

# ICDAR 2021 Competition on Multimodal Emotion Recognition on Comics Scenes

Nhu-Van Nguyen<sup>1,3</sup> (✉)<sup>[0000-0003-2271-6918]</sup>, Xuan-Son Vu<sup>2</sup><sup>[0000-0001-8820-2405]</sup>, Christophe Rigaud<sup>1</sup><sup>[0000-0003-0291-0078]</sup>, Lili Jiang<sup>2</sup><sup>[0000-0002-7788-3986]</sup>, and Jean-Christophe Burie<sup>1</sup><sup>[0000-0001-7323-2855]</sup>

<sup>1</sup> L3i, La Rochelle University, France

{nhu-van.nguyen, christophe.rigaud, jcburie}@univ-lr.fr

<sup>2</sup> Department of Computing Science, Umeå University, Sweden

{sonvx, lili.jiang}@cs.umu.se

<sup>3</sup> INSA-Lyon, France

**Abstract.** The paper describes the “Multimodal Emotion Recognition on Comics scenes” competition presented at the ICDAR conference 2021. This competition aims to tackle the problem of emotion recognition of comic scenes (panels). Emotions are assigned manually by multiple annotators for each comic scene of a subset of a public large-scale dataset of golden age American comics. As a multi-modal analysis task, the competition proposes to extract the emotions of comic characters in comic scenes based on visual information, text in speech balloons or captions and the onomatopoeia. Participants were competing on CodaLab.org from December 16<sup>th</sup> 2020 to March 31<sup>th</sup> 2021. The challenge has attracted 145 registrants, 21 teams have joined the public test phase, and 7 teams have competed in the private test phase. In this paper we present the motivation, dataset preparation, task definition of the competition, the analysis of participant’s performance and submitted methods. We believe that the competition have drawn attention from the document analysis community in both fields of computer vision and natural language processing on the task of emotion recognition in documents.

**Keywords:** Multimodal fusion, Emotion recognition, Multi-label classification

## 1 Introduction

Comics is a *multi-billion dollar industry* which is very popular in North America, Europe, and Asia. Initially, comics were printed on paper books, but nowadays, digitized and born-digital comic books become more and more popular and spread culture, education and recreation all over the world even faster.

However, they suffer from a limited automatic content understanding tools which restricts online content search and on-screen reading applications. To deliver digital comics content with an accurate and user-friendly experience on all mediums, it is often necessary to slightly or significantly adapt their content [1].

These adaptations are quite costly if done manually at large scale, so automatic processing are helpful to keep cost acceptable. This is one of the reasons why the comic book image analysis has been studied by the community of document analysis since about a decade. However, there still exist many challenges to be solved in this domain. While the comic elements such as scenes (panels), balloons, narrative and speech text are quite well detected and segmented now, the character (protagonist) detection, text recognition and element relationship analysis are still challenging, and it is important to draw more efforts from the research community to address these challenges. Moreover, complex tasks such as story understanding or scene analysis have not been well studied yet [1].

### 1.1 Human Emotions

“Master the human condition through word and image in a brilliantly minimalistic way” is one of important requirements in making comics [12], in order to engage readers. Here we look at how to model human emotions to better analyze and understand emotions in comics in a reversed manner.

Researchers on human emotions have approached the classification of emotions from one of two fundamental viewpoints [24]: (1) discrete model where emotions are discrete arise from separate neural system [4,19]; (2) dimensional model where emotions can be characterized on a dimensional basis in groupings [17,11]. Table 1 presents the four popular models for basic emotions [24]. Researchers have debated over whether a discrete or dimensional model to emotion classification was most appropriate, and studies showed that one model may not apply to all people. *Discrete model* uses a limit set of basic emotions, while *dimensional model* emphasizes the co-occurrence of the basic emotions to contribute to *the individual’s affective state*. In this competition, we preferred discrete model by considering the diverse background of crowdsourcing annotators, and the purpose of emotion recognition in comic scene. Further motivated by this Kaggle challenge<sup>4</sup>, we decided to add ‘*neutral*’ to the label list since we believe there does not always exist emotions in any given context in daily life as well in comic scene. Additionally, by considering the challenges of explicitly recognizing emotions in comics, we added a label ‘*others*’. From the above investigations, we finally came out an eight-class label list for this competition including *angry, disgust, fear, happy, sad, surprise, neutral, and others*.

