*1 https://www.nicovideo.jp/ (Accessed: June 4, 2022)
*2 https://www.bilibili.com/ (Accessed: June 4, 2022)

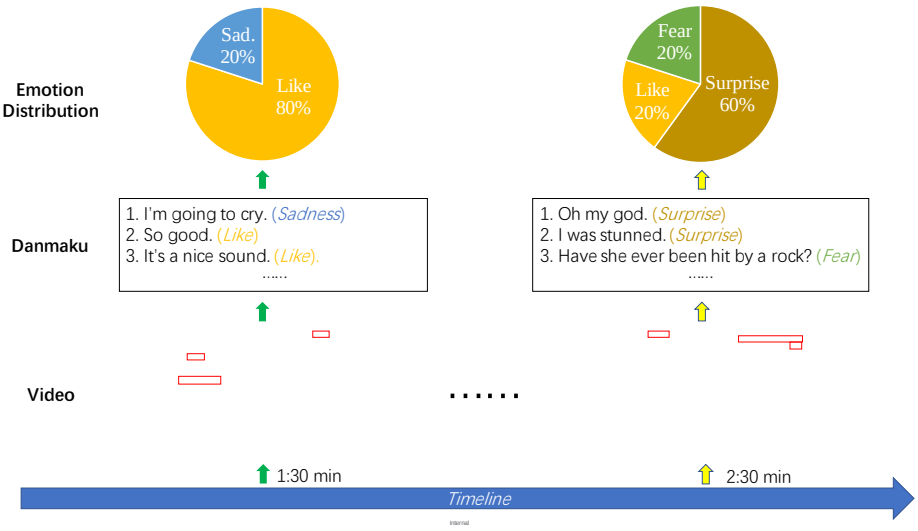

Fig. 1 Example of *Danmaku*. Through Danmaku, viewers share their feelings towards each scene. Danmaku will appear on the corresponding timestamp of the video, so other viewers can see the previous viewers' textual reactions at the same time of watching a scene. We estimate the emotions of Danmaku at a timestamp and output the estimated emotion distribution.

maku emotion analysis corpus to help predict Danmaku emotions. To the best of our knowledge, there is no research on using Danmaku emotion as the ground truth to help train models on predicting evoked emotions on corresponding videos. Our contributions are summarized as follows:

- We build and annotate a Danmaku emotion analysis corpora, which can be used to train the emotion analysis model of short social media texts.
- We automatically label the evoked emotions on videos by modeling the emotion distribution of its Danmaku as a crowd-sourced annotation.
- Label distribution learning is introduced into the evoked expression field, and the results using label distribution learning model prove the feasibility of this method.

2. Related Work

2.1 Affective Video Content Analysis

When watching a video, people may show a variety of emotions, which are often directly related to the content of the video. A multimedia evaluation benchmark MediaEval has a challenge on assessing the emotional impact of movies. Given the LIRIDS [4] dataset, different teams analyze the audio-visual information in videos to predict viewers' emotions while watching. This dataset is one of the most popular datasets used in affective video content analysis. More recently, attention-mechanism is becoming popular in this field, Thao et al. [13] proposed a self-attention mechanism-based method to consider the relation between different modalities. Mittal et al. [14] gave a solution by combining Granger causality and attention mechanism. Both of them achieved a good performance on the LIRIDS dataset.

2.2 Label Distribution Learning

Previous research typically consider Single Label Learning (SLL) and Multiple Label Learning (MLL) to answer the question: *"Which label can describe the instance?"* However, they can only predict the dominant class but not the probability of each label. Meanwhile, Label Distribution Learning (LDL) [2] is a learning paradigm designed to answer the question: *"How much can a label describe the instance?"* There are a lot of work [5,6] on emotion recognition using LDL. Different people feel differently to the same stimulus, and similar emotions always have some intrinsic relationship. Therefore, it is natural and intuitive to use the LDL learning paradigm in emotion recognition.

Table 1 Dataset comparison.

