一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IMQ2022-62,IE2022-139,MVE2022-92(2023-03)

Evoked Emotion Distribution Learning through Analysis of Temporal User Comments in Social Media Videos

Yiming WANG[†], Marc A. KASTNER[††], Da HUO[†], Takahiro KOMAMIZU[†††,†],

Takatsugu HIRAYAMA[††††,†], and Ichiro IDE[†]

† Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
††† Mathematical and Data Science Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
†††† Department of Environment Science, University of Human Environments
Motojuku-cho, Okazaki, Aichi 444-3505, Japan
E-mail: † {wangy, huod}@cs.is.i.nagoya-u.ac.jp, ide@i.nagoya-u.ac.jp
†† mkastner@i.kyoto-u.ac.jp ††† taka-coma@acm.org †††† t-hirayama@uhe.ac.jp

Abstract The field of *affective video content analysis*, which aims to estimate viewers' emotions evoked from a input video, is growing as the amount of online video content increases. However, annotating videos with emotions is challenging due to the subjective and ambiguous nature of emotions. This research introduces the Label Distribution Learning (LDL) paradigm to limit the impact of subjectivity by modeling the label of evoked emotions as a distribution rather than a single dominant emotion. In addition, an approach to automatically annotate the viewers' emotion distribution based on user-generated comments instead of annotating them manually is proposed. A video dataset with emotion distribution annotations is composed using this method. An Evoked Emotion Distribution Learning (EEDL) model is adopted to estimate the emotion distribution evoked from social media videos. Experiments using the proposed EEDL model on the composed dataset show promising prospect for using LDL in this task.

Key words Affective video content analysis, Label Distribution Learning, social media

1. Introduction

Recently, videos on the Internet have become one of the most commonly viewed media. This makes the search, recommendation, and analysis of different kinds of videos in social media increasingly difficult. For example, for users, the abundance of videos can make it hard for them to find videos relevant and interesting to them. They may also feel overwhelmed by the sheer amount of content available [7]. Therefore, research on understanding a viewer's experience or the viewer's perception towards different content is needed, in order to provide some insights that can be applied to address the aforementioned challenges. By analyzing the emotions that viewers experience when watching a video, systems are more likely to recommend content that aligns with users' interests and preferences.

However, measuring the evoked emotion is a complex task as emotion is subjective and often ambiguous. This also leads to a problem that annotating datasets related to human emotions is quite time-consuming and cost-intensive because more annotators are needed to reach agreements on the annotation. Therefore, in this report, the Label Distribution Learning (LDL) paradigm [4] is introduced for the annotation of video frames. User-generated comments are utilized as a prior to automatically annotate viewers' evoked emotion distribution to videos.

Specifically, the LDL paradigm allows models to predict the distribution over different classes rather than a single dominant class label explicitly to limit the subjectivity of human emotions. In order to solve the problem that manual labeling of datasets is infeasible, this report analyzes Danmaku, a special type of comments with timestamps, as a means of automatic annotation. Incorporating these, an approach of automatically annotating viewers' emotion distribution to videos is proposed.

At last, a dataset of social media videos with automatically annotated labels of the emotion distribution of viewers when watching these videos is composed using the approach. Also, to evaluate this dataset, an Evoked Emotion Distribution Learning (EEDL) model is proposed as a baseline method for predicting viewers' evoked emotions from an input video.

The contributions of this work are as follows:

- To limit the subjectivity and ambiguity of human emotions, the Label Distribution Learning (LDL) paradigm is introduced to the field of affective video content analysis.
- Different from traditional datasets using manual annotation, an approach is proposed to automatically annotate emotion distribution evoked from videos by analyzing user-generated Danmaku.
- An Evoked Emotion Distribution Learning (EEDL) model is adopted to predict the emotion distribution evoked from social media videos.