Table 1: Four popular basic emotion models (Yadollahi et al. [24])

Study	Basic emotions	Model types
Ekman[4]	anger, disgust, fear, joy, sadness, surprise	discrete
Plutchik[17]	anger, anticipation, disgust, fear, joy, sadness, surprise, trust	dimensional
Shaver[19]	anger, fear, joy, love, sadness, surprise	discrete
Lovheim[11]	anger, disgust, distress, fear, joy, interest, shame, surprise	dimensional

<sup>4</sup> <http://bit.ly/kaggle-challenges-in-representation-learning>

## 1.2 Emotions in EmoRecCom Challenge

In this competition, we propose to tackle one of the challenge of comic scene analysis: the emotion recognition of comic scene. The emotions come from comic characters feelings in the story and are materialized (to be transmitted to the reader) by the visual information, the text in speech balloons or captions and the onomatopoeia (comic drawings of words that phonetically imitates, resembles, or suggests the sound that it describes), see Fig. 1. While emotion recognition are widely studied in other domains and data such as natural images and multi-modal data issues from social networks, it is not yet exploited in comics images which contain both image and text. Motivated by the value of multi-modality based approaches, the competition task encourages the participants to take advantage of multiple representation of resources to infer the emotions. The task hence is a multi-modal analysis task which can take advantages from both fields: computer vision and natural language processing which are one of the main interests of the document analysis community.



Fig. 1: An example of a comic character in a panel with visual emotion and caption text. It is noted that resulted texts from the automatic transcription method may contain errors (e.g., red underlined words in the example).

For this competition, we crowd-sourced the image annotation step to get eight binary label for eight target emotions associated to each image of the ground-truth (angry, disgust, fear, happy, sad, surprise, neutral, others). The competition participants were asked to propose up to any number of positive labels for each image. The statistics of emotions in our EmoRecCom dataset are shown in Table 2.

Table 2: Statistics of the EmoRecCom dataset with the number of images for each emotion in the ground truth.

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Others
#	4005	3678	3485	4197	1525	3435	6914	670

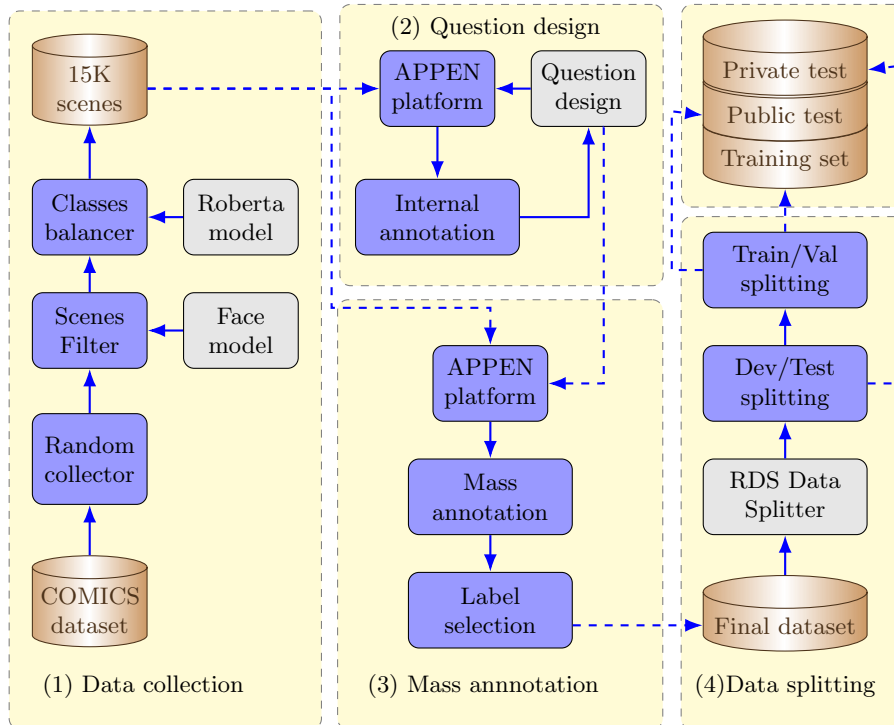


Fig. 2: Our data preparation process has 4 phases: (1) Data collection; (2) Question design (internal annotation); (3) Mass annotation; and (4) Data splitting.