Dataset	Source	Annotation type	Number of videos
LIRIS [4]	Movies	Valence & Arousal	66
EEV [3]	Online videos	15 evoked expressions	23,574
Ours	Online videos	LDL of 6 emotions	208

3. Approach

3.1 Framework

Given a social media video accompanied with Danmaku, a distribution $d = \{d_1, d_2, ..., d_n\}$ is assigned to describe the degree of each evoked emotion from this video, where n is the number of emotion categories. All the degrees d_i should meet the following two constraints: $d_i \in [0,1]$ and $\sum_{i=1}^{n} d_i = 1$. In our data, we use the proportion of a certain emotion in the Danmaku as the degree of that emotion, so the two constraints can be met (see Fig. 1).

We select social media videos with plenty of Danmaku available from a Chinese video-sharing website *Bibilibi*[*2], and then crawl the Danmaku of these videos. A comparison with existing datasets is shown in Table 1.

We estimate the emotions of Danmaku using the trained emotion classifiers. For every 5 seconds interval, we calculate the total number of emotions contained in the Danmaku, and the proportion of each emotion against the total is calculated as the degree of the corresponding emotion label; then, we obtain the ground-truth label distribution of a video. We use the label distribution and the audio-visual feature to train an LDL model, which is used for predicting the degree of each emotion in an arbitrary social media video. This framework is illustrated in Fig. 2.

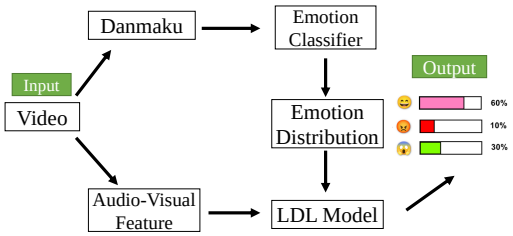

Fig. 2 Framework of the proposed approach. A part of Danmaku used for training emotion classifiers were annotated manually, and the emotion of the other Danmaku was estimated using trained emotion classifiers. After making the label from emotion classifiers and extracting audio-visual features from the video, an LDL model was trained to predict the distribution of the evoked emotion.

3.2 Emotion Model

Since there are no Danmaku emotion corpora publicly available for training emotion classifiers, we decided to construct one. By comparing different discrete emotional models, we chose the Parrot model [12]. There are six emotions defined in the Parrot model, including three positive emotions; *Like*, *Joy*, and *Surprise*, and three negative emotions; *Fear*, *Disgust*, and *Sadness*. Danmaku is usually very short, so its emotion is quite straightforward. Therefore, the emotion should be handled differently from other types of texts. For example, *Disgust* and *Anger* are considered as different categories in many emotion models, but it is difficult to distinguish between them in a Danmaku, because the two emotions in Danmaku usually overlap. Therefore, we chose the Parrot model as the emotional model of this study, which considers *Anger* as a subset of *Disgust*.

3.3 Data Labelling

We semi-automatically annotate Danmaku emotions in a data augmentation manner. In detail, five annotators label a part of the Danmaku manually; then, we use these labeled data to train the emotion classifiers. With these trained models, we can get an estimation for unlabeled Danmaku. Here, we set two thresholds. If the probability is larger than the *high* threshold, we set it as a positive sample, and if the probability is below the *low* threshold, we set it as a negative sample; otherwise, we annotate it manually. Then we add them into the corpus and retrained the emotion classifier. In our experiment, we set the high threshold as 0.8 and the low threshold as 0.4. This is mainly because we use many Danmaku with other emotions as negative samples, and then the accuracy of detecting negative samples is higher than positive ones.

After constructing the Danmaku emotion corpora for training emotion classifiers, we annotate videos using the predicted emotions of Danmaku corresponding to them and then calculate the probability of each emotion at a timestamp as the ground truth for training an LDL model.

3.4 Danmaku Emotion Prediction

We divide the Danmaku emotion corpora into 8:1:1 as the training set, validation set, and test set. Most of these Danmaku are in Chinese, so we use Chinese BERT-wwm [15] as a pretrained model and finetune it with our Danmaku emotion corpus. Then we obtain six binary emotion classifiers, each of which corresponds to one emotion in the Parrot model.

Table 2 Number of labels and the prediction results of emotion classifers.