2. Related Work

2.1 Danmaku

Danmaku (also known as *bullet screens* in English and referred to as *comments* in Japanese) are real-time messages that appear on the screen during the playback of a video. Danmaku was firstly introduced on the Japanese video sharing Website NicoNico(*1) in 2006 [1]. The feature allows users to post comments or messages that appear on the screen in real-time as a video is played. Many Chinese video sharing Websites, such as Bilibili(*2), have adopted Danmaku and developed their own unique communities around their users. Danmaku have also been adopted by other video Websites around the world, and even in YouTube(*3) a beta version of real-time comments is currently available.

2.2 Label Distribution Learning

Previous research typically considers Single Label Learning (SLL) and Multiple Label Learning (MLL) when predicting classes. However, these paradigms can only predict a dominant class but not the probability of mixed labels. Meanwhile, Label Distribution Learning (LDL) [4] is a learning paradigm designed to model a probability distribution. Different people feel differently to the same stimulus, and similar emotions can have some intrinsic relationships. Therefore, it is natural and intuitive to use the LDL learning paradigm in emotion recognition. There are some previous works on emotion recognition using LDL. Yang et al. [12] proposed a model for joint image emotion classification and distribution learning, which incorporates LDL to enhance performance. Another research [11] suggests a circular-structured representation for emotion distribution learning to improve the accuracy of emotion classification.

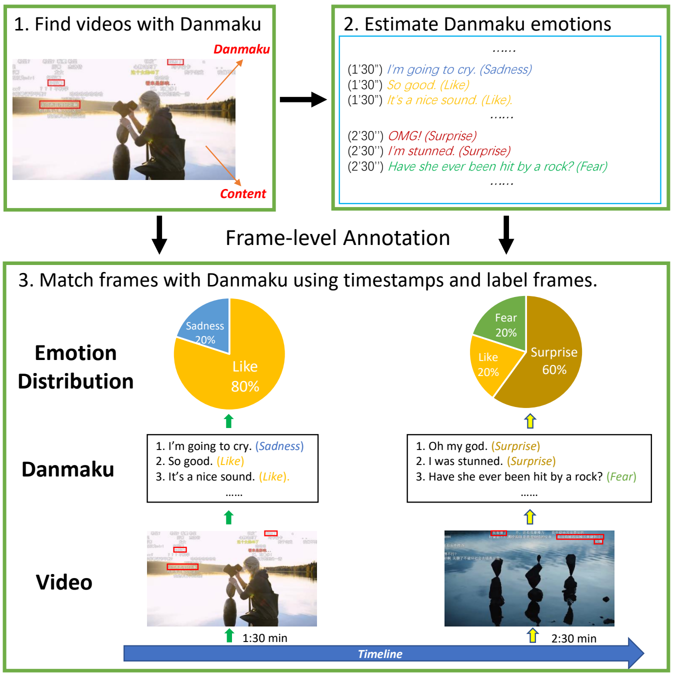

1. Find videos with Danmaku
Danmaku
Content
2. Estimate Danmaku emotions
......
(1'30") I'm going to cry. (Sadness)
(1'30") So good. (Like)
(1'30") It's a nice sound. (Like).
......
(2'30'') OMG! (Surprise)
(2'30'') I'm stunned. (Surprise)
(2'30'') Have she ever been hit by a rock? (Fear)
......
Frame-level Annotation
3. Match frames with Danmaku using timestamps and label frames.
Emotion Distribution
Sadness 20%
Like 80%
Fear 20%
Like 20%
Surprise 60%
Danmaku
1. I'm going to cry. (Sadness)
2. So good. (Like)
3. It's a nice sound. (Like).
......
1. Oh my god. (Surprise)
2. I was stunned. (Surprise)
3. Have she ever been hit by a rock? (Fear)
......
Video
1:30 min
2:30 min
Timeline

Figure 1 Flowchart of the automatic annotation.

3. Dataset Creation

In this section, the methodology to collect the videos for a video dataset containing evoked emotion distribution labels is introduced. Furthermore, the method to annotate videos automatically using Danmaku is proposed.

