An animated picture says at least a thousand words: Selecting Gif-based Replies in Multimodal Dialog

Xingyao Wang
University of Michigan
xingyaow@umich.edu

David Jurgens
University of Michigan
jurgens@umich.edu

Abstract

Online conversations include more than just text. Increasingly, image-based responses such as memes and animated gifs serve as culturally recognized and often humorous responses in conversation. However, while NLP has broadened to multimodal models, conversational dialog systems have largely focused only on generating text replies. Here, we introduce a new dataset of 1.56M text-gif conversation turns and introduce a new multimodal conversational model PEPE THE KING PRAWN for selecting gif-based replies. We demonstrate that our model produces relevant and high-quality gif responses and, in a large randomized control trial of multiple models replying to real users, we show that our model replies with gifs that are significantly better received by the community.

1 Introduction

Conversations are central to many online social platforms. While most conversations are text-based, computer mediated dialog also affords alternative forms of communication, such as emoji or stickers like bitmoji, that allow users to express themselves (Tang and Hew, 2019; Konrad et al., 2020). Increasingly, these visual forms of communication have become common in social media (Bourlai and Herring, 2014; Highfield and Leaver, 2016), with a notable use of the reaction gif (Bakhshi et al., 2016; Miltner and Highfield, 2017). These gifs are short video sequences that depict a particular scene and sometimes contain text that acts as a meta-commentary (Eppink, 2014). As a result, conversations become multimodal where individuals reply to one another using combinations of text and gifs (Figure 1). While conversational AI systems have been developed in a purely text-based setting, such systems do not capture the full multimodal behavior seen online. Here, we study multimodal conversation by introducing new dialog models for selecting gif replies in conversation.

PizzaMagic: Ahhhhh!!! The EMNLP deadline is in 24 hours!!
CasualModel:

Figure 1: Gif responses in conversation like the one shown above are embodied dialog that use visual imagery to convey reactions and emotions. This paper develops a system to select the appropriate gif response to messages. (PDF best viewed with Adobe Acrobat)

Conversation analysis is central to NLP and multiple approaches have analyzed this dialog structure (Jurafsky et al., 1998; Pareti and Lando, 2018; Cohn et al., 2019) and developed conversational agents to engage with people (e.g., Fang et al., 2018; Xu et al., 2020; Hong et al., 2020). Recent work has focused on generating open domain social chatbots that engage in sustained conversations in a natural way (Ram et al., 2018). Because many of these systems are designed to support voice-based dialog, they overlook non-textual forms of interaction used in social media conversations. In parallel, multimodal NLP systems have been developed for image data, often focusing on image-to-text tasks such as image captioning (Melas-Kyriazi et al., 2018; Sharma et al., 2018) and visual question answering (Antol et al., 2015; Huang et al., 2019; Khademi, 2020). More recent work has focused on the reverse text-to-image dimension, such as generating an image from a description (Niu et al., 2020; Ramesh et al., 2021). Our work unites these two strands of research by integrating image-based communication into conversational agents.

Our paper offers three main contributions. First,
we propose the new task of selecting gif responses in multimodal conversation analysis and introduce a new dataset of 1,562,701 real-world conversation turns with gif replies. Second, we introduce a new model PEPE THE KING PRAWN that fuses image and text-based features to select a relevant gif response. In in-house experiments, we show that our model substantially outperforms strong baseline models at selecting the exact gif used in real data and, in a manual test of the quality of the best responses, achieves an nDCG of 0.8145 on the annotated test set. Third, in a real-world test, we deploy our model as a part of a large-scale randomized controlled trial and show that the gif replies produced by our model are more highly voted by the community. Data, code, and models are available at https://github.com/xingyaoww/gif-reply.

2 GIF Communications

Gifs have been widely adopted in communication as a natural form of embodied speech where the visual imagery conveys emotions or a reaction as a response (Bakhshi et al., 2016; Tolins and Samermit, 2016). These gifs commonly come from widely-known cultural products, such as movies or television shows, which provides common knowledge for how they could be interpreted (Eppink, 2014; Miltner and Highfield, 2017). However, a single gif may have multiple interpretations, depending on the context, cultural knowledge of its content, and the viewer (Jiang et al., 2017). As a result, a single gif can serve multiple functions in communication (Tolins and Samermit, 2016).

Gifs have grown in their use through increasing affordances by platforms like Tumblr, Reddit, Imgur, and Twitter that allow gifs to be natively displayed like text in conversation threads (Jiang et al., 2018). Further, gif-based keyboards have been introduced that allow users to search for gifs that have been tagged with keywords or other metadata (Griggio et al., 2019). Yet, these technologies require that gif data be prepared with sufficient tags to be searchable or to have sufficient data to use collaborative filtering techniques for recommendations (Jiang et al., 2018, p.9). As a result, there is a clear gap in identifying appropriate response gifs directly from the text, which this work fills.

3 Data

Despite the widespread use of gifs, no standard dataset exists for text and gif replies. Further, although platforms like Twitter support gif replies, these gifs are not canonicalized to identify which responses correspond to the same gif. Therefore, we construct a new dataset for this task by collecting responses, matching their images, and augmenting this data with metadata about the gif, where possible. A visual description of the whole procedure can be found in Appendix Figure 7.

3.1 Gif Response Data

Gifs have many uses (Miltner and Highfield, 2017) and so we use a two-step approach to collect data that focus specifically on those likely to be used in conversation. First, gif responses are collected from Twitter by identifying all replies to English-language tweets containing animated_gif as embedded media. Tweets were collected from a ~10% sample of Twitter from March 13th, 2019 to Jan 24th, 2020, totaling 42,096,566 tweets with a gif that we were able to retrieve. Twitter does not canonicalize its gifs so two separate gif files may actually have the same imagery. Further, these files may not be identical due to small differences such as color variations or aspect ratios. To identify uses of the reference gifs, we use Average Hash from the imagehash library to create low-dimensional representations of each gif where hash distance corresponds to perceptual distance. Since gifs are animated and may contain varying scenes, we compute the hash for the first, middle, and final frames, concatenating these into a single hash. Two gifs are considered the same if (i) they have identical hashes or (ii) their hamming distance is < 10 and gifs with that hash have been used more than 500 times in Twitter. This latter condition was selected after manual evaluation of thresholds to trade-off between increasing the size of the training data and reducing potential noise caused by matching error. A visual example of this process can be found in Appendix Figure 8.

Not all gif responses in the Twitter data are conversational or appropriate for wider re-use. Therefore, we filter these responses to only those gifs whose imagery matches gifs hosted by the Giphy website, which is the backend for many gif-based keyboards. Giphy contains a wide collection of gifs that are curated to remove content inappropriate for general use (e.g., violent or sexual imagery). Gifs on the platform are categorized (e.g., “reaction” or “celebrities”) and we identify 28 categories containing 972 keywords likely to contain gifs used
in conversation. A total of 2,095,993 gifs linked to those keywords were ultimately retrieved and stored as image hashes. Additional details of categories and keywords are in Appendix B.

After the matching image hashes to filter replies, we identify 115,586 unique gifs, referred to as reference gifs, and 1,562,701 tweet replies using one of these gifs, which forms our official dataset. Figure 2 shows these gifs’ frequency in the data; much like words, a handful of gifs receive widespread use, while a long tail of gifs are rarely used.

3.2 Gif Metadata

We augment our gif data with information about their content. Some gifs have text that transcribes what a person is saying in the gif’s scene or is a meta-commentary on the content. This text is extracted using paddleOCR (Du et al., 2020). Since some gifs are long enough to contain multiple utterances, we run OCR on four frames sampled from each quartile of the gif’s length. Roughly 50% (58,020) of gifs contain at least one extracted word from the selected frames, with an mean of 5.5 extracted words per gif across the dataset.

Second, some gif repositories like Giphy allow users to tag gifs with information on their content or theme, e.g., “face palm” or “movie.” We collect tags for the 115K reference gifs used in Twitter, obtaining 39,651 unique tags. These user-generated tags were moderately noisy due to orthographic variations like spelling, capitalization, and spacing. Therefore, we merge tags by (i) lower-casing the text and (ii) performing a manual merge for similar word forms (e.g., “excited” and “exciting”). To minimize noise, we retain only tags that have been used with at least five gifs and where those gifs have been used at least 1000 times in total; this process removes many low-frequency tags that are either overly-specific or idiosyncratic in their use.

