Donate or Create? Comparing Data Collection Strategies for Emotion-labeled Multimodal Social Media Posts

Christopher Bagdon[1], **Aidan Combs**[1,2,3], **Carina Silberer**[3], and **Roman Klinger**[1]
[1]Fundamentals of Natural Language Processing, University of Bamberg, Germany
[2]Department of Sociology, The Ohio State University, USA
[3]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
{christopher.bagdon,roman.klinger}@uni-bamberg.de
combs.494@osu.edu; carina.silberer@ims.uni-stuttgart.de

Abstract

Accurate modeling of subjective phenomena such as emotion expression requires data annotated with authors' intentions. Commonly such data is collected by asking study participants to *donate* and label genuine content produced in the real world, or *create* content fitting particular labels during the study. Asking participants to create content is often simpler to implement and presents fewer risks to participant privacy than data donation. However, it is unclear if and how study-created content may differ from genuine content, and how differences may impact models. We collect study-created and genuine multimodal social media posts labeled for emotion and compare them on several dimensions, including model performance. We find that compared to genuine posts, study-created posts are longer, rely more on their text and less on their images for emotion expression, and focus more on emotion-prototypical events. The samples of participants willing to donate versus create posts are demographically different. Study-created data is valuable to train models that generalize well to genuine data, but realistic effectiveness estimates require genuine data.

1 Introduction

Emotions play a fundamental role in communication (Chen et al., 2022; Chung and Zeng, 2020), particularly in online settings (Derks et al., 2008). On contemporary social media sites, authors often express emotion through a combination of text and visual content (Illendula and Sheth, 2019; Li and Xie, 2020). Modeling the emotions expressed in social media posts therefore requires multimodal datasets labeled for author emotion. This is, however, a challenging task: Emotions are internal psychological states and external annotators can therefore only approximate the correct labels (Troiano et al., 2023; Nakagawa et al., 2022).

One approach to mitigate this annotator–author label mismatch is to ask study participants to create content fitting provided labels (Troiano et al., 2023, "Write a text that caused emotion X",). While this is simple to implement, the resulting data may differ from real social media data. Such lack of generalizability may lead to limited model robustness (Degtiar and Rose, 2023; Elangovan et al., 2024; Ribeiro et al., 2020; Yang et al., 2023).

An alternative approach is to ask social media users to donate and label their real social media posts (Oprea and Magdy, 2020). While this approach may provide more realistic data, it requires more precautions to protect participant privacy (Keusch et al., 2024; Gomez Ortega et al., 2023).

Little is known about what precisely the differences between study-created and donated author-labeled content may be, or how significant differences may be for modeling. We provide a better understanding of the tradeoffs of these corpus collection methods with the goal of informing future author-labeled corpora collection efforts. To do this, we collect study-created and genuine multimodal social media posts labeled by their authors for emotion. We analyze differences between the events that inspire the posts, how they are labeled for emotion, and sample characteristics. Finally, we explore the impact of these differences on emotion modeling and prediction.

We implement three collection procedures:

1. CREATION: Study-created data. Participants are asked to create posts about an event they experienced that elicited an emotion for which we prompt. This approach is clear-cut to conduct but potentially lacks generalizability.
2. DONATION: Genuine data. Participants provide posts from their social media accounts about an event they experienced that elicited

The work has been conducted at the University of Bamberg and the University of Stuttgart. The Ohio State University is the new affiliation for Aidan Combs.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17307–17330

a prompted emotion. This method yields real-world data balanced across emotions, but might come with privacy issues and participant's self filtering, as well as potentially limited availability of posts.

3. RECENT: Genuine data. Participants submit their five most recent posts and then annotate each for emotion. While this method avoids potential experiment bias in emotion annotation, it may underrepresent emotions that are rarely shared on social media.

We find that (1) study-created data differs from genuine data in several ways. Notably, it is dominated by prototypical emotion triggers, while genuine data is more diverse. (2) The data collection procedures lead to different samples of participants. (3) Models trained on CREATION generalize well to genuine data, but DONATION test data is required to realistically estimate their effectiveness.

2 Related Work

Asking authors to label text is useful in areas where author intent is both important and unclear from the text alone. For example, author-annotated corpora exist for deception detection (Capuozzo et al., 2020; Velutharambath et al., 2024), sarcasm detection (Oprea and Magdy, 2020; Abu Farha et al., 2022), and, of particular interest to us, emotion detection (Kajiwara et al., 2021; Troiano et al., 2019; Scherer and Wallbott, 1997; Troiano et al., 2023). In this section, we review common methods for collecting annotations of authors' internal states.

2.1 Genuine Data Collection

The standard approach in natural language processing and computer vision is to acquire genuine data from the world and request annotations from external annotators. These annotators may be trained experts or recruited through crowdsourcing platforms. However, since annotators do not have access to the original author's internal state, their annotations are often inaccurate (Kajiwara et al., 2021; Troiano et al., 2023; Nakagawa et al., 2022).

To circumvent this issue, researchers may try to indirectly acquire labels from authors. For example, in social media data, author labels can sometimes be inferred using corresponding hashtags (Mohammad, 2012; Abbes et al., 2020). While such approaches are useful when reliable markers are available, they are also subject to error: Annotating texts with hashtags as an indicator of a particular label makes other such markers redundant.

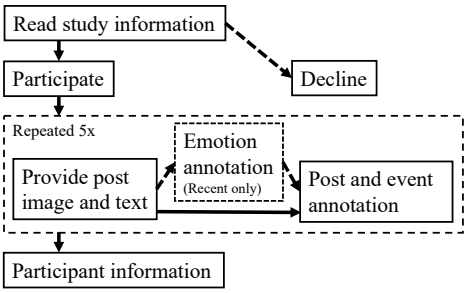


Figure 1: Data collection process.

If author intent is difficult to access in an indirect, rule-based manner, an alternative is to ask social media users to donate and label their own data (Oprea and Magdy, 2020; Kajiwara et al., 2021; Razi et al., 2022; Pfiffner et al., 2024). Directly obtaining labels from authors eliminates incorrect inferences as sources of error.

Collecting genuine social media data annotated by authors is challenging (van Driel et al., 2022). This is because of privacy concerns (Boeschoten et al., 2022; Carrière et al., 2024; Gomez Ortega et al., 2023), the required technical skill set of participants (Keusch et al., 2024), and the potentially limited availability of requested types of data.

2.2 Study-created Data Collection

An alternative to collecting real data is to construct author-labeled datasets by asking participants to role-play or write according to specific instructions that provide target labels. For example, participants may be asked to tell truths or lies (Ott et al., 2013; Capuozzo et al., 2020; Velutharambath et al., 2024; Lloyd et al., 2019), write about events that elicited certain emotions (Troiano et al., 2019, 2023), respond with specific coping strategies (Troiano et al., 2024), or act out hypothetical emotional scenarios (Busso et al., 2008, 2016).

Such approaches for obtaining study-created data have various advantages. They avoid the difficulties of finding a sufficient amount of content fitting uncommon labels "in the wild". Participants have more control over what they contribute, and can better guard their privacy. Finally, the data can be less prone to recall bias as large time gaps between content creation and labeling are avoided.

Such methods may, however, be susceptible to other experiment effects (Vania et al., 2020). For instance, people behave differently when they know they are participating in a study, often to confirm

Label	Question Text	Options
Emotion annotation (RECENT only)		
Emotion	Please select the emotions that you felt as a result of this event [multiple]	[Emo.]
Intensity	Please rate how intensely you felt each of these emotions as a result of this event [Emo.]	1...5
Text–image relationship: How much do these statements apply?		
Text describes image	The text directly describes the image.	1...5
Text → image	The text is required to understand the image.	1...5
Image → text	The image is required to understand the text.	1...5
Image conveys emotion	The image explicitly conveys the emotion you posted about.	1...5
Text conveys emotion	The text explicitly conveys the emotion you posted about.	1...5
Event experience		
Event Duration	How long did the event last?	[Time]
Emotion Duration	How long did you experience emotion as a result of the event?	[Time]
Event Intensity	How intense was your experience of the event?	1...5
Emotion Intensity	How intense was your experience of this emotion?	1...5
Appraisal: Think back to when the event happened and recall its details. Take some time to remember it properly. How much do these statements apply? Some statements might not fit the event exactly, please answer to the best you can.		
Familiarity	The event was familar.	1...5
Predictability	I could have predicted the occurrence of the event.	1...5
Attention	I had to pay attention to the situation.	1...5
Notconsider	I tried to shut the situation out of my mind.	1...5
Pleasantness	The event was pleasant for me.	1...5
Unpleasantness	The event was unpleasant for me.	1...5
Goalrelevance	I expected the event to have important consequences for me.	1...5
Ownresponsibility	The event was caused by my own behavior.	1...5
Goalsupport	I expected positive consequences for me.	1...5
Anticipconseq	I anticipated the consequences of the event.	1...5
Owncontrol	I was able to influence what was occurring during the event.	1...5
Otherscontrol	Someone other than me was influencing what was occuring.	1...5
Acceptconseq	I anticipated that I would easily live with the unavoidable consequences of the event.	1...5
Effort	The situation required me a great deal of energy to deal with it.	1...5
Internalstandards	The event clashed with my standards and ideals.	1...5

Table 1: Wording and response options for survey questions used in the analysis. [Emo.] refers to Anger, Disgust, Fear, Joy, Sadness, Surprise. [Time] refers to one of Seconds, Minutes, Days, Weeks, Months.

the hypotheses they think the researchers have (Mummolo and Peterson, 2019; Nichols and Maner, 2008). More emotionally intense events are remembered more easily, potentially leading participants to preferentially select them when writing posts about them after-the-fact (Kensinger, 2009). Writing differs based on audience, for example varying in the degree to which it is stereotype-consistent (Lyons and Kashima, 2006, 2003). Differences between study-created data and actual social media content are likely to harm the generalizability of models trained on study-created data (see Elangovan et al., 2024; Ribeiro et al., 2020; Yang et al., 2023, for a discussion of differences between hold-out data and real-world data).