3.1 Parrot's Emotion Model

Parrot's emotion model [9] is chosen to describe human emotions for the composed dataset due to its uncomplicated and straightforward nature. The model divides emotions into six categories, including three positive (*like*, *joy*, and *surprise*) and three negative (*fear*, *disgust*, and *sadness*) ones. This simplicity is considered more appropriate for describing the emotions contained in colloquial Danmaku. Additionally, the model handles the overlap of emotions well, as it treats *anger* as a subset of *disgust*, which meets the needs of the study for analyzing the emotions in Danmaku.

3.2 Data Collection and Annotation

To analyze the evoked emotion distribution in social media videos, label distribution for each video is needed to train a supervised model for predicting evoked emotion. However, since manual annotation is extremely time-consuming and labor-intensive for video datasets especially with LDL anno-

(*1): https://www.nicovideo.jp/ (Accessed Jan. 26th, 2023)
(*2): https://www.bilibili.com/ (Accessed Jan. 26th, 2023)
(*3): https://www.youtube.com/ (Accessed Jan. 26th, 2023)

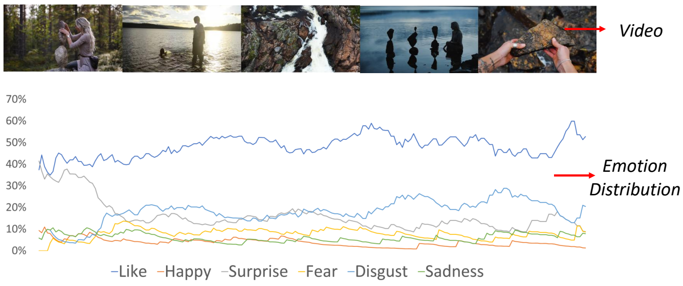


Figure 2 Example of the annotation result.

tation, this work considers using Danmaku, a special kind of comments which contain temporal information, to automatically annotate social media videos to compose a video dataset on LDL evoked emotion as shown in Fig. 1.

The first step is to find videos with a sufficient number of Danmaku available and obtain all Danmaku commented on these videos. Then as the second step, several Danmaku emotion classifiers are applied to detect whether corresponding emotions appear in a Danmaku. As the third step, to automatically annotate the evoked emotions to a video, a Danmaku pool, which contains several Danmaku attached to the video, is created for each frame. Since the emotion classifiers have estimated the emotions contained for all the Danmaku in the second step, the proportion of each emotion contained in the Danmaku pool can be calculated. At last, the frame is annotated with the distribution of these proportions as LDL annotation.

In the following discussions, $P_c(t)$ will be used to represent the label, which is the probability that a viewer will experience the specific emotion c in Parrot's emotion model, where c can take on any value from the set $\{0, 1, 2, 3, 4, 5\}$ corresponding to *like*, *happy*, *surprise*, *fear*, *disgust*, and *sadness*, respectively, while watching the video at timestamp t. Accordingly, it can be defined as:

$$P_c(t) = N_c(t) \Bigg/ \sum\nolimits_{c=0}^{5} N_c(t), \qquad (1)$$

where $N_c(t)$ is the count of emotion c in the Danmaku pool corresponding to the frame at timestamp t.

Fig. 2 shows the distribution of labels $P_c(t)$ for a video. LDL models will be trained with these labels that contain the distribution of different kinds of evoked emotions to social media videos.

Video Collection. For the proposed method, it is important to identify social media videos with a sufficient number of Danmaku. The videos in the composed dataset are collected from the Chinese video-sharing Website Bilibili[*2], where Danmaku is one of the most iconic features of the service. The number of videos is 1,400 and the average duration is 220 sec.

Danmaku Emotion Estimation. To automatically annotate the emotion distribution for videos using Danmaku, the key aspect of the proposed method is to estimate the emotion for each Danmaku. Since there are over 17 million Danmaku corresponding to the collected videos, it is infeasible to annotate them manually. Therefore, Natural Language Processing (NLP)-based emotion classifiers are applied to estimate emotion expressed by Danmaku automatically. Firstly, a Danmaku emotion estimation corpus is constructed manually. As the data is dominantly Chinese, a pretrained Chinese BERT-wwm [3] is employed and fine-tuned on this corpus. At last, six emotion classifiers corresponding to six emotions in Parrot's emotion model are obtained, which are used to estimate the emotion expressed by all the Danmaku attached to the collected videos.