Finally, we performed a manual inspection of all remaining tags to remove tags that are too general (e.g., “emotion”) and retain only noun, adjective, and verb tags (words or multi-word expressions) that describe specific emotions or actions. A total of 241 unique tags were retained (Appendix C). 6.0% of gifs have at least one tag associated with them (mean 1.9 tags). However, these tagged gifs account for 38.7% of the replies in our dataset, suggesting tags are only available for more-popular gifs. Our dataset represents roughly an order of magnitude more data and more tags than the closest related dataset of Chen et al. (2017) that contained 23K gifs with 17 manually-curated emotions.

4 Gif Reply Models

We introduce a series of models for producing a gif response in conversation. Each model will select a gif from the 115K gifs in our dataset as a response to a text-based message. This task is related to but distinct from work on image-text matching (Lee et al., 2018), which aims to find an image describing a piece of text, or text-to-image (e.g., Wen et al., 2015; Xu et al., 2018), which generates an image from a text description. Here, we aim to select gifs that reflect natural continuations or reactions to a message in a dialog, akin to how gifs are used in social media. For all models, additional details on the training procedures and hyperparameters are provided in Appendix A. The three models that follow use varying degrees of information about the gifs and text to select a response.

4.1 Tag-based Predictions

The first model uses tags as a shared representation for characterizing gifs and text. Analogous to how object tags are used as anchor points for image-text matching (Li et al., 2020) and pivot languages are used in machine translation (Cheng et al., 2017), we use tags to bridge information between the text in a tweet and the visual content of a gif. Here, each gif becomes associated with a set of tags describing its conversational functions and for each text, we predict the set of tags for gifs responses to it—in essence, predicting what types of responses are most appropriate. We describe both of these processes next and how gifs are ultimately selected.
**Estimating Gif Tags** Only 6.0% of the gifs in our data have associated tags. Therefore we train a neural model to predict tags using known tags as training data. To capture any changes in emotion or imagery across the gif, we make separate predictions for four frames sampled across the gif (the same used in §3.2). Each frame is passed through an EfficientNet-based (Tan and Le, 2019) GIF encoder, shown in Figure 3, to extract a low-dimensional feature vector from each frame. These frame embeddings are fused using the attention mechanism from a transformer encoder layer. The output of the transformer feeds into a fully connected layer, which is trained as a multi-label classifier using binary cross-entropy to predict which tags should be present.

**Predicting Response Tags for Text** For each message, we predict the $k$-hot distribution of tags for a gif response by training a BERTweet model (Nguyen et al., 2020), which has been pre-trained on a large corpus of Twitter data (shown as “Tweet Encoder” in Figure 3). The model with an additional fully connected layer is trained as a multi-label classifier using binary cross-entropy, using the tags for the gifs used in reply (if known).

**Tag-based Gif Selection** At inference time, given a message, we use the text-to-tag model to predict a $k$-hot distribution over tags. Then, we select the gif whose estimated tag distribution is closest in Euclidean distance.

### 4.2 CLIP variant

The second model uses an end-to-end training approach based on the architecture of OpenAI CLIP (Radford et al., 2021). The architecture features two encoders, one for text and one for images. During training, the encoders are updated using contrastive loss that maximizes the cosine similarity of paired image-text representations and minimizes the cosine similarity of random pairs of images and texts. We replicate the CLIP architecture and training procedure, using BERTweet to encode text and EfficientNet (Tan and Le, 2019) to encode a composite image of four frames from the gif (compared with BERT and ResNet in their implementation). While originally designed to select an image for a text description, our model is trained to select a gif reply for a text message—a more challenging task than the image retrieval task used in the original CLIP setup, as the message may not contain words describing elements of the gif. At inference time, given a tweet, we use the trained tweet encoder to extract its representation and compute its cosine similarity with each encoded representation for our gifs. The gif with the highest cosine similarity is returned as the best response.

### 4.3 PEPE THE KING PRAWN

Our final model, **King Prawn**\(^1\) (referred to as “PEPE”) selects gif responses by using a richer set of multimodal features to create a gif representation. Rather than encode the gif solely from its image content, we use a multimodal encoder that captures (i) any text it might have, (ii) the types of objects present in the gif, and (iii) object regions as visual features. We encode these gif aspects using an OSCAR transformer (Li et al., 2020) to create a unified representation, shown in Figure 3 (bottom). Object names and regions of interest feature vectors are extracted using a pre-trained bottom-up attention model (Anderson et al., 2018).

As input to the OSCAR encoder, the captions to each of the gif’s four frames are concatenated together with an “[INTER_FRAME_SEP]” separator token. We filter object areas detected by the bottom-up attention model (Anderson et al., 2018) and we keep all objects with probability >0.5. We then concatenate object names together with the same inter-frame separator between names of different frames. Together, the caption text, object names, and image-region features are fed into the OSCAR transformer encoder to generate a GIF feature vector; the transformer is initialized with the default OSCAR weights. We use BERTweet to encode text. The entire PEPE model is trained end-to-end using contrastive loss, similar to the CLIP model.

### 5 Evaluation

We initially evaluate the methods in two ways. First, we use traditional classification-based evaluation, testing whether the models can reproduce the observed gif replies. However, some messages could have multiple valid gif responses. Therefore, as a second test, we evaluate the model in a retrieval setting, measuring whether its most-probable responses are good quality for a message.

**Experimental Setup** Models are trained and tested on a dataset containing 1,562,701 Tweet-

\(^1\) **King Prawn** refers to “selectKing INteresting Gifs for Personal RespAWNses.” In this crazy muppet-name-land-grab world we live in, our only regret is that we couldn’t get “Pepino Rodrigo Serrano Gonzales” to fit as a bacronym, which we leave to future work.
GIF pairs associated with 115,586 unique gifs, where 605,063 tweet-gif pairs are associated with at least one tag. Using the finalized 241 unique tags as classes for multi-label classification, we split the dataset by stratify on tags using the iterative train-test split method provided by scikit-multilearn library (Sechidis et al., 2011; Szymański and Kajdanowicz, 2017) to create a 80:10:10 train, dev, and test split which is finalized to train the models described in §4. Following BERTweet (Nguyen et al., 2020), we preprocess tweets in our dataset using NLTK TweetTokenizer for tokenization, emoji package to translate emotion icons, and converted mentions and links to special “@USER” and “HTTPURL” tokens.

**Annotated Data** To test whether each model’s predictions are valid responses, we annotate the ten most-probable gif predictions for a subset of the tweets in our test data. Many tweets in our test set require substantial context to understand due to having few tokens, linking to URLs that provide extra knowledge, mentioning other users in directed communication. These factors suggest social context or general knowledge aids in the recipient’s understanding of the gif’s intentions. While the model can still benefit from training on such examples, judging the appropriateness of response is difficult without access to the social context. Therefore, to reduce interpretation ambiguity, we annotate only tweets without URLs or user mentions and having at least 10 tokens. This process selects tweets with sufficient content to judge appropriateness independent of the larger social context.

Two annotators (the authors) were shown a list of potential gif responses for a tweet and asked to judge whether this is an appropriate gif response (a binary rating). Gifs were selected from the ten most-probable replies for each system and collectively shown in random order to prevent knowing which system generated each reply. A total of 2,500 gif-tweet pairings were annotated. Annotators attained a Krippendorf’s \(\alpha\) of 0.462; while moderate agreement, this value is expected given known differences in how people interpret and value gif responses based on their familiarity with its content, message interpretation, and life-experience (Jiang et al., 2018). We follow the evaluation setup from other retrieval-based dialog systems (e.g. Yu et al., 2021; Kumar and Callan, 2020) and use normalized Discounted Cumulative Gain (nDCG), which measures whether more appropriate gif responses are ranked higher. A gif’s appropriateness score is the sum of annotators’ ratings.

