On the Complementarity of Images and Text for the Expression of Emotions in Social Media

Anna Khlyzova[1,2]**, Carina Silberer**[2] and **Roman Klinger**[2]
[1]Computer Science and Engineering, University of South Florida, USA
[2]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
{anna.khlyzova,carina.silberer,roman.klinger}
@ims.uni-stuttgart.de

Abstract

Authors of posts in social media communicate their emotions and what causes them with text and images. While there is work on emotion and stimulus detection for each modality separately, it is yet unknown if the modalities contain complementary emotion information in social media. We aim at filling this research gap and contribute a novel, annotated corpus of English multimodal Reddit posts. On this resource, we develop models to automatically detect the relation between image and text, an emotion stimulus category and the emotion class. We evaluate if these tasks require both modalities and find for the image–text relations, that text alone is sufficient for most categories (complementary, illustrative, opposing): the information in the text allows to predict if an image is required for emotion understanding. The emotions of anger and sadness are best predicted with a multimodal model, while text alone is sufficient for disgust, joy, and surprise. Stimuli depicted by objects, animals, food, or a person are best predicted by image-only models, while multimodal models are most effective on art, events, memes, places, or screenshots.

1 Introduction

The main task in emotion analysis in natural language processing is emotion classification into predefined sets of emotion categories, for instance, corresponding to basic emotions (fear, anger, joy, sadness, surprise, disgust, anticipation, and trust, Ekman, 1992; Plutchik, 1980). In psychology, emotions are commonly considered a reaction to an event which consists of a synchronized change of organismic subsystems, namely neurophysiological changes, reactions, action tendencies, the subjective feeling, and a cognitive appraisal (Scherer et al., 2001). These theories recently received increasing attention, for instance, by comparing the way how emotions are expressed, based on

these components (Casel et al., 2021), and by modelling emotions in dimensional models of affect (Buechel and Hahn, 2017) or appraisal (Hofmann et al., 2020). Further, the acknowledgment of emotions as a reaction to some relevant event (Scherer, 2005) leads to the development of stimulus detection systems. This task is formulated in a token-labeling setup (Song and Meng, 2015; Bostan et al., 2020; Kim and Klinger, 2018; Ghazi et al., 2015; Oberländer and Klinger, 2020, i.a.), as clause classification (Gui et al., 2017, 2016; Gao et al., 2017; Xia and Ding, 2019; Oberländer and Klinger, 2020, i.a.), or as a classification task into a predefined inventory of relevant stimuli (Mohammad et al., 2014).

In social media, users express emotions including text and images. Most attention has been devoted to Twitter, due to its easy-to-use API and popularity (Mohammad, 2012; Schuff et al., 2017; Wang et al., 2012). However, this platform has a tendency to be text-focused, and has therefore not triggered too much attention towards other modalities. Although text may be informative enough to recognize an emotion in many cases, images may modulate the meaning, or sometimes solely convey the emotion itself (see examples in Figure 1). The growing popularity of vision-centered platforms like TikTok or Instagram, and lack of research on multimodal social media constitute a research gap.

With this paper, we study how users on social media make use of images and text jointly to communicate their emotion and the stimulus of that emotion. We assume that linking depictions of stimuli to the text supports emotion recognition across modalities. We study multimodal posts on the social media platform Reddit[1], given its wide adoption, the frequently found use of images and text, and the available programming interfaces to access the data (Baumgartner et al., 2020a). Our goal is to understand how users choose to use an image

---

[1]https://www.reddit.com/

(a) joy/complementary/animal. https://www.reddit.com/r/happy/ comments/j76dog/my_everyday_joy_ is_to_see_my_adorable_cat_smiles/

(b) fear/complementary/animal. https://www.reddit.com/r/WTF/ comments/k2es5l/dont_move_ to_australia_unless_you_can_ handle/

(c) surprise/complementary/object. https://www.reddit.com/r/What/comments/ exh0ms/why_didnt_it_fall/

Figure 1: Example of posts from Reddit (annotation are emotion/relation/stimulus category).

in addition to text, and the role of the relation, the emotion, and the stimulus for this decision. Further we analyze if the classification performance benefits from a joint model across modalities. Figure 1 shows examples for Reddit posts. In Figure 1a, both image and text would presumably allow to infer the correct emotion even when considered in isolation. In Figure 1b, additional knowledge of the complementary role of the picture depicting an animal can inform an emotion recognition model. In Figure 1c the image alone would not be sufficient to infer the emotion, but the text alone is.

We therefore contribute (1) a new corpus of multimodal emotional posts from Reddit, which is annotated for authors' emotions, image–text relations, and emotion stimuli. We (2) analyze the relations of the annotated classes and find that certain emotions are likely to appear with certain relations and emotion stimuli. Further, we (3) use a transformer-based language model (pretrained RoBERTa model, Liu et al., 2019) and a residual neural network (Resnet50, He et al., 2016) to create classification models for the prediction of each of the three classes mentioned above. We analyze for which classification tasks multimodal models show an improvement over unimodal models. Our corpus is publicly available at https://www.ims.uni-stuttgart.de/data/mmemo.

2 Related Work

Emotion Analysis. Emotion analysis has a rich history in various domains, such as fairy tales (Alm et al., 2005), email writing (Liu et al., 2003), news headlines (Strapparava and Mihalcea, 2007), or

blog posts (Mihalcea and Liu, 2006; Aman and Szpakowicz, 2008; Neviarouskaya et al., 2010). The focus of our study is on emotion analysis in social media, which has also received considerable attention (Purver and Battersby, 2012; Wang et al., 2012; Colnerič and Demšar, 2018; Mohammad, 2012; Schuff et al., 2017, i.a.). Twitter[2] is a popular social media platform for emotion analysis, in both natural language processing (NLP) and computer vision. We point the reader to recent shared tasks for an overview of the methods that lead to the current state-of-the-art performance (Klinger et al., 2018; Mohammad et al., 2018).