3 Data Acquisition Methods

In this section we discuss the data acquisition methods and their advantages and disadvantages.[1]

3.1 Process Overview

The data collection process is illustrated in Figure 1. Potential participants see the study details and choose to participate or decline. Those who accept are asked to provide a social media post using one of three data collection strategies – CREATION, DONATION, and RECENT – and then annotate it. Each participant provides five posts. Finally, participants answer questions about themselves. Participants may complete the study as many times as they wish, up to once per emotion.

3.2 Data Collection Strategies

We request posts that contain both text and an image and are authored by the participant. Other instructions vary by collection strategy.

CREATION. We ask participants to recall an event in which they felt a particular emotion – anger, disgust, joy, fear, sadness, or surprise – and which they remember well. They then write a social media post about it. Participants must select an image from the Flickr database of Creative Commons licensed

[1] All data, code, and surveys are available at https://www.uni-bamberg.de/en/nlproc/projects/item/

images[2] that is similar to the one they would have used if they had created the post naturally. This approach diminishes the risk to participant privacy, but posts may differ from genuine data.

DONATION. We prompt participants for an emotion and ask them to share a genuine post from their timeline. They do so by copy–pasting the text and uploading the image from that post. This approach yields genuine posts while equally representing emotions regardless of their prevalence on social media. However, participants may struggle to find posts representing uncommon target emotions. This approach also raises greater privacy concerns.[3]

RECENT. RECENT is similar to DONATION, but we do not prompt participants for particular emotions. Instead, we ask them to share their five most recent multimodal posts and annotate each with all emotions, and the associated intensities, they felt in response to the event that inspired the post. We adapt this emotion annotation approach from Rhodes-Purdy et al. (2021).[4] RECENT diminishes concerns about the accuracy of emotion annotations at the cost of potentially underrepresenting uncommon emotions. The same privacy concerns that affect DONATION also apply to RECENT.

3.3 Annotation Details

Post and Event Annotations. Table 1 shows the questions participants answer about each of their posts. To understand the roles of images and text, we ask about the relationship between the modalities. To understand the link between the event and the emotion category, we request appraisal labels (Scarantino, 2016; Scherer, 2005). Appraisals are a psychological theory of how events induce emotion (Troiano et al., 2023; Stranisci et al., 2022).

Participant Information. After participants annotate their posts, we ask about their age, gender, education, ethnicity, frequency of social media use and posting, and choice of social media platform. We additionally use demographic information provided by the research platform we use – namely, employment and student status. We summarize these questions in Table 5 in the Appendix.

Study Details. We recruit participants using Prolific. We require that they reside in the United Kingdom or Ireland, have resided there for at least five years, are 18 or older, and be native English speakers. We restrict participation to these demographics to avoid confounding variables, as emotion expression can differ from culture to culture. The survey is conducted via Google Forms. CREATION and DONATION take 30 minutes and we pay participants £4.50. RECENT takes 40 minutes and we pay participants £6.00. After removing posts that do not meet our requirements, our dataset contains 2,507 posts authored by 522 participants.

	Study					
Emotion	CREATION		DONATION		RECENT	
Anger	193	17%	174	15%	16	8%
Disgust	202	17%	195	17%	10	5%
Fear	193	17%	189	16%	4	2%
Joy	189	16%	196	17%	158	79%
Sadness	185	16%	197	17%	16	8%
Surprise	197	17%	198	17%	30	15%
Total	1159		1149		199	

Table 2: Distribution of posts in the dataset. In RECENT, posts are categorized into emotions based on which emotion the participant reported feeling most intensely. In the case of ties, the post is counted in both categories.

4 Analysis and Modeling

We answer research questions about differences in the data between collection strategies (**RQ1**), differences in the events that lead to the posts (**RQ2**), differences between samples (**RQ3**) and in how participants assign labels (**RQ4**), and the impact on model performance and generalization (**RQ5**).

4.1 RQ1: Are there differences in posts between data collection strategies?

Table 2 shows the post and label distributions. We see that CREATION and DONATION represent emotions nearly evenly by design. In RECENT, we obtain posts dominated by joy.

Text. Figure 2 shows the average lengths in character counts. CREATION posts are longer than DONATION and RECENT posts, especially for posts about joy and surprise. Controlling for differences in emotion distribution, CREATION posts are 51% longer than RECENT posts and 26% longer than DONATION posts ($p<0.01$ for both).[5]

Image Style. To investigate differences in image style, we manually label each image as a *meme*, *screenshot*, *graphic*, *professional photo*, *personal*

[2] https://www.flickr.com/creativecommons/by-2.0/

[3] Participants are informed of potential privacy risks, and consent to them before beginning the study.

[4] We did not ask participants to declare a primary emotion in the case of ties for highest intensity. We further discuss this in the Limitations section.

[5] Examples can be seen in Appendix A.2.2.

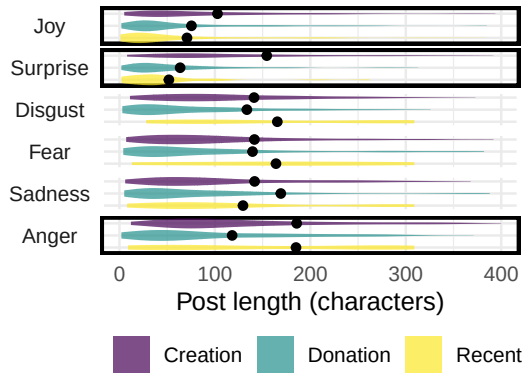


Figure 2: Length of the posts, in characters, by emotion and study. Means are represented by points. Outlined boxes indicate significant differences (one-way ANOVA, $p<0.05$ after Bonferroni correction).

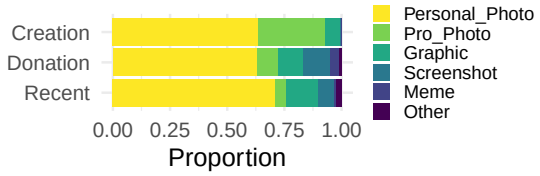


Figure 3: Distribution of image style labels.

photo, or *other*.[6] Figure 3 shows the distribution of image styles across the three data collection strategies. They differ significantly between studies (χ^2 test $p<0.001$). Personal photos are most prominent across studies. The remaining images for CREATION are dominated by professional photos. DONATION and RECENT have fewer professional photos and instead more screenshots, graphics, memes, and other images, which are less prevalent in the study-created data.

The lack of screenshots in CREATION is a consequence of requiring participants to select images from an existing database. Figure 4 illustrates this difference. This is backed by feedback from study participants who found the database limiting. Further, it shows that study-created data might not include all real-world triggers for social media posts.

Image Content. We label the images with GPT-4o to analyze the content.[7] Table 3 shows the 10 most frequent labels for each study (Table 7 in the Appendix shows the 50 most frequent labels). Despite the clear differences in image style that we observed, the word lists indicate that the images show comparable content across all collection

[6]Details and examples are in Appendix A.2.

[7]https://openai.com/gpt-4, prompt and validation details are in Appendix A.2.

Creation		Donation		Recent	
text	43%	text	55%	text	58%
background	12%	tree	11%	tree	14%
tree	12%	background	9%	woman	13%
sky	11%	woman	9%	man	12%
grass	11%	sky	9%	sky	12%
building	9%	person	8%	background	11%
water	8%	people	8%	person	10%
people	8%	grass	8%	grass	9%
woman	7%	man	8%	smile	8%
car	7%	smile	7%	building	8%

Table 3: 10 most common image content words per study. Image content word extraction done by GPT-4o.

(a) CREATION post labeled as surprise.

(b) RECENT post labeled as anger.

Figure 4: Comparison of posts submitted for CREATION and RECENT. CREATION is a more general comment, while RECENT is a reaction to a specific news item. The latter style does not occur in the CREATION data.

strategies, but with differences in the order. The most common for all three was text, though this was less frequent for CREATION (43%) than for DONATION (55%) and RECENT (58%). This is consistent with the smaller number of screenshots and lack of memes in CREATION. The image labels are dominated by references to people or scenery.

Text–Image Relation. To understand the role of the image and the text in conveying an emotion, we asked participants questions about the relationship between the two. Figure 6 shows the response distributions. CREATION participants describe their post images as less necessary to understand the text and as conveying emotion less than DONATION and RECENT posts. CREATION post images are not as integral to the post as a whole. In contrast, in RECENT posts, the text and images are more dependent on one another.

4.2 RQ2: Are there differences in the events that inspire participants to write posts?

With this and the following research questions, we aim at understanding potential reasons for the differences that we observe in the posts.