**Results** The PEPE model was able to identify relevant and good-quality gif responses, as shown by its performances on the test data (Table 1) and annotated data (Table 2). Performance on the test set is expected to be low, given the challenge of identifying the exact gif used for a tweet when multiple possible gifs are likely to be equally valid. However, the PEPE model is still able to identify the exact gif (out of 115K) in its top 10 predictions for 3% of the data, substantially outperforming all
Table 1: Models’ precision-at-k on selecting the exact gif used as a response for a tweet in the test set; this performance is an underestimate of each model, as many model-predicted gifs may be appropriate.

| Model            | Top-1       | Top-5       | Top-10      |
|------------------|-------------|-------------|-------------|
| Tag-based        | 0.000000    | 0.000092    | 0.000119    |
| Random           | 0.000020    | 0.000059    | 0.000158    |
| CLIP variant     | 0.000488    | 0.001669    | 0.002783    |
| Distribution sampling | 0.000996 | 0.005098    | 0.009780    |
| PEPE             | 0.005375    | 0.018723    | 0.030918    |

Table 2: Models’ nDCG scores at proposing appropriate gif replies, measured from annotations on the top 10 most probable gif replies of each model.

| Model            | nDCG        |
|------------------|-------------|
| Random           | 0.3273      |
| Tag-based        | 0.4526      |
| Distribution sampling | 0.4969  |
| CLIP variant     | 0.5934      |
| PEPE             | 0.8145      |

Table 3: Results for ablated versions of PEPE where specific input is removed (cf. Table 2) show that all input forms contribute to the ability to select replies.

| Model                        | nDCG        |
|------------------------------|-------------|
| PEPE                         | 0.8145      |
| PEPE without object names    | 0.7665      |
| PEPE without caption         | 0.7559      |
| PEPE without object features | 0.7533      |

Performance on the annotated data (Table 2) provides a more realistic assessment of whether models can generate high-quality replies, as it measures whether the models’ replies themselves were good. The PEPE model attains substantially higher performance (p<0.01) than other models. While the CLIP variant model performs well, the content-agnostic Distribution sampling baseline performs nearly as well. This baseline’s high performance speaks to the multiple interpretations of gifs and the ease at which readers can make connections between a gif and message. Indeed, even the random-gif model has a non-zero nDCG, highlighting the ability for an arbitrary gif to still be considered appropriate. We speculate that popular gifs may be popular because of this ease of multiple interpretations.

Ablation study PEPE fuses multiple types of input, which may uniquely contribute to model’s ability to select gif replies. To understand how these inputs each contribute, we performed an ablation study on the annotated test set by removing one input from Oscar GIF Encoder shown in Figure 3 (i.e., a gif’s caption, object names, or objects’ visual features) and evaluating the model’s resulting gifs on the same test instances.

The ablated model performances, shown in Table 3, reveal that each input is useful for selecting gifs. Object features capture visual information about what specifically is present in the gif (beyond the discrete names of what is present, e.g., “person” or “building”) and show that multimodality is important for high performance—predicting replies just from a gif’s caption and categorized content are insufficient. Similarly, the caption of a gif (if present) is important, as the text can help make explicit the intended interpretation of a gif.

### 6 Field Experiment

To test the generalizability of our models and quality of their responses, we conduct a large-scale randomized controlled trial (RCT) that has the models respond to real users and measure their perception of reply quality.²

#### 6.1 Experimental Setup

Gifs were posted to the Imgur platform, which is a highly active social media community that supports both image and text-based interactions. On Imgur, users may create posts, which contain one or more images with optional commentary, or comment on posts or replies. Similar to pre-2018 Twitter, comments are limited to 140 characters. Imgur conversations are threaded and frequently contain both image and text comments. Like Reddit, users may upvote and downvote content, providing a score of how well it was received by the community; we use

²The performance decrease for removing object names is statistically significant (p<0.01, bootstrapped). The decreases for removing captions and objects’ visual features are significant from the name-removal model (p<0.01) but the two models are statistically equivalent (p>0.19).

³This experiment was ruled as Not Regulated by the University of Michigan IRB (HUM00197631). However, IRB approval is not sufficient to prevent harm (Bernstein et al., 2021) and significant precautions were taken to minimize potential risk (See §9).
Table 4: Model-selected replies to messages (paraphrased for privacy). Click an image to view the gif on Giphy.

---

this score in our experiments to evaluate quality.

Our experiment focuses on generating Gif-based replies to top-level text comments (comments made directly to the post). This setup mirrors the conversational data our models were trained on. Imgur supports several ways of filtering its stream of posts. To ensure that our replies have sufficient visibility, we select posts that have already receive 10 comments and appear in the “most viral” sorting. From these posts, we reply to the top-rated text comment. The RCT runs from 8 AM to 8 PM (local time), making at most 10 replies per hour.

Not all topics or comments are suitable for automated responses and great care was taken to prevent potential harm to the community. Through multiple rounds of testing which replies would be responded to, we curated a list of keywords that could lead to potential controversial replies, such as terms about religion or race (full list in Appendix D). Any comment containing a token or lemma matching a word on this list is excluded and not replied to. As a further safeguard, experimenters monitored all replies to remove any that were deemed inappropriate. See the Ethics Section (§9) for a longer discussion of safeguards.

The field experiment consists of five arms, corresponding to the three trained models and the two baseline models. During each trial, one model is selected and generates a response; the trained model replies with the most probable gif.4

Not all models are equally likely to perform well and so to make the most use of our trial budget, we use Thompson sampling (Russo et al., 2018) to randomly select which arm of the trial to use. Thompson sampling builds a probability model for the estimated reward of each arm (here, the score a reply receives) and samples from the model such that higher-rewarding arms are sampled more frequently. As a result, this method can provide tighter estimates for the reward of the most useful arms. Scores in Imgur have a skewed distribution, with few comments receiving very high scores and most receiving near the default score (1). Therefore, we use Poisson Thompson sampling. Some comments may be downvoted to receive scores below zero, so for simplicity, we truncate these scores to 0.

To evaluate the results of the RCT, we construct a Negative Binomial regression on the dependent variable of the score received for a model’s reply, truncating negative scores to zero. The Negative binomial was chosen instead of Poisson due to over-dispersion in the score variable. The models are treated as a categorical variable, using the random model as a reference. Since the score will depend, in part, on the attention received by the parent post and comment (higher-rated comments are displayed first), we include linear effects for the post and parent comment. Finally, we include five text-related variables to control for the con-
tent of the parent comment: the topic distribution (Appendix Table 9) from a 10-topic model (dropping one topic due to collinearity), the sentiment and subjectivity of the message estimated using TextBlob library, the length of the comment, and whether the comment contained a question.

6.2 Results

The field experiment demonstrates that the PEPE model is able to generate significantly higher-scoring responses. Figure 4 shows the Negative Binomial regression coefficients for the three models and empirical distribution baseline, with the random gif model as a reference; full regression results are shown in Appendix Table 6. The PEPE model substantially outperforms all other models \((p < 0.01)\) in this real-world setting. Surprisingly, despite performing second-best in our annotated evaluations, the CLIP model performs worst, with its replies receiving fewer upvotes than the two baselines that randomly select gifs. We investigate potential explanations for these performances next.

The Random and Distributional-sampling baseline models perform surprisingly well relative to models that take the text and gif content into account, with only the PEPE model outperforming them. The performance of the random baselines matches prior work showing people are still able to draw some connection between their interpretation and the reply (Madden, 2018, p.29). Further, we observed that, when the model’s reply truly seemed random, some users replied say they upvoted solely because they enjoyed the gif.

As a follow-up experiment, we tested whether models could be getting higher (or lower) scores by repeatedly picking the same gifs that are skewed towards a positive or negative reaction. Figure 5 shows the score distribution for the top ten most frequently used gifs (visual examples in Appendix Table 7) for each of the three trained models and reveals surprisingly divergent behavior for how the community reacts. Each model had a different set of most-used gifs, indicating the models did not converge to a universal set of common replies. Indeed, a gif’s frequency-of-use and mean reply score were uncorrelated in all three models \((r \approx -0.01, p > 0.73\) for all models). The most-used gifs for each model had average scores that were positive, but the distributions for each gif show that some uses were occasionally downvoted. This high variance in scores indicates that a gif’s intrinsic qualities are not solely responsible for the received score and, instead, appropriate use in context plays a significant part in community reception.

We examined whether models relied on the same set of gifs. Figure 6 shows the distribution of gif
uses by each model, indicating that the tag-based model relied frequently on a small set of gifs. However, the PEPE and CLIP variant models were substantially more varied, indicating they draw from the long-tail of possible gifs.

