Supplementary Material for Uncertain Multimodal Intention and Emotion Understanding in the Wild

Qu Yang[1], Qinghongya Shi[1], Tongxin Wang[1] and Mang Ye[1,2] ✉
[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Taikang Center for Life and Medical Sciences, Wuhan University, Wuhan, China
{yangqu, shiqinghongya, wangtx, yemang}@whu.edu.cn
https://github.com/yan9qu/CVPR25-MINE

A. Keywords and Taxonomy Definition

We follow the previous work [6] that studied the top 100 themes commonly discussed by humans on social media platforms. These 100 themes are analyzed at a fine-grained level as shown in Sec. A, with scenarios of overlap between themes and topics based on the keywords associated with them.

For taxonomy, we argue that a social process (*e.g.*, a tweet posted on a social platform) should have both intention and emotion properties, from both a philosophical perspective [12] and a common understanding of intention and emotion. Therefore, in our work, we label each sample with both intention and emotion labels, where there are 20 categories of intention labels and 11 categories of emotion labels. The categories and definition are shown in Sec. B.

We evaluate the quality of our dataset labeling by analyzing the distribution of labels. Fig. A show the number of categories for the intention and sentiment labels, respectively. The label information in the bar charts is consistent with our general understanding of social life.

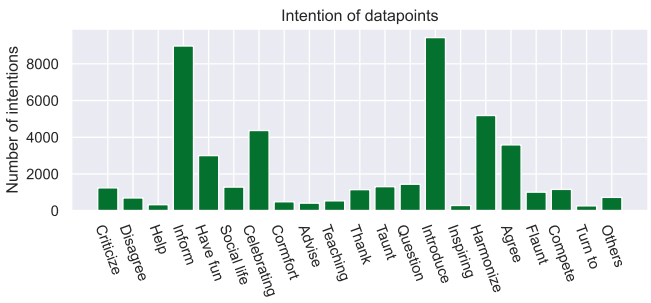

Figure A. Counts of different intentions in MINE.

B. Relation of Intention, Emotion and Topic

As we discussed earlier, we claim that intention, emotion and topic are correlated. Jia *et al*. [5] use object detection techniques [3] to correlate subjects with intention in image data and explore their potential co-occurrence. However, this process depends on the visual performance of the object detector. We go a step further and use the given topic keywords in the single data with the intention, emotion labels and topics in MINE to analyze the possible correlation, as shown in Figs. B and C.

Here, we first analyze the interrelationship between intentions and emotions. Fig. B shows the proportions of each cross-sectional sample to the total number of each column (intention). We obtain the following conclusions: 1. There is alignment in the polarity of the categories of intention and emotion. We can see that some intention classes tend to co-occur with certain emotion classes. For example, aggressive intention classes such as "Criticize", "Disagree" and "Taunt" are often associated with negative emotions such as "Angry" and "Disgusted". On the other hand, positive emotions such as "Happy", "Excited" and "Proud" are strongly associated with intention categories such as "Celebrating" and "Harmonize". 2. There is complementarity of intentions to categories of emotions. When the emotion is "Sad", the user posting social media has a low level of positive emotions, while the intention of "Comfort" provides complementary positive emotions. Intentions may introduce positive emotions when there is an extreme lack of positive emotions or abundance of negative emotions. 3. There is neutrality of intention and emotion. Some stable intention categories show a high tendency towards neutral emotion, such as those related to "Teach", "Agree", and "Introduce". These intention categories do not express strong affective states.

Then we discuss the relationship between intention and topic shown in Fig. C. The correlation between intention and topic is weaker than that between intention and emotion, but we can still observe some patterns, such as the

✉ Corresponding author.