One of the questions that needs to be answered when developing an emotion classification system is that of the appropriate set of emotions. There are two main theories regarding emotion models in psychology that found application in NLP: discrete sets of emotions and dimensional models. Psychological models that provide discrete sets of emotions include Ekman's model of basic emotions (anger, disgust, surprise, joy, sadness, and fear, Ekman, 1992) and Plutchik's wheel of emotions (adding trust and anticipation, Plutchik, 1980, 2001). Dimensional models define where emotions lie in a vector space in which the dimensions have another meaning, including affect (Russell, 1980; Bradley et al., 1992) and cognitive event appraisal (Scherer, 2005; Hofmann et al., 2020; Shaikh et al., 2009). In our study, we use the eight emotions from the Plutchik's wheel of emotions.

Multimodal Analyses. The area of emotion analysis also received attention from the computer vi-

---

[2]https://www.twitter.com/

sion community. A common approach is to use transfer learning from general image classifiers (He and Ding, 2019) or the analysis of facial emotion expressions, with features of muscle movement (De Silva et al., 1997) or deep learning (Li and Deng, 2020). Dellagiacoma et al. (2011) use texture and color features to analyze social media content. Other useful properties of images for emotion analysis include the occurrence of people, faces, shapes of objects, and color distributions (Zhao et al., 2018; Lu et al., 2012).

Such in-depth analyses are related to stimulus detection. Peng et al. (2016) detect emotion-eliciting image regions. They show, on a Flickr image dataset, that not only objects (Wu et al., 2020) and salient regions (Zheng et al., 2017) have an impact on elicited emotions, but also contextual background. Yang et al. (2018), inter alia, show that it is beneficial for emotion classification to explicitly integrate visual information from emotion-eliciting regions. Similarly, Fan et al. (2018) study the relationship between emotion-eliciting image content and human visual attention.

Image–Text Relation. A set of work aimed at understanding the relation between images and text. Marsh and White (2003) establish a taxonomy of 49 functions of illustrations relative to text in US government publications. The relations contain categories like "elicit emotion", "motivate", "explains", or "compares" and "contrasts". Martinec and Salway (2005) aim at understanding both the role of an image and of text.

In contrast to these studies which did not develop machine learning approaches, Zhang et al. (2018) develop automatic classification methods for detection of relations between the image and a slogan in advertisements. They detect if the image and the text make the same point, if one modality is unclear without the other, if the modalities, when considered separately, imply opposing ideas, and if one of the modalities is sufficient to convey the message. Weiland et al. (2018) focus on detecting if captions of images contain complementary information. Vempala and Preoţiuc-Pietro (2019) infer relationship categories between the text and image of Twitter posts to see how the meaning of the entire tweet is composed. Kruk et al. (2019) focus on understanding the intent of the author of an Instagram post and develop a hierarchy of classes, namely *advocative*, *promotive*, *exhibitionist*, *expressive*, *informative*, *entertainment*, *provoca-*

tive/discrimination, and *provocative/controversial*. They also analyze the relation between the modalities with the classes *divergent*, *additive*, or *parallel*. Our work is similar to the two previously mentioned papers, as the detection which emotion is expressed with a post is related to intent understanding.

3 Corpus Creation

To study the roles of images in social media posts, we create an annotated Reddit dataset with labels of emotions, text–image relations, and emotion stimuli. We first discuss our label sets and then explain the data collection and annotation procedures.

3.1 Taxonomies

We define taxonomies for the emotion, relation, and stimulus tasks.

Emotion Classification. To classify social media posts in terms of what emotion the author likely felt when creating the post, we use the Plutchik's wheel of emotions as the eight labels in our annotation scheme, namely *anger*, *anticipation*, *joy*, *sadness*, *trust*, *surprise*, *fear*, and *disgust*.

Relation Classification. To develop a classification scheme of relations of emotion-eliciting image–text pairs, we randomly sampled 200 posts, and created a simple annotation environment for preliminary annotation that displayed an image–text pair next to questions to be answered (see Figure 6 in the Appendix). Based on the preliminary annotation, we propose the following set of relation categories. 1. *complementary*: the image is necessary to understand the author's emotion; the text alone is not sufficient but when coupled with the image, the emotion is clear; 2. *illustrative*: the image illustrates the text but the text alone is enough to understand the emotion; the image does not communicate the emotion on its own; 3. *opposite*: the image and the text pull in different directions; they are contradicting when taken separately, but when together, the emotion is clear; 4. *decorative*: the image is used for aesthetic purposes; the emotion is primarily communicated with the text while the image may seem unrelated; 5. *emotion is communicated with image only*: the text is redundant for emotion communication.

We show examples for the *complementary* and *illustrative* relations in Figure 2. An example for the *opposite* relation could be an image with an ugly creature with a text "isn't he the prettiest thing
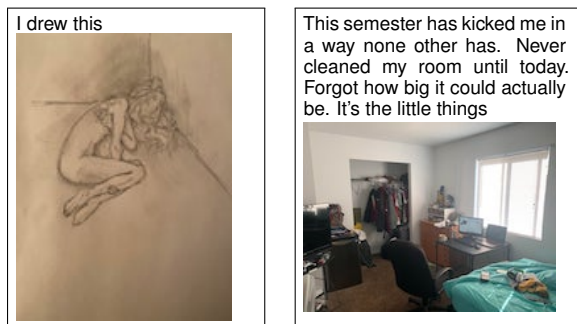
(a) Relation: complementary. https://www.reddit.com/r/sad/comments/jxgoxj/i_drew_this/

(b) Relation: illustrative. https://www.reddit.com/r/happy/comments/jwje64/this_semester_has_kicked_me_in_a_way_none_other/

Figure 2: Example of image–text relationships in posts.

in the world". Posts in which the text and the image are essentially unrelated fall into the *decorative* category. Posts where images have inspirational texts like "No Happiness is Ever Wasted" and the text contains the same words would fall into the last category (*image-only*).