We setup a public website<sup>5</sup> to centralize all related information and participants were competing on the CodaLab platform<sup>6</sup>, an open source framework for running competitions, from December 16<sup>th</sup> 2020 to March 31<sup>th</sup> 2021. The challenge has attracted 145 registrants, 21 teams have joined the public test phase, and 7 teams have competed in the private test phase.

In the following, we will describe data preparation, proposed challenges, participant’s proposed methods and discuss the results.

## 2 Data Preparation

In the competition, we propose a comic scenes dataset which is composed of comic images from the comic books COMICS public dataset [7]. The COMICS dataset<sup>7</sup> includes over 1.2 million scenes (120 GB) paired with automatic textbox transcriptions (the transcriptions are done by Google Vision OCR, which includes some recognition errors, see Fig. 1). For the overall data preparation

<sup>5</sup> <https://emoreccom.univ-lr.fr>

<sup>6</sup> <https://competitions.codalab.org/competitions/27884>

<sup>7</sup> <https://obj.umiacs.umd.edu/comics/index.html>

process, Fig. 2 shows the workflow of the four phases which are described in next four sections.

## 2.1 Data Collection

First, we random select 50K images of comic scenes from COMICS dataset. Then we filter out the scenes which do not contain any textbox transcriptions or person faces. To detect faces in scenes, we use the comic face recognition model from the work in [14]. Afterwards, we train a multi-class emotion classification model based on RoBERTa [10], which help to select the potential scenes for the mass annotation phase. The model is trained using EmotionDataset [15] available online<sup>8</sup>. We used this text-based classification model on the transcription to predict the eight-class emotion of 50K comic scenes. We selected up to 2K scenes for each emotion detected by fine-tuning the *RoBERTa-Large* model (in which “disgust” were detected in 5,017 images and other emotions were detected in less than 2K images for each of them). After the previous step, we overall had 8,000 comic scenes. To ensure the balance of scenes containing different emotions, we randomly selected 4,500 scenes where the RoBERTa could not detect any emotion from the remaining scenes of the 50K set. These 12.5K (4.5K+8K) scenes were proposed to the annotators on the crowd-sourcing platform for annotation. To be more precise, at least three annotators were assigned for each scene.

## 2.2 Question design and mass annotation

We chose to annotate the dataset using crowd-sourcing service platform in order to easily annotate several times each image by different person in order to reduce the subjectivity bias. We experimented different designs and questionnaire to select the most suitable approach for the annotation process. We compared platforms like “Amazon Mechanical Turk” (AMT) - mturk.com, “Appen” - appen.com, and “Toloka” - toloka.ai, which all provide ready-to-use web-based interface for image classification tasks. We selected “Appen”, who bought out CrowdFlower few years ago, for its renown quality and to avoid any ethical issues with AMT.

**Annotator selection.** Since all comic scene images are originated from American comics, we did not ask to limit the geographical origin of the annotator but instead we required experienced and accurate annotators with good English skills. We remunerate each annotated image 3 ¢ and ask annotator to spend at least 10s on each image.


**Annotation tool.** We customized the default web tool proposed by *Appen* in order to provide custom guidelines and conditional answers. The basic annotation sequence was as follows:

<sup>8</sup> <https://github.com/collab-uniba/EmotionDatasetMSR18>

- Read ALL TEXTS in the given image.
- Check FACE expressions of all characters.
- Connecting (mentally) TEXTS with the FACE expressions.
- Decide labels: angry, disgust, fear, happy, sad, surprise, neutral, other.

To let new annotators get familiar with the task, we provide six must-read examples of correct/wrong emotion annotations. After reading these instructions, an image from a particular scene is displayed with a questionnaire to fill up and submit. The questionnaire is composed of a main question: “How many actors are in the scene (visible or not)?” with the possibility to answer between 0 and 3 actors (comic character). Then, word sentiment and face emotions are asked for each actor as shown in Fig. 3. The final question is: “Based on the above answers, which emotions are in the scene?”. All question is requiring an answer.

