Examining Mobs Images Using Vision-Language Models

Ariana Mondiri and Samer Al-khateeb[0000−0001−6327−5720]

Department of Computer Science, Design and Journalism
Creighton University
{ArianaMondiri, SamerAl-khateeb1}@Creighton.edu

Abstract. A mob is a type of event that is usually coordinated through social media, where individuals can gather, engage in an activity with a specific goal, and then disperse. Event-based social media (EBSM) platforms provide a fertile ground for groups to coordinate and organize events online, offline, or both, such as mobs. A mob has a "mob organization process" with five phases, and during the "replaying and republishing" phase, organizers and participants share various content on social media—including text, URLs, images, and videos across multiple platforms. In this research, we examine the images shared on one of the most popular EBSM platforms, Meetup.com. Our goal is to generate accurate captions for these shared images, detect emotions shown in the images, and explore whether these images reveal anything about who posted them. Results show that some images can have a poor caption while others have a very good one. Additionally, using a visual-textual multi-modal approach performed very well in detecting the emotions expressed in the images. Finally, images shared by event organizers seem to have distinct characteristics from those shared by regular participants.

Keywords: CLIP · BLIP · GPT-4o-Mini · Vision-Language Models · Mobs Case Study.

1 Introduction

Mobs are events typically coordinated and organized online through social media platforms [3]. These events bring people together—either online or in person—to carry out a specific action, after which they disperse rapidly (e.g., within 20 minutes [2]) or over an extended period (e.g., lasting days or even months [2]). Such events may be recreational, community-oriented, supportive of a cause, aimed at demanding social or political change, or in some cases, deviant (i.e., harmful and illegal) [2]. Mobs, regardless of type, often follow a sophisticated process referred to as the "mob organization process" [2], which consists of five key phases: planning, recruitment, execution, replaying and republishing, and evaluation. During the *replaying and republishing* phase, various artifacts—including text, URLs, images, and videos—are created and shared online by both the event organizers and participants across multiple social media platforms [2].

In our previous studies [3], we analyzed the textual content shared during these phases. In this research, however, we focus on the visual artifacts—specifically the images—shared on Meetup.com, a widely used "Event-Based Social Network" (EBSN), with the goal of answering the following research questions:

- How can we generate accurate captions for images posted on an event-based social media site?
- How can we detect emotions from images, and how accurately can these emotions be identified?
- What can the sentiments and emotions extracted from images reveal about the user who posted them?

The rest of the paper is organized as follows: in Section 2, we review related work regarding image captioning and sentiments extraction from images; in Section 3, we describe the method used to address the research questions; in Section 4 we explain our analysis; in Section 5, we highlight the results. Finally, in Section 6, we conclude the study and suggest possible future research directions.

2 Related Section

Image Captioning helps encode meaningful and descriptive semantic information in an image. In his survey, Zou et al. [12] identifies several categories of image captioning techniques. Methods include leveraging semantic cues to improve caption quality. Multi-caption approaches generate several captions per image to capture richer information. Cross-modal embedding maps different types of data, like images and text, into a shared embedding space to improve alignment. Finally, few-shot image captioning addresses the challenge of training models with only a small number of labeled examples.

Building on these approaches, recent work has focused on generating meaningful captions for social media content. Bharne and Bhaladhare [4] developed a multimodal system to enhance the authenticity of user profiles using automatic image captioning. They pretrained an image encoder and a text model on a dataset of over $12,000$ real and scam profiles, then fine-tuned the combined model on image–text pairs. They used the generated captions as inputs for several classifiers, achieving F1 scores up to 0.8587 and accuracy up to 0.8373. This demonstrates the potential of using generated captions as semantic signals for downstream social media profiling tasks.

Saini et al. [11] aimed to enhance the generation of dynamic social media captions. They first leveraged CLIP (Contrastive Language-Image Pre-Training) with predefined textual prompts to extract mood and aesthetics, then used BLIP (Bootstrapping Language-Image Pretraining) to iteratively generate descriptive captions through self-supervised learning. The final optimized caption was generated by Qwen2-7B using the labels obtained from CLIP, description generated by BLIP, and original caption from the dataset. They demonstrated the superiority of their method through a user study, where 85.4% of participants preferred

captions generated by their model over a baseline. This highlights the effectiveness of multi-stage caption refinement for producing richer and more descriptive captions.

