 1-12 seconds in general. However, considering the resonance of danmu users for video cultural content, narration soundtrack, and other video content and elements, although danmu sentiment marks different entities, the emotional polarity of the same clip is almost continuous. At the same time, due to the large variance in the duration of the selected micro films, the selection of a fixed time window will result in sentimental time series of some videos that are too short. After careful consideration, 50 time windows had been selected at last.

Pandas module was used to divide time ordinal numbers according to the total duration of videos and the relative time of danmu released by users. We also used comprehensive weights to quantify the single series of sentiment values. The Matplotlib module was also involved in drawing sentiment time series, part of which were shown in the **Fig.3**.

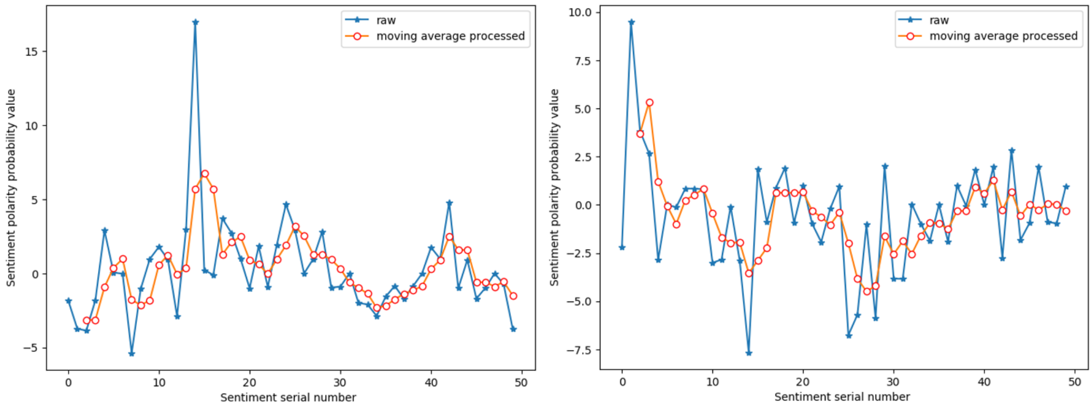

Fig.3 Partial Sentiment Curves(The source video of the left picture is *100% Salt* and the source video of the right picture is *The Coming Day*)

By scrutinizing the danmu contents in the turning interval reflected by the wave crest and trough through the **Fig.3**, we can test whether the sentiment curve accurately reflects the change of audience emotions in the clip. Specifically, In the initial phase of the left picture, the overall trend of the curve is upward. According to the bullet screen recordings, the audience viewed the setting, actors, and other video content positively during this time period, and there are several turning points when the video was played for the ordinals 13,19,26, and 42. To sum up, the emotional curve can well reflect the sentimental dynamics of danmu users.

There may be two adjacent extreme points in the emotion curve with different sign symbol. Considering the continuity of the emotion curve, the time window formed by two extreme points is bound to cross the sentimental polarity. Such time windows can represent large emotional transitions. In addition to the shift of sentimental polarity caused by camera transition, the danmu sentimental fluctuation is greatly affected by the Key Opinion Leader (KOL) phenomenon, and will last for several ordinals.

To sum up, the sentimental curve has strong interpretability, and curves share commonalities. From above paragraphs we can see that the highlights within a film mainly depend on ’disputed points’ and ’highlights’. Disputed points are mainly triggered by controversial comments sent from KOL, while highlights are formed by friendly discussion and consensus among the audience.

4.2 Clustering of sentiment series

This paper applies Z-score normalization before calculating distance, which standardizes the mean and variance of each time series to 0 and 1 respectively. In this paper, Elbow's criterion is used to determine the optimal number of clusters in K-shape and FCM algorithms and aim to minimize clustering distortion. When the number of cluster k is lower than that of the real cluster, the aggregation of each cluster will enhance as k increases. While when k reaches the real clustering number, the descent range will decrease sharply, and then it will be flat with the continuous increase of k value.

The K-shape model in Tslearn is used to test the number of clusters from 2 to 6, and the curve of distortion degree basically declines with the number of clusters is drawn, as shown in **Fig.4**. As the number of clusters is between 3 and 4, the distortion value drops sharply, and when it is between 4 and 5, the decrease of the distortion value slows down, the critical value of 4 can be taken as the optimal number of clusters.

