Happy Dance, Slow Clap: Using Reaction GIFs to Predict Induced Affect on Twitter

Boaz Shmueli[1,2,3] , **Soumya Ray**[2], and **Lun-Wei Ku**[3]
[1]Social Networks and Human-Centered Computing, TIGP, Academia Sinica
[2]Institute of Service Science, National Tsing Hua University
[3]Institute of Information Science, Academia Sinica
shmueli@iis.sinica.edu.tw soumya.ray@iss.nthu.edu.tw lwku@iis.sinica.edu.tw

Abstract

Datasets with *induced emotion* labels are scarce but of utmost importance for many NLP tasks. We present a new, automated method for collecting texts along with their *induced reaction* labels. The method exploits the online use of reaction GIFs, which capture complex affective states. We show how to augment the data with *induced emotions* and *induced sentiment* labels. We use our method to create and publish ReactionGIF, a first-of-its-kind affective dataset of 30K tweets. We provide baselines for three new tasks, including induced sentiment prediction and multilabel classification of induced emotions. Our method and dataset open new research opportunities in emotion detection and affective computing.

1 Introduction

Affective states such as emotions are an elemental part of the human condition. The automatic detection of these states is thus an important task in affective computing, with applications in diverse fields including psychology, political science, and marketing (Seyeditabari et al., 2018). Training machine learning algorithms for such applications requires large yet task-specific emotion-labeled datasets (Bostan and Klinger, 2018).

Borrowing from music (Gabrielsson, 2001) and film (Tian et al., 2017), one can distinguish between two reader perspectives when labeling emotions in text: *perceived* emotions, which are the emotions that the reader recognizes in the text, and *induced* emotions, which are the emotions aroused in the reader. However, with the exception of Buechel and Hahn (2017), this distinction is mostly missing from the NLP literature, which focuses on the distinction between author and reader perspectives (Calvo and Mac Kim, 2013).

The collection of perceived emotions data is considerably simpler than induced emotions data, and presently most human-annotated emotion datasets are labeled with perceived emotions (e. g., Strapparava and Mihalcea, 2008; Preoţiuc-Pietro et al., 2016; Hsu and Ku, 2018; Demszky et al., 2020). Induced emotions data can be collected using physiological measurements or self-reporting, but both methods are complex, expensive, unreliable and cannot scale easily. Still, having well-classified induced emotions data is of utmost importance to dialogue systems and other applications that aim to detect, predict, or elicit a particular emotional response in users. Pool and Nissim (2016) used distant supervision to detect induced emotions from Facebook posts by looking at the six available emoji reactions. Although this method is automatic, it is limited both in emotional range, since the set of reactions is small and rigid, and accuracy, because emojis are often misunderstood due to their visual ambiguity (Tigwell and Flatla, 2016).

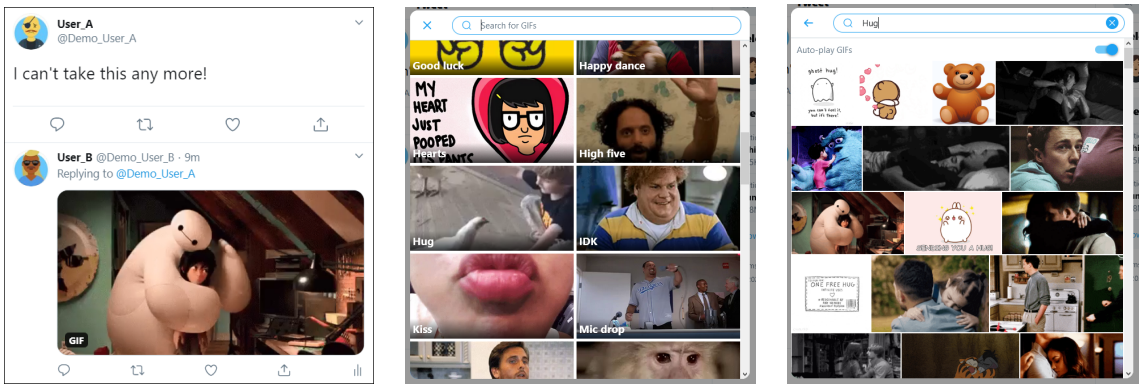
To overcome these drawbacks, we propose a new method that innovatively exploits the use of reaction GIFs in online conversations. Reaction GIFs are effective because they "display emotional responses to prior talk in text-mediated conversations" (Tolins and Samermit, 2016). We propose a fully-automated method that captures in-the-wild texts, naturally supervised using *fine-grained*, *induced reaction* labels. We also augment our dataset with sentiment and emotion labels. We use our method to collect and publish the ReactionGIF dataset.[1]