Sentiments Detection from images has traditionally followed a single model framework with a visual sentiment analysis-based approach [10] [5] or a textual sentiment analysis-based approach [7]. However, relying solely on images results in a lack of contextual information. Thus, visual-textual multi-modal analysis bridges the gap and improves the reliability of the model. Notably, Lyu et al. [6] explored the use of GPT-4V for social multimedia analysis tasks including sentiment analysis. They found that GPT-4V demonstrates strong multimodal reasoning capabilities, including linking images and text, and applying commonsense. GPT-4V not only performs competitively on benchmark sentiment datasets but also extends sentiment detection beyond coarse categories (positive, neutral, negative) by handling nuanced sentiment and even sarcasm [6].

Some of the *challenges* with *image captioning* and *generative AI sentiment detection from images* are that, despite these advances, Nguyen et al. [9] highlight that pre-trained models such as CLIP and BLIP are often trained on large-scale web datasets containing noisy and uninformative captions. To address this, they replaced or augmented the raw text captions in web datasets with synthetic ones and trained CLIP models on the resulting datasets. Their approach led to improved model performance across multiple benchmarks, demonstrating the importance of high-quality captions for effective vision-language training. This suggests that directly using pre-trained CLIP and BLIP models without further adaptation on domain-specific, high-quality data may limit accuracy and reliability. Ensuring higher-quality textual grounding is therefore essential for robust multimodal analysis.

Another important challenge in generative AI is emotional bias. Mehta and Buntain [8] found that generative AI models tend to favor negative emotions, consistently producing outputs skewed towards negative affect across multiple evaluation methods. They used three distinct mechanisms including zero-shot image classification models, deep learning-based image classifiers fine-tuned for emotion recognition, and auto-captioning pipelines. The results revealed a consistent emotional skew, with generative image models disproportionately producing outputs associated with negative emotions.

3 Methodology

3.1 Data Collection

Data has been collected using Meetup.com's GraphQL API. We used a keyword-based search for public groups featuring the topic "Flash Mobs". This search identified 27 public Meetup groups that collectively hosted $3,536$ public events. Of these events, only 210 included images, resulting in $1,375$ photos tagged with 42 different topics. We assigned each image the same topics as its corresponding event. Note that a single event could be tagged with multiple topics. The data

also contains mobbers social factor (traits), summarized in Table 1. Each mobber may be associated with these traits at an intensity ranging from 0 to 1.

3.2 Emotions and Sentiments Extraction

Figure 1 describes our pipeline to generate *image captions* and detect *emotions from images*. We start our data preprocessing by feeding the images, the topic of the image (which is the same as it's corresponding event), to CLIP and BLIP to extract the semantics needed to construct a multimodal input prompt. This prompt was given to GPT-4o-Mini to generate the needed image captions and detect the image sentiments.

4 Analysis

To answer the *first* and *second* research questions, we sought to enhance multimodal sentiment analysis of GPT-4o-Mini with added semantic information extracted from BLIP captions and CLIP prompts both used in a zero shoot manner. To address the challenge of noisy caption without relying on retraining or fine-tuning, we leverage model confidence scores to weight the reliability of BLIP and CLIP outputs. All confidence scores reported range from 0 to 1. From a curated list of prompts we extract the top 2 CLIP prompt matches. Gpt-4o-mini model receives as input an image, the topic name, BLIP caption with associated confidence score, and CLIP top 2 prompts with associated confidence scores.

We set gpt-4o-mini temperature to 0 to obtain deterministic result suitable for further analysis. To avoid bias due to noisy CLIP and BLIP output we explicitly tell the model to consider confidence score during analysis. Additionally, to prevent biased results towards one category of emotions or sentiments, we provide the model with a diverse, predefined list of emotions and sentiments. This ensures comprehensive coverage, avoids overlap, and allows the model to consider all relevant emotional and sentiment categories that might be encountered.