Stimulus Classification. Based on the preliminary annotation procedure described for the relation taxonomy, we further obtain the following categories for emotion stimuli in images of multimodal posts: *person/people*, *animal*, *object*, *food*, *meme*, *screenshot/text in image*, *art/drawing*, *advertisement*, *event/situation*, and *place*. We provide examples of all stimuli in the Appendix in Figure 5.

3.2 Data Collection

We collect our multimodal data from Reddit, where posts are published under specific subreddits, user-created areas of interest, and are usually related to the topic of the group. Our data comes from 15 subreddits which we found by searching for emotion names. These subreddits are "happy", "happiness", "sad", "sadness", "anger", "angry", "fear", "disgusting", "surprise", "what", "WTF", "Cringetopia", "MadeMeSmile", "woahdude", which we complement by "r/all".

We collect the data from the Pushshift Reddit Dataset, a collection of posts and comments from Reddit from 2015 (Baumgartner et al., 2020b), with the help of the Pushshift-API[3]. We only consider posts which have both text and an image. From the initial set of instances that we collected (5,363) we manually removed those with images of low

quality, pornographic and sexually inappropriate content, spam, or in a language other than English.

3.3 Data Annotation

We developed the annotation task with a subsample of 400 posts in a preliminary experiment. It was performed by two groups of three students and with a direct interaction with the authors of this paper, to obtain an understandable and unambiguous formulation of the questions that we used for the actual crowdsourcing annotation. The actual annotation of 1,380 randomly sampled posts was then performed with Amazon Mechanical Turk (AMT[4]) in two phases. In the first phase, we identify posts which likely contain an emotion by asking

1. Does the author want to express an emotion with the post?

In the second phase, we collect annotations for posts which contain an emotion (we accept a post if 1/3 of the annotators marked it as emotional) and ask

2. What emotion did the author likely feel when writing this post?
3. What is the relation between the image and the text regarding emotion communication?
4. What is it in the image that triggers the emotion?

For both phases/experiments, we gather annotations by three annotators. All questions allow one single answer. We show the annotation interface on Amazon Mechanical Turk for the second phase in the Appendix in Figure 7.

For the modelling which we describe in Section 4, we use a union of all labels from all annotators, acknowledging the subjective nature of the annotation task. This leads to multi-label classification, despite the annotation being a single-label annotation task.

Quality Assurance and Annotator Prescreening. Each potential annotator must reside in a predominantly English-speaking country (Australia, Canada, Ireland, New Zealand, United Kingdom, United States), and have an AMT approval rate of at least 90 %. Further, before admitting annotators to each annotation phase, we showed them five manually selected posts that we considered to be straightforward to annotate. For each phase, annotators needed to correctly answer 80 % of the questions associated with those posts. Phase 1 had a 100 % acceptance rate; in Phase 2 this qualifica-

---

[3] https://www.github.com/pushshift/api

[4] https://www.mturk.com/

4

	Label	≥ 1	≥ 2	= 3	κ
Emo.	Yes	1,061	670	333	0.3
	No	1,047	710	319	0.3
Which emotion?	Anger	138	41	8	.26
	Anticipation	85	12	1	.11
	Disgust	268	127	57	.45
	Fear	64	15	5	.28
	Joy	585	444	329	.67
	Sadness	103	52	27	.56
	Surprise	435	221	84	.38
	Trust	54	6	1	.11
	Overall	*1732*	*918*	*512*	*.47*
Relation?	Complementary	1042	773	388	.02
	Decorative	124	6	0	.01
	Illustrative	476	152	4	.07
	Image only	142	27	0	.11
	Opposite	28	0	0	−.01
	Overall	*1812*	*958*	*392*	*.04*
Stimulus?	Advertisement	23	4	0	.14
	Animal	146	112	83	.79
	Art/drawing	157	58	33	.46
	Event/situation	132	27	2	.15
	Food	78	56	36	.74
	Meme	129	58	8	.34
	Object	211	102	51	.50
	Person	260	168	91	.61
	Place	46	12	5	.34
	Screenshot	528	351	195	.53
	Overall	*1710*	*948*	*504*	*.53*

Table 1: Corpus statistics for emotions, relations, and stimuli. "≥ 1", "≥ 2", "= 3" means that at least one, at least two, and all three annotators labeled the post with the respective emotion respectively. The overall number of posts that were annotated in Phase 1 is 1,380, and 1,054 for Phase 2. κ refers to Fleiss' kappa.

tion test had a 55 % acceptance rate. We summarize participation and qualification statistics in Tables 4 and 5 in the Appendix.

Annotators and Payment. Altogether, 75 distinct annotators participated in Phase 1, and 38 annotators worked in Phase 2. We paid $0.02 for each post in Phase 1, and $0.08 for each post in Phase 2. The average time to annotate one post was 16 and 38 seconds in Phase 1 and 2, respectively. This leads to an average overall hourly wage of $7. Overall, we paid $337.44 to annotators and $105.06 for platform fees and taxes.

3.4 Statistics of Annotated Dataset

In total, 1,380 posts were annotated via AMT (we do not discuss the preliminary annotations here). All results are summarized in Table 1.