	Like	Joy	Sur.	Fear	Dis.	Sad.
Annotations	3,343	3,889	2,113	3,215	3,160	3,224
Acc	0.9366	0.9444	0.8700	0.9036	0.8754	0.8833
AUC	0.9365	0.9380	0.8709	0.9039	0.8758	0.8738

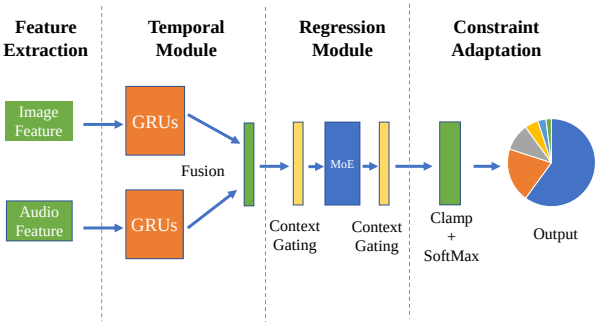

Fig. 3 Proposed LDL evoked emotion prediction model.

We use Accuracy (Acc) and Area Under Curve (AUC) to test the performance of these emotion classifiers and choose the ones with the highest Accs as the final emotion classifiers to predict whether a Danmaku has an emotion. The prediction results in the test set and the number of labels for each emotion are shown in Table 2. Besides, 7,329 neutral Danmaku were annotated and set as negative samples when training emotion classifiers.

3.5 Evoked Emotion Prediction

Recently, multimodal data fusion models are used for affective video content analysis. We also use a similar model. After extracting the audio-visual features from a video, we use a temporal model to learn the temporal information, and then fuse the vectors from different modalities together as one vector before feeding them into the next module. Then we use a regression model for continuous value prediction. Considering the constraints of LDL models, there is another clamp layer and a softmax layer modifying the output to meet the constraints of LDL models. So the final output of this model is $d_p = \{d_1^p, d_2^p, ..., d_n^p\}$, which contains the predicted degrees of each class in the emotion model. Details will be explained in Section 4.

4. Base Model

The proposed LDL evoked emotion prediction model as illustrated in Fig. 3, consists of four parts: feature extraction, temporal module, regression module, and constraint adaptation module.

4.1 Feature Extraction

We extract features from videos at 1 Hz. In detail, for visual features, we choose EfficientNet [7], which is pretrained on ImageNet for image classification and leads to a 512-dimension vector. The audio segment is also extracted at 1 Hz with a maximum 0.5 second padding on both sides of the video. We use the Vggish [8] which is pretrained on a subset of YouTube-8M and gives an 128-dimension vector. Since our labeling was performed at 0.2 Hz, we give all frames among every 5 seconds interval the same ground truth.

Table 3 Modality comparison. We use information from different modalities to train LDL models. In this table (↑) higher is better and (↓) lower is better.

Modality	Chebyshev (↓)	Clark (↓)	Canberra (↓)	Cosine (↑)	Intersection (↑)	KL (↓)
Audio	0.225	1.060	2.187	0.8183	0.696	0.348
Visual	0.236	1.063	2.205	0.8114	0.685	0.363
Audio + Visual	**0.220**	**1.055**	**2.148**	**0.8566**	**0.710**	**0.333**

4.2 Model Architecture and Implementation

We choose the 2-layer GRUs [9] as a temporal module to learn the temporal characteristics of videos. For each modality, there is a 2-layer GRUs corresponding to it. After the GRUs, we concatenate the outputs of different modalities as inputs to the regression module. In the regression module, we use the structure of a combination of two Context Gatings [10] and a Mixture of Expert (MoE) [11]. Context gating helps select important features by giving them a high weight and discard trivial features by giving them a low weight. In general, a neural network performs well in some parts of a dataset but not so well in others. MoE trains several neural networks at the same time and gives suitable ones higher weights when dealing with different parts of the dataset. In our experiments, we set the number of experts as 8.

Since LDL requires the prediction to meet two constraints $d_i \in [0,1]$ and $\sum_{i=1}^{n} d_i = 1$. We apply a post-process to the output. Concretely, we clamp the predicted continuous value after the regression module to $[0,1]$ by setting the value of less than 0 to 0 and greater than 1 to 1 to satisfy the first constraint. Then we normalize the clamped results through a softmax layer to meet the second constraint. We use the Kullback-Leibler (KL) loss to train the model, and the dropout is set as 0.3. The batch-size is 16 and the learning rate is 5e-4 when training.