**1** Read comic scene



**2** Answer questions related to each character

How many actors are in the scene (visible or not)? (required)

0

1

2

3

1st actor: word sentiment level (required)

Very Positive

Slightly Positive

Neutral

Slightly Negative

Very Negative

NOT APPLICABLE

1st actor: face emotions (required)

Angry

Disgust

Fear

Happy

Sad

Surprise

Neutral

Other

NOT APPLICABLE

**3** Decide final label set

Based on above answers, which emotions are in the scene? do not miss any emotions) (required)

(please do not miss any emotions) (required)

Angry

Disgust

Fear

Happy

Sad

Surprise

Neutral

Other

Fig. 3: Interactive annotation tool for the emotion annotation process on *Appen* with three main parts: (1) the scene, (2) questions, and (3) final label set. The number of questions changes dynamically based on the number of actors. For each actor, there will be two corresponding questions: (a) emotion based on textual information, and (b) emotion based on visual information. After answering emotions for all actors, annotators decide final label set of the given scene.

**Annotation quality.** We create the final labels by majority voting method, which is straightforward and meaningful. We take common emotions which are

chosen by at least 2 annotators. Images that do not have emotion in common (that the three annotators does not have the consent on any labels) are ignored. In total of 2,031 images were ignored, so the EmoRecCom dataset contains 10,199 comic scenes at the end. The final label is binary, an emotion is present in the scene (value 1) or not (value 0).

In addition, we create the probability label (or label certainty) based on the frequency an emotion is chosen by annotators for each scene. This probability are provided to participants but we not used for evaluation.

### 2.3 Data Splitting

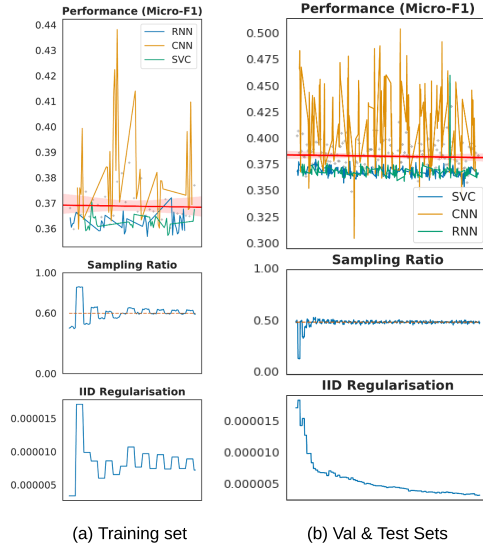


Fig. 4: Learning Dynamics for splitting the data of EmoRecCom into 3 sets (public training, public testing, and private testing) using  $RDS^{STO}$  [13].

act as base-learners on textual data of the task as follows.

1. **RNN**. Recurrent neural network (RNN) is a strong baseline for sequence classification tasks. Here the network is designed with 1 embedding layer, 1 LSTM layer of 64 units, and an output layer of 8 neuron units.
2. **CNN-Text**. The CNN network consists of 1 embedding layer, 4 pairs of Conv2D + MaxPool2D layers, and an output layer of 8 neuron units.

Data splitting process is a non-trivial task for linear regression tasks [22]. The same challenge appears in multi-label classification tasks as using random selection method does not guarantee a fair number of samples across multiple labels. Therefore, we apply the reinforced data splitting method (RDS) [13] to result three sets for the data challenge. Moreover, using the fair splitting approach of RDS, we reduce the race for a small gain in performance, but poor in generalization - i.e., try to improve performance on majority labels to gain better performance overall, but not consider minority labels.

**Baseline models.** To apply RDS [13], it is required to implement baseline learning models for the given task. Here we choose three different baseline models to

3. **SVC**. This model is built based on *OneVsRestClassifier*<sup>9</sup> of Scikit-Learn [16]. For each of 8 classifiers, data of a label is fitted against all the other labels.

Figure 4 displays the learning patterns of the splitting process. Similar to [9,22], we applied RDS two times to split the data into three sets including *public train*, *public test*, and *private test* sets.

Table 3: Simple statistics of the EmoRecCom dataset with the number of comic scenes for each competition phase.

	Warm-Up	Public Training	Public Testing	Private Testing
#	100	6,112	2,046	2,041

The final dataset composes of training set, public test set and private test set (see Table 3). There are 6,112 training examples with the respective annotated labels, 2,046 examples (transcriptions + images) of public test set without labels. The private test set contains 2,041 examples without labels. Participants can evaluate the results on the public test set and the private test set by uploading it to the competition website hosted by <https://codalab.org>.

### 3 EmoRecCom Challenge

#### 3.1 Multi-label emotion classification task

In this competition, participants designed systems to learn two modalities of data: images and text (automatic transcriptions). The objective is to assign multiple emotions for each data sample. At test phase, their system is presented with a set of comic scenes (each scene is a pair of image and text), and must determine the probability of the 8 emotions to appear in each scene. This is a classic multi-label classification problem.