2 Automatic Supervision using GIFs

Figure 1a shows a typical Twitter thread. User *A* writes *"I can't take this any more!"*. User *B* replies with a reaction GIF depicting an embrace. Our method automatically infers a *hug* reaction, signaling that *A*'s text induced a feeling of love and caring. In the following, we formalize our method.

[1] github.com/bshmueli/ReactionGIF

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 395–401
August 1–6, 2021.



(a) A root tweet ("I can't take...") with a *hug* reaction GIF reply.

(b) The reaction categories menu offers 43 categories (*hug*, *kiss*, ...).

(c) The top reaction GIFs offered to the user from the *hug* category.

Figure 1: How reaction GIFs are used (left) and inserted (middle, right) on Twitter.

2.1 The Method

Let (t, g) represent a 2-turn online interaction with a root post comprised solely of text t, and a reply containing only reaction GIF g. Let $R = \{R_1, R_2, ..., R_M\}$ be a set of M different *reaction categories* representing various affective states (e. g., *hug*, *facepalm*). The function $\mathfrak{R}$ maps a GIF g to a reaction category, $g \mapsto \mathfrak{R}(g)$, $\mathfrak{R}(g) \in R$. We use $r = \mathfrak{R}(g)$ as the label of t. In the Twitter thread shown in Figure 1a, the label of the tweet *"I can't take this any more!"* is $r = \mathfrak{R}(g) =$ *hug*.

Inferring $\mathfrak{R}(g)$ would usually require humans to manually view and annotate each GIF. Our method automatically determines the reaction category conveyed in the GIF. In the following, we explain how we automate this step.

GIF Dictionary We first build a dictionary of GIFs and their reaction categories by taking advantage of the 2-step process by which users post reaction GIFs. We describe this process on Twitter; other platforms follow a similar approach:

Step 1: The user clicks on the GIF button. A menu of reaction categories pops up (Figure 1b). Twitter has 43 pre-defined categories (e. g., *high five*, *hug*). The user clicks their preferred category.

Step 2: A grid of reaction GIFs from the selected category is displayed (Figure 1c). The user selects one reaction GIF to insert into the tweet.

To compile the GIF dictionary, we collect the first 100 GIFs in each of the $M = 43$ reaction categories on Twitter. We save the 4300 GIFs, along with their categories, to the dictionary. While in general GIFs do not necessarily contain affective information, our method collects *reaction* GIFs that depict corresponding affective states.

Computing $\mathfrak{R}(g)$ Given a (t, g) sample, we label text t with reaction category r by mapping reaction GIF g back to its category $r = \mathfrak{R}(g)$. We search for g in the GIF dictionary and identify the category(ies) in which it is offered to the user. If the GIF is not found, the sample is discarded. For the small minority of GIFs that appear in two or more categories, we look at the positions of the GIF in each of its categories and select the category with the higher position.

2.2 Category Clustering

Because reaction categories represent overlapping affective states, a GIF may appear in multiple categories. For example, a GIF that appears in the *thumbs up* category may also appear in the *ok* category, since both express approval. Out of the 4300 GIFs, 408 appear in two or more categories. Exploiting this artefact, we propose a new metric: the pairwise *reaction similarity*, which is the number of reaction GIFs that appear in a pair of categories.