To answer the *third* research question, we aggregate features at the user level by averaging the intensity of each feature, including emotions, sentiments, and scalar traits per user (see Table 1). We then used the Pearson Correlation coefficient to measure the linear correlation between user who are organizers vs. non-organizers and the various user traits, emotions, as well as sentiment categories.

5 Results

Since our images lack ground-truth captions, we relied on confidence scores to perform a quantitative analysis to answer the *first* research question. Table 2 summarizes the CLIP Top-1 and Top-2 concept. Top-1 predictions are generally

Table 1. This table shows the individual (mobber)-related social factors (trait) and their estimation methods. Note: we took the average of all mobber's scores to calculate the event (mob) score, except for the factor with an asterisk (*).

Mobber Trait	**Estimation Method Using Data From Meetup.com**
Utility	The *Utility* of the individual is the average of $money\_score$ + $reward\_score$ of the comments and replies made by each individual (mobber). Then, we took the average of all mobber's scores to get the utility gained by mobbers in the mob.
Interest	The *Interest* of the individual is the number of comments & replies of that mobber divided by the number of users that commented and/or replied in that mob.
Control*	Since we do not have information about the number of connections the mobber has in a mob, we used *Control* = 1/the number of invited mobbers.
Power	The *Power* of a mobber is calculated by multiplying the mobber's *Interest* by the mobber's *Control* [1].
Enthusiasm	The *Enthusiasm* of a mobber is calculated by taking the average of the $tone\_pos$ score + emo_pos score + $loyaltyVirtue$ score of their comments and/or replies on that mob.
Skeptic	The *Skepticism* score of the mobber is calculated by taking the average of the emo_anx score of their comments and/or replies on that mob.
Social Status	The *Social Status* score of a mobber was calculated by taking the average *social* score of their comments and/or replies on that mob.
Anger	The *Anger* score of a mobber was calculated by taking the average emo_anger score of their comments and/or replies on that mob.
Motivation	The *Motivation* score of a mobber was calculated by adding the following scores and taking their average: *allure* score + *curiosity* score + $risk$ score + $reward$ score. These scores were calculated by LIWC of the mobber's comments and/or replies on that mob.
Risk Taking	The *Risk Taking* score of a mobber was calculated by taking the average $risk$ score of their comments and/or replies on that mob.
Lacking Academic Skills	The *Lacking Academic Skills* score of a mobber was calculated by adding the following scores and taking their average: total words count (WC score) + words per sentence (WPS score) + $bigWords$ score + dictionary word count (*dic* score) + *analytic* score + *clout* score + *authentic* score + *tone* score + *linguistic* score + punctuation marks (*allpunc* score). These scores were calculated by LIWC of the mobber's comments and/or replies on that mob. The assumption here is comments and replies written by educated mobbers will have higher scores than uneducated mobbers.
Careless Apathy Indifference	This score was calculated by taking the average $careVirtue$ score of the mobber's comments and/or replies on that mob.
Exposure*	To measure the time the mobber was exposed to the mob, we calculated the time difference (in minutes) between the mob creation time and the mobber's response time (the mobber's RSVP time).

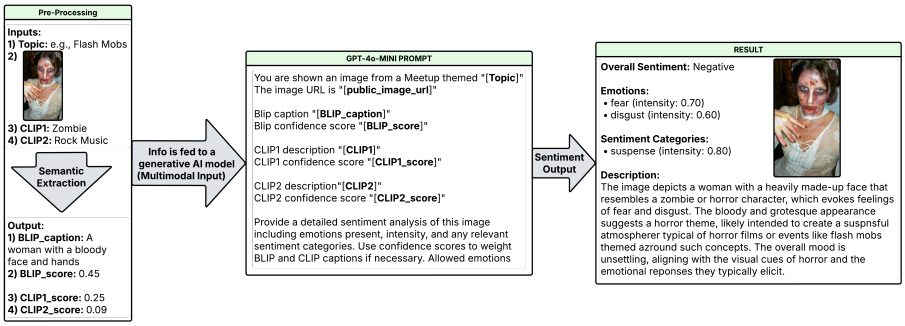

Fig. 1. The method used to generate image captions and extract image sentiments.