Did the author want to express an emotion with the post? The total agreement of all three annotators (=3) was achieved in 47 % of the time (652 posts out of 1380). The overall inter-annotator agreement for this question is fair, with Fleiss κ=.3. We consider this value to be acceptable for a pre-filtering step to remove clearly non-emotional posts for the actual annotation in the next phase.

Of the 1,380 posts in Phase 1, 1,061 were labeled as "emotion", of which seven were flagged as being problematic by annotators (see Figure 7 in Appendix). Therefore, in total, 1,054 posts are considered for Phase 2.

What emotion did the author likely feel when writing this post? Table 1 gives the individual counts of instances that received a particular emotion label by at least one, two, or all three annotators. Note that the overall number of instances can be greater than the number of instances in the case that annotators disagree. *Joy*, *surprise* and *disgust* are the more frequent classes, with 585, 435, and 268 posts that received this label by at least one annotator. The number of posts in which at least two annotators agreed is considerably higher for *joy* than for the other emotions, which is also reflected in the moderate overall inter-annotator agreement with Fleiss κ=.47. For most classes, the agreement is moderate, with some exceptions (*anger* is often conflated with *disgust* as we will see below, and *anticipation*, and *trust*).

The agreement, however, can be considered to be similar to what has been achieved in other (crowdsourcing-based) annotation studies. As examples, Purver and Battersby (2012) report an agreement accuracy of 47 %. Schuff et al. (2017) report an agreement of less than 10 % when a set of 6 annotators needed to label an instance with the same emotion (but higher agreements for subsets of annotators).

What is the relation between the image and the text regarding emotion communication? The most dominant relations in our dataset are *complementary* (1,042 instances in which one annotator decided for this label) and *illustrative* (476). There are fewer instances in which annotators marked the relation *opposite* (28), *decorative* (124) and that the text is not required to infer the emotion (142).

The inter-annotator agreement is low, due to the skewness of the dataset and a therefore high expected agreement: overall, we only achieve κ=.04. Note that this inbalanced corpus poses a challenge in the results described in Section 5.

What is it in the image that triggers the emotion? The emotion stimuli categories are more balanced: Most frequently, people comment on

what we classify as screenshots (528 out of 1054 received this label by at least one annotator), followed by depictions of people (260), objects (211), pieces of art (157), and depictions of animals (146). The agreement is moderate with an overall κ=.53. The labels *place* and *advertisement* are underrepresented in the dataset.

Cooccurrences. We now turn to the question which of the variables of the emotion category, the relation, and the stimulus category cooccur. Figure 3 shows the results with absolute counts above the diagonal, and odds-ratio values for the cooccurrence of multiple emotions annotated by different annotators below the diagonal (details regarding the calculation can be found in Schuff et al., 2017). The emotion combinations of *joy–surprise* (150 times), *surprise–disgust* (126), *surprise–anger* (63), and *disgust–anger* (62 times) are most often used. This is presumably an effect of the fact that people share information on social media that they find newsworthy. Further, this shows the role of *surprise* in combination with both positive and negative emotions—as common in emotion annotations to limit ambiguity, we modelled the task in a single-label annotation setup. Therefore, this shows that different interpretations of the same post are possible.

The odds-ratio values point out the specificity of the combination of *disgust–anger*. This could be explained with the difference of these emotions regarding their motivational component, namely to tackle a particular stimulus or to avoid it (known as the fight-or-flight response). The combination of *sadness–fear* can be explained with the importance of the confirmation status of a stimulus (future or past) which distinguishes these two emotions. This property might be ambiguous in depictions in social media. The combinations of *fear–anticipation* and *fear–trust* might be considered surprising. Such combinations of positive and negative emotions frequently occur in motivational text depictions, for instance "don't be afraid of your fears".

We show the cooccurrence counts and odds ratios for the stimulus and the emotion in Figure 4. For the emotions *anger*, the stimuli of *advertisments* and *screenshots* are outstanding. *Anticipation* has the highest value for art. *Disgust* is particularly specific for *food* and *advertisement*. This shows the metaphoric use of the term (in the sense of repugnance) and a more concrete use (in the sense of revulsion). Interestingly, *fear* is spe-
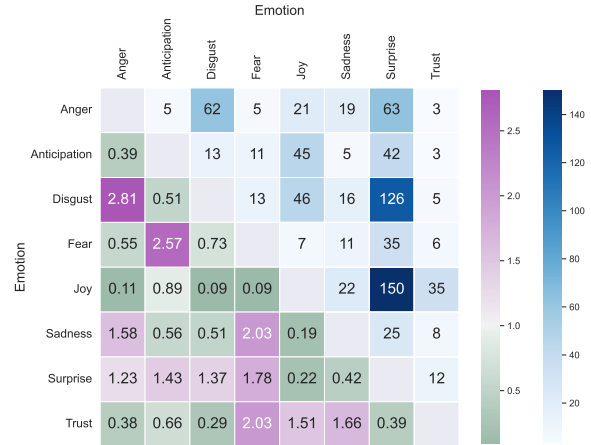

Figure 3: Emotion-Emotion cooccurrences. The values above the diagonal are absolute counts, while the numbers below the diagonal are odds ratios. I higher value denotes that the combination is particular specific.

cific for stimuli of *animals*, *art*, and *memes*. *Joy* is the only emotion that has a high odds ratio with *places*, and *persons*, but also with *animals*. *Sadness* and *trust* have the highest value for *memes*.

We do not discuss the *relation* category further, given the predominance of the *complementary* class and its limited inter-annotator agreement.

4 Methods

In the following, we present the models that we used to predict (1) each variable (emotion, stimulus, relation) separately in each modality, and (2) across modalities with joint models.