Topics	Fine-grained keywords
Family	family/lover/marriage/relationship/home/parent/children/relatives/birthday/celebration
Education	education/learning/teaching/student/teacher/campus/school/college/university/science culture/ knowledge/skill/course/technology
Politics	politics/election/elect/vote/speech/congress/power/authority/government/statesmanship council/democracy
Law	law/legal/statute/principle/rule/constitution/court/fairness/justice/legislation/system convict/prison/jail
Economy	economy /business/finance/product/brand/stock market/investing/advertisement advertising/advertising agency
Entertainment	entertainment/fun/amusement/enjoyment/vacation/holiday/travel/park/game/party socializing/friend/friendship/celebration
Art	design/art/craft/creation/music/cartoon/dance/movie/film/building architecture/sculpture/painting/photography
Sport	exercise/sport/athletics/competition/gymnastics/basketball/ football/soccer volleyball/hockey/running/swimming/diving/tennis/golf/surfing/sailing
Public service	public service/service/environment/welfare/charity/ ambulance/news/broadcast weather/report
Others	working/meeting/conference/social life/daily life/feedback/assessment/evaluation discussion/analysis/emotion/emotional/feeling/mood/thought/health/eating/living

Table A. Keywords used in our data collection process.

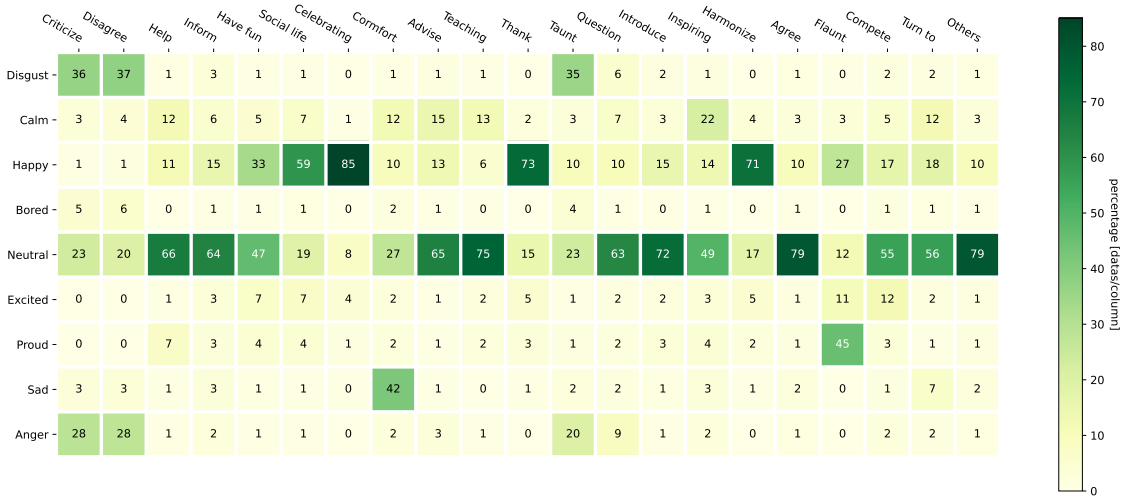

Figure B. There is a significant intrinsic correlation between intention and emotion.

correlation between “Inspire”, “Teach” and “Education” (topic), “Advise” and “Economy”, “Comfort” and “Family”, and so on. Unlike emotion, where a category corresponds to a category of intentions, the correspondence between intention and topic is one-to-one, i.e., an intention corresponds to a specific topic more clearly. In summary, all three aspects can help machines understand human social life, but the intention-emotion correlation is stronger, and one aspect can serve as a strong indicator of the other.

C. Feature Extraction Details

This section provides detailed procedures for feature extraction from each modality in the MINE dataset. All extracted features are aligned to a dimension of 768 to facilitate consistent multimodal fusion.