more reliable than Top-2. The minimum Top-1 score is 0.040 and the maximum is 1.000, showing that some images are classified with very low confidence while others are nearly certain. For Top-2, scores range from 0.000 to 0.410, confirming that the second concept rarely approaches the confidence of the top prediction. The standard deviation of 0.207 for CLIP Top-1 scores compared to 0.059 for CLIP Top-2 scores highlights the consistently lower and less variable confidence of Top-2 predictions. The BLIP confidence scores, summarized in Table 3, show an average value of 0.505 with a standard deviation of 0.093, indicating moderate variability. The scores range from a minimum of 0.211 to a maximum of 0.758. Most predictions are centered around the mid-confidence range.

Table 2. CLIP Top-1 and Top-2 Concept Score Summary

Statistic	Top-1 Score	Top-2 Score
Count	1375	1375
Mean	0.313	0.110
Standard Deviation	0.207	0.059
Minimum	0.040	0.000
25% Percentile	0.150	0.070
Median (50%)	0.250	0.100
75% Percentile	0.410	0.140
Maximum	1.000	0.410

For the *second* research question, we found that the emotions expressed in the images closely align with the topics of the mobs and the images themselves. For instance, *happiness* was the most prominent emotion in images related to "park games" and "flash mobs," while *calmness* ranked among the top three emotions in mobs with the topic "make new friends". *Curiosity* was the second most common emotion overall, followed by *excitement* (see Figure 2). A similar pattern was observed in the sentiment categories: *fun* ranked the highest, fol-

Table 3. BLIP Confidence Score Summary

Statistic	Value
Count	1375
Mean	0.505
Standard Deviation	0.093
Minimum	0.211
25% Percentile	0.439
Median (50%)	0.514
75% Percentile	0.573
Maximum	0.758

lowed by *creativity* and *purpose*. Additionally, *suspense* was notably associated with images tagged under the topic "zombies" (see Figure 3).

For the *third* research question, we found that in general, organizers of the events who share images tend to have power, interest, control, and utility (positive correlation), they also tend to be lacking academic skills, motivation, enthusiasm, and exposure (negative correlation) (see Figure 4). Also, these organizers tend to post images with excitement emotion (positive correlation) while non-organizers tend to post images with curiosity emotions (negative correlation) (see Figure 5). Finally, organizers who post images tend to post images that are categorized as fun, suspense, creative, and solitude (see Figure 6)

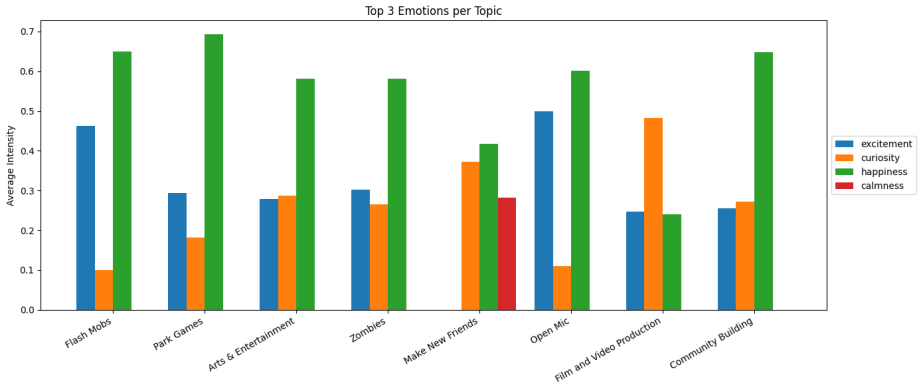

Fig. 2. The top three emotions of the ten most popular topics.

6 Conclusion and Future Research Directions

In this working paper, we examine images shared during events organized and coordinated through an event-based social media platform called Meetup.com.