4.1 Text

For the text-based model, we fine-tune the pre-trained RoBERTa model[5] (Liu et al., 2019). We perform multi-task learning for emotion, stimulus and relation by adding a fully connected layer (for each set of labels), on top of the last hidden layer. The model combines the loss for all three sets of labels and updates the weights accordingly during the training phase.[6] We use a learning rate of $3 \cdot 10^{-5}$ for all layers, except for the top three fully connected ones ($3 \cdot 10^{-3}$). We use the learning rate scheduler with a step size of 5 and train for maximally 20 epochs, but perform early stopping if the validation loss does not improve by more than 0.005%.

---

[5]https://huggingface.co/transformers/model_doc/roberta.html
[6]Our first choice of only one layer performed en par to multiple stacked layers.
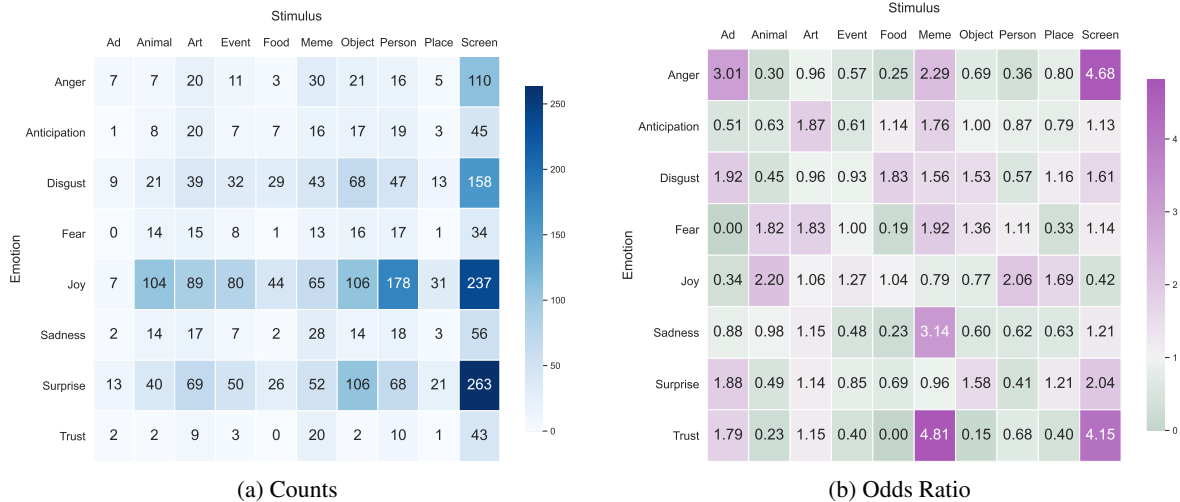
(a) Counts

(b) Odds Ratio

Figure 4: Emotion-Stimulus Cooccurrences.

4.2 Images

We built the image-based model on top of a pre-trained deep residual network model with 48 convolutional layers and 2 pooling layers (ResNet, He et al., 2016). We use the ResNet50 that is provided by PyTorch[7] and was pretrained on 1,000 ImageNet categories (Russakovsky et al., 2015; Deng et al., 2009). As with the text-based model, we add three fully connected layers on top of the fully connected layer of the ResNet50 model, with the sigmoid activation function. Unlike RoBERTa, we do not fine-tune the convolutional layers to prevent the pre-trained weights to change.[8]

4.3 Joint Models

We evaluate three simple multimodal methods which combine the information from the text and the image modality on the traditional three different stages: early, late, and model-based fusion (Snoek et al., 2005).

In early (feature-based) fusion, the features extracted from both modalities are fused at an early stage and passed through a classifier. As the input, our early-fusion model takes the tokenized text and preprocessed image (images are resized, converted to tensors, and normalized by the mean and standard deviation[9]), and concatenates them into one vector to pass through the final classifier, that consists of several layers (three linear, dropout, and three fully connected layers) with the input size

depending on the longest text in the training set and output size depending on the task. The activation function is, as in all our models, a sigmoid function.

In late (decision-based) fusion, classification scores are obtained for each modality separately. These scores are then fed into the joint model. In our late-fusion model, we pass the text and image through the text-based and image-based models respectively, and concatenate the output probabilities of these models.[10] We then pass this vector through a fully connected layer with twice the number of classes from the two models as input and output, and apply sigmoid for prediction. That is, for the emotion classification, the vectors of eight labels from RoBERTa and ResNet50, summing up to 16, are passed to the fully connected layer.

For model-based fusion, we extract text and image features from our unimodal text and image-based classifiers, respectively (from the last hidden layers before the fully conntected ones), and feed these to a final classifier.[11]

5 Results

We evaluate our models on predicting emotions, text–image relations, and emotion stimuli using unimodal and multimodal models, based on the F_1 measure. We use the dataset of 1054 instances in which we aggregate the labels from the three annotators by accepting a label if one annotator assigned

---

[7]https://pytorch.org/hub/pytorch_vision_resnet/
[8]We performed experiments with unfreezing several top convolutional layers, however, it did not lead to better results.
[9]https://pytorch.org/vision/stable/transforms.html

[10]Experiments with summed vectors did not improve results.
[11]Experiments with more complex models with multiple top layers did not improve results, thus, we chose a single-layer-on-top model for the experiments.