3.3 Emotion Distribution Generation by Analyzing Danmaku

In the automatic annotation process, a Danmaku pool, which contains several Danmaku, is created for each frame to calculate the proportions of each emotion category in it, represented as $P_c(t)$ for the frame calculated by Eq. 1. To obtain the label distributions calculated for a video over time, it is necessary to decide how the count of emotion $N_c(t)$ is represented, that is, how to create a Danmaku pool for each frame. One intuitive approach is to simply create a Danmaku pool that contains all Danmaku in the following timestamp. This method will be referred to as *Timestamp-Matching Method (TMM)* hereafter, as a baseline. Here, $n_c(t)$ is named to represent the number of times emotion c appears in the following one timestamp t. So if the TMM is applied, $N_c(t)$ in Eq. 1 and $n_c(t)$ are identical and set as:

$$N_c(t) = n_c(t). \qquad (2)$$

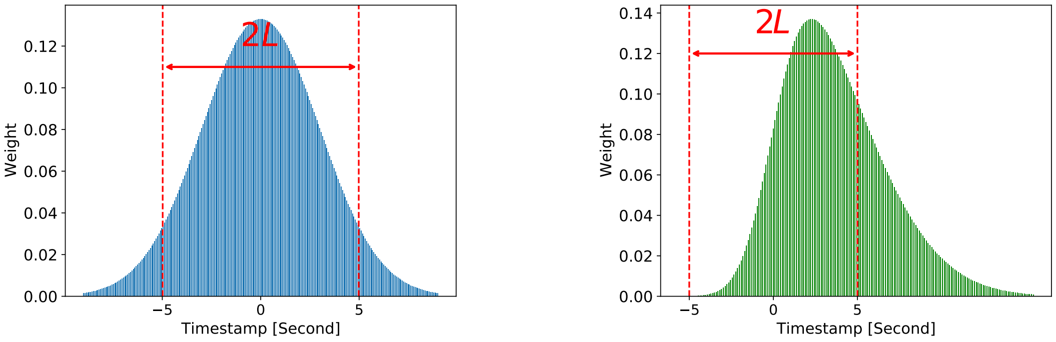
3.3.1 Sliding Window Method

The proposed *Sliding Window Method (SWM)* aims to take into account not only the Danmaku in the following timestamp, but also those from timestamps before and after when creating a Danmaku pool. Specifically, for each frame, a certain length of clip is cut before and after the current timestamp in a sliding window manner. In this case, $N_c(t)$ will be described as:

$$N_c(t) = \sum\nolimits_{i=t-L}^{t+L} n_c(i), \qquad (3)$$

where $\pm L$ sec. is the duration of the sliding window.

Although the SWM makes use of the context information in the video, the difference in timestamps between Danmaku and video frames is also a concern, as a large difference can reduce the relevance of the Danmaku to the video frame and increase noise. Therefore, after creating a Danmaku pool using the SWM, the idea is to give different weights to each



(a) Gaussian distribution weighting

(b) Skewed normal distribution weighting

Figure 3 Distributions used for weighting. The x-axis with a value of 0 represents the timestamp of the video frame to be annotated. *2L* refers to the width of the sliding window. The y-axis refers to the corresponding weight.

Danmaku in the pool according to the difference between the timestamps of Danmaku and video frame.

Gaussian Distribution Weighting. With Gaussian Distribution Weighting (GDW) as shown in Fig.3(a), Danmaku closer to the frame can be given higher weights, while those farther away can be given lower weights. This can help to ensure that the label for each frame is not solely influenced by faraway Danmaku with little relevance to the frame, but rather is a more balanced representation of the emotions of Danmaku at different distances from the frame. The formula for $N_c(t)$ using Gaussian is:

$$N_c(t) = \sum\nolimits_{i=t-L}^{t+L} n_c(i) \cdot f_{\text{Gaussian}}(i; t, \sigma), \tag{4}$$

where $f_{\text{Gaussian}}(i; t, \sigma)$ is the Gaussian function:

$$f_{\text{Gaussian}}(i; t, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(i-t)^2}{2\sigma^2}}, \tag{5}$$

where σ^2 is the variance.