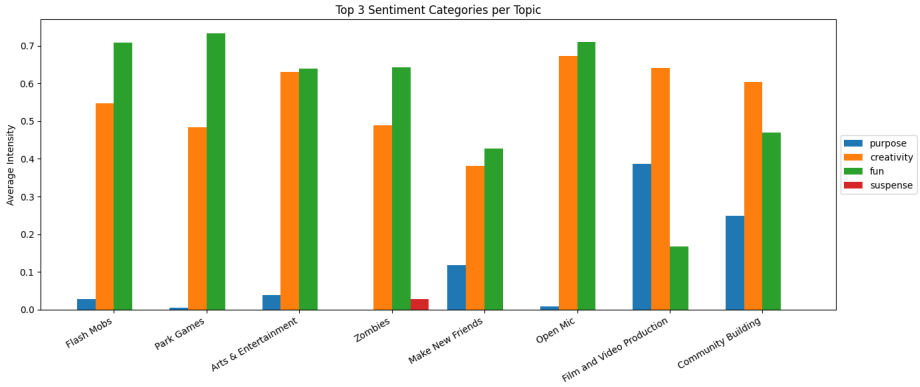

Fig. 3. The top three sentiments categories of the ten most popular topics.

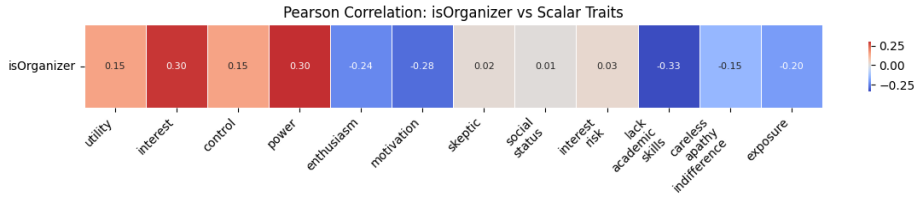

Fig. 4. The Pearson correlation coefficient values for the event organizers and participant traits.

This platform has many active groups that organize various types of events held online, offline, or both. Analyzing the images shared during different types of mob events is an important research problem, as it give us a better idea of what's happening at the mob in general, and can help us better understand the mob type (e.g., deviant vs. non-deviant); reveal the narratives of the mob by tracking cultural, social, or political symbols and their meaning; and public emotions and sentiments that are not expressed in text. It can also be used to develop predictive models of mobbers' traits and detect mob crises, as the shared images can provide real-time situational awareness.

"We are just testing the water" here and have so much more to do. One possible future research direction is to predict the emotions and sentiments expressed in images posted by a user. Another possible future direction is to develop a full evaluation protocol to assess the accuracy of the generated captions. Finally, we can develop a predictive model that classifies the type of mob (e.g., deviant vs. non-deviant) or the outcome of the mob (successful vs. unsuccessful) based on the provided images.

Some limitations of this work arise from the collected data, as most of it comes from public groups and public events/mobs, and not all mobs had shared images. Additionally, the majority of the images analyzed in this study are from

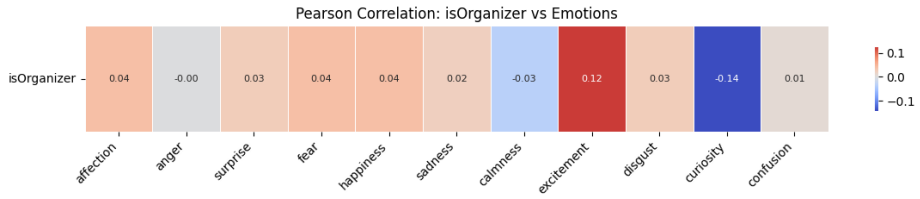

Fig. 5. The Pearson correlation coefficient values for the event organizers and emotions shown in images.