4.3 Expriment Results

The summarized experimental results of different modalities are in Table 3. We compare the use of visual information only, audio information only, and combined audio-visual information. We use six metrics including Chebyshev distance, Clark distance, Canberra distance, Cosine similarity, Intersection similarity, and KL loss, to test the performance of our model.

The results show that audio modality contains more information that affects human emotions. In addition, when using the audio-visual information together, more information brings improvement. The difference is small, which may be because we use the MoE, which performs well in a large dataset, in the regression module and the still restricted amount of data used for training. We will increase the video length in the future to see if the results will improve.

5. Conclusion and Future Work

We analyzed the correlation between the viewers' expressed emotions in Danmaku and the evoked emotions on the corresponding social media videos. The results of the preliminary model prove the feasibility of the idea and are an initial step for future research in this direction. We believe this work shows us that in addition to audio-visual features of the video, user feedbacks could help us in affective video content analysis through crowd-sourced annotation.

In the future, we will try the following ways to improve our work. First, we will increase the number of videos. Second, we will try to use the internal relationship between different emotions to make a new loss function to train the model in a multi-task way. Third, we will also try different temporal modules and regression modules. We hope that this work can help with the analysis of subjectivity about evoked emotions of videos.

Acknowledgement

Parts of this research were supported by JSPS KAKENHI 22H03612, and also a joint research with NII, Japan.

References

[1] J.A. Russell, A circumplex model of affect, Personality & Social Psychology, 39(6):1161–1178, 1980.

[2] X. Geng, Label distribution learning, IEEE Trans. on Knowledge & Data Engineering, 28(7):1734–1748, 2016.

[3] J.J. Sun, T. Liu, A.S. Cowen, F. Schroff, H. Adam, and G. Prasad, EEV: A large-scale dataset for studying evoked expressions from video, Computing Research Repository, arXiv Preprint, arXiv:2001.05488, 2021.

[4] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, LIRIS-ACCEDE: A video database for affective content analysis, IEEE Trans. on Affective Computing, 6(1):43–55, 2015.

[5] 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. on Artificial Intelligence, pages 3266–3272, 2017.

[6] 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. on Computer Vision and Pattern Recognition, pages 4235–4244, 2021.

[7] M. Tan and Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, In Proc. 36th Int. Conf. on Machine Learning, pages 6105–6114, 2019.

[8] S. Hershey, S. Chaudhuri, D.P. W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, and K. Wilson, CNN architectures for largescale audio classification, In Proc. 42nd IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pages 131–135, 2017.

[9] K. Cho, B.V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, Computing Research Repository, arXiv Preprint, arXiv:1406.1078, 2014.

[10] A. Miech, I. Laptev, and J. Sivic, Learnable pooling with context gating for video classification, Computing Research Repository, arXiv Preprint, arXiv:1706.06905, 2017.

[11] M.I. Jordan and R.A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, Neural Computation, 6(2):181–214, 1994.

[12] W.G. Parrott, Emotions in social psychology: Essential readings, Psychology Press, Philadelphia, PA, USA, 2001.

[13] H.T.P. Thao, B. Balamurali, D. Herremans, and G. Roig. AttendAffectNet: Self-attention based networks for predicting affective responses from movies, In Proc. 25th Int. Conf. on Pattern Recognition, pages 8719–8726, 2021.

[14] T. Mittal, P. Mathur, A. Bera, and D. Manocha, Affect2MM: Affective analysis of multimedia content using emotion causality, In Proc. 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pages 5657–5667, 2021.

[15] Y. Cui, W. Che, T. Liu, B. Qin and Z. Yang, Pre-training with whole word masking for Chinese BERT, IEEE/ACM Trans. on Audio, Speech, and Language Processing, 29:3504–3514, 2021.

[16] Y. Ma, X. Liang, and M. Xu, THUHCSI in MediaEval 2018 emotional impact of movies task, In Working Notes Proc. MediaEval 2018 Workshop, pages 149–152, 2018.

[17] P. Philippot, Inducing and assessing differentiated emotion-feeling states in the laboratory, Cognition & Emotion, 7(2):171–193. 1993.