For participants, we give the access to the private test set only one week to upload the results, before the close of the competition. To be fair between participants, all participants have to register any pre-trained models or external datasets that they have used in their system and they must demonstrate that they did not manually label the private test set. We reserve the right to disqualify entries that use any unregistered models/datasets or that may involve any manual labeling of the test set.

#### 3.2 Competition platform

Participants competed on CodaLab.org<sup>10</sup> from 16<sup>th</sup> December 2020 to 31<sup>th</sup> March 2021. There are three phases:

<sup>9</sup> <http://bit.ly/scikit-learn-multilabel-clf>

<sup>10</sup> <https://competitions.codalab.org/competitions/27884>



1. **Warm Up:** 16<sup>th</sup> December 2020 to 10<sup>th</sup> January 2021, participants are given a warm-up dataset of 100 samples to get used with the dataset.
2. **Public data:** From 10<sup>th</sup> January 2021 to 24<sup>th</sup> March 2021, participants are provided 6,112 training examples with the respective annotated labels and a testing set consists of 2,046 examples without labels. Participants can submit their prediction results to the platform and see the results as well as their ranking in the public leader board.
3. **Private Test** From 24<sup>th</sup> March 2021 to 31<sup>th</sup> March 2021, participants are provided 2,041 examples without labels. They must submit the prediction results for this private dataset to the platform before the deadline.

### 3.3 Evaluation protocol

Evaluation scripts are made and setup in the platform Codalab.org so participants can upload the predictions and get the evaluation automatically. As mentioned earlier, there are 8 emotion classes including: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral, 7=Others. Participants must submit the result in the same order as the testing set, with the score (probability) indicating the presence of each emotion as the following format:

image_id	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Others
0_27_5	0.5351	0.0860	0.0693	0.1317	0.0443	0.00883	0.2858	0.1947

### 3.4 Evaluation metric

The submissions have been evaluated based on the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) score. The ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. The Area Under the ROC Curve (AUC) summarizes the curve information in one number.

The ROC-AUC was calculated between the list of predicted emotions for each image given by the participants and its corresponding target in the ground truth (as described in Section 2.2). This is a multi-class classification where the chosen averaging strategy was one-vs-one algorithm that computes the pairwise ROC scores and then the average of the 8 AUCs, for each image [5]. In other words, the score is the average of the individual AUCs of each predicted emotion. To compute this score, we use the Scikit-learn implementation<sup>11</sup>.

## 4 Baselines and Error Analysis

### 4.1 Baselines

We implemented a text-only baseline model with two variants to act as baselines for the task. The baseline and its variants are implemented with two intuitions

<sup>11</sup> <http://bit.ly/scikit-learn-auc>

in mind: (1) they should not be the same as those models which were previously used as base-learners in the Data Splitting Process to avoid any biases, (2) they should leverage powerful pre-trained language or language-vision models such as DistilBert [18] or Roberta [10]. Thus, a custom multilabel model was built on top of following language models including Roberta<sub>Base</sub> and Roberta<sub>Large</sub>. The baseline and its variants were trained using BCEWithLogitsLoss.

## 4.2 Error Analysis



(a) One modality is not enough. In this example, the text-based baseline model cannot detect the emotion “fear”, which is present mostly in the visual modality.



(b) Bad accuracy for “sad, other” : due to the imbalance issue, the accuracy of the two emotions “sad, other” are the worst. Here “sad” was not detected.



(c) Different perceptions about emotions presented in the scenes: this scene can be perceived as either “disgust” (ground truth) or “angry” (prediction).



(d) Complicated scene due to multimodality: human annotators proposed “angry, disgust, sad, surprise”, whereas the baseline model predicted “angry”.

Fig. 5: The four most common errors of the baseline.

Our baselines reached the AUC score of 0.5710 for Roberta<sub>Base</sub> and 0.5812 for Roberta<sub>Large</sub> for the public test. Therefore, we used the Roberta<sub>Large</sub> for the private test which reached the AUC score of 0.5867. We showed in Fig. 5 the four most common errors of the baseline model.

## 5 Submissions and Results

### 5.1 Participation

Within three and a half months, the challenge has attracted 145 registered participants. During the competition, 21 teams have submitted their results and recorded nearly 600 submission entries. In final, there are 5 top teams that submitted their methods to the competition organizers. Table 4 summarizes the approach and the results obtained from the top 5 teams who submitted documents describing their approach for the final evaluation.