To automatically discover affinities between reaction categories, we use our similarity metric and perform hierarchical clustering with average linkage. The resulting dendrogram, shown in Figure 2, uncovers surprisingly well the relationships between common human gesticulations. For example, *shrug* and *idk* (**I d**on't **k**now) share common emotions related to uncertainty and defensiveness. In particular, we can see two major clusters capturing negative sentiment (left cluster: *mic drop* to *smh* [**s**hake **m**y **h**ead]) and positive sentiment (right cluster: *hug* to *slow clap*), which are useful for downstream sentiment analysis tasks. The two rightmost singletons, *popcorn* and *thank you*, lack sufficient similarity data.

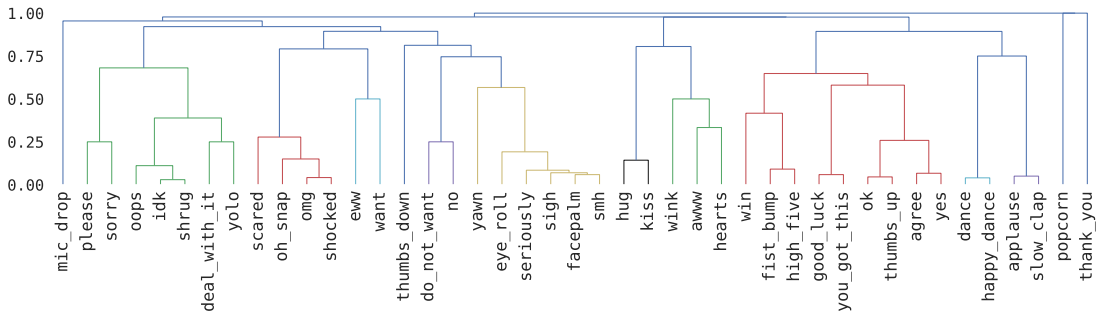


Figure 2: Hierarchical clustering (average linkage) of reaction categories shows relationships between reactions.

3 ReactionGIF Dataset

We applied our proposed method to 30K English-language (t, g) 2-turn pairs collected from Twitter in April 2020. t are text-only root tweets (not containing links or media) and g are pure GIF reactions. We label each tweet t with its reaction category $r = \mathfrak{R}(g)$. See Appendix A for samples. The resulting dataset, ReactionGIF, is publicly available.

Figure 3 shows the category distribution's long tail. The top seven categories (*applause* to *eye-roll*) label more than half of the samples (50.9%). Each of the remaining 36 categories accounts for between 0.2% to 2.8% of the samples.

Label Augmentation Reaction categories convey a rich affective signal. We can thus augment the dataset with other affective labels. We add **sentiment labels** by using the positive and negative reaction category clusters, labeling each sample according to its cluster's sentiment (§2.2). Furthermore, we add **emotion labels** using a novel reactions-to-emotions mapping: we asked 3 annotators to map each reaction category onto a subset of the 27 emotions in Demszky et al. (2020) — see Table 1. Instructions were to view the GIFs in each category and select the expressed emotions. Pairwise Cohen's kappa indicate moderate interrater agreements with $\kappa_{12} = 0.512$, $\kappa_{13} = 0.494$, $\kappa_{23} = 0.449$, and Fleiss' kappa $\kappa_F = 0.483$. We use the annotators' majority decisions as the final many-to-many mapping and label each sample according to its category's mapped emotions subset.

GIFs in Context As far as we know, our dataset is the first to offer reaction GIFs with their eliciting texts. Moreover, the reaction GIFs are labeled with a reaction category. Other available GIF datasets (TGIF by Li et al., 2016, and GIFGIF/GIFGIF+, e. g., Jou et al., 2014) lack both the eliciting texts and the reaction categories.