	Model	Emo.	Rel.	Stim.
	Majority Baseline	.22	.56	.21
uni-modal	Text	**.53**	**.77**	.45
	Image	.41	.67	.59
multi-modal	Early fusion	.40	.72	.33
	Late fusion	.47	.72	.41
	Model-based fusion	**.53**	.76	**.63**

Table 2: Experimental results in predicting emotions, relations, and stimuli using unimodal and multimodal models. The results are presented in weighted F_1 score. Bold face indicates the highest value in each column/task.

it (this approach might be considered a "high-recall" aggregation of the labels, similar to Schuff et al. (2017)). Despite being a single-label annotation task, this leads to a multi-label classification setup. In other words, the annotation process requires annotators to select a single label (for each set of labels), e.g. one emotion per post; however, the experiments are conducted using multiple labels per set, depending on how many labels are given by three annotators for each set of labels. The data is randomly split into 853 instances for training, 95 instances for validation, and 106 test instances.

Table 2 summarizes the results, averaging across the values for each class variable. We observe that the emotions and the relations can be predicted with the highest F_1 with the text-based unimodal model. The discrepancy to the image-based model is substantial, with .53 to .41 for the emotions and .77 to .67 for the relations. The stimulus detection benefits from the multimodal information from both the image and the text—the highest performance, .63, is achieved with the model-based fusion approach. From the unimodal models, the image-based model is performing better than the text-based model. This is not surprising—in multimodal social media posts that express an emotion, the depictions predominantly correspond to a stimulus, or their identification is at least important. The corpus statistics show that: posts in which the image is purely used decoratively are the minority.

Table 3 shows detailed per-label results. For the **emotion classification** task, we see that for three emotions, the text-only model leads to the best performance (*disgust*, *joy*, *trust*, while the latter is too low to draw a conclusion regarding the importance of the modalities). The other emotions benefit from a multimodal approach. Overall, still, the text-

	Label	Unimodal		Multimodal		
		Txt	Img	Early	Late	Mb.
Emotions	Anger	.08	.04	.03	0	**.14**
	Anticipation	0	0	**.12**	0	0
	Disgust	**.47**	.23	.26	.27	.39
	Fear	0	0	**.04**	0	0
	Joy	**.84**	.66	.64	.85	.78
	Sadness	.28	0	.09	0	**.37**
	Surprise	.61	.57	.52	.57	**.70**
	Trust	**.04**	0	0	0	0
Relations	Compl.	**.99**	**.99**	.98	**.99**	**.99**
	Decorative	.05	0	.11	0	**.20**
	Illustrative	.65	.46	.50	**.66**	.61
	Image-only	**.38**	0	.24	0	.34
	Opposite	0	0	0	0	0
Stimuli	Advert.	0	0	0	0	0
	Animal	.60	**.78**	.28	.52	.74
	Art/drawing	.02	.27	.26	0	**.45**
	Event/sit.	.20	.16	.13	0	**.29**
	Food	.22	**.63**	.19	0	.62
	Meme	.40	.35	.17	0	**.57**
	Object	.32	**.68**	.17	.25	.65
	Person	.43	**.59**	.21	.52	**.59**
	Place	0	.06	.04	0	.19
	Screenshot	.72	.78	.63	.78	**.79**

Table 3: Experimental results for all labels in predicting emotions, relations, and stimuli using the text-based and image-based unimodal models, and fusion models. The results are presented in F1 score.

based model shows highest average performance, given the dominance of the emotion *joy*.

For most **stimulus** categories, either the image or a multimodal model performs best. This is not surprising, given that the stimulus is often depicted in the visual part of a multimodal post. More complex depictions that could receive various evaluations, like *art*, *events/situations*, and *memes* require multimodal information. In those, the image information alone is not sufficient—the performance difference is between 22pp and 13pp in F_1. For those stimuli, in which the text-based model outperforms the multimodal models, the difference is lower. The text-based model is never performing best, but shows acceptable performance for *animals*, *memes*, *screenshots* and *person* depictions.

Regarding the **relations**, the *complementary* class is predicted with the best performance; which is due to the frequency of this class. The label *decorative* can only be predicted with a (slightly) acceptable performance with the multimodal approach, while *illustrative* predictions based on text-only are nearly en par with a multimodal model.

From the three multimodal fusion approaches,

early fusion performs the worst, followed by late fusion. Model-based fusion most often leads to the best result. We show examples for instances in which the multimodal model performs better than unimodal models in Table 6 in the Appendix.

6 Conclusion and Future Work

With this paper, we presented the first study on how users in social media make use of text and images to communicate their emotions. We have seen that the number of multimodal posts in which the image does not contribute additional information over the text is in the minority, and, hence, interpretation of images in addition to the text is important. While the inter-annotator agreement for relation was not reliable enough to draw this conclusion, prediction of stimulus correlates with prediction of emotion due to the information that is present in the image but missing in the text, and thus makes images play a significant role in analysis of social media posts. This is also the first study on stimulus detection in multimodal posts, and we have seen that for the majority of stimulus categories, the information in the text is not sufficient.

In contrast to most work on emotion stimulus and cause detection in NLP, we treated this task as a discrete classification task, similar to early work in targeted sentiment analysis. An interesting step in the future will be to join segment-based open domain stimulus detection, as it is common in text analysis, with region-based image analysis, and ground the textual references in the image. This will allow to go beyond predefined categories.

Acknowledgements

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Saima Aman and Stan Szpakowicz. 2008. Using Roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I.*

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020a. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020b. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Margaret M Bradley, Mark K Greenwald, Margaret C Petry, and Peter J Lang. 1992. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition*, 18(2):379.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Niko Colnerič and Janez Demšar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*, 11(3):433–446.

Liyanage C. De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat.*, volume 1, pages 397–401 vol.1.

Michela Dellagiacoma, Pamela Zontone, Giulia Boato, and Liliana Albertazzi. 2011. Emotion based classification of natural images. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, DETECT '11, page 17–22, New York, NY, USA. Association for Computing Machinery.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. 2018. Emotional attention: A study of image sentiment and visual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531.