Table 4: Top 5 teams on private test data with submitted papers describing their final approaches.

No	Team Name	Team members	Affiliation	AUC
1	S-NLP	Quang Huu Pham, Viet-Hoang Phan, Viet-Hoang Trinh, Viet-Anh Nguyen, and Viet-Hoai Nguyen	RnD Unit, Sun Asterisk Inc	0.6849
2	DeepblueAI	Chunguang Pan, Manqing Dong, Zhipeng Luo	DeepblueAI Technology	0.6800
3	NELSLIP	Xinzhe Jiang*, Chen Yang*, Yunqing Li*, Jun Du (* equal contribution)	NET-SLIP lab, University of Science and Technology of China	0.6692
4	DETA	Quang-Vinh Dang, Guee-Sang Lee	Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea.	0.6324
5	Gululu	Xiaoran Hu, Masayuki Yamamura	Department of Computer Science Tokyo Institute of Technology Tokyo, Japan	0.5716
-	Baseline	EmoRecCom organizers	-	0.5867

*1<sup>st</sup> position: S-NLP - team from Sun Asterisk Inc, Japan.* The team experiments different approaches to fuse both image and text modalities. Using specialized multi-modal framework such as MMBT (MultiModal BiTransformer) framework [8] does not give the best results, this team uses the conventional multimodal fusion architectures for this task. Three fusion levels are performed: the feature level or early fusion, the decision level or late fusion, and the mid-fusion. This method used EfficientNetB3 [20], Resnet [6] as the backbone for visual feature extraction, and RoBERTa [10] for textual features extraction. At the end, the average of prediction scores from 5 different models (image only, text only, early fusion, mid-fusion, late fusion) is used as the final score. Different experiments of this team are shown in Table 5.

*2<sup>nd</sup> position: DeepblueAI - team from DeepblueAI Technology, China.* This method uses the average prediction score from two models as the final prediction. The first model leverages the BERT-base [3] for the textual information only. The second model takes both image and textual information as the inputs. This method integrates the image embedding (Resnet50 [6]) with the textual information as the inputs of the BERT-based model, where the image embedding is considered

as a special token of the BERT-based model. The average of the token’s embedding is further processed with a multi-sample dropout module for getting the prediction. The early fusion architecture is illustrated in Fig. 6.

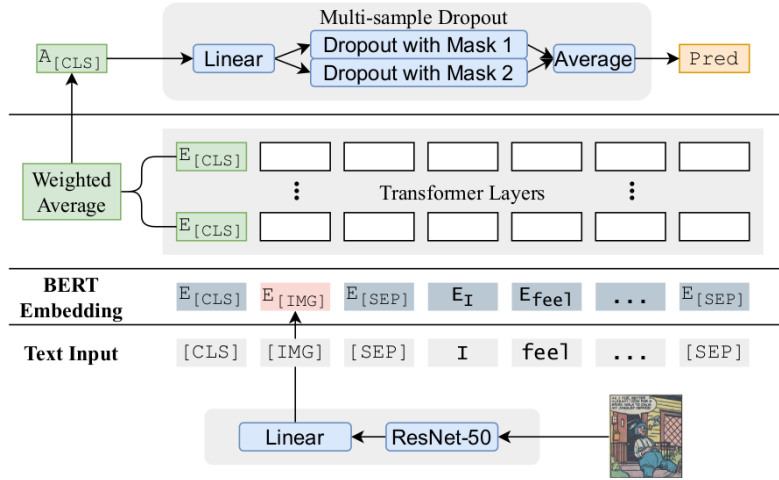


Fig. 6: Early fusion approach by the 2<sup>nd</sup> team - DeepblueAI.

3<sup>rd</sup> position: *NELSLIP* - team from *NEL-SLIP lab, University of Science and Technology of China*. This team uses pretrained models like BERT[3], DistilBERT [18] and XLNet [25] to extract the feature of text modality, and ResNet50 to extract the feature of visual modality. The concatenation of a visual feature and a text feature were then processed by a 4-layer self-attention module and a Fully Connected (FC) classifier. Finally, the overall performance is boosted by the stacking ensemble of 100 different models (see Table 6).

4<sup>th</sup> position: *DETA* - team from *Department of Artificial Intelligence Convergence, Chonnam National University, Korea*. This method fuses the text features

Table 5: Different experiments of the winner team - S-NLP.