Categories		Interpretations
Express emotions	Surprised	When someone encounters sudden and unexpected sounds or movements
	Excited	Enthusiasm, eagerness or anticipation, and general arousal
	Happy	Feel pleasure or satisfaction with something
	Proud	Exhilarated pleasure and a feeling of accomplishment
	Calm	Peace of mind being free from agitation, excitement, or disturbance
	Neutral	Feeling indifferent, nothing in particular
	Bored	Doing something that doesn't give someone satisfaction
	Disgust	A feeling of aversion towards something offensive
	Fear	Emotional reaction to something that seems dangerous
	Anger	Feel a strong resentment or hostility for something
	Sad	Feel unhappy or sorrowful for something
Implicit intentions	Flaunt	Show off something in a proud or boastful way
	Competition	Try to be better or more successful than someone or something else
	Comfort	Make someone feel less worried, unhappy, or upset
	Advise	Give someone an opinion or suggestion about what they should do
	Introduce	Make someone or something known to someone else for the first time
	Turn to	Seek help or support from someone or something
	Have fun	Enjoy oneself or do something that makes one happy
	Social life	The activities and relationships with other people outside of family
	Celebrating	Mark a special occasion or achievement by doing something enjoyable or festive
	Harmonize	Make something fit well or agree with something else
	Question	Ask someone for information or clarification about something
	Help	Assist someone or do something that makes their situation easier or better
	Criticize	Express disapproval or find fault with someone or something
	Inspiring	Motivate someone or make them feel enthusiastic or hopeful about something
	Thank	Express gratitude or appreciation to someone for something they have done or given
	Teach	Instruct someone or impart knowledge or skills to them
	Inform	Tell someone something that they need or want to know.
	Taunt	Mock or provoke someone in a cruel or insulting way
	Agree/Support	Have the same opinion as someone or show approval of their actions or ideas
	Disagree/Oppose	Have a different opinion from someone or show disapproval of their actions or ideas

Table B. Emotion and intention taxonomies of our MINE dataset with brief interpretations.

C.1. Text Feature Extraction

For text feature extraction, we employ the pre-trained BERT language model [7], renowned for its efficacy in various NLP tasks. Given a text utterance, we derive token embeddings $\mathbf{z}^T \in \mathbb{R}^{L_T \times H_T}$, where L_T represents the sequence length and H_T is the feature dimension, set at 768. We use the [CLS] token embedding from the last layer of BERT as the text representation.

C.2. Image Feature Extraction

We utilize ViT-S [2] for image feature extraction, chosen for its proficiency in capturing global information critical for intention and emotion discrimination. A user-posted image $\in \mathbb{R}^{H \times W \times C}$ is segmented into patches $N \times (P^2 \times C)$, with P as the patch size (16×16 pixels) and N denoting the number of patches. The attention module then generates the image representation $\mathbf{z}^I \in \mathbb{R}_I^L$, where L_I is the visual dimension. For instances with multiple images, we aggregate the visual features using element-wise addition,

ensuring consistency for subsequent multimodal fusion.

C.3. Video Feature Extraction

Video feature extraction begins with the FFmpeg toolkit [13], where we sample frames at 1 Hz, considering the average video duration of 24.4 seconds. The Video Swin Transformer [10] is then employed to process these frames sequentially, yielding video feature vectors L_V with a dimension of 768, aligning with the image feature dimensions for consistency in subsequent processing stages.

C.4. Audio Feature Extraction

Audio features are initially acquired through the librosa toolkit [11] at a sampling rate of 16,000 Hz, following the protocol from [14]. Subsequently, the pre-trained wav2vec 2.0 model [1] is utilized, known for its robust acoustic representation learning. The resultant audio feature embeddings $\mathbf{z}^A \in \mathbb{R}^{L_A \times H_A}$ are obtained from the last hidden layer, with L_A as the sequence length and a dimension of 768 to match other modalities.

All extracted features are stored in our preprocessed dataset format, enabling efficient loading and processing for multimodal fusion experiments.