Emotion Appraisals. Figure 5 shows the distribu-

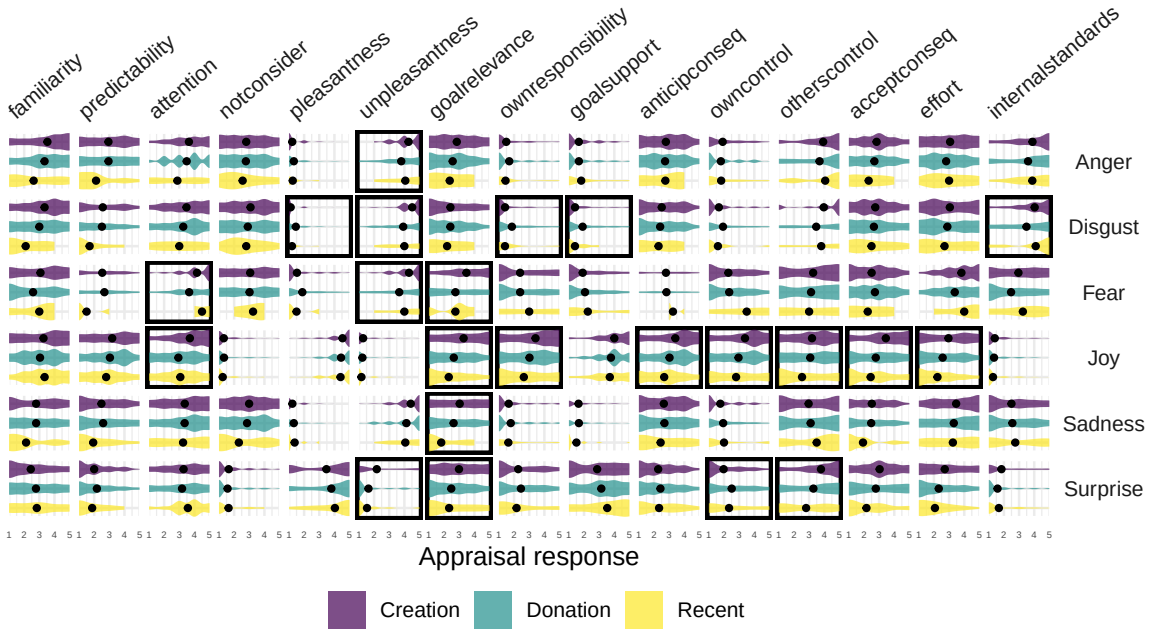


Figure 5: Responses to event appraisal questions (columns) across emotions (rows) and studies (color). Responses were provided on 5-point Likert scales. Means are represented by points. Outline: Significant at p<0.05 according to an ANOVA with Bonferroni correction.

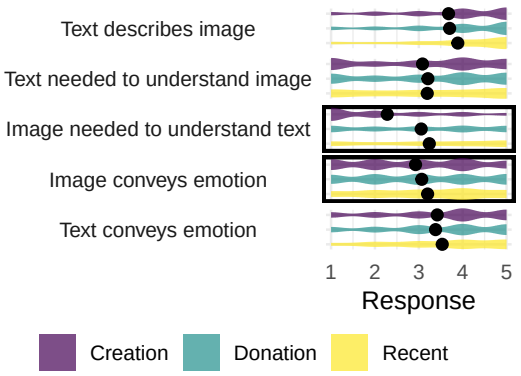


Figure 6: Participant ratings of the relationship between post text and images, on five point Likert scales. Means are represented by points. Outline indicates significance (one-way ANOVA, p<0.05 after Bonferroni correction).

tion of responses for the event emotion appraisal survey questions by study and post emotion. On 22 of the 90 question–emotion combinations (15 questions × 6 post emotions), we find statistically significant differences between studies. Joy shows significant differences in more questions (8 of 15 questions) than any other emotion.[8]

[8]The reason is, presumably, that the imbalanced corpus in RECENT leads to more power to detect significant differences for Joy. When RECENT is excluded from consideration, there are similarly many significant differences between CREATION and DONATION for Joy as for the other emotions.

In most cases, the events that inspired study-created posts are rated more highly on the appraisal dimensions than the events that inspire genuine posts. The cases which break this pattern tend to be those appraisal dimensions which are negatively associated with the emotion at hand. For example, participants rate disgust-causing events as less pleasant and less likely to have positive consequences in CREATION. RECENT posts, in contrast, tend to rate lowest on these appraisal dimensions, particularly in joy and surprise, where they are best represented. This suggests that participants use more prototypical events for particular emotions in CREATION than in genuine posts.

Duration and Intensity. Events that are recalled by prompting for a specific emotion may be dominated by the emotion and the duration and intensity of the event. To analyze this, we estimate mixed-effects models that include the study type and emotion as independent variables and random intercepts to account for grouping by participant. We find that, on a 5-point Likert scale, controlling for emotion, participants rate CREATION events as 0.34 points more intense than DONATION events, and their emotional responses to CREATION events as 0.36 points more intense ($p < 0.001$ for both). There are no intensity differences between DONATION and RECENT events or emotional responses.

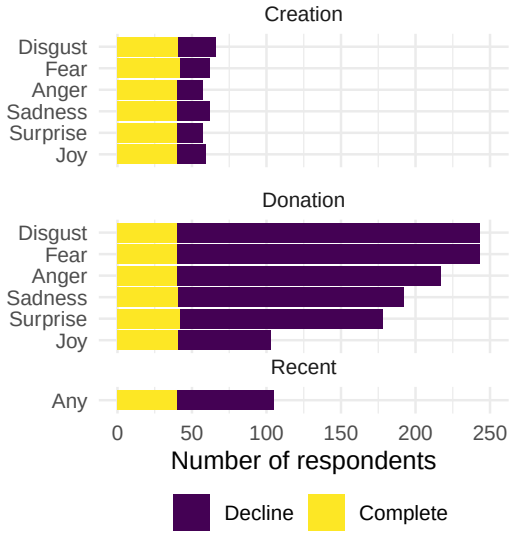


Figure 7: Differences in decline rates between studies and emotions. "Decline" means that a participant opted not to participate after reading the study information.

Participants report that emotion responses to CREATION events last significantly longer than responses to DONATION events ($p < 0.001$), and responses to DONATION events last significantly longer than responses to RECENT events ($p < 0.01$). There are no significant differences between CREATION and DONATION in the length of the events themselves. However, RECENT events are significantly shorter than DONATION events ($p < 0.01$).

It is possible that these observations are a consequence of the role that emotions play in memorizing events. In CREATION, participants may preferentially select events that are more memorable and more emotionally prototypical than those they actually post about on social media.

4.3 RQ3: Are there differences in participant characteristics?

We find that participants who completed CREATION are significantly older, less likely to be students, and more likely to be European than those who completed DONATION and RECENT.[9]

RECENT and DONATION require that participants be comfortable sharing their real social media posts with researchers, and, additionally, DONATION requires that those posts be about specific emotions. These differences in study requirements may change who is willing to participate. Figure 7 reports the decline rates after having read the instructions. Decline rates are considerably higher in DONATION and RECENT than in CREATION (χ^2 test, $p < 0.001$). Participants are indeed hesitant to provide researchers with their social media posts.

In DONATION we find differences between decline rates across emotions (χ^2 test $p < 0.001$). Participants finish the study more often when asked for joy-inducing event posts. The fact that the CREATION decline rate does not vary by emotion suggests that people do not find it particularly challenging to create posts about emotions other than joy, but rather they struggle to find appropriate posts in their social media feeds. Participants who declined DONATION and RECENT commonly cited privacy concerns and, in DONATION, they noted a lack of posts that reflected the desired emotion.

4.4 RQ4: Are there differences in how participants label posts for emotion?

Some emotions are harder to find on social media, and some posts may be about events which evoke more than one emotion. This may lead DONATION participants searching for uncommon emotions to submit posts which also or better represent common emotions. In such cases, the labels of the posts would be affected by the target labels presented to the participants, which is an undesirable potential source of bias. To investigate the potential scope of this issue, we allowed participants to label their posts freely with any number of emotions in RECENT. We presume that the annotations of multi-emotion posts are likely to be more affected by the target label presented to the participant.

Figure 8 shows the fractions of posts to which multiple emotions were assigned. Multi-emotion posts are common. Most anger, fear disgust, and surprise posts were additionally labeled with another emotion. We conclude that bias in DONATION emotion annotations is a possibility that should be investigated and mitigated. One approach would be to allow participants to label their posts with multiple emotion labels and the intensity of those emotions, as we did in RECENT.

4.5 RQ5: Do data differences affect model performance?

Our experiments aim to assess if CREATION and DONATION are equally suitable as training and as testing data for predictive models[10]. Specifically,

[9] Table 6 in the Appendix shows the sample demographics.

[10] We do not train or test on RECENT because of its smaller size and unbalanced emotion distribution.

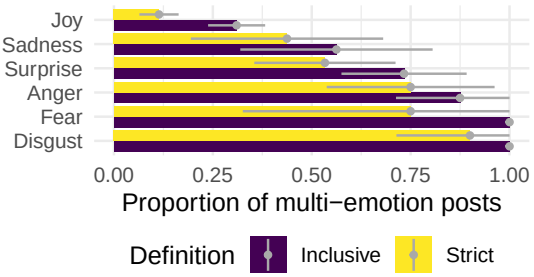


Figure 8: Prevalence of multi-emotion posts in the RECENT study. Lines represent 95% confidence intervals. Posts were counted as multi-emotion under the Inclusive definition if the participant indicated the event evoked two or more emotions, and under the Strict definition if the participant indicated that two or more emotions were tied for greatest intensity.