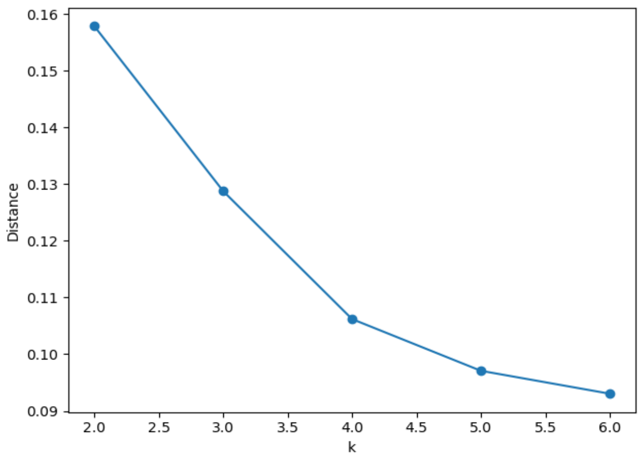

Fig.4 Elbow Criterion for Determining the Optimal Number of Clusters

Based on the optimal number of clusters obtained, 126 micro films are clustered into 4 groups. To better examine the effect of different models, three algorithms including SBD-K-shape, FCM and agglomerative hierarchical clustering are tested according to the silhouette coefficient (SI) and the Davidson-Boding Index (DBI) provided by Scikit-learn (a machine learning framework). The evaluation indicators of clustering effect are shown in **Table 6**.

Table 6 Evaluation Metrics for Time Series Clustering Models

Clustering metrics	SBD-K-shape	DTW-FCM	DTW-Agglomerative hierarchical clustering
DBI_Score	2.13	2.91	2.76
Silhouette_Score	0.35	0.27	0.32

It can be seen from **Table 6** that the SBD-K-shape is superior to the FCM and agglomerative hierarchical clustering both in the two evaluation indicators, with a silhouette coefficient of 0.35 and a Davidson's Boding index of 2.13, which shows its superiority in clustering mining and better homogeneity within the same group. The algorithm mainly focuses on morphological similarity between different samples, so all series within a cluster can be regarded as the contraction and expansion of cluster center, showing a relatively parallel trend with the cluster center. The cluster centers were highlighted in blue and the rest of grey lines represented other movie samples.

As can be seen from **Fig.5**, the sentiment curves of micro films can fall into four categories, named as ’V-opening type’, ’middle inverted V type’, ’inverted V-opening type’ and ’fluctuating type’, respectively.

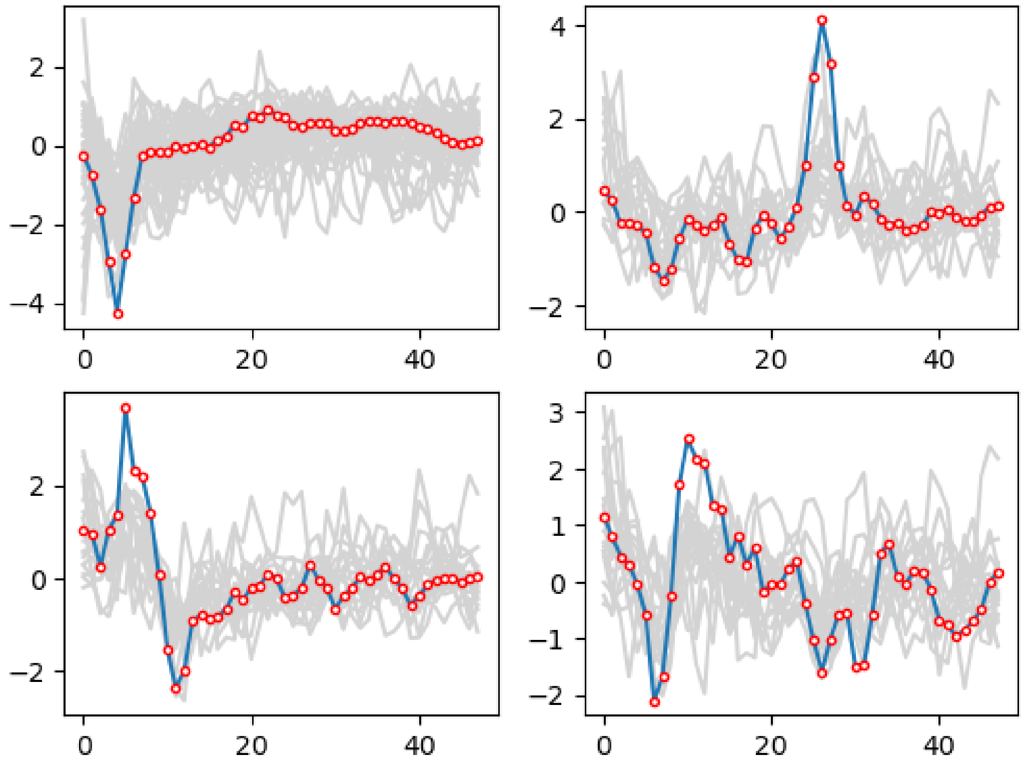

Fig.5 Sentiment Curve K-shape Clustering Results

A. For the sentiment curve with the basic form of ’V-opening type’, there is a high proportion of flat intervals in the whole duration, which does not include obvious highlights. There is a disputed point at the beginning of the films that drops sharply, starting with a V shape.