D. Privacy and Ethic Concerns

Privacy and ethics are important aspects to consider when dealing with real tweets as data sources. To address possible concerns and issues, we have taken the following three measures:

- First, we protect users' privacy at the acquisition level by only providing pre-extracted features instead of the original tweets. Access to the source data requires formal application via email and signing a data usage agreement.
- Second, we ensure that the pre-extracted features are only used for academic research by implementing a strict questionnaire system for academics who need to use the dataset. The questionnaire system requires information about the purpose of use, affiliation, etc.
- Third, we respect the wishes of individual users who do not want their personal tweets to be included in the dataset for research use. We have designed a questionnaire system and regular feedback system to address this issue. Users can fill out a reluctant sharing questionnaire using a tweets id list (Will be released soon) provided by us. We then update the dataset version periodically to filter out such information that may contain sensitive or reluctant sharing.

E. Formulations of Multi-label Metrics

The formulations of evaluation metrics are as follows:

$$\text{Macro } F1 = \frac{\sum_{l=1}^{L} F1_l}{L},$$

$$\text{Micro } F1 = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L} TP_l + \frac{1}{2}\sum_{l=1}^{L}(FP_l + FN_l)},$$

$$\text{Samples } F1 = \frac{\sum_{i=1}^{N} F1_i}{N},$$

where true positive (TP), false positive (FP), and false negative (FN) are used to calculate the precision (Pre) and recall (Rec). L and N represent the number of categories and the total number of instances in a mini-batch, respectively. F1 score represents a harmonic mean of precision and recall. Different scenarios use different calculation methods of F1 scores, including Macro F1, Micro F1, and Samples F1.

- Macro F1 is the unweighted average of each category's F1 scores over all instances. It is usually dominated by categories that are easily identifiable.
- Micro F1 counts the sums of the TP, FN, and FP across all categories to acquire a global F1 score. It is susceptible to the influence of a large number of categories under the extreme imbalanced data distribution.
- Samples F1 computes the F1 score for each instance on multi-labels and returns the average value of all instances.

F. Experimental Setups

As sequence lengths of the segments in each modality need to be fixed, we use zero-padding for shorter sequences. For all methods, the training batch size is 16, the number of training epochs is 100. For MISA, MulT, and MMIN, we use Adam optimizer [8] with learning rates of 3×10^{-5}, 3×10^{-5}, and 1×10^{-5}, respectively. It takes about 20 minutes to obtain the result (MISA [4]) on MINE, via $1\times$ RTX 3090 GPU. We run every method with three different random seeds to obtain the average results. We adjust the hyper-parameters with macro F1-score. For a fair comparison, we report the average performance over three runs of experiments with different random seeds. We adopt bert-base-uncased as backbone to use for the text modality. In the experimental data splitting, we partition the data into three sets: training, validation, and testing. The training set consisted of 16,134 samples, while the validation and testing sets each had 2,017 samples.

G. Artificial Missing Modality

As discussed in the main text, artificially creating missing modal data by manually masking off some modalities can lead to potentially erroneous data. In this section, we provide specific examples and analyze the causes of the errors.