Do any of our models spark more subsequent conversation? We fit a separate Negative Binomial regression on the total number of comments made to our reply, using the same IVs as the score regression and include the reply’s score itself as another IV. This model (Appendix Table 8) shows that both the distributional-sampling baseline and PEPE models produced replies that led to fewer subsequent comments (p<0.01)—despite the PEPE model having the most-upvoted replies. However, the score of the gif reply was positively associated (p<0.01) indicating that more appropriate replies do receive more subsequent conversation. We speculate that the random models may have led to more conversation due to users replying to express confusion about why the particular gif was used. This result points to a need to understand what text and visual factors in gifs influence the volume of subsequent dialog and an opportunity to optimize gif models for both quality and number of conversation turns.

7 Related Work

This work draws upon two strands of research from dialog systems and multimodal NLP. Conversational dialog systems have traditionally been built upon large-scale dialog corpora from social media platforms (Bessho et al., 2012) such as Twitter. Our approaches are fundamentally information retrieval based systems that mirror the approach by text-based conversational systems that retrieve existing messages from a large social media corpus as potential replies and rank these to select a response. Our work mirrors models that use neural networks for ranking (Yan et al., 2016; Inaba and Takahashi, 2016; Penha and Hauff, 2021, e.g.); however, we note that many recent knowledge-grounded and open domain models use encoder-decoder methods to improve versatility and applicability (e.g., Ghazvininejad et al., 2018; Gao et al., 2019; Zhou et al., 2020). Generative approaches are likely inappropriate for gif-based conversation as gifs are more akin to mimetic artifacts that build on cultural knowledge (Eppink, 2014), making synthesizing a new gif from scratch likely less effective.

All three models used here rely on joint embedding spaces for gif and text. Multiple works in NLP have been proposed to align these representations (Kiros et al., 2014; Wang et al., 2016), often for particular applications such as visual question answering (Antol et al., 2015). Recent work has focused on embeddings these media with a single encoder that takes both text and images as input (e.g., Wang et al., 2019; Chen et al., 2020), in contrast to our model that uses separate image and text encoders (Figure 3); these multimodal encoders are prohibitively computationally expensive to use in our setting during inference time, as the model would need to be run on each gif (and message) to rank replies, compared with our model that only needs to encode text. However, performance and efficiency improvements in aligning image and text representations would likely benefit our task.

8 Conclusion

People like using gifs in online conversations—gifs are a fun and playful way to communicate. However, modern NLP conversational agents operate only by text. Here, we introduce a new dataset of 1.56M conversation turns using gifs, including captions and metadata, and develop a new conversational model PEPE THE KING PRawn that selects appropriate gif responses for messages through comparing encoded gif and text representations. In two evaluations, we show that PEPE is able to generate highly-relevant gif responses and in a large-scale RCT, we show that the gif replies from the PEPE model received significantly higher scores from the general public. Our work demonstrates the opportunity for using NLP methods to successfully engage in multimodal conversations.
The interactive nature of the RCT necessitated a close consideration of ethical issues (Thieltges et al., 2016). Prior to beginning the RCT, the study team obtained IRB approval to interact with users. While necessary in the legal sense, IRB approval is not sufficient to justify the ethical grounds of the study. The primary risks of the study are if the automated models respond with an inappropriate gif or respond to a message that is not suitable for automated response (e.g., discussing the death of a loved one or making an offensive statement). These risks were mitigated in multiple ways throughout the dataset construction and field experiment.

First, the selection criteria for which comments we reply to was designed to only reply to content that was already deemed appropriate by the community. By selecting only posts that had received sufficient upvotes to be called “viral” and were already receiving comments, we mitigate the risk of engaging in topics or conversations that are inappropriate according to the norms of the Imgur community, as these posts would be removed by moderators or would have received sufficient downvotes to stay in obscurity.

Second, by focusing on the top-voted comment to these posts, we again reply to content that has already been deemed high-quality by the comment. This comment-level criteria substantially lowers the risk of our models commenting on inappropriate comments (e.g., a comment insulting another user), as these comments are readily downvoted by the community prior to our intervention.

Third, we employed extensive filtering to avoid replying to any comment containing a potentially sensitive topic, e.g., a discussion of race or trauma. The initial set of keywords was developed through examining potentially sensitive topics and then iteratively added to by simulating which messages our RCT would reply to and examining whether it would be appropriate. During the field RCT, experimenters continuously monitored the comments to ensure no harm was being done. Ultimately, only three comments were removed during the initial two days, which was due to a bug in the lemmatization and these comments should have been filtered out by our earlier criteria; these comments were removed quickly and we did not observe any notable response from the community.

Fourth, one risk is replying with an inappropriate gif, which is mitigated by the use of Giphy to seed our initial gifs. As this platform is curated and does not host objectively offensive gifs (e.g., overly-violent content), our initial gif set is relatively free of objectionable gifs. Because our model learns directly from gifs’ frequency of use, unless objectively offensive gifs are widely used, they are unlikely to be deployed from our RCT; we speculate that few objectively offensive gifs are widely used and, in practice, we have not identified any during the study period or when examining hundreds of random gifs in our data (or used in the RCT).

Finally, one risk is that by learning gif responses from observed data, our models may reinforce cultural stereotypes that are encoded in the gifs themselves (Erinn, 2019), e.g., the association of African American individuals with strong emotions. While our gif data is relatively clean of overtly offensive gifs, we acknowledge that our model likely does inadvertently perpetuate some of these latent biases in the data. However, the success of our model suggests a future mitigation strategy for platforms suggesting gifs: as biases become known, our approach can be used to suggest less-biased gifs as potential responses to mitigate future harm.

Acknowledgments

We thank the reviewers, area chairs, and senior area chairs for their thoughtful comments and feedback. We also thank the Blablablab for helpful feedback and letting us deploy PEPE to the group’s Slack and putting up with the ridiculous gif replies and Imgur for being a wonderful community. This material is based upon work supported by the National Science Foundation under Grant No. 2007251.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 6077–6086. IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 2425–2433. IEEE Computer Society.
Saeideh Bakhshi, David A. Shamma, Lyndon Kennedy, Yale Song, Paloma de Juan, and Joseph Joish Kaye. 2016. Fast, cheap, and good: Why animated gifs engage us. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016, pages 575–586. ACM.

Michael S Bernstein, Margaret Levi, David Magnus, Betsy Rajala, Debra Satz, and Charla Waeiss. 2021. Esr: Ethics and society review of artificial intelligence research. ArXiv preprint, abs/2106.11521.

Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. 2012. Dialog system using real-time crowdsourcing and Twitter large-scale corpus. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 227–231, Seoul, South Korea. Association for Computational Linguistics.

Elli Bourlai and Susan C Herring. 2014. Multimodal communication on tumblr: “i have so many feels!” In Proceedings of the 2014 ACM conference on Web science, pages 171–175.

Weixuan Chen, Ognjen Oggi Rudovic, and Rosalind W Picard. 2017. Gifgif+: Collecting emotional animated gifs with clustered multi-task learning. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 510–517. IEEE.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal Image-TExt representation learning. In ECCV.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pages 3974–3980. ijcai.org.

Michelle Cohn, Chun-Yen Chen, and Zhou Yu. 2019. A large-scale user study of an Alexa Prize chatbot: Effect of TTS dynamism on perceived quality of social dialog. In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, pages 293–306, Stockholm, Sweden. Association for Computational Linguistics.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weivei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. PP-OCR: A Practical Ultra lightweight OCR system. ArXiv preprint, abs/2009.09941.

Jason Eppink. 2014. A brief history of the gif (so far). Journal of visual culture, 13(3):298–306.

Wong Erinn. 2019. Digital blackface: How 21st century internet language reinforces racism.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 96–100, New Orleans, Louisiana. Association for Computational Linguistics.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots. Now Foundations and Trends.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 5110–5117. AAAI Press.

Carla F Griggio, Joanna McGregor, and Wendy E Mackay. 2019. Customizations and expression breakdowns in ecosystems of communication apps. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–26.

Tim Highfield and Tama Leaver. 2016. Instagrammatics and digital methods: Studying visual social media, from selfies and gifs to memes and emoji. Communication research and practice, 2(1):47–62.