Skew Normal Distribution Weighting. One of the characteristics of Danmaku is that they often appear after the corresponding frame in the video because viewers tend to comment after watching the video frame. In this case, the Gaussian distribution may not take into account the skewed nature of the time distribution of Danmaku. Therefore, the Skew Normal Distribution Weighting (SNDW) as shown in Fig.3(b), which is a variant of the Gaussian distribution and allows for skewness or asymmetry in data, is also considered, which may capture the skewed time distribution of Danmaku and the lag between Danmaku and video frames.

The formula for $N_c(t)$ with skew normal distribution weighting is:

$$N_c(t) = \sum\nolimits_{i=t-L}^{t+L} n_c(i) \cdot f_{\text{skew}}(i; t, \sigma). \tag{6}$$

where $f_{\text{skew}}(i; t, \sigma)$ is the skew normal distribution function:

$$f_{\text{skew}}(i; \mu, \sigma) = \frac{2}{\sqrt{2\pi\sigma^2}} e^{-\frac{(i-t)^2}{2\sigma^2}} \cdot \Phi(\alpha(i-t)), \tag{7}$$

where the function $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. The skewness parameter α controls the degree of asymmetry in the distribution.

4. Evoked Emotion Distribution Learning Model

Multimodal data fusion models are usually used to analyze affective video content by combining information from multiple modalities, such as audio and visual data. Here, a similar framework named Evoked Emotion Distribution Learning (EEDL) model is adopted to predict the emotion distribution evoked from a given video.

The proposed EEDL model is illustrated in Fig. 4, which consists of four parts: feature extraction, temporal module, regression module, and constraint adaptation module.

4.1 Feature Extraction

Both audio and visual features are extracted from the video content to analyze the emotions that tend to be evoked. By extracting both features, multiple dimensions of this information are expected to be captured.

Audio Feature. VGGish [5] is used to extract audio features from the video. The audio is first preprocessed by being segmented into non-overlapping clips with a length of 1 sec., and in case that it is shorter than 1 sec., it is padded with silence. These segments are then passed through the VGGish model, which produces a 128-dimensional feature vector for each 1 sec. clip.

Visual Feature. EfficientNet [10] is chosen as the pre-trained model to extract visual features from videos at 1 frame per sec., which produces a 512-dimensional feature vector.

4.2 Temporal Module

The temporal module in the proposed EEDL model is

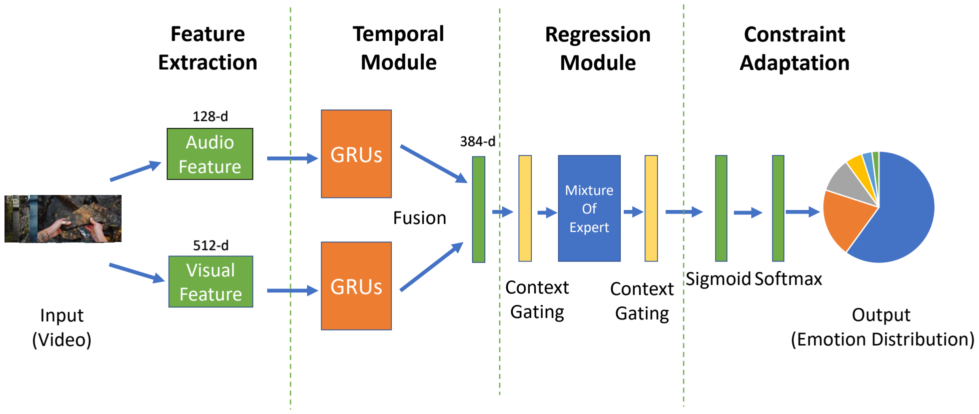


Figure 4 Proposed Evoked Emotion Distribution Learning (EEDL) model.

responsible for learning the temporal characteristics of the video content. In the EEDL model, one separate 2-layer Gated Recurrent Unit (GRU) [2] is trained for each modality (audio and visual) to take into account the temporal characteristics of videos. The final state output of each GRU is a 256-dimensional vector and a 128-dimension vector. They are then concatenated to one vector and fed to the subsequent regression model.