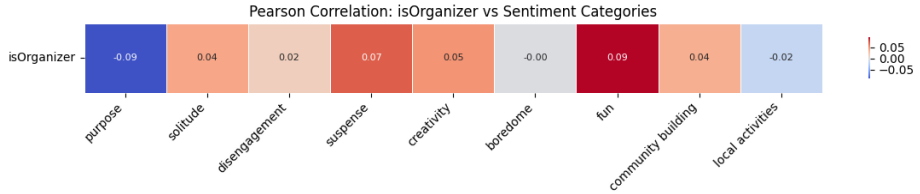

Fig. 6. The Pearson correlation coefficient values for the event organizers and sentiment categories shown in images.

benign mobs, such as dance mobs, musical theatre mobs, and singing lessons mobs, resulting in a lack of deviant mob cases. To address this limitation, we are currently gathering data from other social media platforms, including Facebook Events, Reddit, and BlueSky, which we plan to use for training our model on a wider variety of event types.

Acknowledgments. This work is based upon work supported in part by the Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332) and the U.S. Army Research Laboratory (W911NF2510147). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article. The authors declare the funding received to conduct this research in the acknowledgments section.

References

1. Al-khateeb, S., Agarwal, N.: Analyzing deviant cyber flash mobs of isil on twitter. In: Agarwal, N., Xu, K., Osgood, N. (eds.) Social Computing, Behavioral-Cultural Modeling, and Prediction. pp. 251–257. Springer International Publishing, Cham (2015)
2. Al-Khateeb, S., Agarwal, N.: Flash mob: a multidisciplinary review. Social Network Analysis and Mining **11**(1), 97 (2021)

3. Al-khateeb, S., Burright, J., Fernandes, S.L., Agarwal, N.: Analyzing and predicting meetup mobs outcome via statistical analysis and deep learning. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. pp. 164–173. Springer (2024)
4. Bharne, S., Bhaladhare, P.: Enhancing user profile authenticity through automatic image caption generation using a bootstrapping language–image pre-training model. In: RAiSE-2023. vol. 182. MDPI (2024). https://doi.org/10.3390/engproc2023059182
5. Chen, T., Borth, D., Darrell, T., Chang, S.F.: DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. arXiv preprint arXiv:1410.8586 (2014). https://doi.org/10.48550/arXiv.1410.8586, https://arxiv.org/abs/1410.8586
6. Lyu, H., Huang, J., Zhang, D., et al.: Gpt-4v(ision) as a social media analysis engine. ACM Transactions on Intelligent Systems and Technology **16**(3), 1–54 (2025). https://doi.org/10.1145/3709005, https://doi.org/10.1145/3709005
7. Mardjo, A., Choksuchat, C.: HyVADRF: Hybrid VADER–Random Forest and GWO for Bitcoin Tweet Sentiment Analysis. IEEE Access **10**, 101889–101897 (2022). https://doi.org/10.1109/ACCESS.2022.3209662, https://doi.org/10.1109/ACCESS.2022.3209662, received 25 August 2022, accepted 16 September 2022, published 26 September 2022, current version 30 September 2022
8. Mehta, M., Buntain, C.: Emotional images: Assessing emotions in images and potential biases in generative models (2024). https://doi.org/10.48550/arXiv.2411.05985
9. Nguyen, T., Gadre, S.Y., Ilharco, G., Oh, S., Schmidt, L.: Improving multimodal datasets with image captioning, n.d.
10. Ortis, A., Farinella, G.M., Battiato, S.: Survey on visual sentiment analysis. IET Image Processing **14**(8), 1440–1456 (2020). https://doi.org/10.1049/iet-ipr.2019.1270, https://doi.org/10.1049/iet-ipr.2019.1270
11. Saini, H., Bhandari, P., Bhattacharya, S., Jain, K.: Captify: A deep learning-based social media caption generator using multimodal data, n.d.
12. Zou, B., Thobhani, A., Kui, X., Abdussalam, A., Asim, M., Shah, S., ELAffendi, M.: A survey on enhancing image captioning with advanced strategies and techniques. Computer Modeling in Engineering & Sciences **142**(3), 2247–2280 (2025). https://doi.org/10.32604/cmes.2025.059192