Intentonomy: a Dataset and Study towards Human Intent Understanding Supplementary Material

Menglin Jia[1,2] Zuxuan Wu[2,3] Austin Reiter[2] Claire Cardie[1] Serge Belongie[1] Ser-Nam Lim[2]
[1]Cornell University [2]Facebook AI [3]Fudan University

Appendix

The aim of our work is to investigate the complex psycho-emotional landscape hidden behind social media posts, and to lay the groundwork for the research in this domain. Such research can foster the development of systems to identify harmful posts and to reduce social media abuse and misinformation. In our work we proposed to explore human intent understanding by introducing a new image dataset along with a new annotation process. We conduct an extensive analysis on the relationship between *content* and *intent*. We also presented a framework with two complementary modules for the task. In the supplemental material, we provide the following items that shed further insight on these contributions:

- Details for reproduce our results (A);
- An extended discussion of hashtag experiments (B);
- Information about data collection process (C);
- Intentonomy data analysis (D);
- A datasheet for our motive taxonomy (E);
- Additional related work (F) and other questions regarding our work (G).

A. Experiment Details

A.1. Experimental setup

Training details To extract visual information, we use a ResNet50 [19] model which is pretrained on ImageNet [8]) as the backbone of our framework. We use Pytorch [35] to implement and train all the models on a single NVIDIA V100 GPU. We adopt standard image augmentation strategy during the training(randomly resize crop to 224×224, horizontal flip). We use stochastic gradient descent with 0.9 momentum with batch size as 128. The learning rate is warmed up linearly from 0 to base learning rate ($1\mathrm{e}{-3}$ for image only models, $5\mathrm{e}{-4}$ for the rest) during the first five epochs. Since the dataset is not balanced, we follow [30, 7] to stabilize the training processing by initializing the the bias for the last linear classification layer with $b = -\log\left((1-\pi)/\pi\right)$, where the prior probability π is set to 0.01.

Localization loss For the $\mathcal{L}_{loc}$, we conducted grid search for λ with the range $\{0.5, 0.1, 0.01, 0.001\}$. We set $\lambda = 0.1$ in the end, which is also consistent with the parameter used in previous work [40].

Hashtags To obtain hashtags, we index the Unsplash photos using KNN [23], and retrieved a total of 661,505 Instagram images with associated hashtags. We experiment with a range of k for the nearest neighbor search: further details are shown in Sec. B and D. We also compare four different word embeddings [4, 37, 6, 11], which all utilize wiki data for pretraining. The hashtag features are followed by a 2 layer MLP $[1024, 2048]$, with a ReLU activation using a dropout of 0.25, before concatenated with image feature.

Intent *vs.* content study To obtain $Mask^{\mathcal{O}}(I)$, we use a pretrained mask-RCNN (X101 32x8d FPN 3x) model[1] [18] trained on COCO dataset [31] to obtain objects' segmentation masks with a threshold of 0.6. Multiple objects are merged together. $Mask^{\mathcal{C}}(I)$ is defined as the pixel area in an image I that does not belong to $Mask^{\mathcal{O}}(I)$. A ResNet50 [19] model, pretrained on ImageNet [8] and finetuned on each variation of the dataset. All images are resized to longest side of 1280 before processing.

To analyze the relations between content disruption levels and intent recognition scores, we fit a line $\alpha \boldsymbol{X} + \beta = \boldsymbol{y}_{F1}$, and define the correlation $\rho(\boldsymbol{X}, \boldsymbol{y}_{F1}) \in \{\text{positive}, \text{neutral}, \text{negative}\}$ based on the normalized slope values ($\bar{\alpha} = \alpha / |\boldsymbol{X}| \times 10$). The value of $\bar{\alpha}$ and $\rho(\boldsymbol{X}, \boldsymbol{y}_{F1})$ are used to group intent classes as described in Sec. A.2.

To investigate the relationship between intent and specific thing and stuff classes, we use a pretrained panoptic FPN segmentation model [27] trained on COCO panoptic dataset and obtain masks for both thing and stuff classes

[1]detectron2 model zoo https://github.com/facebookresearch/detectron2/blob/master/MODEL_ZOO.md

	Classes	Frequency	Definition
Content	$\mathcal{O}$-classes	7| 24.9%	$\bar{\alpha}_{\mathcal{O}} > \bar{\alpha}_{\mathcal{C}}, \rho_{\mathcal{C}} \neq$ positive
	$\mathcal{C}$-classes	2| 11.1%	$\bar{\alpha}_{\mathcal{O}} < \bar{\alpha}_{\mathcal{C}}, \rho_{\mathcal{O}} \neq$ positive
	Others	19 | 64.0%	o.w.
Difficulty	Easy	3|23.8%	$D \leq 5$
	Medium	15| 51.6%	$D \in (5, 15]$
	Hard	10| 24.6%	$D > 15$

Table 1. Intent classes categorization. We propose to group 28 classes based on two criteria and report the definition, frequency (in the forms of [number of classes | training image percentage]). See text for definition of D.