Admiration	Curiosity	Fear	Pride
Amusement	Desire	Gratitude	Realization
Anger	Disappointment	Grief	Relief
Annoyance	Disapproval	Joy	Remorse
Approval	Disgust	Love	Sadness
Caring	Embarrassment	Nervousness	Surprise
Confusion	Excitement	Optimism	

Table 1: The 27 emotions in Demszky et al. (2020).

4 Baselines

As this is the first dataset of its kind, we aim to promote future research by offering baselines for predicting the reaction, sentiment, and emotion induced by tweets. We use the following four models in our experiments:

- **Majority**: A simple majority class classifier.
- **LR**: Logistic regression classifier (L-BFGS solver with $C = 3$, maximum iterations 1000, stratified K-fold cross validation with $K = 5$) using TF-IDF vectors (unigrams and bigrams, cutoff 2, maximum 1000 features, removing English-language stop words).
- **CNN**: Convolutional neural network (100 filters, kernel size 3, global max pooling; 2 hidden layers with 0.2 dropout; Adam solver, 100 epochs, batch size 128, learning rate 0.0005) with GloVe embeddings (Twitter, 27B tokens, 1.2M vocabulary, uncased, 100d) (Pennington et al., 2014).
- **RoBERTa**: Pre-trained transformer model (base, batch size 32, maximum sequence length 96, 3 training epochs) (Liu et al., 2019).

We hold out 10% of the samples for evaluation. The code is publicly available along with the dataset for reproducibility. The experiment results are summarized in Table 2.

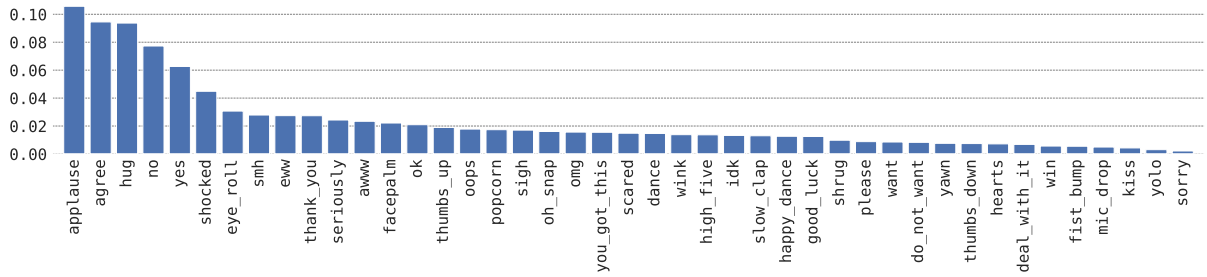


Figure 3: Distribution of the 43 reaction categories in ReactionGIF

Task →	Reaction				Sentiment				Emotion
Model ↓	Acc	P	R	F_1	Acc	P	R	F_1	LRAP
Majority	10.4	1.1	10.4	2.0	58.0	33.7	58.0	42.6	0.445
LR	22.7	19.5	22.7	18.0	64.7	64.4	64.7	62.4	0.529
CNN	25.5	17.3	25.5	19.1	67.1	66.8	67.1	66.3	0.557
RoBERTa	**28.4**	**23.6**	**28.4**	**23.9**	**70.0**	**69.7**	**70.0**	**69.8**	**0.596**

Table 2: Baselines for the reaction, sentiment, and emotion classification tasks. All metrics are weight-averaged. The highest value in each column is emboldened.

Affective Reaction Prediction is a multiclass classification task where we predict the reaction category r for each tweet t. RoBERTa achieves a weight-averaged F_1-score of 23.9%.

Induced Sentiment Prediction is a binary classification task to predict the sentiment induced by tweet t by using the augmented labels. RoBERTa has the best performance with accuracy 70.0% and F_1-score of 69.8%.

Finally, **Induced Emotion Prediction** uses our reaction-to-emotion transformation for predicting emotions. This is a 27-emotion *multilabel* classification task, reflecting our dataset's unique ability to capture complex emotional states. RoBERTa is again the best model, with Label Ranking Average Precision (LRAP) of 0.596.