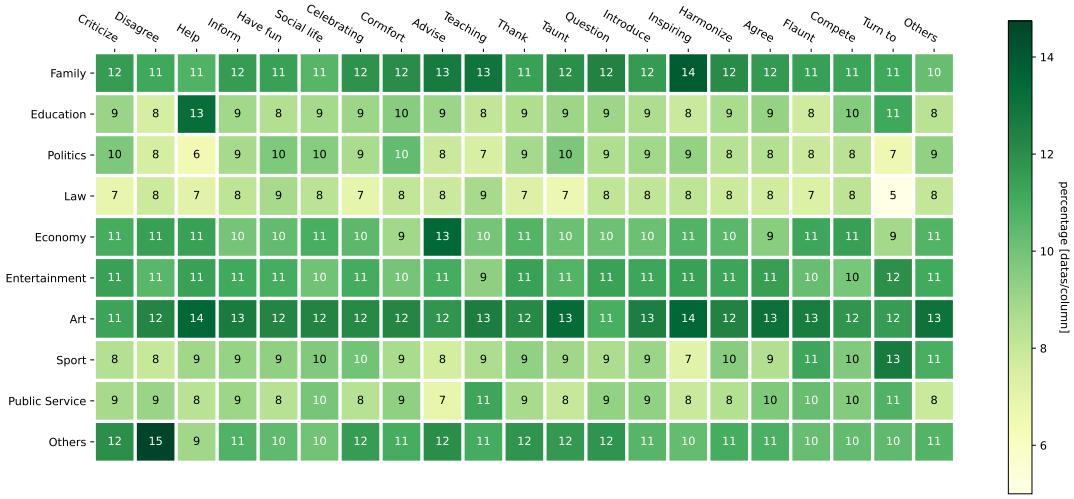

Figure C. The relationship between intention and topic.

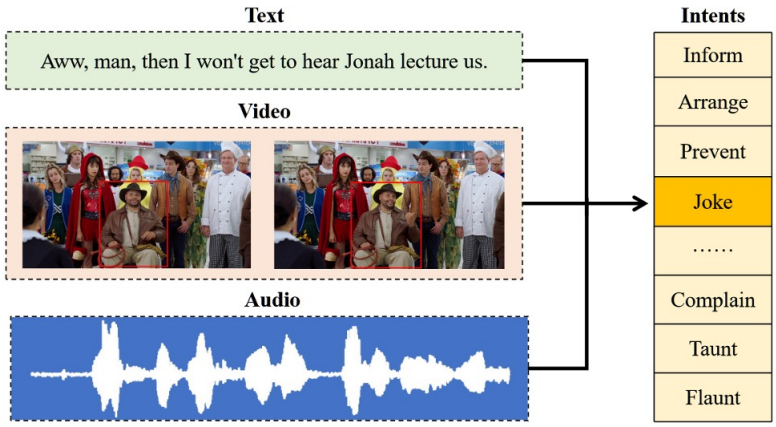

Figure D. An example from [14], the intention is "Joke".

Taken an example in [14], the conveyed intention in Sec. G is "Joke" based on the combination of three modalities. However, following the missing modality setting of [9], we can obtain missing modality data by filtering out one or two modalities of the original data while keeping the same intention label. In this case, the data become Sec. G. If we only consider the text modality, we may infer that the intention of the pure text is "Complain", which is inconsistent with "Joke".

In multimodal emotion or intention understanding, we often need to combine multiple sources of information to make a comprehensive judgment. There are two kinds of effects among different modalities: complementary and exclusive. In the process of masking, if the missing modality is complementary, the final emotion or intention will not change significantly. However, if the missing modality is exclusive, it may cause completely opposite emotions and intentions. In this case, wrong labels will be generated.

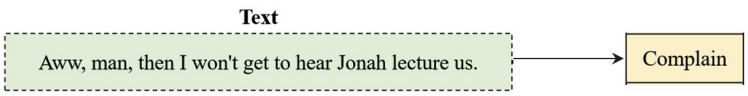

Figure E. Missing modality data (text) of Sec. G, the intention should be "Complain".

In MINE, we exploit the natural characteristics of social media to preserve the original modal information of the data. This natural missing modality problem avoids the occurrence of wrong labels compared to manual ones. It provides the first reliable platform for studying the missing modality problem.

H. Annotation Protocol Details

This section details our annotation methodology and quality control procedures for the MINE dataset. Given the subjective nature of intention and emotion labeling, we implemented a rigorous multi-stage annotation process.

The annotation process was structured into manageable batches of approximately 1,500 samples each. We employed two independent annotation groups (15 and 13 members respectively) composed of trained annotators. Domain experts from social sciences and affective computing supervised the process and provided regular feedback.

The workflow consisted of several key phases:

Training Phase: The first batch served as a training set, where annotators received detailed guidelines and expert feedback. Initial annotation discrepancies were significant (27.13% and 24.13% in the first two batches), prompting additional training sessions. This iterative training continued until the discrepancy rate stabilized below 3%.