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Gui Lin, Yulan He, Qin Lu, and Kam-Fai Wong. 2017. Overview of NTCIR-13 ECA task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pages 361–366, Tokyo, Japan.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A question answering approach for emotion cause extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602, Copenhagen, Denmark. Association for Computational Linguistics.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yuwei He and Guiguang Ding. 2019. Deep transfer learning for image emotion analysis: Reducing marginal and joint distribution discrepancies together. *Neural Processing Letters*, 3:2077–2086.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? Annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359. Association for Computational Linguistics.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, page 125–132, New York, NY, USA. Association for Computing Machinery.

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. *arXiv preprint arXiv:1907.11692*.

Xin Lu, Poonam Suryanarayan, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. 2012. On shape and the computability of emotions. In *ACM International Multimedia Conference*, volume 2012, pages 229–238.

Emily E Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6):647–672.

Radan Martinec and Andrew Salway. 2005. A system for image–text relations in new (and old) media. *Visual communication*, 4(3):337–371.

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144.

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.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 2014. Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Baltimore, Maryland. Association for Computational Linguistics.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. *Recognition of Fine-Grained Emotions from Text: An Approach Based on the Compositionality Principle*, pages 179–207. Springer Berlin Heidelberg, Berlin, Heidelberg.

Laura Ana Maria Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70, Barcelona, Spain (Online). Association for Computational Linguistics.

Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. 2016. Where do emotions come from? Predicting the emotion stimuli map. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 614–618.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491, Avignon, France. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

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

Klaus R. Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Mostafa Al Masum Shaikh, Helmut Prendinger, and Mitsuru Ishizuka. 2009. A linguistic interpretation of the occ emotion model for affect sensing from text. *Affective Information Processing*, pages 45–73.

Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402.

Shuangyong Song and Yao Meng. 2015. Detecting concept-level emotion cause in microblogging. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 119–120, New York, NY, USA. Association for Computing Machinery.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of Twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy. Association for Computational Linguistics.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing Twitter "Big Data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Lydia Weiland, Ioana Hulpuş, Simone Paolo Ponzetto, Wolfgang Effelsberg, and Laura Dietz. 2018. Knowledge-rich image gist understanding beyond literal meaning. *Data & Knowledge Engineering*, 117:114–132.

Lifang Wu, Mingchao Qi, Meng Jian, and Heng Zhang. 2020. Visual sentiment analysis by combining global and local information. *Neural Process. Lett.*, 51(3):2063–2075.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. 2018. Weakly supervised coupled networks for visual sentiment analysis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7584–7592.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference*.

Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W Schuller, and Kurt Keutzer. 2018. Affective image content analysis: a comprehensive survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Survey track*, pages 5534–5541.

Honglin Zheng, Tianlang Chen, Quanzeng You, and Jiebo Luo. 2017. When saliency meets sentiment: Understanding how image content invokes emotion and sentiment. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 630–634.