Method		Macro F1	
WordBreak	Embeddings	All	Hard
	fastText [4]	19.92 $\pm$ 0.86	6.47 $\pm$ 0.93
✓	BERT [11]	6.58 $\pm$ 0.13	0.0 $\pm$ 0.0
✓	fastText [4]	20.04 $\pm$ 0.53	6.63 $\pm$ 1.45
✓	GloVe [37]	21.37 $\pm$ 0.19	6.64 $\pm$ 0.83
✓	static BERT [6]	18.97 $\pm$ 0.23	**7.47 $\pm$ 0.86**

Table 2. Model performance with $\boldsymbol{HT}$ feature only on val set. Static BERT with our proposed WordBreak method gives best result.

in the images (with a threshold of $\tau_p = 0.7$, p with area less than 10% of the whole image are ignored). The CAM heatmaps are averaged over all five trained model results with $\tau_{cam} = 0.4$. All images are resized to longest side of 1280 before processing.

A.2. Identifying intent classes

To quantify and analyze the experimental results, we group 28 classes into subsets based on two different criteria, *i.e.*, content and difficulty. Table 1 shows a summary.

By content Intent categories are grouped into object-dependent ($\mathcal{O}$-classes), context-dependent ($\mathcal{C}$-classes), and *Others* which depends on both foreground and background information.

By difficulty Based on random guessing and standard classification results using full content information, we categorize classes based on how far the CNN model achieves than the random results. Formally, given a random guessing score r and model result s for a class m, the information gain is defined as $D(m) = rlog(s/r)$. $D(m)$ takes both the value of r and the relative gain from s to r into considerations. The larger D is, the easier the class m is for a standard CNN model to learn.

B. Additional Hashtags Results

Separating hashtags benefits *hard* classes In Table. 2, we report performance using $\boldsymbol{HT}$ only and compare different hashtag representation methods. Hashtags, despite not constituting a natural language, are compact by definition. We observe that separating hashtags into phrases outperforms subword-level embedding for the whole hashtag. FastText embedding [4] utilizes sub-word information and usually works well with rare words. Yet separating hashtags is able to achieve a 15.5% gain on *hard* classes, and 7% on overall macro F1 score. We use static-BERT [6] for all the other experiments since improving hard classes is the reason why we propose using multiple modalities. Note that BERT [11] yields results comparable to random guessing. A possible reason is that the average token length for a hashtag is 4.7 (std = 3.5), which suggests a low level of contextual information within any given hashtag.

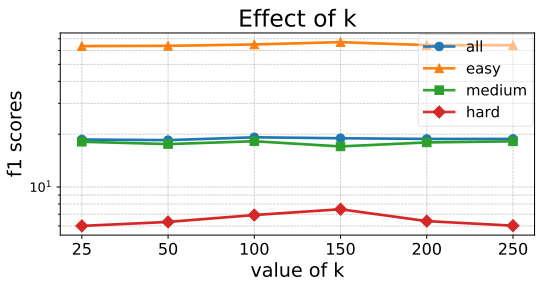
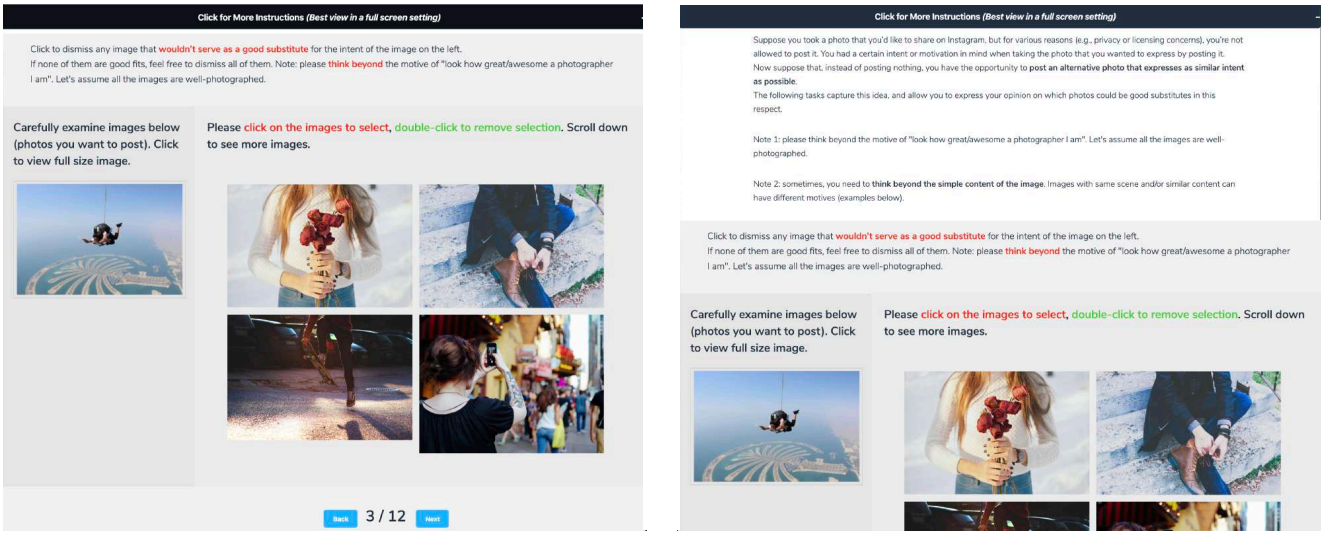