Chung Hoon Hong, Yuan Liang, Sagnik Sinha Roy, Arushi Jain, Vihang Agarwal, Ryan Draves, Zhizhuo Zhou, William Chen, Yujian Liu, Martha Miracky, et al. 2020. Audrey: A personalized open-domain conversational bot. In Alexa Prize Proceedings.

Pingping Huang, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. 2019. Multi-grained attention utterance ranking model for conversational dialogue answering. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3595–3600, Florence, Italy. Association for Computational Linguistics.

Michimasa Inaba and Kenichi Takahashi. 2016. Neural utterance ranking model for conversational dialogue systems. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 393–403, Los Angeles. Association for Computational Linguistics.

Jialun “Aaron” Jiang, Jed R Brubaker, and Casey Fiesler. 2017. Understanding diverse interpretations of animated GIFs. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pages 1726–1732.
Jialun “Aaron” Jiang, Casey Fiesler, and Jed R Brubaker. 2018. “The Perfect One” Understanding Communication Practices and Challenges with Animated GIFs. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–20.

Daniel Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. *Lexical, prosodic, and syntactic cues for dialog acts*. In *Discourse Relations and Discourse Markers*.

Mahmoud Khademi. 2020. Multimodal neural graph memory networks for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7177–7188, Online. Association for Computational Linguistics.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *ArXiv preprint*, abs/1411.2539.

Artie Konrad, Susan C Herring, and David Choi. 2020. Sticker and emoji use in facebook messenger: implications for graphicon change. *Journal of Computer-Mediated Communication*, 25(3):217–235.

Vaibhav Kumar and Jamie Callan. 2020. Making information seeking easier: An improved pipeline for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3971–3980, Online. Association for Computational Linguistics.

Kuang-Huei Lee, X. Chen, G. Hua, H. Hu, and Xiaodong He. 2018. Stacked cross attention for imagetext matching. *ArXiv preprint*, abs/1803.08024.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

John Savery Madden. 2018. *The Phenomenological Exploration of Animated GIF Use in Computer-Mediated Communication*. Ph.D. thesis, University of Oklahoma.

Luke Melas-Kyriazi, Alexander Rush, and George Han. 2018. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 757–761. Brussels, Belgium. Association for Computational Linguistics.

Kate M Miltner and Tim Highfield. 2017. Never gonna GIF you up: Analyzing the cultural significance of the animated GIF. *Social Media+ Society*, 3(3):2056305117725223.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. Online. Association for Computational Linguistics.

Tianrui Niu, Fangxiang Feng, Lingxuan Li, and Xiaojie Wang. 2020. Image synthesis from locally related texts. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 145–153.

Silvia Pareti and Tatiana Landó. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gustavo Penha and Claudia Hauff. 2021. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 160–170, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *ArXiv preprint*, abs/2103.00020.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. *Conversational ai: The science behind the alexa prize*. *ArXiv preprint*, abs/1801.03604.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Mark Chen, Rewon Child, Vedant Misra, Pamela Mishkin, Gretchen Kruegerand Sandhini Agarwal, and Ilya Sutskever. 2021. DALL-E: Creating images from text. https://openai.com/blog/dall-e/.

Daniel J Russo, Benjamin Van Roy, Abbas Kazeroni, Ian Osband, and Zheng Wen. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, pages 145–158.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In First International Workshop on Learning with Imbalanced Domains: Theory and Applications, pages 22–35. PMLR.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 6105–6114. PMLR.

Ying Tang and Khe Foon Hew. 2019. Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. International Journal of Communication, 13:27.

Andree Thieltges, Florian Schmidt, and Simon Hegelich. 2016. The devil’s triangle: Ethical considerations on developing bot detection methods. In 2016 AAAI Spring Symposium Series.

Jackson Tolins and Patrawat Samermit. 2016. Gifs as embodied enactments in text-mediated conversation. Research on Language and Social Interaction, 49(2):75–91.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 5005–5013. IEEE Computer Society.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: cross-modal adaptive message passing for text-image retrieval. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 5763–5772. IEEE.

Miaomiao Wen, Nancy Baym, Omer Tamuz, Jaime Teevan, Susan T Dumais, and Adam Kalai. 2015. Omg ur funny! computer-aided humor with an application to chat. In ICCC, pages 86–93.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1835–1845, Online. Association for Computational Linguistics.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 1316–1324. IEEE Computer Society.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016, pages 55–64. ACM.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. ArXiv preprint, abs/2105.04166.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. Computational Linguistics, 46(1):53–93.
| Category             | Subcategory          |
|----------------------|----------------------|
| Cartoons & Comics    | aqua teen hunger force |
| Celebrities          | richard pryor       |
| Reactions            | angry                |
| Emotions             | happy                |
| Anime                | bleach               |
| Art & Design         | psychedelic          |
| Nature               | sunrise              |
| Transportation       | bicycle              |

Table 5: Examples of GIF categories on GIPHY

A Additional Details on Model Training

Following, we provide additional details on how each of the three models was trained.

A.1 Tag-based Model

**EfficientNet-based Tag Classifier** Gifs are reshaped to 224 by 224 pixel while keeping the aspect ratio by padding and normalized to a mean of 0.5 and standard deviation of 0.5 for each channel before feeding into the EfficientNet-based model. We selected unique GIFs from the finalized dataset that has at least one associated tag and using the iterative train test split on k-hot tag representation to select 5% of those GIFs for validation. The EfficientNet tag classifier was trained for 100 epochs on a batch size of 32, using AdamW optimizer with learning rate 1e-5 and weight decay 1e-3. The best validation performance was achieved at the 40th epoch with macro-f1 of 0.30 in predicting 241 multi-label classes. Early experiment shows that transformer encoder layer (macro-f1 of 0.30) outperforms linear layer (macro-f1 of 0.19) in fusing multi-frame gif features on the development set, therefore transformer encoder layer is used to fuse features of different frames in our implementation.

**Tweet-to-tag classifier** Using the finalized dataset mentioned in §3, we use tweet as input, and the k-hot tag representation of that tweet instance as ground truth label to train the multi-label classifier along with the tweet encoder for 241 multi-label classes. Early experiment shows that transformer encoder layer (macro-f1 of 0.30) outperforms linear layer (macro-f1 of 0.19) in fusing multi-frame gif features on the development set, therefore transformer encoder layer is used to fuse features of different frames in our implementation.

A.2 CLIP variant

The evaluation performance for model selection is measured by nDCG. For every tweet-gif pair in the validation set, we measure the top 30 predicted GIFs from the model using the tweet as input. The relevance of an occurring ground truth gif in the top 30 predictions given a tweet is set 1 for the nDCG calculation.

CLIP variant is trained on the same finalized dataset using contrastive loss. It was trained for 16 epochs with a batch size of 16 using AdamW optimizer of learning rate 1e-5 and weight decay 1e-3. Best validation performance is achieved at epoch 6 with an nDCG value of 0.015.

We replace the Transformer encoder layer with a linear Layer on Efficient GIF Encoder from Figure 3, and use this as our GIF Encoder for the CLIP variant. Image inputs to the GIF encoder are normalized following the official CLIP implementation.

A.3 PEPE

The PEPE model follows most configurations from the CLIP variant model, but replace the EfficientNet GIF encoder with an Oscar GIF encoder based on Oscar pre-trained multi-modal transformer (Li et al., 2020).

Extra metadata are extracted from GIFs in the finalized dataset for further training. Captions within the GIF are extracted using PaddleOCR (Du et al., 2020), and only extracted text with probability greater than 0.9 are kept as caption metadata.

Object tags and their corresponding features are extracted with bottom-up attention (Anderson et al., 2018) using py-bottom-up-attention package. Object instances are filtered to only keep instances that have a score higher than 0.5, then object tags and their corresponding features are extracted from these instances. Final object features of dimension 2054 are obtained by concatenating feature output with dimension 2048 from Faster-RCNN with scaled box position coordinates of the object following (Li et al., 2020).

The PEPE model is trained on the finalized dataset with extracted caption and object metadata. It was trained for 16 epochs with a batch size of 8 using AdamW optimizer of learning rate 1e-6 and weight decay 1e-3. Preprocessing for GIFs is validation macro-f1 was 0.07 achieved at the 70th epoch.
Figure 7: A diagram of the pipeline used to collect, canonicalize, and filter gif-reply data from Twitter.
the same as the Tag-based model. Max sequence length is set to 256 tokens for the Oscar transformer. Best evaluation performance is achieved at epoch 12 with an nDCG score of 0.007.