			Training		
		Mod.	DONATION	CREATION	Zero-shot
Test F_1	Creatn	V	.16	.18	.24[1]
		T	.49	.58	.61[1]
		T+V	.60	.62	.56[2]
	Donatn	V	.19	.18	.19[3]
		T	.41	.42	.45[1]
		T+V	.38	.40	.43[2]

Table 4: Performance of models predicting emotion in multimodal social media posts using text alone (T), image alone (V), and both modalities combined (T+V). We report macro F_1 scores over 5 runs. Zero-shot models are chosen according to the best individual performance on development data. [1]llama3.2-vision. [2]minicpm-v. [3]llava-llama3.

do data differences between CREATION and DONATION have an effect on model performance when used as training data? Are there differences in effectiveness when testing on CREATION vs. DONATION? We fine-tune unimodal and multimodal models separately on CREATION and DONATION data subsets. We also evaluate multimodal foundation models in a zero-shot setup.

Setup. We divide the data such that the development and test sets each have 25 posts per emotion (300 posts per set) and use the remaining data for training (800 posts per strategy).[11] As unimodal models, we use RoBERTa for text (Liu et al., 2019) and ViT for images (Dosovitskiy et al., 2021). For multimodal models, we use the dual encoder CLIP (Radford et al., 2021), applying early fusion by concatenating text and image embeddings, then adding a classification head on top. Appendix A.3 gives model and training details. All models are fine-tuned and tested five times.

For the zero-shot setup, we prompt llama3.2-vision, llava-llama3, and minicpm-v. We prompt each model five times and report results averaged across runs. We select the best model per modality on the development data and report the corresponding result on the test data[12]. We provide model details and prompts in Appendix A.3.

Results. Table 4 shows the main results.[13] Models trained on CREATION and DONATION data perform equally well when tested on DONATION data (bottom block in Table 4). This suggests that the differences between CREATION and DONATION content may not be very important for model training. Performance scores on CREATION test data are higher (top block), and likely unrealistically optimistic, in comparison to scores on DONATION test data. This suggests that genuine data is required to reliably estimate model effectiveness. The zero-shot models' results, which by design solely reflect differences in the test sets, underpin this finding.

Influence of Post and Respondent Features. Our analyses in Sections 4.1, 4.2, 4.3, and 4.4 show that collection strategies lead to differences in samples, and our analyses show that these differences carry over to model performances. We now investigate this relationship between post and respondent features and model performance further.

We fit three separate logistic regression models to test for predictors of correct emotion classification with three multimodal models whose performance is shown in Table 4. CLIP trained on DONATION and CREATION, and minicpm-v as a zero-shot approach. Our dependent variable is a binary variable indicating whether the model predicted the emotion in a given test set post correctly on a particular run. To account for non-independence between predictions for the same post from each of the five runs, we use clustered standard errors.

Independent variables are text length, image style, and text–image relation variables (cf. RQ1), emotion appraisals and event and emotion duration and intensity (RQ2), and participant sociodemographics (RQ3). Numeric and ordinal variables are scaled to $[0; 1]$. For categorical variables, we

[11] RECENT is not included in the test set for reasons detailed in the Limitations Section.

[12] Three other zero-shot models were also tested, but returned results that were either worse in all instances or unparseable. See Appendix A.3 for details on these models.

[13] Full results including precision, recall, and F_1-scores for individual emotions are in Appendix A.4, Tables 9 and 10.

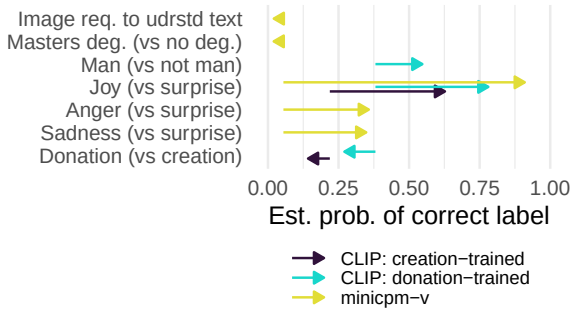


Figure 9: Influence of post and respondent features on T+V model accuracy. Arrows represent predicted change in the probability of a correct post classification when the indicated variable is changed from its reference value (categorical variables) or from its minimum to its maximum (numeric variables). Only effects significant at the p<.05 level or better are shown.

choose reference levels (which serve as our baseline for each variable) following Johfre and Freese (2021).[14] We control for the study the post is from and the emotion it shows. We omit 27 posts for which at least one independent variable is missing.

Figure 9 shows the effects of the statistically significant variables on the probability that a post will be classified correctly.[15] Arrows begin at the probability of correct classification indicated by the model intercept: that is, when all independent variables are either 0 (numeric/ordinal) or at their reference values (categorical). Arrow heads indicate the change in this probability when the indicated independent variable is changed, either to its maximum value (numeric/ordinal) or to the indicated category (categorical).

Minicpm-v is less likely to predict emotion accurately when authors report the image is necessary to understand the post text. Figure 6 shows that CREATION posts have images that are less necessary for text understanding. Together, these two results suggest performance on CREATION data may be better in part because CREATION posts have images that are less integral to text interpretation.

Minicpm-v accuracy is also lower for posts authored by people with advanced degrees compared to those who have not gone to college. The CLIP model trained on DONATION is more likely to predict emotion correctly for posts authored by men. Notably, these effects are net of the effects of emotion and all other variables in the regression. These sociodemographic effects underscore the importance of being attentive to sample composition.

Consistent with our results above, Figure 9 shows that some emotions, and particularly Joy, are predicted more accurately than others. CLIP is less likely to accurately predict emotions for DONATION posts as compared to CREATION posts. That this effect is significant despite the presence of the other independent and control variables indicates there are further study differences that are not accounted for here.

We did not observe statistically significant impacts on classification outcomes for other variables not included in Figure 9. We note that our test set size is relatively small and effects are therefore conservative. The relationship between post/author features and model performance is an important avenue for future research.

5 Conclusion

In this paper, we compared methods for collecting study-created and genuine multimodal social media posts labeled by their authors for emotion. Our work is the first to directly compare these methods of collecting author-labeled corpora, and the first multimodal social media corpus labeled by its authors for emotion.

Our results show that CREATION posts are different in content and style and represent more prototypical events than genuine posts collected in DONATION and RECENT. RECENT leads to a more realistic emotion distribution, but does not evenly represent all emotions, which is a challenge for model development. More participants are comfortable participating in CREATION, suggesting that this corpus may better represent social media users. We note that all presented approaches may be easily scaled up to large quantities of posts. Corpus sizes are only limited by available participants on study platforms such as Prolific, and, for genuine data strategies, the amount of data they can provide.

Despite content differences, CREATION training data leads to models that perform on par with models trained on DONATION data. Modeling results on DONATION test data show worse – and likely more realistic – performance than on the CREATION test data. Therefore, we suggest that future studies consider using a strategy similar to DONATION for developing test sets and CREATION to collect corpora for model development as needed for model optimization.

[14] Refer to Appendix 11 for additional details.
[15] Full results are reported in Appendix Table 11.

Acknowledgments

We thank all ARR reviewers for their thorough reviews and valuable suggestions. We also thank our study participants for contributing to our work. This work has been supported by the Deutsche Forschungsgesellschaft (DFG) in the project "User's Choice of Images and Text to Express Emotions in Twitter and Reddit" (ITEM, Project KL 2869/11-1, No. 513384754).

Limitations

We took precautions to allow for the best comparability of the data that we obtained by our studies, but some pragmatic decisions were required in the design. Most importantly, the studies were performed sequentially, each with multiple phases. As such, the time periods in which posts were collected differed and could potentially affect the content of posts. For example, there are more posts about the US presidential assassination attempt in the study which occurred days after the event than in studies which occured weeks later. We do however note that we do not find evidence in our analysis for such differences that would influence the conclusions of our work. In all three conditions there were posts on a variety of topics, containing a variety of images – posts about current events are not a dominant portion of our dataset. The image content analysis and our qualitative review of posts supports. The requirement in donation and creation to find or create posts about specific emotions likely pushed respondents to reach further back in their feeds, mitigating recency effects to some extent. For example, participants submitted posts related to COVID lockdowns which occurred 3–4 years prior to data collection. That said, we acknowledge it as a limitation, particularly in recent, and we encourage future collections of this type to bear it in mind and spread out their data collection if possible, as we did with recent.

Another limitation may be that we did not control for the (hypothetical or real) social media platform. Posts from different social media platforms may differ in content and style and may therefore not be directly comparable. While we do ask for the platform a post is from, we do not control for this. We do a short analysis of the distribution of platforms in our collected data in Appendix A.2.3. We consider platform effects to be an important direction for future work.

As mentioned in Section 3.2 and Section 4.4, for the RECENT study we asked participants to label their posts for all emotions they experienced in response to the event that inspired their post. While we do ask for emotion intensity ratings, in the case of ties we did not ask them to select a primary emotion. This is a limitation of our work, as it prevents us from using RECENT posts in our modeling analysis, as in the case of ties we cannot assign a single emotion label to the post, making them incompatible with posts from CREATION and DONATION. We suggest future work, for all collection strategies, both ask participants to label all emotions and intensities and to select a primary emotion in the case of ties.

The training and testing of models is limited by the size of our dataset. These experiments are designed to inform us the best methods to collect data, and as such we will use the findings to expand the dataset, allowing us to confirm our results in future work.