B. For the sentiment curve whose basic form is ’middle inverted V type’, there is a highlight with a great increase in the middle part of the time series, while the other time windows do not contain obvious disputed points and highlights. This can be attributed to the heated discussion generated by the climax of the plot.

C. For the ’inverted V-opening type’ sentiment curve, there is an adjacent highlights and disputed point at the beginning of the films, but there is only minimal fluctuation in the latter part. It could be argued that a wonderful opening was mishandled afterwards or that a controversial part led to a rapid increase in negative attitudes.

D. For the sentiment curve with the basic form of ’fluctuating type’, the overall trend shows a drastic fluctuation, which is different from the other three sentiment curves which are relatively stable. There are several disputed points in the overall curve, located at the beginning, middle and end. What's worth mentioning is that there is an obvious highlight of the curve immediately following the first disputed point.

5 Discussion and conclusion

This study mainly focused on the clustering of sentiment curves of micro films. It aimed to quantify the sentiment value of films by a set of time series sequences and then cluster them into several groups, which try to figure out the underlying common ground within a cluster and contribute to the refined operation of the video market. The findings are as follows:

A. The proposed Sentiment classification model based on BERT tuning can effectively enhance the performance of the sentimental quantification model. This is notably significant for sentiment analysis of danmu comments, as it introduces a new deep learning method in addition to the traditional emotion lexicon. In the empirical phase, we classified the sentiment polarity of randomly selected danmu corpus by refining the pre-training model and compared it with the Hownet sentiment lexicon method. Finally, the accuracy rate of our model reached 81.64%, and the F1-score of positive and negative label comments reached 0.81 and 0.82, respectively, which were significantly higher than the previous methods.

B. The sentiment curve generated by comprehensively weighted sentiment polarity probability value and moving average method can effectively describe the trend of audience's sentiment fluctuation, which smoothens highly discontinuousness of original curves.

C. In this study, broadening the methodology of previous studies, we chose three clustering algorithm which are applicable to short time sequences. From the perspective of clustering indicators such as contour coefficient and DBI, the performance of SBD-K-shape was significantly better than the DTW-FCM and agglomerative hierarchical clustering. The proposed algorithm eliminates the phase shift between sequences, which better reflects the morphological characteristics of cluster centers. The findings also show that the SBD-K-shape algorithm has better performance in the clustering task of sentiment series.

D. The results of clustering indicate that the sentiment curves of micro films on the Bilibili website can be categorized into four groups. "V-opening type", "inverted V-opening type", "middle inverted V type", and "fluctuating type". For the first three types of sentiment curves, there is an obvious relatively stable time window, whereas the fourth type of sentiment curve demonstrates a fluctuating trend. In addition, we discovered that there are likely to be "disputed points" or "highlight" at the beginning and the climax of the film, which result in a significant shift in the sentiment curve. This suggests that filmmakers should place a high priority on the opening scene, as it has a significant impact on the audience's viewing experience.

However, this study has the following limitations:

A. Although the comment data from Bilibili website is employed in our Fine-tuned BERT model, they are actually different from the danmu comments. As a result, there is a gap between the two data sets when it comes to test accuracy, and the sentiment of danmu comments can not be quantified precisely.

B. Due to the large variance in video duration, we did not employ a fixed tine window. In fact, this would result in some films with too-short time windows and danmu comments could not reflect the sentimental situation due to a lack of information.

C. The SBD-K-shape algorithm has advantages in the clustering of time series with displacement similarity, but it also has the problem that similarity calculation is abnormal due to very few synchronous curve displacements.

In the future, we hope to classify films based on the sentiment curves clustering method proposed in this paper. The E-Divisive with Medians algorithm or other methods are then used to divide the stages of the four videos according to the number of danmu comments. After that, the number of danmu comments and sentimental intensity in each stage were calculated and taken as independent variables. The number of likes and coins were taken as dependent variables for regression. Thus, the influence of specific stages of various video types on video popularity can be investigated.

References

[1] H. Wu, F. Pitié and G. J. F. Jones, "Investigating Automated Mechanisms for Multi-Modal Prediction of User Online-Video Commenting Behaviour," 2021 International Conference on Content-Based Multimedia Indexing (CBMI), Lille, France, 2021, pp. 1-6, doi: 10.1109/CBMI50038.2021.9461909.

[2] Yang Y. Making sense of the “raw meat”: A social semiotic interpretation of user translation on the danmu interface[J]. Discourse, Context & Media, 2021, 44: 100550.