B GIF categories on GIPHY

| Category     | Subcategory      |
|--------------|------------------|
| Reactions    | what             |
| Reactions    | hair flip        |
| Reactions    | bored            |
| Reactions    | frown            |
| Reactions    | slow clap        |
| Reactions    | mic drop         |
| Reactions    | goodbye          |
| Reactions    | meh              |
| Reactions    | scared           |
| Reactions    | do not want      |
| Reactions    | confused         |
| Reactions    | drunk            |
| Reactions    | wow              |
| Reactions    | mad              |
| Reactions    | awesome          |
| Reactions    | please           |
| Reactions    | thumbs down      |
| Reactions    | frustrated       |
| Reactions    | oh snap          |
| Reactions    | disgusted        |
| Reactions    | rejected         |
| Reactions    | embarrassed      |
| Reactions    | hug              |
| Reactions    | yolo             |
| Reactions    | interested       |
| Reactions    | thank you        |
| Reactions    | sarcastic        |
| Reactions    | shocked          |
| Reactions    | cool story bro   |
| Reactions    | middle finger    |
| Reactions    | you got this     |
| Reactions    | whatever         |
| Reactions    | omg              |
| Reactions    | deal with it     |
| Reactions    | sigh             |
| Reactions    | oops             |
| Reactions    | angry            |
| Reactions    | finger guns      |
| Reactions    | good luck        |
| Dependent variable: | Gif reply score |
|---------------------|-----------------|
| post score          | $-0.0002^{***}$ (0.00003) |
| comment score       | 0.001*** (0.0001) |
| CLIP variant model  | $-0.161^{***}$ (0.058) |
| Distribution-sampling model | 0.057 (0.056) |
| PEPE model          | 0.223*** (0.051) |
| Tag-based model     | $-0.017$ (0.055) |
| number of days after reply | 0.003*** (0.0005) |
| comment text polarity | $-0.039$ (0.058) |
| comment text subjectivity | $-0.033$ (0.052) |
| topic 0 (Politics related) | 0.078 (0.155) |
| topic 1 (Family & Pets related) | 0.300** (0.148) |
| topic 2 (Employment related) | $-0.119$ (0.184) |
| topic 3 (Social media related) | 0.140 (0.165) |
| topic 4 (Transportation related) | $-0.172$ (0.188) |
| topic 5 (Food related) | 0.133 (0.194) |
| topic 6 (COVID related) | $-0.082$ (0.200) |
| topic 7 (Entertainment related) | $-0.057$ (0.161) |
| topic 8 (People related) | 0.272 (0.198) |
| comment is a question | 0.068 (0.049) |
| length of parent comment | $-0.003$ (0.002) |
| intercept           | 0.231** (0.115) |

Observations 8,369  
Log Likelihood $-14,899.820$  
$\theta$ 0.548*** (0.013)  
Akaike Inf. Crit. 29,841.640

Note: *$p<0.1$; **$p<0.05$; ***$p<0.01$

Table 6: Negative Binomial regression on score of the gif reply. The random-gif baseline is set as the reference category for models.
Reactions abandon thread
Reactions excited
Reactions suspicious
Reactions win
Reactions applause
Reactions popcorn
Reactions sleepy
Reactions nod
Reactions awww
Reactions disappointed
Reactions ugh
Reactions laughing
Reactions oh no you didn't
Reactions smh
Reactions agree
Reactions serious
Reactions party hard
Reactions shut up
Reactions ok
Reactions help
Reactions smile
Reactions incredulous
Reactions yawn
Reactions idk
Reactions sexy
Reactions fist bump
Reactions dancing
Reactions nom
Reactions eww
Reactions hello
Reactions not bad
Reactions success
Reactions burn
Reactions proud
Reactions i give up
Reactions hearts
Reactions pleased
Reactions fml
Reactions sorry
Reactions aroused
Reactions happy dance
Reactions good job
Reactions wtf
Reactions seriously
Reactions want
Reactions rage
Reactions table flip
Reactions love
Reactions amused
Reactions flirt
Reactions judging you
Transportation truck
Transportation spaceship
Transportation van
Transportation submarine
Transportation motorcycle
Transportation bmw
Transportation helicopter
Transportation chevrolet
Transportation volkswagen
Transportation boat
Transportation bus
Transportation porsche
Transportation tank
Transportation audi
Transportation toyota
Transportation airplane
Transportation hovercraft
Transportation nissan
Transportation bicycle
Transportation train
Transportation rocket
Transportation yacht
Transportation ferrari
Transportation honda
Transportation sailboat
Transportation car
Transportation tesla
Holidays mardi gras
Holidays oktoberfest
Holidays kwanzaa
Holidays mothers day
Holidays yom kippur
Holidays st patricks day
Holidays memorial day
Holidays fourth of july
Holidays fathers day
Holidays labor day
Holidays rosh hashanah
Holidays new years
Holidays passover
Science global warming
Science astronomy
Science physics
Science laser
Science stars
Science robot
Science atoms
Science meteor
Science bubbles
Science medicine
Science nebula
Science technology
Science mathematics
Science chemistry
Science biology
Science planets
Science magnets
Science molecules
Science asteroids
Science space
Science bill nye
Science engineering
Science diy
Science nuclear
Science computers
Fashion & Beauty chanel
Fashion & Beauty alexander mcqueen
Fashion & Beauty model
Fashion & Beauty victorias secret
Fashion & Beauty prada
Fashion & Beauty karlie kloss
Fashion & Beauty jessica stam
Fashion & Beauty emily ratajkowski
Fashion & Beauty miranda kerr
Fashion & Beauty kate upton
Fashion & Beauty louis vuitton
Fashion & Beauty makeup
Fashion & Beauty kate moss
Fashion & Beauty cara delevingne
Fashion & Beauty runway
Fashion & Beauty jourdan dunn
Fashion & Beauty julia nobis
Fashion & Beauty jewelry
Fashion & Beauty beauty
Fashion & Beauty christian iman
Fashion & Beauty marc jacobs
Fashion & Beauty shoes
Fashion & Beauty dress
Fashion & Beauty gucci
Greetings get well
Greetings bye
Greetings im out
Greetings sympathy
Greetings thank you
Greetings new baby
Greetings im sorry
Greetings congratulations
Food & Drink pancakes
Food & Drink sandwich
Food & Drink happy hour
Food & Drink sushi
Food & Drink steak
Food & Drink pasta
Food & Drink french toast
Food & Drink mimosa
Food & Drink tea
Food & Drink whiskey
Food & Drink pickle
Food & Drink cake
Food & Drink egg roll
Food & Drink broccoli
Celebrities mr. t
Celebrities danny mcbride
Celebrities michael fassbender
Celebrities seth rogen
Celebrities elijah wood
Celebrities jon hamm
Celebrities tom hanks
Celebrities kate upton
Celebrities arnold schwarzenegger
Celebrities tom hiddleston
Celebrities al pacino
Celebrities sean connery
Celebrities javier bardem
Celebrities ken jeong
Celebrities will smith
Celebrities maya rudolph
Celebrities jack mcbrayer
Celebrities leonardo dicaprio
Celebrities clint eastwood
Celebrities robert downey jr
Celebrities michael ian black
Celebrities adrien brody
Celebrities tom hardy
Celebrities joseph gordon levitt
Celebrities mark ruffalo
Celebrities adam baldwin
Celebrities rebel wilson
Celebrities jim carrey
Celebrities melissa mccarthy
Celebrities julianne moore
Celebrities hayden panettiere
Celebrities anna kendrick
Celebrities will forte
Celebrities ryan gosling
Celebrities andrew garfield
Celebrities nick offerman
Celebrities weird al yankovic
Celebrities will arnett
Celebrities bruce lee
Celebrities christian bale
Celebrities paul dano
Celebrities eddie murphy
Celebrities sam rockwell
Celebrities mike tyson
Celebrities jude law
Celebrities rooney mara
Celebrities adam sandler
Celebrities chris hemsworth
Celebrities kristen wiig
Celebrities james franco
Celebrities adam scott
Celebrities seth green
Celebrities jeremy renner
Celebrities morgan freeman
Celebrities bradley cooper
Celebrities dave chappelle
Celebrities rachel mcadams
Celebrities nicholas cage
Celebrities megan fox
Celebrities robert redford
Celebrities elizabeth banks
Celebrities liam neeson
Celebrities willem dafoe
Celebrities jonah hill
Celebrities michael cera
Celebrities charlie sheen
Celebrities emma roberts
Celebrities jon stewart
Celebrities patton oswalt
Celebrities samuel l jackson
Celebrities alison brie
Celebrities matt lucas
Celebrities ellen page
Celebrities amanda bynes
Celebrities jake gyllenhaal
Celebrities rob lowe
Celebrities steve carell
Celebrities conan obrien
Celebrities cillian murphy
Celebrities mindy kaling
Celebrities ben stiller
Celebrities john travolta
Celebrities gary oldman
Celebrities amy poehler
Celebrities ian somerhalder
Celebrities richard pryor
Celebrities bruce willis
Celebrities daniel day lewis
Celebrities chuck norris
Celebrities ed helms
Celebrities don cheadle
Celebrities michael caine
Celebrities george carlin
Celebrities alia shawkat
Celebrities emma stone
Celebrities adam devine
Celebrities larry david
Celebrities taylor kitsch
Celebrities matthew perry
Celebrities dave franco
Celebrities olivia munn
Celebrities emily blunt
Celebrities mila kunis
Celebrities ru paul
Celebrities jason bateman
Celebrities anne hathaway
Celebrities tracy morgan
Celebrities natalie portman
Celebrities brad pitt
Celebrities tom cruise
Celebrities sylvester stallone
Celebrities tina fey
Celebrities dolph lundgren
Celebrities tony hale
Celebrities donald Glover
Celebrities paul rudd
Celebrities angelina jolie
Celebrities scarlett johansson
Celebrities david cross
Celebrities alec baldwin
Celebrities david duchovny
Celebrities will ferrell
Celebrities chris rock
Celebrities adam brody
Celebrities jennifer lawrence
Celebrities aubrey plaza
Celebrities jackie chan
Celebrities alexa chung
Celebrities ricky gervais
Celebrities jessica walter
Actions cooking
Actions fighting
Actions smiling
Actions laughing
Actions dreaming
Actions crying
Actions spinning
Actions tossing drink
Actions sleeping
Actions eating
Actions sneezing
Actions singing
Actions pout
Actions slapping
Actions finger guns
Actions running
Actions swimming