Main Annotation Phase: Following successful training, the groups worked independently on subsequent batches. We implemented a 'differentiation' metric to measure inter-group agreement, calculating the proportion of differing annotations between groups. Expert review was triggered for any batch exceeding a 3% differentiation rate, ensuring consistent quality throughout the process.

Quality Control: To maintain annotation quality:

- Each data point received dual annotations from different groups
- Weekly calibration meetings were held where annotators discussed challenging cases. These sessions focused on samples that received divergent annotations, allowing annotators to align their understanding of ambiguous cases. For example, posts containing sarcasm or multiple emotions required special attention to establish consistent annotation guidelines
- Expert review for batches exceeding the differentiation threshold

For intention annotation, annotators were instructed to consider multiple applicable intentions, reflecting the multi-label nature of the task. For emotion annotation, they were directed to identify the dominant emotion, following our single-label approach. This methodology ensured both comprehensive coverage of intentions and clear emotional categorization.

To establish a reliable benchmark for evaluating emotion and intention understanding performance, we allocated 20% of the data (4,000 samples) for validation and testing sets, which were carefully annotated by domain experts. This expert-annotated portion was equally divided into validation (2,000 samples) and test sets (2,000 samples), providing a high-quality ground truth for model evaluation.

I. Ethical Considerations

I.1. Consent of data subjects

We use twitter's official Academic API to obtain the data, obtaining consent from the data subject (twitter official).

I.2. Privacy and anonymization

To address the privacy challenges and risks posed by emotion and intention analysis of tweets, we adopted several measures to ensure the quality and validity of our dataset and analysis. First, we provide MINE by using questionnaires, where we asked the users to focuses their purposes on academic. Second, we anonymized the tweets and removed any personal or sensitive information that could identify or harm the tweet authors by feature extractors. Third, we exposed only the features of our dataset, without revealing the source data or the tweet URLs.

I.3. Potential misapplications

Potential misapplications of emotion and intention analysis of tweets:

- Manipulating or influencing public opinion or behavior by generating or spreading fake or biased tweets that express certain emotions or intentions. For example, creating tweets that express anger or fear towards a political candidate or a social issue, or creating tweets that express happiness or satisfaction with a product or a service.
- Exploiting or violating the privacy or dignity of tweet authors by analyzing their emotions or intentions without their consent or awareness. For example, inferring the personal traits, preferences, or vulnerabilities of tweet authors based on their emotions or intentions, or using their emotions or intentions to target them with personalized advertisements or recommendations.
- Misinterpreting or misrepresenting the emotions or intentions of tweet authors by ignoring the context, culture, or language of the tweets. For example, assuming that the emotions or intentions expressed in the tweets are universal and consistent across different situations, groups, or regions, or failing to account for the sarcasm, irony, or humor in the tweets.

To avoid the situations, we provide a clear license and terms of use for our dataset, where we specified the purpose and scope of our dataset, and we warned the users about the potential misuse or abuse of our dataset.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NIPS*, 2020. 4

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 3

[3] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 1

[4] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM MM*, 2020. 4

[5] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Intentonomy: a dataset and study towards human intent understanding. In *CVPR*, 2021. 1

[6] Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P Rana, Pushp Patil, Yogesh K Dwivedi, and Sridhar Nerur. Advances in social media research: Past, present and future. *Information Systems Frontiers*, 2018. 1

[7] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NACCL*, 2019. 3

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 4

[9] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *CVPR*, 2023. 5

[10] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 4

[11] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, 2015. 4

[12] Tobias Schröder, Terrence C Stewart, and Paul Thagard. Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive science*, 2014. 1

[13] Suramya Tomar. Converting video formats with ffmpeg. *Linux journal*, 2006, 2006. 4

[14] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *ACM MM*, 2022. 4, 5