A Appendix

(a) Person/people | (b) Animal | (c) Object | (d) Food | (e) Meme

(f) Screenshot/text in image | (g) Art/drawing | (h) Advertisement | (i) Event/situation | (j) Place
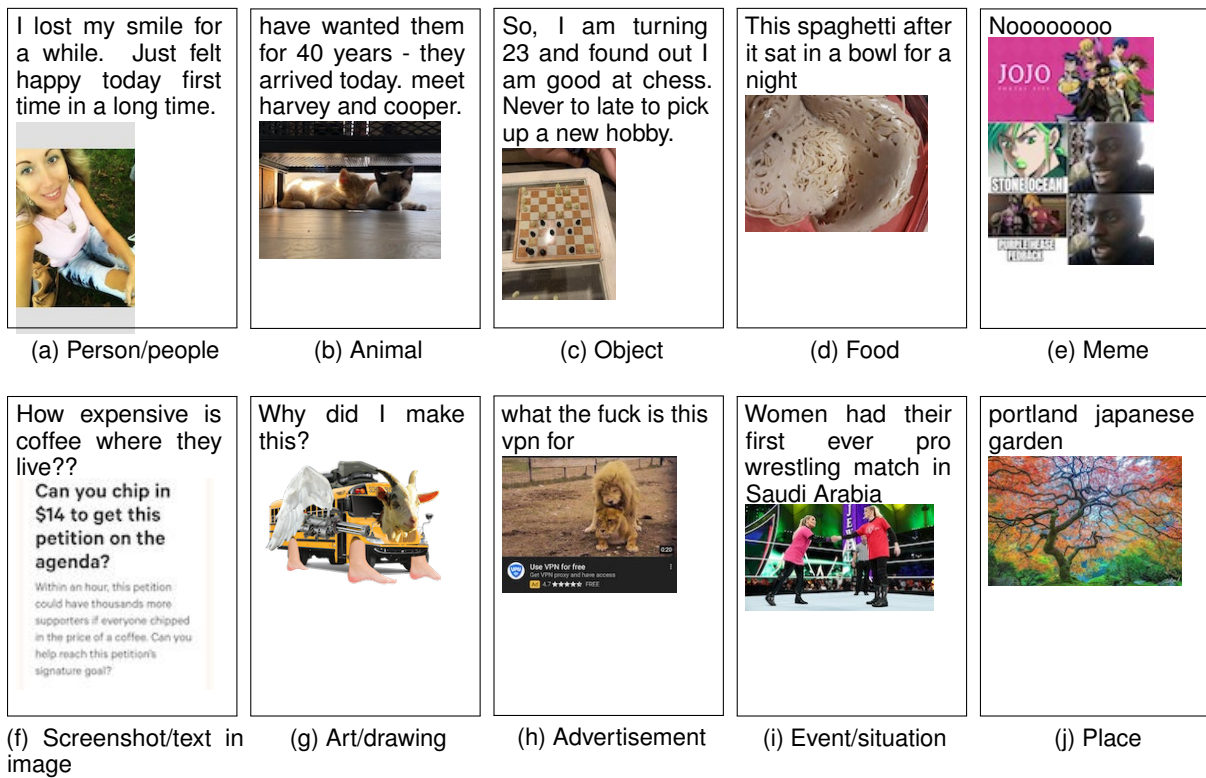
Figure 5: Examples of emotion stimuli in post images.
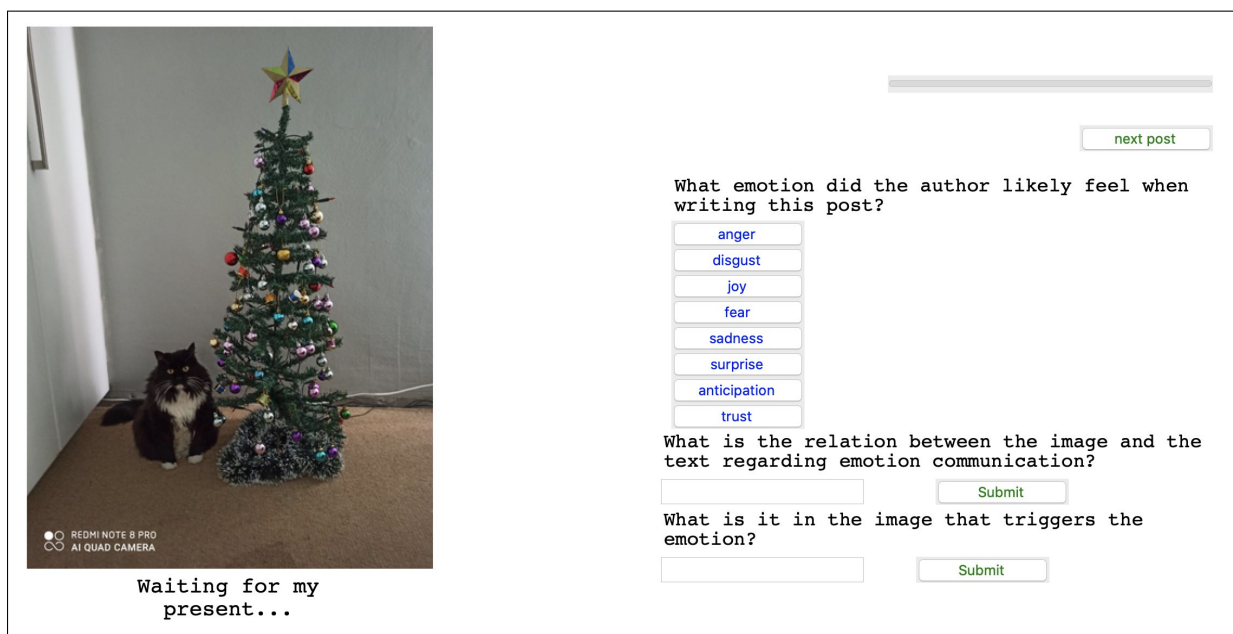
Figure 6: Annotation tool to define taxonomies.

Figure 7: Annotation Environment on Amazon Mechanical Turk

Qualification	Description
Qualification test 1: emotional/non-emotional posts	5 posts presented to annotators to label the post emotional or non-emotional; passing score of 80%
Qualification test 2: emotion, relation, stimulus identification	5 posts presented to annotators to label the post for emotions, relations, and stimuli; passing score of 80%
Region	Annotators must reside in either of the six English-speaking countries (Australia, Canada, Ireland, New Zealand, United Kingdom, United States) to force the task to be done by native speakers.
Human Intelligence Task (HIT) approval rate	The HIT approval rate represents the proportion of completed tasks that are approved by Requesters and ensures the quality of the job workers do on the platform.

Table 4: Qualifications used on AMT for data annotation.

	Qualification		Participation	
	Attempted	Passed	From previous task	New
Task 1	75	75	-	75
Task 2	69	38	17	21

Table 5: Statistics on participation for the two tasks. All numbers are the counts of workers. Qualification tests are described in Table 4. *Attempted* are the number of workers that took the qualification test, while *passed* is the number of workers that answered at least 80% of the questions correctly. *From previous task* refers to the number of workers that participated in Phase 1 as well as Phase 2, while *new* are the participants that have not participated in the previous phase.
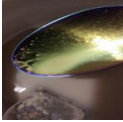
	Text	Image	Gold	Predictions (Emotions/Stimulus)		
				Image-only	Text-Only	Multimodal
Emotion	Found a fly in my tea halfway through it		Disgust	Joy	Disgust/Surprise	Disgust
	Dont know if it has been posted before but here u go		Joy	Joy	Disgust	Joy
	I find a monster under my bed		Sadness	Joy	Fear/Surprise	Surprise
	Definitely stoked with how much weight I've lost since overcoming my alcoholism!		Joy	Fear	Joy	Joy
Stimulus	It causes unnatural amounts of pain to just look at it		Art/Drawing	Art/Drawing	Person	Art/Drawing
	I have no idea the context of this picture from Steam Powered Giraffe but the sheer happiness in it makes me happy. Hope it does for you, too! You see, he never smiles that big!		Person	—	Person	Person
	After multiple tries, my sunflower finally bloomed! What a beauty.		Object	—	—	Object

Table 6: Examples in which the multimodal model-based model returns the correct result, but at least one unimodal model does not. "—" means that the model was not confident enough to predict any of the labels from the set.