The focus of our modeling analysis is to compare training and test sets. As such, the pretrained models we use are selected based on their established performance and reliability, rather than seeking state-of-the-art performance.

Furthermore, for the zero-shot models, we chose to only use models which we could run locally for two reasons: (1) Reproducibility; changes to models such as GPT-4o are out of our control. There is no guarantee that we or other researchers will have access to the exact model used in our experiments. Furthermore, there is no guarentee that changes to the model will be disclosed to users. (2) Costs; When budgeting our research we prioritize increasing data collection efforts over using LLM API services. Especially given the above explanation about our model selection choices.

Ethical Considerations

This study was approved by the ethics review board at the University of Bamberg. In all data acquisition efforts, all participants have been informed about the collection procedure and the use of the data. Nevertheless, we would like to reflect on various potential challenges in this work.

While the participants have been informed about the use of the data, it may sometimes be the case that the content of a post does comprise anonymity, and the study participant may not be aware of the potential impact. This is not an issue with the study-created data, but may be a challenge in the RECENT

and DONATION data. Further, the collected data may contain information about other people not actively participating in the study. Because of these two challenges, we decided to only share the RECENT and DONATION data for research purposes upon request.

We use a Creative Commons licensed image database as to best avoid copyright issues. Copyright regulations vary widely by country. In some countries, fair use rules allow for using images protected by copyright for the purposes of academic research. However, in others research use of copyrighted images may be a copyright infraction. Given the international nature of emotion recognition research, we wish to prioritize data collection methods which produce data that can be used in as many legal contexts as possible. However, teams that wish to use image data must consider their own local context when designing data collections and using available corpora.

VS Code Copilot was used to assist in the writing of the code for the data analysis. Copilot was only used for debugging, documentation, refactoring, and code completion.

References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. DAICT: A dialectal Arabic irony corpus extracted from Twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6265–6271, Marseille, France. European Language Resources Association.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814. Association for Computational Linguistics.

Laura Boeschoten, Jef Ausloos, Judith E. Möller, Theo Araujo, and Daniel L. Oberski. 2022. A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, 4(2):388–423.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Pasquale Capuozzo, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, and Giuseppe Sartori. 2020. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1423–1430, Marseille, France. European Language Resources Association.

Thijs C. Carrière, Laura Boeschoten, Bella Struminskaya, Heleen L. Janssen, Niek C. De Schipper, and Theo Araujo. 2024. Best practices for studies using digital data donation. *Quality & Quantity*.

Junhan Chen, Yumin Yan, and John Leach. 2022. Are emotion-expressing messages more shared on social media? a meta-analytic review. *Review of Communication Research*, 10:–.

Wingyan Chung and Daniel Zeng. 2020. Dissecting emotion and user influence in social media communities: An interaction modeling approach. *Information & Management*, 57(1):103108. Big data and business analytics: A research agenda for realizing business value.

Irina Degtiar and Sherri Rose. 2023. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(Volume 10, 2023):501–524.

Daantje Derks, Agneta H. Fischer, and Arjan E.R. Bos. 2008. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3):766–785. Instructional Support for Enhancing Students' Information Problem Solving Ability.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Aparna Elangovan, Jiayuan He, Yuan Li, and Karin Verspoor. 2024. Principles from clinical research for NLP model generalization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2293–2309, Mexico City, Mexico. Association for Computational Linguistics.

Alejandra Gomez Ortega, Jacky Bourgeois, Wiebke Toussaint Hutiri, and Gerd Kortuem. 2023. Beyond data transactions: A framework for meaningfully informed data donation. *AI & Society*.

Anurag Illendula and Amit Sheth. 2019. Multimodal emotion classification. In *Companion Proceedings of*

The 2019 World Wide Web Conference, WWW '19, page 439–449, New York, NY, USA. Association for Computing Machinery.

Sasha Shen Johfre and Jeremy Freese. 2021. Reconsidering the reference category. *Sociological Methodology*, 51(2):253–269.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104. Association for Computational Linguistics.

Elizabeth A. Kensinger. 2009. Remembering the Details: Effects of Emotion. *Emotion Review*, 1(2):99–113.

Florian Keusch, Paulina K. Pankowska, Alexandru Cernat, and Ruben L. Bach. 2024. Do You Have Two Minutes to Talk about Your Data? Willingness to Participate and Nonparticipation Bias in Facebook Data Donation. *Field Methods*, 36(4):279–293.

Yiyi Li and Ying Xie. 2020. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

E. Paige Lloyd, Jason C. Deska, Kurt Hugenberg, Allen R. McConnell, Brandon T. Humphrey, and Jonathan W. Kunstman. 2019. Miami University deception detection database. *Behavior Research Methods*, 51(1):429–439.

Anthony Lyons and Yoshihisa Kashima. 2003. How Are Stereotypes Maintained Through Communication? The Influence of Stereotype Sharedness. *Journal of Personality and Social Psychology*, 85(6):989–1005.

Anthony Lyons and Yoshihisa Kashima. 2006. Maintaining stereotypes in communication: Investigating memory biases and coherence-seeking in storytelling. *Asian Journal of Social Psychology*, 9(1):59–71.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Jonathan Mummolo and Erik Peterson. 2019. Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, 113(2):517–529.

Tsubasa Nakagawa, Shunsuke Kitada, and Hitoshi Iyatomi. 2022. Expressions Causing Differences in Emotion Recognition in Social Networking Service Documents. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4349–4353. ACM.

Austin Lee Nichols and Jon K. Maner. 2008. The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, 135(2):151–166. PMID: 18507315.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289. Association for Computational Linguistics.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.

Nico Pfiffner, Pim Witlox, and Thomas N. Friemel. 2024. Data Donation Module: A Web Application for Collecting and Enriching Data Donations. *Computational Communication Research*, 6(2):1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. Instagram data donation: A case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

Matthew Rhodes-Purdy, Rachel Navarre, and Stephen M Utych. 2021. Measuring simultaneous emotions: Existing problems and a new way forward. *Journal of Experimental Political Science*, 8(1):1–14.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

Andrea Scarantino. 2016. The philosophy of emotions and its impact on affective science. *Handbook of emotions*, 4:3–48.

Klaus Scherer and Harald Wallbott. 1997. The ISEAR questionnaire and codebook. *Geneva Emotion Research Group*.

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. APPReddit: a corpus of Reddit posts annotated for appraisal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.

Enrica Troiano, Sofie Labat, Marco Antonio Stranisci, Rossana Damiano, Viviana Patti, and Roman Klinger. 2024. Dealing with controversy: An emotion and coping strategy corpus based on role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1634–1658, Miami, Florida, USA. Association for Computational Linguistics.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1).

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Irene I. van Driel, Anastasia Giachanou, J. Loes Pouwels, Laura Boeschoten, Ine Beyens, and Patti M. Valkenburg. 2022. Promises and pitfalls of social media data donations. *Communication Methods and Measures*, 16(4):266–282.

Clara Vania, Ruijie Chen, and Samuel R. Bowman. 2020. Asking Crowdworkers to Write Entailment Examples: The Best of Bad Options. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 672–686. Association for Computational Linguistics.

Aswathy Velutharambath, Amelie Wührl, and Roman Klinger. 2024. Can factual statements be deceptive? the DeFaBel corpus of belief-based deception. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2708–2723, Torino, Italia. ELRA and ICCL.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Survey Details

We provide the detailed questions on the participant information in Table 5 and the details of the participant information collected in Table 6. We allow participants to participate in each emotion/data collection strategy combination once. We ask them to only complete this information the first time. If they provide it more than once anyways, we use their most recent data. For the small number of participants who did not complete the study demographic information, we impute it using demographic data provided by Prolific.

Label	Question Text	Options
Participant information		
Age	How old are you?	$\mathbb{N}^{\geq 18}$
Gender	With which gender(s) do you identify? [multiple]	Woman, Man, Nonbinary, Transgender, Other [write-in]
Education	What is the highest level of education you completed?	No formal qualifications, Secondary education, High school, Undergraduate degree (BA/BSc/other), Graduate degree (MA/MSc/Mphil/other), Doctorate degree (PhD/other)
Ethnicity	With which of the following ethnic groups do you identify the most? [multiple possible]	Australian/New Zealander, North Asian, South Asian, East Asian, Middle Eastern, European, African, North American, South American, Hispanic/Latino, Indigenous, Other [write in]
Social Media Use	Approximately how often do you use social media (browse or participate)?	Every day, 4-6 days a week, 2-3 days a week, Once per week, Occasionally, but less than once per week, Never
Social Media Post	Approximately how often do you post on social media?	Every day, 4-6 days a week, 2-3 days a week, Once per week, Occasionally, but less than once per week, Never
Preferred platform	What is your preferred social media site?	X (Twitter), Facebook, Instagram, Reddit, Tiktok, LinkedIn, Other [write-in]

Table 5: Wording and response options for participant information.