Method	Backbone	ROC-AUC on test
Text Only	RoBerta	0.6423
Image Only	Efficient (B3)	0.5412
Early fusion	RoBerta	0.6358
Late fusion	Efficient (B3) + RoBERTa	0.6288
Mid-fusion	Efficient (B3) + RoBERTa	0.6654

Table 6: Different experiments by the 3<sup>rd</sup> team - NETSLIP

Methods	ROC-AUC on val	ROC-AUC on test
Vision-Text model (ResNet-50 & BERT)	0.6320	-
XLNet	0.6302	-
BERT	0.6478	-
DistilBERT	0.6528	-
Averaging ensemble of 100 models	-	0.6674
Stacking ensemble of 100 models	-	0.6692

and image features from pre-trained models, RoBERTa [10] for text and EfficientNet [20] for images. The concatenation of the text features and the image features goes through a series of  $1 \times 1$  convolutions to decrease channels of the features. It is then processed by the BI-GRU (Bi-directional Gated Recurrent Unit) module [2] before a FC classifier to produce the final prediction.

5<sup>th</sup> position: *Gululu* - team from Department of Computer Science Tokyo Institute of Technology, Tokyo, Japan. This team crop the center of the comic image before extracting the visual feature. The text feature is extracted by pre-trained BERT model. Then the Visual attention network (VAN) [21] is used to learn attention weights for the features. A FC classifier is used at the end to produce the prediction scores.

## 5.2 Discussions

The emotion recognition in comic scenes is a hard problem. We have experienced the difficulty and ambiguity by doing the internal annotation and by observing the external annotation. This is the reason why we asked at-least three annotators to give their decision for each comic scene.

All the methods leverage both image and text modalities to get the final predictions. However, all teams have consent that the text information is dominant, but the visual information can help improve the performance. The common approach of these methods is to extract text feature; visual feature from pre-trained models such as BERT [3], RoBERTa [10] for text and ResNet [6], EfficientNet [20] for image. And then merge those features. Early fusion and late fusion are experimented by some methods, but only one method (the winner) fuses the two features at mid-level and it performed very well compared to the early and late fusion. Early fusion is better than late fusion. This remark is confirmed further by the fact that both the winner and the second team used the early fusion approach while other three teams only used the late fusion.

While the late fusion and mid-level fusion give good performance, as demonstrated in the methods of the first team, we believe that the early fusion is more compelling and have more room to improve. The model can learn more about the underlying structure of the two modalities if they do not yet undergo significant independent processing. Another common technique is the ensemble (average or

stacking) where we combine different models to get the final score. The number of models used among methods varies from 2 to 100 models.

Some flawed approaches have been shared by the teams. First, unsupervised learning on the original COMICS dataset (1.2M comic scenes) based on Masked Language Model (MLM) or comic book classification tasks performs worse than existing pre-trained models. We believe the main reason is that the amount of text in COMICS dataset is still not enough to pre-train a language model. Second, text data augmentation by using transcriptions from nearby comic scenes does not help. Finally, eight independent binary classification models cannot outperform the multilabel classification model.

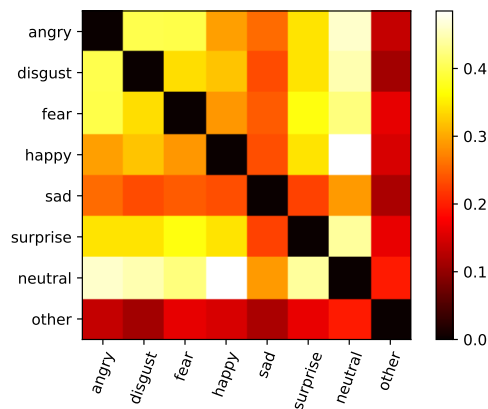


Fig. 7: Cosine similarity correlation of emotions based on multi-hot encoded vectors of 8 emotions in training data.

information from visual features. Moreover, it is needed to have further investigations on how to explore the correlation of emotions (see Figure 7) (e.g., using graph-learning [23]) to enhance knowledge for modeling the interplay between visual and textual modalities.