Figure 1. Effect of k for $\boldsymbol{HT}$ features on val set. In general, F1 score peaks at $k = 150$, $k \in \{25, 50, 100, 150, 200, 250\}$. Y-axis is in log scale.

Hashtags from k nearest neighbours How does the noise in collected Instagram hashtags impact classification results? We collect hashtags by fetching pixel-level similar Instagram posts using KNN. Thus the collected hashtags are less and less relevant to the image, as k increases. As pointed out by [32, 33, 21] and mentioned above, hashtags are prone to noise: one may include irrelevant hashtags for the post (*e.g.* `#likesforlikes`, `#igers`). We study the performance of resulting hashtag features by varying the number of top nearest neighbors for each sample i. Fig. 1 shows that F1 score for "hard" and "easy" classes peak at $k = 150$. "Medium" classes are less sensitive to the value of k and peak at 100. We use $k = 150$ for all the other experiments.

C. Dataset Creation Details

Given the inherent abstract nature of intent understanding, one challenge we are facing is that how to collect reasonable labels in an effective manner. A standard annotation process for image classification task is to ask qualified annotators to select from a list of labels given one image. Annotators become qualified after a series training sessions for the label information [22].This approach would have been time-consuming and highly dependent on the expertise of our annotators. We instead adopt a *game with a purpose* approach to keep annotators engaged and let them focus on the "swapabilities" of image pairs regarding the intent. We



(a) (b)

Figure 2. Annotation interface. We present a story to the workers to put them into the mindset of the imagined user who want to post the image presented. (a) Main annotation page, with probe image and 2×2 image grid displayed side-by-side. (b) Collapsible instruction on the top of the interface.

use relative similarity comparison in batch using grid format following [48]. The annotation task is to select all the images in the gird that clearly have a different intent than the reference image on the left. Note that the resulting labels represent the *perceived* intent: the viewer's opinion of the intent of the image. This section provide more details on the dataset acquisition process.

C.1. Annotation interface

As noted in [43], *games with a purpose* annotation approach, like the ESP Game [45], reCAPTCHA [46] and BubbleBank [9], require some artistry to design tools that keep user engaged. Keeping this principle in mind, we design an interface [2] that displays a probe image and a 2×2 images grid side by side. Amazon Mechanical Turk workers are asked to select all the images in the gird that clearly have a different motive than the reference image on the left. A welcome splash page is shown at the beginning of each annotation task, to briefly introduce or remind the annotators.

Fig. 2 shows the main annotation interface. There is a collapsible section on top of the interface that display instructions. Images inside the grid are sorted dynamically according to the height-to-width ratios, so the interface looks nicer (inspired by [3]). The probe image on the left is always kept shown on the screen throughout scrolling up and down the page.

Since human motives are inherently abstract to understand, we provide a narrative, which is shown below, for the annotators so they could focus on the swapability of images. The narrative presents a story for the workers, which bring them to the scenario of the imagined user who want to post the image presented on the left. We also provided example selections inside the collapsible instructions and the welcome splash page (see Fig. 2(b)).

> Annotation narratives: Suppose you took a photo that you'd like to share on Instagram, but for various reasons (e.g., privacy or licensing concerns), you're not allowed to post it. You had a certain intent or motivation in mind when taking the photo that you wanted to express by posting it. Now suppose that, instead of posting nothing, you have the opportunity to *post an alternative photo that expresses as similar intent as possible*. The following tasks capture this idea, and allow you to express your opinion on which photos could be good substitutes in this respect.

We used 4 images per grid, 12 grids per HITs, including 1 catch trials. We only use annotation results that pass the catch trials. In order to get a richer similarity representations, and to examine the quality of the annotators, we also use 3 annotators for the same HIT.

Annotators' feedbacks of our interface and general annotation system include: "I've been enjoying doing these hits", "I enjoy these tasks, so I would like to keep doing them", "I truly enjoy these hits and always appreciate the feedback!" "I hope to see more from you guys soon i love doing these!" "Thanks for Your HITs, i really enjoyed working on them and i hope i did good." "I enjoy the HITs and am glad to be able to contribute."