4.3 Regression Module

The regression module accepts the fused output vectors from the temporal module, which is responsible for performing continuous value prediction. The regression module of the EEDL model is a combination of a Mixture of Expert (MoE) [6] and two Context Gating [8] layers placed before and after it to predict the degree of a particular emotion being expressed in the video.

4.4 Constraint Adaptation Module

The constraint adaptation module is responsible for modifying the output of the regression module to meet the constraints of Label Distribution Learning (LDL):

（1） The probability of each class or label must fall within the range of 0 to 1.

（2） The sum of the probabilities for all classes must be equal to 1.

The constraint adaptation module uses a Sigmoid layer to satisfy the first constraint. Then a Softmax layer is used to satisfy the second constraint.

4.5 Implementation

The Kullback–Leibler (KL) loss was used when training the EEDL model, and the batch size was set as 128 with 50 epochs. During the first 20 epochs, the learning rate was fixed at 1×10^{-4}. After that, the learning rate was decayed using a fixed decay rate, resulting in a learning rate of 0 at the end. The number of experts in the MoE was set as 10. The length of the sliding window L was set as 5 sec., and the skewness parameter of skew normal distribution α was set as 1, which means the distribution would be right-skewed to match the characteristic of Danmaku.

5. Experiments

The most commonly used Label Distribution Learning (LDL) metrics include *Chebyshev distance*, *Clark distance*, *Canberra distance*, *Cosine similarity*, *Intersection similarity*, *KL divergence*, and *Progressive Circle distance*, and they are adopted to evaluate the performance of the proposed Evoked Emotion Distribution Learning (EEDL) model and Danmaku pool creation method on the composed dataset.

5.1 Experiments on Danmaku Pool Creation Approaches

As mentioned in Section3.3, the Timestamp-Maching Method (TMM) and the Sliding Window Method (SWM) are proposed as the Danmaku pool creation methods and Gaussian Distribution Weighting (GDW) and Skew Normal Distribution Weighting (SNDW) are proposed for weighting each Danmaku in the pool. Four different Danmaku pool creation approaches were tested: TMM, SWM, SWM+GDW, and SWM+SNDW. The EEDL model with fixed parameters was used in all these four settings.

The goal of this experiment is to explore the performance of different approaches for creating the Danmaku pool on the ease of learning for the EEDL model. Table 1 summarizes the results of the experiment, which indicates that using the SWM instead of the TMM facilitates the model to learn. By comparing the results between using only the SWM and SWM+GDW, the latter is more suitable to weight each Danamaku in the Danmaku pool. However, when the SDNW is applied instead of the GDW, the results become worse. This suggests that actually not only Danmaku after the video frame but also before the frame have a connection to the evoked emotion of videos.

Table 1 Experiments on different Danmaku pool creation approaches.

Approach	**Chebyshev** (↓)	**Clark** (↓)	**Canberra** (↓)	**Cosine** (↑)	**Intersection** (↑)	**KL** (↓)	**PC** (↓)
TMM	0.309	1.553	3.256	0.772	0.595	0.573	0.576
SWM	0.203	0.996	2.000	0.871	0.731	0.251	0.369
SWM+GDW	**0.201**	**0.984**	**1.969**	**0.891**	**0.740**	**0.239**	0.436
SWM+SNDW	0.210	1.032	2.016	0.864	0.733	0.262	**0.349**

Table 2 Experiments on number of videos in the dataset.