## 6 Conclusion

We organized the first competition on the multi-modal emotion recognition in images of documents. The problem is challenging and we put a lot of effort into building a high-quality benchmark dataset. The competition has attracted many participants and teams who submitted numerous results of their algorithms to the public leaderboard. Through method descriptions of the top teams on the private leaderboard, we observed many findings which are very important to facilitate future research in this domain. In particular, to combine the image and

One important remark is that, we have found none of the teams exploited separately the two emotions relevant information in the image modality: the face of comic characters and the onomatopoeia. All the team considered the visual modality by the whole image. We believe this is one of the main reason that the visual modality give worse performance compared to text modality. The onomatopoeia and the face are the two most important information in the scenes that describe the emotions of characters, if we can explicitly focus on those two objects, it will be easier to learn the relevant

text modalities, early, mid-level and late fusions have been experimented. Based on the performance of the submissions, we believe that multimodal approaches have great potential to improve the performance of not only emotion recognition but also other tasks in the field of document analysis. The competition dataset and website will be open to the public even after the conference. Also, a legacy version of the competition will remain open without limitation of time to encourage future result comparison<sup>12</sup>. We strongly believe that this competition and the contributed benchmark data will play an important role in the development of future research for the multimodal document analysis community.

## References

1. Augereau, O., Iwata, M., Kise, K.: A survey of comics research in computer science. *Journal of Imaging* **4** (04 2018)
2. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* **abs/1412.3555** (2014)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 NAACL*. pp. 4171–4186. Association for Computational Linguistics (Jun 2019)
4. Ekman, P.: An argument for basic emotions. *Cogn. Emot.* **6**(3-4), 169–200 (1992)
5. Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning* **45**(2), 171–186 (2001)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *2016 IEEE CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pp. 770–778. IEEE Computer Society (2016)
7. Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J., Daumé, H., Davis, L.: The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 6478–6487 (2017)
8. Kiela, D., Bhooshan, S., Firooz, H., Testuggine, D.: Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019)
9. Le, D.T., Vu, X.S., To, N.D., Nguyen, H.Q., Nguyen, T.T., Le, T.K.L., Nguyen, A.T., Hoang, M.D., Le, N., Nguyen, H., Nguyen, H.D.: ReINTEL: A multimodal data challenge for responsible information identification on social network sites. pp. 84–91. Association for Computational Linguistics, Hanoi, Vietnam (Dec 2020)
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019)
11. Lovheim, H.: A new three-dimensional model for emotions and monoamine neurotransmitters. *Med. Hypoth.* **78**(2), 341–348 (2012)
12. McCloud, S.: *Making comics: Storytelling secrets of comics, manga and graphic novels*. Harper New York (2006)
13. Nguyen, H.D., Vu, X.S., Truong, Q.T., Le, D.T.: Reinforced data sampling for model diversification (2020)
14. Nguyen, N., Rigaud, C., Burie, J.: Digital comics image indexing based on deep learning. *J. Imaging* **4**(7), 89 (2018)

<sup>12</sup> <https://competitions.codalab.org/competitions/30954>

15. Novielli, N., Calefato, F., Lanubile, F.: A gold standard for emotion annotation in stack overflow. In: Proceedings of the 15th International Conference on Mining Software Repositories. p. 14–17. MSR '18, Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3196398.3196453>
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
17. Plutchik, R., Kellerman, H.: Emotion: Theory, research and experience. Academic Press **3** (1986)
18. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* **abs/1910.01108** (2019)
19. Shaver, P., Schwartz, J., Kirson, D., O’connor, C.: Emotion knowledge: Further exploration of a prototype approach. *J. Pers. Soc. Psychol.* **52**(6) (1987)
20. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114 (2019)
21. Truong, Q.T., Lauw, H.W.: Vistanet: Visual aspect attention network for multi-modal sentiment analysis. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 305–312 (Jul 2019). <https://doi.org/10.1609/aaai.v33i01.3301305>
22. Vu, X.S., Bui, Q.A., Nguyen, N.V., Nguyen, T.T.H., Vu, T.: Mc-ocr challenge: Mobile-captured image document recognition for vietnamese receipts. RIVF '21, IEEE (2021)
23. Vu, X.S., Le, D.T., Edlund, C., Jiang, L., Nguyen D., H.: Privacy-preserving visual content tagging using graph transformer networks. In: ACM International Conference on Multimedia. ACM MM '20, ACM (2020)
24. Yadollahi, A., Shahraki, A.G., Zaiane, O.R.: Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput. Surv.* **50**(2), Article 25 (2017)
25. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems. pp. 5754–5764 (2019)