[2] Interface is modified based on simpleamt, which use Jinja2 as backend. UI design was adapted from *Snapshot by TEMPLATED, templated.co @templatedco.*

Keywords	# Instagram post	# Unsplash # photos sampled
"people"	39,174,751	8,000
"travel"	479,354,358	4,500
"happy"	564,642,361	5,500
"business"	60,129,975	2,000

Table 3. Keywords and hashtags mapping

C.2. Images selection

Candidate images Our goal is to fetch photos from Unsplash, that is similar to images uploaded to social medias like Instagram. All the images are visually and aesthetically pleasing content generated by users. Each photo of Unsplash has a list of associated keywords, produced by an online deep-learning based API. We use these keywords to query photos from Unsplash. Criterias for the chosen keywords are: 1) it should be reasonable and possible to appear in Instagram; 2) it should cover a wide range of scenarios in everyday life. With such requirement in mind, we chose four keywords by browsing Unsplash website and using common sense: "people", "travel / vacation", "happy", and "business", which were selected according to the popular hashtags on Instagram. Table 3 summarizes the keywords and related number of public instagram posts as of 2020/3/20. A total of $20,000$ images were fetched using these four keywords. During annotation process, our annotators found that around $5K$ images do not have any intent labels, so we discard those in the analysis and experiments.

Probe images We carefully chose probe images that cover a reasonably large range of scenes and objects [13], including both cluttered and relative uniform scenes, and diverse range of objects, colors, textures and shapes. In order to reduce possible ambiguity during annotations, the probe image also uniquely represents one human motive only. The probe image are manually inspected by all the authors.

C.3. Annotators management

To ensure quality, we restrict access to MTurks who pass our qualification task. And we constantly check the performance and send feedback to MTurks. After first 100 annotation tasks (HIT) we launched at MTurk, we limit the annotation task to the top annotators

Each annotator needs to take a qualification test in order to get access to our annotation task. The purposes of qualification test are two folds: firstly, to help us to select qualified workers who understand that we are annotating motives; secondly, to help workers get familiar with our designed narratives in the annotation. A total of four questions are presented to the potential annotators. Aside from the requirement of having an Instagram account, we provide three questions that serve as an introductory training and qualification task. Three image triplets (a probe image, and two substitute options) were carefully curated, and each triplet was presented as three images side by side. We specifically selected images that either has similar content but different motives with the probe image, or similar motive but different motives.

Periodically, we check the annotation progress and send messages to workers to inform how many catch trials they failed. We received positive responses from annotators about such feedback system. One annotator commented that "It's always nice when workers receive feedback from requesters on MTurk about the quality of the work being done, and it was reassuring to receive emails (even if they were more-or-less automated) from your team to let me know I was doing well."

C.4. Annotation methods comparison

Instead of selecting from a list motive labels given each image, we adopted image comparison approach, using "unsatisfactory substitutes" and mental imagery. The average annotation time for one annotation task is 20.60 (± 9.65) minutes. Each annotation task contains 48 images. Therefore, the annotators spend 25.75 second per image on average.

To compare two annotation approaches, two authors of this study annotated 57 random sampled images from our dataset using standard image tagging annotation method (82.2 second) per image. Our image comparison method using "unsatisfactory substitutes" requires less annotation time per image.

C.5. Human-in-the-loop

We adopted a hybrid human-in-the-loop strategy to incrementally learn a motive classifier in the annotation process. Starting from a set of randomly selected images, the dataset is enlarged by an iterative process that utilizes a trained classifier to recommend relevant images to annotators. At each iteration and for each motive label, we train a deep learning classifier using 90% of the labeled data. 10% of the held-out data is always added to the test set. The trained classifier is applied to the rest of unlabeled data, and images with a score larger than 0.35 are sent back to annotators for verification. We applied this method until there is no positive image left in the unlabeled set for each label. See Fig. 3 for examples of our dataset images.

C.6. Inter annotator agreement

As explained in the main text, each image was inspected by three annotators. We use Fleiss' kappa score [14] to measure inter annotator agreements per annotation task. The average score is 59.84%, indicating "moderate" agreement [17]. The inter-annotator agreement score demon-