[3] Yang Yang, Yu Weijie. A study on user portraits integrating pop-up content features and behavioral features--Taking Bilibili teaching videos as an example[J]. Intelligence Science,2022,40(12):161-169.DOI:10.13833/j.issn.1007-7634.2022.12.020.

[4] Zhang Baosheng, Yang Xiaoting, Wang Xiaohong. Research on user emotion early warning mechanism of short video pop-up platform based on interactive ritual chain theory--Bilibili danmu website as an example[J]. Intelligence Science,2023,41(01):100-109.DOI:10.13833/j.issn.1007-7634.2023.01.012.

[5] Nickerson R S. Confirmation bias: A ubiquitous phenomenon in many guises[J]. Review of general psychology, 1998, 2(2): 175-220.

[6] Hong Q, Wang SY, Zhao QP, Li JF, Rao WX. Classification of video user groups based on pop-up sentiment analysis and clustering algorithm[J]. Computer Engineering and Science,2018,40(06):1125-1139.

[7] Desai P S. News Sentiment Informed Time-series Analyzing AI (SITALA) to curb the spread of COVID-19 in Houston[J]. Expert Systems with Applications, 2021, 180: 115104.

[8] Bhullar G, Khullar A, Kumar A, et al. Time series sentiment analysis (SA) of relief operations using social media (SM) platform for efficient resource management[J]. International Journal of Disaster Risk Reduction, 2022, 75: 102979.

[9] Huang Lihe, Shi Yingxin. Research on network public opinion event monitoring for video barrage [J].Information Journal, 2022,41 (02) : 146-154.

[10] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[J]. arXiv preprint cs/0205070, 2002.

[11] Jiang H X. The research of barrage video from communication vision[D]. Master thesis submitted to School of Journalism and Communication, Central China Normal University. Retrieved from http://d. wanfangdata. com. cn/Thesis, 2014.

[12] Chen Mo. Application and Development of Barrage Technology in TV Industry[J]. Journalism Lover, 2015(10): 83-85.

[13] Zhan Xuemei. Development of Barrage of Video Website in China[J]. Popular Science & Technology, 2014,16(10): 232-233.

[14] Zhang Yining. Research on Barrage Video from Interactive Perspective[D]. Nanjing: Nanjing Normal University, 2015.

[15] Murakami N, Ito E. Emotional Video Ranking Based on User Comments[C]//Proceedings of

the 13th International Conference on Information Integration and Web-based Applications and Services. 2011: 499-502.

[16] Yamamoto T, Nakamura S. Leveraging Viewer Comments for Mood Classification of Music Video Clips[C]//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA:ACM, 2013: 797.

[17] Su B, Peng J. Sentiment Analysis of Comment Texts on Online Courses Based on Hierarchical Attention Mechanism[J]. Applied Sciences, 2023, 13(7): 4204.

[18] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2019.

[19] Zhang T, Wu F, Katiyar A, et al. Revisiting few-sample BERT fine-tuning[J]. arXiv preprint arXiv:2006.05987, 2020.

[20] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002: 102-111.

[21] Tan S C, Lau P S, Yu X W. Finding similar time series in sales transaction data[C]//Current Approaches in Applied Artificial Intelligence: 28th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2015, Seoul, South Korea, June 10-12, 2015, Proceedings 28. Springer International Publishing, 2015: 645-654.

[22] D'Urso P, De Giovanni L, Massari R. Trimmed fuzzy clustering of financial time series based on dynamic time warping[J]. Annals of operations research, 2021, 299: 1379-1395.

[23] Abbasimehr H, Shabani M. A new framework for predicting customer behavior in terms of RFM by considering the temporal aspect based on time series techniques[J]. Journal of ambient intelligence and humanized computing, 2021, 12(1): 515-531.

[24] Zhang T, Wu F, Katiyar A, et al. Revisiting few-sample BERT fine-tuning[J]. arXiv preprint arXiv:2006.05987, 2020.] and Xiong et al.([4])[Xiong R, Yang Y, He D, et al. On layer normalization in the transformer architecture[C]//International Conference on Machine Learning. PMLR, 2020: 10524-10533.

[25] Xiong R, Yang Y, He D, et al. On layer normalization in the transformer architecture[C]//International Conference on Machine Learning. PMLR, 2020: 10524-10533.

[26] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the royal statistical society. series c (applied statistics), 1979, 28(1): 100-108.

[27] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & geosciences, 1984, 10(2-3): 191-203.

[28] Bouguettaya A, Yu Q, Liu X, et al. Efficient agglomerative hierarchical clustering[J]. Expert Systems with Applications, 2015, 42(5): 2785-2797.