Actions falling
Actions smoking
Actions flirting
Actions dancing
Actions breaking up
Actions drinking
Actions fainting
Actions shocked
Actions bored
Actions unimpressed
Actions sick
Actions stressed
Actions nervous
Actions sad
Actions relaxed
Actions sassy
Actions tired
Actions reaction
Actions hungry
Actions scared
Actions angry
Actions drunk
Actions lonely
Actions pain
Actions excited
Actions happy
Actions surprised
Actions inspired
Actions suspicious
Actions frustrated
Actions love
Actions embarrassed
Actions disappointed
Actions hockey
Actions rugby
Actions nhl
Actions rock climbing
Actions diving
Actions formula one
Actions rowing
Actions skydiving
Actions mma
Actions lacrosse
Actions ufc
Actions volleyball
Actions softball
Actions mlb
Actions martial arts
Actions horse racing
Actions skiing
| Sports          | swimming       | Adjectives    | slow motion   |
|----------------|----------------|---------------|---------------|
| Sports         | roller skating | Adjectives    | cute          |
| Sports         | football       | Adjectives    | cold          |
| Sports         | tennis         | Adjectives    | funny         |
| Sports         | nba            | Adjectives    | weird         |
| Sports         | boxing         | Adjectives    | trippy        |
| Sports         | parkour        | Adjectives    | black and white|
| Sports         | nascar         | Adjectives    | pretty        |
| Sports         | golf           | Adjectives    | scary         |
| Art & Design   | art            | Adjectives    | creepy        |
| Art & Design   | typography     | Adjectives    | hd            |
| Art & Design   | illustration   | Animals       | lizard        |
| Art & Design   | transparent    | Animals       | meerkat       |
| Art & Design   | glitch         | Animals       | otter         |
| Art & Design   | pixel          | Animals       | cow           |
| Art & Design   | morph          | Animals       | caterpillar   |
| Art & Design   | black and white| Animals       | koala         |
| Art & Design   | geometry       | Animals       | corgi         |
| Art & Design   | collage        | Animals       | penguin       |
| Art & Design   | architecture   | Animals       | duck          |
| Art & Design   | psychedelic    | Animals       | elephant      |
| Art & Design   | 3d             | Animals       | raccoon       |
| Art & Design   | mash up        | Animals       | hippo         |
| Art & Design   | photography    | Animals       | kangaroo      |
| Art & Design   | loop           | Animals       | chicken       |
| Art & Design   | cinemagraph    | Animals       | monkey        |
| Art & Design   | sculpture      | Animals       | ferret        |
| Art & Design   | timelapse      | Animals       | seal          |
| Art & Design   | design         | Animals       | owl           |
| Art & Design   | animation      | Animals       | jellyfish     |
| Memes          | sips tea       | Animals       | bulldog       |
| Memes          | steal yo girl  | Animals       | crab          |
| Memes          | arthur         | Animals       | butterfly     |
| Memes          | crying dawson  | Animals       | giraffe       |
| Memes          | confused       | Animals       | panda         |
| Memes          | deal with it   | Animals       | pig           |
| Memes          | like a boss    | Animals       | red panda     |
| Memes          | hair flip      | Animals       | grumpy cat    |
| Memes          | forever alone  | Animals       | sheep         |
| Memes          | look at all the fucks i give | Animals | turtle |
| Memes          | cuca           | Animals       | wolf          |
| Memes          | judge judy     | Animals       | lion          |
| Memes          | feels          | Animals       | bird          |
| Memes          | fail           | Animals       | hamster       |
| Memes          | dank memes     | Animals       | polar bear    |
| Adjectives     | vintage        | Animals       | goat          |
| Adjectives     | sexy           | Animals       | whale         |
| Adjectives     | bright         | Animals       | mouse         |
| Adjectives     | dark           | Animals       | camel         |
| Adjectives     | hot            | Animals       | chihuahua     |
| Animals  | Movies          |
|----------|-----------------|
| skunk    | the dark knight |
| squirrel | citizen kane    |
| frog     | edward scissorhands |
| horse    | kill bill       |
| pug      | casablanca      |
| tiger    | pulp fiction    |
| unicorn  | terminator      |
| bear     | zoolander       |
| poodle   | bridesmaids     |
| the fifth element |      |
| the breakfast club |      |
| addams family |      |
| breakfast at tiffany |      |
| cry baby | the goonies     |
| donnie darko |      |
| waynes world | native american |
| say anything |      |
| the godfather |      |
| blue velvet |      |
| the princess bride |      |
| clueless |      |
| ghostbusters |      |
| spiderman |      |
| sixteen candles |      |
| ace ventura |      |
| the blues brothers |      |
| fight club |      |
| indiana jones |      |
| the notebook |      |
| get out |      |
| the matrix |      |
| star wars |      |
| night of the living dead |      |
| the shining |      |
| 500 days of summer |      |
| bladerunner |      |
| elf |      |
| the big lebowski |      |
| some like it hot |      |
| american psycho |      |
| easy rider |      |
| reservoir dogs |      |
| texas chainsaw massacre |      |
| the avengers |      |
| beetlejuice |      |
| labyrinth |      |
| scarf |      |
| spring breakers |      |
| rocky | Cartoons & Comics |
| pretty in pink |      |
Decades 80s Nature crystals
Decades vintage Nature forest
Decades 30s Nature sunset
Decades 60s Nature fire
Decades 50s Nature lava
Decades 70s Nature reef
Decades 40s Nature tornado
Decades 90s Nature northern lights
Decades 20s Nature landscape
Weird 80s Nature prairie
Weird vintage Nature night
Weird ghost Nature plants
Weird zombie Nature cave
Weird morph Nature trees
Weird psychedelic Nature constellations
Weird vampire Nature clouds
Weird alien Nature hurricane
Weird 90s Nature sand
Weird robot Nature mushrooms
Weird clown Nature snow
Stickers cat stickers Nature geyser
Stickers excited stickers Nature lake
Stickers love stickers Nature mountains
Stickers animated text stickers Nature smoke
Stickers emoji stickers Nature rainbow
Stickers weird stickers Music action bronson
Stickers high five stickers Music adele
Stickers birthday stickers Music frank ocean
Stickers party stickers Music kendrick lamar
Stickers cheeseburger stickers Music the beatles
Stickers happy stickers Music mc hammer
Stickers dinosaur stickers Music zayn malik
Nature sun Music nicki minaj
Nature waves Music backstreet boys
Nature wind Music lizzo
Nature river Music cl
Nature mist Music snoop dogg
Nature desert Music madonna
Nature moon Music usher
Nature waterfall Music vampire weekend
Nature stars Music the rolling stones
Nature tsunami Music g dragon
Nature coral Music jennifer lopez
Nature glacier Music janet jackson
Nature weather Music destinys child
Nature beach Music lady gaga
Nature sunrise Music jay z
Nature comet Music elvis presley
Nature ocean Music bruno mars
Nature ice Music cardi b
**C  List of selected tags from GIPHY**