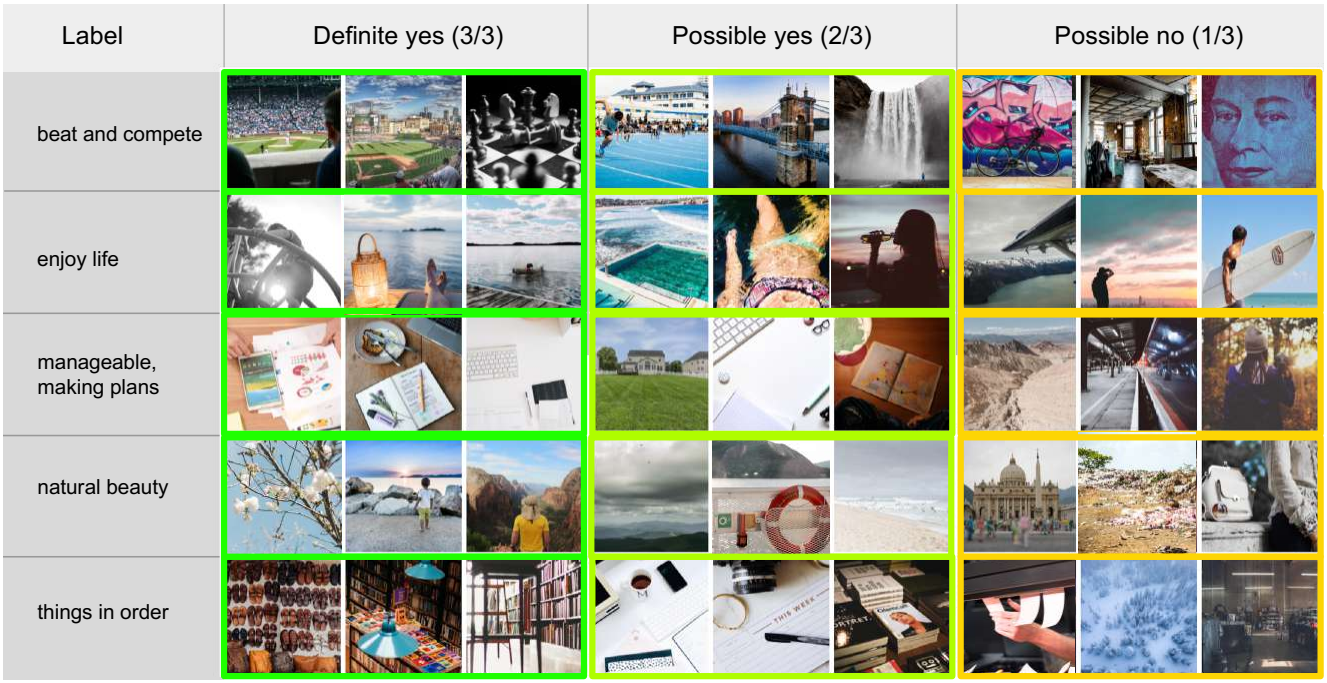
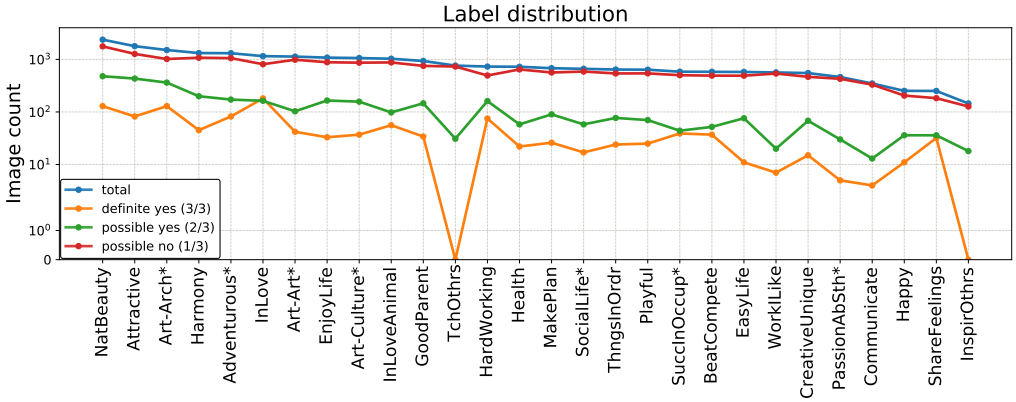
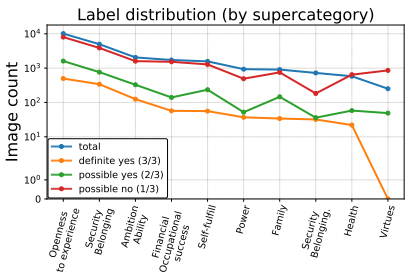


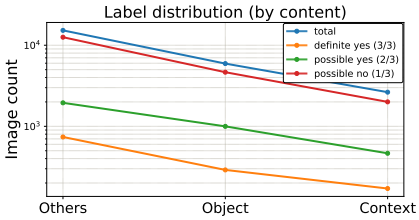
Figure 3. Sample motive labels, and images that are respectively marked as definite yes (3 out of 3 annotators agree), possible yes (2 out of 3 agree), and possible no (1 out of 3 agree). Images that belong to "definite yes" and "possible yes" can have completely different objects, scenes. This further illustrate the high intra-class variance nature of intent classification.



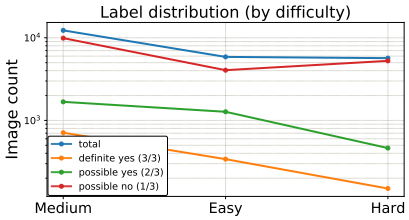
(a) Per-class distribution.



(b) Distribution by supercategories.



(c) By content groups.