Dataset Size	**Chebyshev** (↓)	**Clark** (↓)	**Canberra** (↓)	**Cosine** (↑)	**Intersection** (↑)	**KL** (↓)	**PC** (↓)
20%	0.231	1.042	2.123	0.835	0.700	0.290	0.399
40%	0.214	**0.992**	2.023	0.845	0.721	0.267	0.432
60%	0.204	1.005	2.057	0.847	0.723	0.280	**0.343**
80%	0.216	1.007	2.036	0.858	0.720	0.258	0.390
100% (Original)	**0.203**	0.996	**2.000**	**0.871**	**0.731**	**0.251**	0.369

5. 2 Experiments on Number of Videos in the Dataset

One feature of the proposed dataset is its scalability because it is automatically annotated. In order to verify the benefit of increasing the size of the dataset, this experiment tested a variety of dataset sizes and compared the performance of the same EEDL model across all of them. The labels of the dataset were all generated using the SWM only.

Various proportions of the dataset (20%, 40%, 60%, 80%, and 100%) were selected and divided into training, validation, and testing sets in the same proportion as 8:1:1. Table 2 shows that as the dataset size increases, the metrics move in the direction of high performance, which might be because with the increase of training videos, the model is more generalized and not easy to be disturbed by noisy data.

6. Conclusion

This report proposed a method that can annotate viewers' evoked emotions in social media videos automatically by analyzing Danmaku, that are real-time user comments including timestamps of the video. The Label Distribution Learning (LDL) paradigm [4] was introduced to affective video content analysis, to deal with the subjectivity and ambiguous of emotion. The experimental results demonstrated that utilizing user feedback as a form of crowd-sourced annotation can assist in affective video content analysis, and also revealed a correlation between the emotions expressed by the audience in Danmaku and the emotions evoked in the corresponding social media videos.

Possible future works for the further improvement of the method are as follows:

- Use speech-to-text models to obtain linguistic information from video narrations to achieve a higher performance.
- Use more sophisticated distribution learning models.

Acknowledgement Part of this work was supported by JSPS KAKENHI 22H03612, and a joint research project with NII.

Reference

[1] Q. Bai, Q.V. Hu, L. Ge, and L. He. Stories that big Danmaku data can tell as a new media. *IEEE Access*, Vol. 7, pp. 53509–53519, 2019.

[2] K. Cho, V.B. Merriënboer, C. Gulcehre, D.B. Bahdanau, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Comput. Res. Reposit., arXiv Preprint, arXiv:1406.1078*, 2014.

[3] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Trans. Audio Speech Lang. Process.*, Vol. 29, pp. 3504–3514, 2021.

[4] X. Geng. Label distribution learning. *IEEE Trans. Knowl. Data Eng.*, Vol. 28, No. 7, pp. 1734–1748, 2016.

[5] S. Hershey, S. Chaudhuri, D.P. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, and B. Seybold. CNN architectures for large-scale audio classification. In *Proc. 42nd IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 131–135, 2017.

[6] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.*, Vol. 6, No. 2, pp. 181–214, 1994.

[7] J. Lee and A.E.H. Sami. Large-scale content-only video recommendation. In *Proc. 2017 IEEE Int. Conf. Comput. Vis. Workshops*, pp. 987–995, 2017.

[8] A. Miech, I. Laptev, and J. Sivic. Learnable pooling with context gating for video classification. *Comput. Res. Reposit.„ arXiv Preprint, arXiv:1706.06905*, 2017.

[9] W.G. Parrott. *Emotions in social psychology: Essential readings*. New York, NY, USA: Psychology Press, 2001.

[10] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. 36th Int. Conf. Mach. Learn.*, pp. 6105–6114, 2019.

[11] J. Yang, J. Li, L. Li, X. Wang, and X. Gao. A circular-structured representation for visual emotion distribution learning. In *Proc. 2021 IEEE/CVF Conf. Comput. Vis. and Pattern Recongnit.*, pp. 4237–4246, 2021.

[12] J. Yang, D. She, and M. Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *Proc. 26th Int. Joint Conf. Artif. Intell.*, pp. 3266–3272, 2017.