| Music                  | tlc                  |
|------------------------|----------------------|
| Music                  | david bowie          |
| Music                  | coldplay             |
| Music                  | kpop                 |
| Music                  | missy elliott        |
| Music                  | solange              |
| Music                  | whitney houston      |
| Music                  | carrie underwood     |
| Music                  | shakira               |
| Music                  | britney spears       |
| Music                  | lil nas x            |
| Music                  | mariah carey         |
| Anime                  | samurai champloo     |
| Anime                  | fullmetal alchemist   |
| Anime                  | bleach               |
| Anime                  | spaceship battleship yamato |
| Anime                  | manga                |
| Anime                  | hetalia              |
| Anime                  | princess mononoke    |
| Anime                  | my neighbor totoro   |
| Anime                  | cowboy bebop         |
| Anime                  | kawaii               |
| Anime                  | kiba                 |
| Anime                  | berserk              |
| Anime                  | evangelion           |
| Anime                  | black lagoon         |
| Anime                  | inuyasha             |
| Anime                  | ninja scroll         |
| Anime                  | sakura               |
| Anime                  | hayao miyazaki       |
| Anime                  | cardcaptor sakura    |
| Anime                  | rock lee             |
| Anime                  | code geass           |
| Anime                  | kakashi hatake       |
| Anime                  | hinata hyuga         |
| Anime                  | death note           |
| Anime                  | gundam               |

**D  List of filtering keywords on Imgur experiment**

depression, depressing, mental, health, death, dead, alcohol, alcoholism, weed, drugs, addiction, covid, beer, stoned, black, white, arabic, hispanic, latino, latina, latinx, police, cop, racism, racists, race, sexism, sexist, sexy, armed, overthrow, government, republican, democrats, maga, liberal, liberals, conservative, conservatives, offender, victim, disability, disabled, jerking, PD, gun, shots, fired, cops, officer, officers, killing, murder, murdered, kill, kills, killed, murders, shoot, taser, bystander, trigger, handgun, pansexual, sexuality, homosexual, gay, lesbian, corona, virus, coronavirus, vaccine, vaccinated, viruses, vaccination, die, fascist, fascists, antifa, sharia, islam, islamic, christian, jewish, muslim, blasphemy, blasphemous, death, conviction,
church, priest, pastor, religious, religion, sharia, shia, sunni, judge, bible, quran, torah, hindu, hindus, christians, jew, jews, muslims, islamist, execute, murder, captive, captives, malpractice, insurance, insured, threat, threatening, war, troops, violence, fighting, conflict, medicine, prescription, drug, dying, hospice, life, doctor, hospital, nurse, pedophiles, pedophile, bitch, republicans, democrat, coup, tax, recession, pedo, criminal, criminals, politician, politicians, health, healthcare, america, american, voter, voting, votes, vote, voters, citizen, immigrants, immigrant, citizens, candian, canada, eu, european, trump, red, blue, cancer, slavery, slaves, slave, disease, sickness, sorry, nazi, nazis, death, pro-death, pro-life, profile, abortion, aborted, aborting, victims, jail, whore, slut, rape, raped, raping, behead, beheadings, beheaded, torture, tortured, torturing, taliban, afghanistan, soldier, soldiers, kabul
| Tag based             | CLIP variant              | PEPE                  |
|----------------------|---------------------------|-----------------------|
| 2gG2xiMTtFwsg        | IfesfEtobCSbsHzC8d        | tnYri4n2Frmig         |
| fnjvxV295sWEJywXU    | m9d3Xi3ShZ42CxlWP         | 5wWf7GR2nhgamhRaEuA   |
| BAPSj0xM1cFe8        | f9k1tV7HyORengKF8v        | 5gw0VWGbgNm8w         |
| iJsvRxNTacup6DVfLP   | loithnzQ1JQ8fizx8w        | iXTrbbYMQBCMM         |
| 3oEjHLcg4QMUMumb9m   | bfrlODg5LqXxS             | 65ODCwM00NVmEyLsX3    |
| aKrTvuo4hlKM         | 4HmjGg306HiLHWlm2f        | 26AHLBZUCln53ozzi8   |
| 3oKIPiiDN24q8Awtwc   | 7J26CGAhos6d5S1A6         | 3o8doT9BL7dgtolp7O   |
| Jhu44mYwUmSHTTW3tj   | 8hZ9FMolyKc0X8BSr7        | Fq6Bdkj3coEWQ         |
| jTrWazIFGfvYY34PSJ   | iqkHA3DmB8GjORY030        | 3oEjHAU0qG3ISS0f1C    |
| 1396L17pwHWOJtTG     | OOzcnk3pLHDHqW6s6Tb       | KzyMeEfDh4Jtw         |

Table 7: Examples of top 10 most frequently used gifs across all models in the RCT. Click an image to view the gif on Giphy. Images are ordered from most-used (top) to tenth-most (bottom).
Dependent variable: Cumulative number of replies received

gif reply score 0.096*** (0.010)
post score −0.0004*** (0.0001)
comment score 0.0002 (0.0002)
CLIP variant model −0.196 (0.152)
distribution-sampling model −0.664*** (0.160)
PEPE model −0.450*** (0.138)
Tag-based model −0.195 (0.146)
number of days after reply −0.001 (0.001)
comment text polarity 0.048 (0.164)
comment text subjectivity −0.055 (0.147)
topic 0 (Politics related) −0.275 (0.430)
topic 1 (Family & Pets related) −0.264 (0.412)
topic 2 (Employment related) −1.182** (0.549)
topic 3 (Social media related) 1.381*** (0.421)
topic 4 (Transportation related) −0.021 (0.514)
topic 5 (Food related) −0.896 (0.567)
topic 6 (COVID related) −0.459 (0.564)
topic 7 (Entertainment related) −0.529 (0.452)
topic 8 (People related) −1.776*** (0.647)
comment is a question 0.114 (0.133)
length of parent comment 0.0003 (0.007)
intercept −1.877*** (0.313)

Observations 8,369
Log Likelihood −2,466.965
θ 0.143*** (0.013)
Akaike Inf. Crit. 4,977.930

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Negative Binomial regression on cumulative number of replies received. The random-gif baseline is set as the reference category for model comparison.

| Topic | Dirichlet parameter | Keywords |
|-------|---------------------|----------|
| 0     | 0.1172              | people fuck trump shit make thing country n’t vote fucking |
| 1     | 0.20164             | good time love kid make cat dog day year guy |
| 2     | 0.09554             | pay work money people make job year buy time company |
| 3     | 0.11245             | post make read people good time thing imgur video work |
| 4     | 0.06541             | car live year drive day place time road city back |
| 5     | 0.05672             | eat make food good water drink taste cheese pizza coffee |
| 6     | 0.06662             | people covid die vaccine life make work problem mask n’t |
| 7     | 0.0888              | movie play game good watch show love great time song |
| 8     | 0.02752             | wear mask red shirt woman hair white man hat black |
| 9     | 0.14292             | back make put hand time guy car head thing big |

Table 9: Topic modeling keywords for Imgur Comments