(d) By difficulty groups.

Figure 4. Training data distribution. Class names ends with "*" are abbreviated.

Intent classes	Top hashtags
Attractive	#portrait #fashionblogger #womenempowerment #makeup
SocialLife Friendship	#family #sun #sea #beach
NaturalBeauty	#mountains #landscape #sunrise #sunset #naturelovers
Playful	#travel #guitar #lifestyle #puppy #livemusic
Happy	#smile #newbornphotography #mood #headshot #vibes
WorkILike	#entrepreneur #smallbusiness #motivation #marketing

Table 4. Common hashtags for six intent classes.

Dataset	Intent type	# Intent classes	# Images
MDID [29]	Textual	8	1299
Intentonomy	Visual	28	14455

Table 5. Comparison with prior work.

strates the complexity of the annotation task, and the inherently abstract nature of human intent understanding.

C.7. Test set annotation

We ask one author, as chief executive to annotate the validation and test set. The annotation process took three weeks. We found that there are more images per label in the resulting annotation, comparing to the MTurk result. This further demonstrate the MTurks are able to identify correct motive labels using our *game with a purpose* approach. Yet in general MTurks tend to miss some of the labels.

D. Dataset Analysis

In this section, we analyze the properties of the dataset in more detail, and examine the inter-annotator agreement.

Dataset statistics Fig. 4 shows the label distribution of whole training data, over 28 classes, 9 super categories, 3 content classes, and 3 difficulty classes. It shows there is class imbalance in our dataset, which is the property of datasets in the real world [42].

Hashtags We also fetched hashtags from Instagram, with the hope of further capturing the semantics of human intents. In total, we fetched 1,700,915 unique hashtags. Each Unsplash photo has an average of 457.6 ($\pm$317.512) hashtags. As noted in [44], hashtags serve as a medium of self-expression that not limited to objective descriptions of image content. Table 4 lists a collections of top hashtags for selected intent classes

Lexical statistics We fetch the accompanying text description with the images found on the website. These descriptions are generated by a deep-learning based API and verified by human. We report the lexical (word-level) statistic of the dataset. Specifically, the top words occurred in the descriptions of validation images are presented. Table 6 shows top 10 frequent non-stopping words per class, shedding light on the properties of the images. Although the descriptions can be heavily biased, Table 6 illustrates that, as they should, the occurrences of image objects and properties are relatively balanced across all the classes, indicating that most of the frequent words are not necessarily directly predictive of the intent label. However, we do admit that there are exceptions. Certain words can be correlated to certain human intents. For example, "face" occurs frequently in the class "CreativeUnique". "Smiling" is one of the top 10 frequent words in "Playful", "Happy", and "InspirOthrs".

Note that there are 985 "man" and 1714 "woman" in total in the test set, indicating the existence of gender bias in our dataset, which is a common issue in nowadays machine learning systems [5, 16, 10, 15, 38]. "Woman" occurs 74% more than "man". We observe that female gender word tend to associates with classes like "attractive", "happy". "enjoy life". Male gender, on the other hand, associates with "exciting life", "health", "beat and compete". As pointed out in [16], such gender-specific associations, even with subjectively positive words such as the intent labels presented, are *benevolent sexism*. We would like to raise the awareness of such phenomenon. Any machine learning down-streaming tasks should always apply fairness into consideration during algorithm development.

E. Intent Taxonomy

Table 7 lists the detailed taxonomy and explanation for each intent class. Note that there are similarities between emotions and motives. For example, the category "happy/joy" appears in both emotion recognition [36, 28, 20, 47] and intent recognition [24]. Indeed, the common Latin root word of "emotion" and "motivation" is "movere" (to move) [39]. Young [50] argues both emotion and motivation influence human behavior, and that emotion arises from the interplay (*e.g.* conflict, frustration, satisfaction) of motives. Emotions can also be viewed as a reward or punishment for a specific motivated behavior [41].

F. Additional Related Work

Comparison with prior work Table 5 summarizes the differences between Intentonomy and prior work that focuses on social media intent. Other discussions can be found in the main text .

Subjective attributes Recently, there are some progress in building datasets describing the subjective perspective of images [1]. For example, [51] studies visual commonsense