	Creation		Donation		Recent	
Gender identity						
Man	109	(0.46)	119	(0.5)	22	(0.56)
Woman	126	(0.53)	114	(0.48)	17	(0.44)
Other	2	(0.01)	3	(0.01)	0	
Ethnicity *						
European	196	(0.84)	178	(0.76)	29	(0.72)
Non-European or multiethnic	38	(0.16)	57	(0.24)	11	(0.28)
Education						
No college degree	81	(0.34)	73	(0.31)	12	(0.32)
Bachelor's degree	97	(0.4)	99	(0.42)	15	(0.39)
Master's degree or higher	63	(0.26)	62	(0.26)	11	(0.29)
Employment status						
Full-Time	136	(0.59)	140	(0.6)	27	(0.68)
Part-Time	40	(0.17)	41	(0.18)	7	(0.17)
Not employed (unemployed, homemaker, retired, disabled)	45	(0.19)	40	(0.17)	5	(0.12)
Other	10	(0.04)	12	(0.05)	1	(0.03)
Student *						
Yes	32	(0.14)	60	(0.26)	9	(0.22)
No	199	(0.86)	175	(0.74)	31	(0.78)
Social media use						
Every day	203	(0.85)	202	(0.86)	35	(0.88)
2–6 days a week	30	(0.13)	29	(0.12)	3	(0.07)
Once per week or less	6	(0.03)	5	(0.02)	2	(0.05)
Social media post						
Every day	19	(0.08)	32	(0.14)	2	(0.05)
2–6 days a week	65	(0.27)	63	(0.27)	15	(0.38)
Once per week or less	157	(0.65)	141	(0.6)	23	(0.58)
Preferred platform						
Facebook	86	(0.35)	78	(0.33)	16	(0.4)
Instagram	79	(0.32)	78	(0.33)	14	(0.35)
X (Twitter)	37	(0.15)	52	(0.22)	7	(0.17)
Other	42	(0.17)	30	(0.13)	3	(0.07)
Mean age (std. dev) *						
	38.4	(12.3)	33.2	(10.7)	36	(11.8)

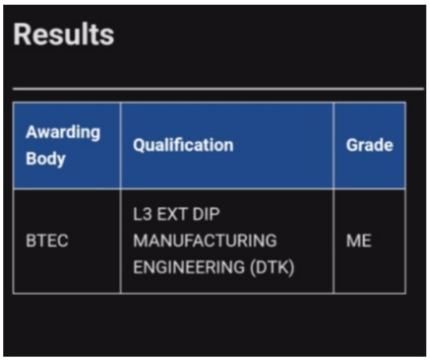
Table 6: Sample composition of the three studies. For categorical variables, values are counts and proportions are shown in parentheses. For age, values are sample means and standard deviations are shown in parentheses. Proportions sum to less than one due to small numbers of participants missing information on given variables. Starred categories are those for which a χ^2 test (categorical variables) or one-way ANOVA (age) shows statistically significant study differences at at least the p<0.05 level.

A.2 Analysis Details

A.2.1 Image Label and Content Analysis

Image Labels: To investigate differences in image style, we manually label each image as a *meme*, *screenshot*, *graphic*, *professional photo*, *personal photo*, or *other*. We inductively developed this list from our observations of our data. The list of labels is hierarchical, such that if the image is a meme and a personal photo, it will be labeled as meme. Examples of each can be seen in Figure 10. *Memes* use an established meme format edited to the authors' specific use. *Screenshots* are images of the authors' computer or smart phone screen. These includes shots of programs, websites, news stories, other social media posts, etc. *Graphic* refers to any image drawn, painted, or otherwise created and is not a photograph, but includes photographs which have been edited with graphic overlays. *Professional photos* are photographs which are **clearly** a stock photo or created for professional purposes such as news, sports media, or advertisements. All other photographs are labeled as *personal photo*. Images which do not fit into any of these are labeled as *other*. This includes images with inspirational quotes and still shots of movies or TV shows.

(a) Meme

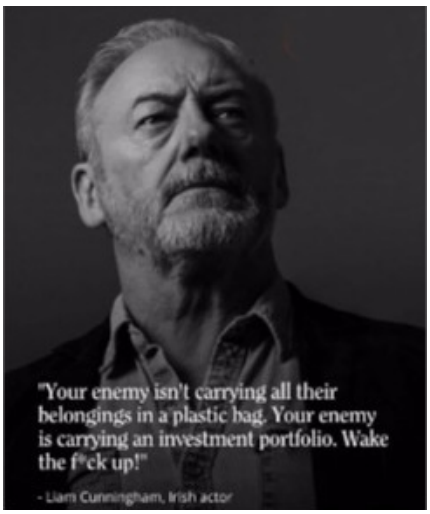


(b) Screenshot

(c) Graphic

(d) Professional Photo

(e) Personal photo



(f) Other

Figure 10: Examples of image labels

Image Content: To analyze image content we use the following prompt with GPT-4o to extract a list of words describing the objects and scenes in each image:

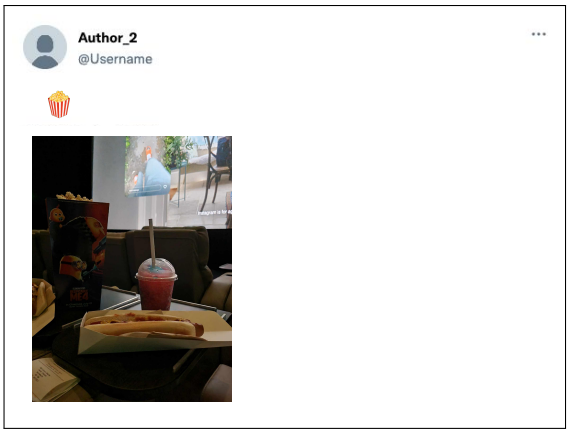
"Please analyze the attached image and list the contents of the image using 1-2 word phrases. If the image contains text, only list 'text', do not repeat the text contained in the image. Output the list one content per line, with no other information but the 1-2 word description.'"

We use GPT-4o for its combination of ease of use, quality of output, and low cost. We validate the results by manually checking the output for 100 randomly selected images. Of 673 content words only 10 were false positives, 98.5% precision. The top 50 most common content words for each study can be seen in Table 7.

Creation		Donation		Recent	
text	496	text	627	text	115
background	140	tree	125	tree	27
tree	137	background	108	woman	26
sky	128	woman	105	man	23
grass	127	sky	98	sky	23
building	100	person	94	background	21
water	97	people	93	person	20
people	94	grass	92	grass	18
woman	77	man	89	smile	16
car	77	smile	84	building	16
light	72	light	74	people	15
blue	71	building	72	dog	14
cloud	71	blue	63	sunglasses	14
black	69	water	62	shirt	14
wall	67	table	59	wall	13
man	64	crowd	53	flower	13
green	64	red	53	water	13
table	61	number	52	table	12
red	61	car	50	blue	12
smile	54	wall	49	chair	11
person	53	cloud	47	cloud	11
chair	52	flower	46	sofa	9
street	50	child	42	crowd	9
White	48	clothing	42	light	9
sign	46	shirt	39	red	9
rock	44	black	39	cat	8
floor	43	hand	38	floor	8
hand	43	sign	37	clothing	8
flower	37	street	37	pavement	7
child	37	floor	37	mountain	7
face	36	green	36	male_child	6
hair	36	expression	36	hand	6
expression	35	drink	36	rock	6
flag	35	dog	36	hat	6
dog	34	hair	34	sunset	6
shirt	33	rock	34	number	6
leaf	33	chair	32	drink	6
plant	33	White	30	orange	5
yellow	32	decoration	29	child	5
clothing	32	glass	26	spectacles	5
drink	31	smoke	25	black	5
eyes	30	bag	25	top	5
hat	29	food	25	hair	5
suit	29	dress	25	two	5
hands	28	blanket	25	outdoor	5
window	28	window	25	window	5
field	27	plant	24	White	5
beach	27	player	24	beach	5
crowd	26	orange	24	colorful	5
bag	25	suit	24	shadow	5

Table 7: Top 50 most common image content words for each approach, generated by GPT-4o.

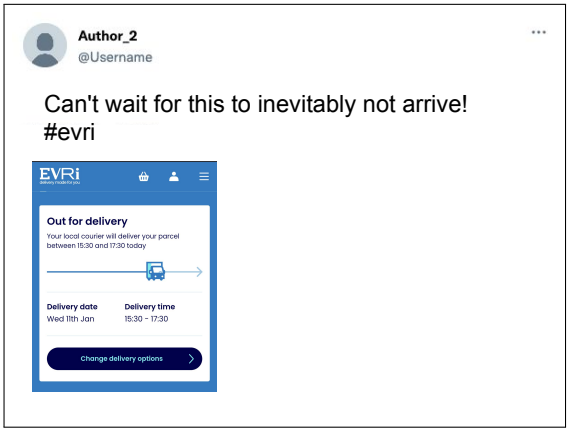
(a) CREATION post labeled as surprise.



(b) RECENT post labeled as joy.

Figure 11: Comparison of posts submitted for CREATION and RECENT. CREATION gives a long detailed explanation of the event, including descriptions of their emotions, while RECENT uses only a single emoji to describe a similar event, spending time with their dad.

(a) CREATION post labeled as anger.



(b) DONATION post labeled as anger.

Figure 12: Comparison of posts submitted for CREATION and DONATION. CREATION uses a stock photo, while DONATION demonstrates a reaction post in which the author uses a screenshot of an event they experienced to illustrate their emotion.

A.2.2 Examples of Posts

Here we look at specific examples of posts which highlight the differences described in Section 4.

Figure 11 shows how length and detail of description vary by collection strategy. In post (a) from CREATION you can see how the author explicitly describes both the event and the emotion they experienced. In post (b) from RECENT the author uses only a single emoji to describe a similar event, spending time with their Dad. The event can only be gleaned from the additional annotation in which the author describes it as "I felt happy because I was spending time with my dad".