Class	Top words
Attractive	woman (257), wearing (100), white (56), standing (55), black (52), photography (41), man (37), near (36), top (35), holding (33)
BeatCompete	man (48), person (29), woman (25), daytime (24), black (21), white (21), holding (16), photography (14), standing (14), riding (13)
Communicate	woman (29), black (16), sitting (13), photography (13), person (11), brown (10), holding (10), man (10), white (10), two (9)
CreativeUnique	woman (36), man (20), holding (16), person (14), photography (14), white (14), face (12), black (12), green (11), blue (11)
CuriousAdventurousExcitingLife	man (62), person (57), woman (55), daytime (51), standing (42), white (39), photography (36), near (34), wearing (29), gray (27)
EasyLife	woman (51), white (20), person (18), daytime (18), sitting (18), photography (17), man (14), standing (13), near (13), wearing (12)
EnjoyLife	woman (86), daytime (45), standing (35), near (35), person (34), man (33), holding (30), sitting (30), white (29), water (28)
FineDesignLearnArt-Arch	photography (54), white (47), building (46), woman (43), daytime (41), near (39), photo (36), concrete (32), people (30), brown (27)
FineDesignLearnArt-Art	woman (61), person (40), daytime (40), man (39), white (38), black (31), holding (29), photography (29), standing (28), near (28)
FineDesignLearnArt-Culture	woman (89), man (47), wearing (37), standing (37), daytime (29), holding (27), white (26), black (25), near (24), photography (23)
GoodParentEmoCloseChild	woman (71), man (42), wearing (37), daytime (33), white (32), holding (28), black (27), near (26), standing (26), photography (23)
Happy	woman (94), wearing (48), standing (28), man (27), smiling (23), black (20), shirt (17), white (16), brown (14), photography (13)
HardWorking	macbook (14), person (13), book (10), holding (9), woman (8), white (8), man (7), brown (7), using (6), sitting (6)
Harmony	woman (94), standing (62), man (52), person (50), near (47), daytime (43), white (33), sitting (33), photo (30), photography (30)
Health	man (44), woman (32), person (20), daytime (18), people (18), white (17), photography (16), body (15), near (15), water (15)
InLove	woman (97), man (68), wearing (36), standing (35), near (34), daytime (33), white (29), person (28), photography (28), sitting (26)
InLoveAnimal	woman (53), white (45), man (36), standing (33), person (32), photography (28), near (26), black (26), daytime (25), brown (24)
InspirOthrs	man (11), person (9), standing (8), holding (7), woman (7), black (5), stage (5), playing (4), wearing (3), smiling (3)
ManagableMakePlan	white (37), black (28), person (22), near (20), woman (20), brown (15), photo (15), holding (13), book (13), macbook (12)
NatBeauty	woman (107), standing (98), man (96), daytime (76), mountain (72), near (65), person (64), photography (64), water (57), white (56)
PassionAbSmthing	woman (27), wearing (17), man (16), white (15), standing (13), black (13), daytime (12), near (12), photography (12), brown (11)
Playful	woman (69), wearing (29), man (26), black (23), holding (19), white (16), smiling (14), standing (14), near (14), daytime (13)
ShareFeelings	people (16), man (10), person (8), group (8), holding (7), woman (7), black (7), smartphone (5), focus (4), photography (4)
SocialLifeFriendship	woman (46), photography (22), wearing (22), man (20), black (20), person (16), standing (16), people (15), daytime (15), sitting (14)
SuccInOccupHavGdJob	woman (43), man (31), black (24), white (23), wearing (22), standing (18), photo (13), person (13), holding (12), near (11)
TchOthrs	woman (50), man (37), person (23), white (22), black (21), photography (21), near (21), wearing (21), standing (19), daytime (18)
ThngsInOrdr	white (25), woman (23), brown (19), black (19), standing (18), man (18), top (15), person (12), near (12), daytime (12)
WorkILike	woman (49), man (34), person (25), black (21), wearing (20), sitting (18), near (18), holding (18), daytime (18), white (15)

Table 6. Lexical statistics of the image descriptions in the validation set. Top 10 most frequent non-stopping words per class. The numbers next to each word is the count within that specific class

Class	Descriptions
Attractive	Being good looking, attractive.
BeatCompete	Beat people in a competition.
Communicate	To communicate or express myself.
CreativeUnique	Being creative (e.g., artistically, scientifically, intellectually). Being unique or different.
CuriousAdventurousExcitingLife	Exploration - Being curious and adventurous. Having an exciting, stimulating life.
EasyLife	Having an easy and comfortable life.
EnjoyLife	Enjoying life
FineDesignLearnArt-Arch	Appreciating fine design (man-made wonders like architectures)
FineDesignLearnArt-Art	Appreciating fine design (artwork)
FineDesignLearnArt-Culture	Appreciating other cultures
GoodParentEmoCloseChild	Being a good parent (teaching, transmitting values). Being emotionally close to my children.
Happy	Being happy and content. Feeling satisfied with one's life. Feeling good about myself.
HardWorking	Being ambitious, hard-working.
Harmony	Achieving harmony and oneness (with self and the universe).
Health	Being physically active, fit, healthy, *e.g.* maintaining a healthy weight, eating nutritious foods. To be physically able to do my daily/routine activities. Having athletic ability.
InLove	Being in love.
InLoveAnimal	Being in love with animal
InspirOthers	Inspiring others, Influencing, persuading others.
ManagableMakePlan	To keep things manageable. To make plans
NatBeauty	Experiencing natural beauty.
PassionAbSmthing	Being really passionate about something.
Playful	Being playful, carefree, lighthearted.
ShareFeelings	Sharing my feelings with others.
SocialLifeFriendship	Being part of a social group. Having people to do things with. Having close friends, others to rely on. Making friends, drawing others near.
SuccInOccupHavGdJob	Being successful in my occupation. Having a good job.
TeachOthers	Teaching others.
ThngsInOrdr	Keeping things in order (my desk, office, house, etc.).
WorkILike	Having work I really like.