Figure 12 highlights how image types can affect posts: post (a) uses a stock photo to stand in as a representation of the event which triggered anger in them. The specific event is not detailed and the generic image does not help give any more clues to the reader. The post from DONATION, however, uses a screenshot that directly references the event which triggered the author' anger. These reaction posts are not possible in CREATION, since the participants only have access to the image database, and not their own personal images. This is also another example of CREATION posts being about general events, where DONATION tend to be about specific events.

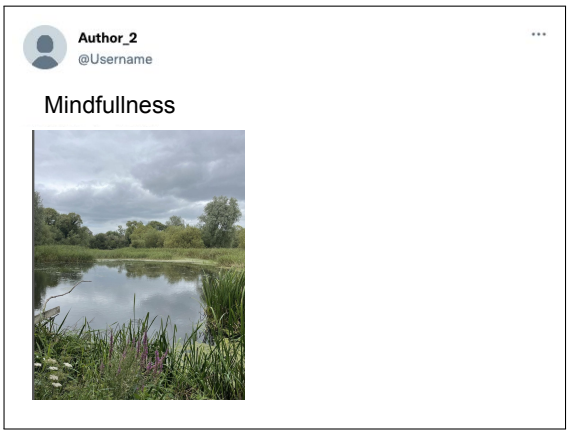
Figure 13 illustrates how posts can rely more on the text or the image to convey emotion. In post (a), a CREATION post, the image conveys almost no emotion, while the text explicitly describes the author's

(a) CREATION post labeled as fear.

(b) DONATION post labeled as fear.

Figure 13: Comparison of posts submitted for CREATION and DONATION. CREATION's image conveys no emotion and relies entirely, on the text to convey emotion, while DONATION's text is emotion neutral and relies on the image to convey the experience of fear. experienced to illustrate their emotion.

(a) CREATION post labeled as sadness.



(b) DONATION post labeled as sadness.

Figure 14: Comparison of posts submitted for CREATION and DONATION. CREATION is a prototypical event for sadness, while DONATION is a more general sense of sadness.

emotion. Post (b), a DONATION post, on the other hand, conveys no emotion in the text, relying on the image to convey the author's emotion of fear.

In Figure 14 we can see that post (a) was inspired by a prototypical event for sadness, the death of a loved one. Post (b) is less obvious what event inspired the author to post, or indeed which emotion they are expressing. The author describes the event as "I was feeling upset that day and so was on a walk around a local nature reserve to unwind". In this case the event is more difficult to define.

A.2.3 Platform Distribution between Approaches

Figure 15 shows the distribution of posts between social media platforms. Participants in DONATION and RECENT were asked which platform their donated post was originally posted on. Participants in CREATION were asked which platform(s) they would have posted their study-created post on.

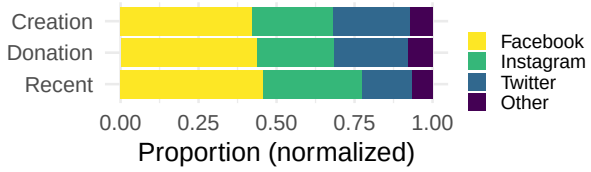


Figure 15: Distribution of post platforms under the CREATION, DONATION, and RECENT approaches. As multiple responses were allowed in CREATION, proportions are normalized such that they sum to one.

A.3 Model Details

All models are trained using four NVIDIA L40 GPUs. Each supervised model is fine-tuned five times, using the exact same setup and environment. For zero-shot models, every prompt is run five times per model. The average and range of scores is reported. All code can be found in the supplementary materials.[16]

A.3.1 Fine-tuned Models

Text-only: We use RoBERTa-base pretrained model for both tokenizer and model, via the transformers Python package.[17] All text-only models are trained using 10 epochs, 1e-5 training rate, 16 batch size, 0.01 weight decay, and early stopping.
Vision-only: We fine-tune the vit-base-patch16-224 model using a combination of transformers and torchvision Python packages.[18] Images are preproccessed by resizing them to 224x224. All image-only models are trained using 5 epochs, 5e-5 learning rate, 32 batch size, cross entropy loss, and early stopping.
Text+Vision: We fine-tune the clip-vit-base-patch32 model using a combination of transformers and torchvision Python packages.[19] Both images and text are encoded via the CLIP processor, and then fused using a simple concatenation method. All text+vision models are trained using 10 epochs, 1e-5 learning rate, batch size 8, cross entropy loss, and early stopping.

A.3.2 Zero-shot Models

We use six multimodal models made available through the Ollama Python package[20]. These are llama3.2-vision[21] (11b, ID 085a1fdae525), minicpm-v[22] (8b, v2.6, ID c92bfad01205), llava[23] (7b, v1.6, ID 8dd30f6b0cb1), llava-llama3[24] (8b, ID 44c161b1f465), bakllava[25] (7b, ID 3dd68bd4447c), and llava-phi3[26] (3.8b, ID c7edd7b87593). Bakllava and llava-phi3 returned nearly entirely unparseable results and were hence eliminated from further analyis.

We perform basic cleaning on responses (converting to lowercase, removing punctuation and whitespace, removing the word "emotion"). We continue to prompt the model until we received five valid emotion predictions, with "valid" defined as equalling one of the six provided emotion options after cleaning. We prompt models up to a maximum of 30 times for each post. In the small number of cases in which we did not receive five valid predictions in 30 tries, we count missing responses as incorrect for the purposes of performance statistic calculations.

[16] https://www.uni-bamberg.de/en/nlproc/projects/item/
[17] https://huggingface.co/transformers/v2.9.1/model_doc/roberta.html
[18] https://huggingface.co/google/vit-base-patch16-224
[19] https://huggingface.co/openai/clip-vit-base-patch32
[20] https://pypi.org/project/ollama/0.4.6/
[21] https://ollama.com/library/llama3.2-vision
[22] https://ollama.com/library/minicpm-v
[23] https://ollama.com/library/llava
[24] https://ollama.com/library/llava-llama3
[25] https://ollama.com/library/bakllava
[26] https://ollama.com/library/llava-phi3

Section	Text	Vision	Text + Vision
Task Descr.	Which emotion does the following text from a social media post convey most strongly?	Which emotion does the following image from a social media post convey most strongly?	Which emotion does the following text and image from a social media post convey most strongly?
Labels	Please choose one of the set [anger, disgust, fear, joy, sadness, surprise].	Please choose one of the set [anger, disgust, fear, joy, sadness, surprise].	Please choose one of the set [anger, disgust, fear, joy, sadness, surprise].
Format Instr.	Only provide a single word indicating the emotion. Do not provide explanation or analysis. Only provide plain text.	Only provide a single word indicating the emotion. Do not provide explanation or analysis. Only provide plain text.	Only provide a single word indicating the emotion. Do not provide explanation or analysis. Only provide plain text.
Data Input	Post text: {text}		Post text: {text}
Image		this_image.{image type}	this_image.{image type}

Table 8: Prompts for text, vision, and text + vision modalities. Variables are typeset in {curly brackets}. Image filenames were changed to a default name before sending to the LLM to avoid the model being able to gain additional information.

A.4 Additional Modeling Results

			llama3.2-vision		minicpm-v		llava		llava-llama3	
		Modality	F_1	Range	F_1	Range	F_1	Range	F_1	Range
Test	Creatn	Vision	.24	.20–.30	.21	.20–.24	.25	.21–.31	.25	.21–.33
		Text	.61	.60–.63	.58	.54–.65	.55	.52–.58	.49	.41–.55
		T+V	.53	.50–.57	.56	.52–.60	.44	.42–.48	.43	.37–.49
	Donatn	Vision	.21	.17–.25	.25	.21–.30	.23	.19–.28	.19	.14–.24
		Text	.45	.41–.48	.45	.42–.49	.42	.39–.45	.44	.41–.47
		T+V	.39	.35–.43	.43	.41–.45	.33	.28–.38	.26	.18–.30

Table 9: Macro F_1 of multimodal zero-shot models for predicting emotion on CREATION and DONATION posts using post text (T), post image (V), and both text and image (T+V).