Table 7. The taxonomy for our Intentonomy dataset.

reasoning, requiring computational model to answer challenging questions about an image and provide a rationale justification. Some prior work studies visual rhetoric from different perspectives: 1) protest activity from social media images [49]; 2) memorability [26]; 3) personality [34]; 4) evoked emotions and sentiment [36, 28, 20, 2, 25, 47]. Our work focuses on the *perceived* intent recognition, which is another psychological feature [3].

[3]The definition of "motives" according to Merriam Webster [12] is that something (such as a need or desire) that causes a person to act.

G. Other Concerns and Thoughts

Comparison with human performance A proper human experiment involves careful experimental design accounting for variables including demographic information, life experience, and cultural background. At present, such an effort is out of the scope of our study. We believe, however, that our project provides a starting point for future studies with human subjects.

References

[1] Xavier Alameda-Pineda, Miriam Redi, Nicu Sebe, and Shih-Fu Chang. *Understanding Subjective Attributes of Data*, 2019 (accessed October 07, 2020). 6

[2] Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe. Recognizing emotions from abstract paintings using non-linear matrix completion. In *CVPR*, pages 5240–5248, 2016. 8

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)*, 32(4), 2013. 3

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 1, 2

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016. 6

[6] Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *ACL*, pages 4758–4781, Online, July 2020. Association for Computational Linguistics. 1, 2

[7] Y. Cui, M. Jia, T.Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[9] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, pages 580–587, 2013. 3

[10] Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887, 2019. 6

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NACCL*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 2

[12] Merriam-Webster Dictionary. Merriam-webster. *On-line at http://www. mw. com/home. htm*, 2002. 8

[13] Marin Ferecatu and Donald Geman. A statistical framework for image category search from a mental picture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1087–1101, 2008. 4

[14] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013. 4

[15] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. 6

[16] Peter Glick and Susan T Fiske. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In *Social Cognition*, pages 116–160. Routledge, 2018. 6

[17] Lisa Hartling, Michele Hamm, Andrea Milne, Ben Vandermeer, P Lina Santaguida, Mohammed Ansari, Alexander Tsertsvadze, Susanne Hempel, Paul Shekelle, and Donna M Dryden. *Validity and inter-rater reliability testing of quality assessment instruments*. Agency for Healthcare Research and Quality (US), 2012. 4

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[20] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *CVPR*, pages 1705–1715, 2017. 6, 8

[21] Hamid Izadinia, Bryan C. Russell, Ali Farhadi, Matthew D. Hoffman, and Aaron Hertzmann. Deep classifiers from image tags in the wild. In *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*, MMCommons '15, page 13–18, New York, NY, USA, 2015. Association for Computing Machinery. 2

[22] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *ECCV*, 2020. 2

[23] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 1

[24] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *CVPR*, pages 216–223, 2014. 6

[25] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 159–168, 2015. 8

[26] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *ICCV*, pages 2390–2398, 2015. 8

[27] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019. 1

[28] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *CVPR*, pages 1667–1675, 2017. 6, 8

[29] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *EMNLP*, pages 4614–4624, 2019. 6

[30] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *CVPR*, 2018. 1

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[32] D.K. Mahajan, R.B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 2

[33] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *CVPR*, 2016. 2

[34] Nils Murrugarra-Llerena and Adriana Kovashka. Cross-modality personalization for retrieval. In *CVPR*, pages 6429–6438, 2019. 8

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 1

[36] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, pages 860–868, 2015. 6, 8

[37] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 1, 2

[38] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19, 2019. 6

[39] Sarah M Sincero. *Motivation and Emotion*, 2012 (accessed October 07, 2020). 6

[40] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *CVPR*, 2020. 1

[41] Julian F Thayer and Richard D Lane. A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of affective disorders*, 61(3):201–216, 2000. 6

[42] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, Salt Lake City, UT, 2018. 6

[43] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, Boston, MA, 2015. 3

[44] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *CVPR*, Salt Lake City, UT, 2018. 6

[45] Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006. 3

[46] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008. 3

[47] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning visual emotion representations from web data. In *CVPR*, pages 13106–13115, 2020. 6, 8

[48] Michael Wilber, Sam Kwak, and Serge Belongie. Cost-effective hits for relative similarity comparisons. In *Human Computation and Crowdsourcing (HCOMP)*, Pittsburgh, 2014. 3

[49] Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 786–794, 2017. 8

[50] Paul Thomas Young. *Emotion in man and animal; its nature and relation to attitude and motive.* Wiley, 1943. 6

[51] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 6