Test	Modality	Emotion	Training Data: CREATION F_1	CREATION P	CREATION R	DONATION F_1	DONATION P	DONATION R	Zero-shot F_1	Zero-shot P	Zero-shot R
Creation	Vision	anger	.25	.27	.21	.17	.17	.14	.01[1]	.10[1]	.01[1]
	Vision	disgust	.16	.07	.16	.15	.10	.16	.19[1]	.30[1]	.14[1]
	Vision	fear	.12	.12	.12	.18	.10	.25	.14[1]	.31[1]	.09[1]
	Vision	joy	.11	.20	.12	.06	.09	.05	.39[1]	.26[1]	.78[1]
	Vision	sadness	.23	.29	.25	.20	.14	.19	.26[1]	.26[1]	.26[1]
	Vision	surprise	.22	.23	.24	.17	.14	.17	.18[1]	.17[1]	.20[1]
	Text	anger	.46	.42	.40	.47	.50	.49	.65[1]	.60[1]	.71[1]
	Text	disgust	.57	.41	.67	.45	.31	.60	.55[1]	.76[1]	.43[1]
	Text	fear	.50	.52	.49	.34	.33	.29	.59[1]	.89[1]	.44[1]
	Text	joy	.73	.67	.77	.70	.58	.65	.76[1]	.63[1]	.95[1]
	Text	sadness	.61	.57	.64	.57	.41	.66	.57[1]	.44[1]	.83[1]
	Text	surprise	.57	.79	.50	.28	.17	.22	.20[1]	.59[1]	.12[1]
	Text + Vision	anger	.55	.62	.51	.50	.45	.50	.58[2]	.51[2]	.70[2]
	Text + Vision	disgust	.55	.55	.61	.53	.42	.58	.53[2]	.66[2]	.44[2]
	Text + Vision	fear	.57	.52	.60	.56	.65	.54	.56[2]	.75[2]	.45[2]
	Text + Vision	joy	.79	.74	.82	.77	.83	.78	.71[2]	.56[2]	.97[2]
	Text + Vision	sadness	.65	.63	.66	.62	.59	.65	.60[2]	.54[2]	.68[2]
	Text + Vision	surprise	.59	.77	.50	.61	.71	.55	.20[2]	.39[2]	.13[2]
Donation	Vision	anger	.13	.28	.10	.22	.12	.21	.34[3]	.34[3]	.34[3]
	Vision	disgust	.14	.25	.14	.16	.27	.14	.03[3]	.08[3]	.02[3]
	Vision	fear	.19	.24	.18	.23	.15	.28	.07[3]	.10[3]	.06[3]
	Vision	joy	.11	.20	.12	.03	.06	.03	.36[3]	.23[3]	.92[3]
	Vision	sadness	.21	.15	.19	.26	.22	.26	.00[3]	.00[3]	.00[3]
	Vision	surprise	.27	.25	.33	.19	.13	.21	.10[3]	.22[3]	.06[3]
	Text	anger	.36	.41	.31	.47	.43	.44	.54[1]	.57[1]	.52[1]
	Text	disgust	.48	.42	.50	.34	.27	.41	.30[1]	.44[1]	.22[1]
	Text	fear	.35	.33	.36	.18	.10	.13	.40[1]	.68[1]	.29[1]
	Text	joy	.55	.38	.81	.61	.48	.74	.53[1]	.40[1]	.79[1]
	Text	sadness	.43	.50	.39	.45	.39	.44	.39[1]	.30[1]	.54[1]
	Text	surprise	.20	.30	.14	.33	.31	.32	.27[1]	.53[1]	.18[1]
	Text + Vision	anger	.34	.37	.34	.33	.25	.34	.51[2]	.48[2]	.54[2]
	Text + Vision	disgust	.29	.32	.27	.28	.24	.28	.27[2]	.41[2]	.21[2]
	Text + Vision	fear	.37	.36	.38	.35	.44	.34	.33[2]	.51[2]	.25[2]
	Text + Vision	joy	.57	.52	.68	.57	.47	.67	.50[2]	.35[2]	.87[2]
	Text + Vision	sadness	.42	.48	.42	.38	.45	.38	.44[2]	.43[2]	.46[2]
	Text + Vision	surprise	.36	.38	.32	.34	.32	.30	.25[2]	.53[2]	.17[2]

Table 10: Model predictive performance by emotion. Results shown are averaged over 5 runs. Zero-shot models are chosen for each data-test set combination based on overall performance on development data, and are the same here a reported in the main text. [1]llama3.2-vision. [2]minicpm-v. [3]llava-llama3.

Table 11: Coefficients for logistic regressions predicting correct model emotion classification, in log odds units. Numeric and ordinal variables are scaled to range from 0–1. Reference levels – which serve as baselines for each variable – are indicated with hyphens in the table. They are chosen according to Johfre and Freese (2021), and are in general either the lowest category conceptually (in the case of inherently ordered variables like education) or lowest category numerically in terms of their effect on the dependent variable (in the case of unordered variables like emotion). Clustered standard errors (reported in parentheses) are used to account for modeling multiple predictions of the same posts. $^{+}$p<.1. *p<.05. **p<.01. *** p<.001.

	CLIP: creation-trained		CLIP: donation-trained		minicpm-v	
	Coef	SE	Coef	SE	Coef	SE
(Intercept)	−1.27	(0.98)	−0.49	(0.94)	−2.85	(1.49)+
Emotion						
surprise	———		———		———	
anger	0.42	(0.54)	−0.19	(0.54)	2.26	(0.70)**
disgust	0.64	(0.51)	0.26	(0.48)	1.02	(0.67)
fear	0.85	(0.53)	0.40	(0.53)	0.87	(0.67)
joy	1.79	(0.45)***	1.76	(0.43)***	5.16	(0.89)***
sadness	1.05	(0.56)+	0.47	(0.55)	2.22	(0.74)**
Study						
creation	———		———		———	
donation	−0.53	(0.26)*	−0.51	(0.25)*	−0.59	(0.37)
Text-image relation						
Text describe image	−0.30	(0.37)	0.13	(0.34)	0.31	(0.50)
Text understand image	−0.28	(0.36)	−0.36	(0.34)	−0.49	(0.46)
Image understand text	−0.50	(0.36)	−0.46	(0.33)	−1.02	(0.45)*
Image conveys emotion	−0.14	(0.43)	−0.65	(0.47)	0.08	(0.53)
Text conveys emotion	0.41	(0.47)	0.67	(0.47)	−0.03	(0.51)
Emotion appraisals						
Familiarity	0.18	(0.32)	−0.01	(0.32)	0.17	(0.39)
Predictability	−0.20	(0.40)	0.07	(0.40)	−0.74	(0.58)
Pleasantness	0.30	(0.77)	−0.20	(0.75)	−0.49	(1.04)
Unpleasantness	0.34	(0.66)	0.28	(0.69)	1.07	(0.78)
Goalrelevance	0.09	(0.44)	0.00	(0.44)	−0.14	(0.58)
Ownresponsibility	−0.04	(0.44)	0.01	(0.41)	0.10	(0.54)
Anticip.conseq	0.24	(0.44)	0.09	(0.44)	0.69	(0.59)
Goalsupport	0.27	(0.49)	−0.01	(0.50)	0.28	(0.72)
Own control	−0.08	(0.49)	0.05	(0.47)	−0.38	(0.62)
Others control	0.25	(0.33)	−0.12	(0.32)	−0.07	(0.42)
Accept conseq.	−0.35	(0.37)	−0.48	(0.39)	−0.52	(0.50)
Internal standards	−0.73	(0.44)+	−0.70	(0.42)+	−0.99	(0.51)+
Attention	0.41	(0.46)	0.60	(0.46)	0.52	(0.56)
Not consider	0.18	(0.44)	0.09	(0.41)	0.62	(0.52)
Effort	0.09	(0.45)	−0.29	(0.46)	−0.58	(0.55)
Participant gender identity						
Nonman	———		———		———	
Man	0.40	(0.27)	0.68	(0.27)*	0.63	(0.36)+
Age	1.43	(0.77)+	1.22	(0.73)+	0.70	(0.96)
Participant ethnicity						
Non-European or multiethnic	———		———		———	
European	0.16	(0.34)	−0.03	(0.32)	0.08	(0.39)

continued . . .

Table 11: Coefficients for logistic regressions predicting correct model emotion classification, in log odds units. Numeric and ordinal variables are scaled to range from 0–1. Reference levels – which serve as baselines for each variable – are indicated with hyphens in the table. They are chosen according to Johfre and Freese (2021), and are in general either the lowest category conceptually (in the case of inherently ordered variables like education) or lowest category numerically in terms of their effect on the dependent variable (in the case of unordered variables like emotion). Clustered standard errors (reported in parentheses) are used to account for modeling multiple predictions of the same posts. $^{+}$p<.1. *p<.05. **p<.01. *** p<.001.

	CLIP: creation-trained		CLIP: donation-trained		minicpm-v	
	Coef	SE	Coef	SE	Coef	SE
Participant education						
No college	———		———		———	
B.Sc.	0.20	(0.28)	0.10	(0.28)	0.07	(0.33)
Masters or more	−0.33	(0.35)	0.15	(0.35)	−1.03	(0.42)*
Participant employment status						
Not full-time	———		———		———	
Full-Time	−0.06	(0.26)	−0.06	(0.26)	−0.01	(0.33)
Participant is a student						
No	———		———		———	
Yes	0.02	(0.32)	−0.39	(0.30)	0.21	(0.43)
Participant social media frequency						
Use social media	−0.04	(0.44)	0.06	(0.46)	1.16	(0.86)
Post on social media	−0.31	(0.42)	−0.21	(0.38)	−0.03	(0.52)
Participant preferred platform						
Facebook	———		———		———	
Instagram	0.63	(0.33)+	0.50	(0.31)	0.67	(0.45)
X (Twitter)	0.30	(0.40)	−0.17	(0.38)	0.45	(0.54)
Other platform	0.33	(0.41)	0.19	(0.38)	0.31	(0.50)
Image style						
Personal Photo	———		———		———	
Pro Photo	0.10	(0.31)	0.02	(0.33)	−0.02	(0.43)
Other image style	−0.33	(0.36)	−0.28	(0.36)	0.02	(0.42)
Post text length	−0.11	(1.12)	−0.82	(1.09)	0.63	(2.31)
Duration and intensity						
Event duration	−0.94	(0.64)	−0.55	(0.63)	−0.11	(0.75)
Event intensity	0.14	(0.78)	−0.08	(0.80)	1.59	(1.06)
Emotion duration	0.37	(0.75)	0.12	(0.77)	−0.58	(0.84)
Emotion intensity	−0.04	(0.73)	0.38	(0.72)	−1.37	(1.06)
Num.Obs.	1365		1365		1362	
RMSE	0.45		0.45		0.39	