5 Discussion

Reaction GIFs are ubiquitous in online conversations due to their uniqueness as lightweight and silent moving pictures. They are also more effective and precise[2] when conveying affective states compared to text, emoticons, and emojis (Bakhshi et al., 2016). Consequently, the reaction category is a new type of label, not yet available in NLP emotion datasets: existing datasets use either the discrete emotions model (Ekman, 1992) or the dimensional model of emotion (Mehrabian, 1996). The new labels possess important advantages, but also present interesting challenges.

Advantages The new reaction labels provide a rich, complex signal that can be mapped to other types of affective labels, including sentiment, emotions and possibly feelings and moods. In addition, because reaction GIFs are ubiquitous in online conversations, we can automatically collect large amounts of inexpensive, naturally-occurring, high-quality affective labels. Significantly, and in contrast with most other emotion datasets, the labels measure *induced* (as opposed to *perceived*) affective states; these labels are of prime importance yet the most difficult to obtain, with applications that include GIF recommender systems, dialogue systems, and any other application that requires predicting or inducing users' emotional response.

Challenges The large number of reaction categories (reflecting the richness of communication by gestures) makes their prediction a challenging task. In addition, the category distribution has a long tail, and there is an affective overlap between the categories. One way to address these issues is by accurately mapping the reactions to emotions. Precise mapping will require a larger GIF dictionary (our current one has 4300 GIFs), a larger dataset, and new evaluation metrics. A larger GIF dictionary will also improve the *reaction similarity*'s accuracy, offering new approaches for studying relationships between reactions (§2.2).

[2]For example, the *facepalm* reaction is "a gesture in which the palm of one's hand is brought to one's face, as an expression of disbelief, shame, or exasperation.", Oxford University Press, lexico.com/en/definition/facepalm

6 Conclusion

Our new method is the first to exploit the use of reaction GIFs for capturing in-the-wild *induced* affective data. We augment the data with induced sentiment and emotion labels using two novel mapping techniques: reaction category clustering and reactions-to-emotions transformation. We used our method to publish ReactionGIF, a first-of-its-kind dataset with multiple affective labels. The new method and dataset offer opportunities for advances in emotion detection.

Moreover, our method can be generalized to capture data from other social media and instant messaging platforms that use reaction GIFs, as well as applied to other downstream tasks such as multimodal emotion detection and emotion recognition in dialogues, thus enabling new research directions in affective computing.

Acknowledgements

We thank the anonymous reviewers for their comments and suggestions. Special thanks to Thilina Rajapakse, creator of the elegant Simple Transformers package, for his help.

This research was partially supported by the Ministry of Science and Technology in Taiwan under grants MOST 108-2221-E-001-012-MY3 and MOST 109-2221-E-001-015- [*sic*].

Ethical Considerations and Implications

Data Collection

The ReactionGIF data was collected from Twitter using the official API in full accordance with their Development Agreement and Policy (Twitter, 2020). Similar to other Twitter datasets, we include the tweet IDs but not the texts. This guarantees that researchers who want to use the data will also need to agree with Twitter's Terms of Service. It also ensures compliance with section III (Updates and Removals) of the Developer Agreement and Policy's requirement that when users delete tweets (or make them private), these changes are reflected in the dataset (Belli et al., 2020).

Annotation

Annotation work was performed by three adult students, two males and one female, who use social media regularly. The labeling involved viewing 43 sets of standard reaction GIFs, one for each reaction category. These reaction GIFs are the standard offering by the Twitter platform to all its users. As a result, this content is highly familiar to users of social media platforms such as Facebook or Twitter, and thus presents a very low risk of psychological harm. Annotators gave informed consent after being presented with details about the purpose of the study, the procedure, risks, benefits, statement of confidentiality and other standard consent items. Each annotator was paid US$18. The average completion time was 45 minutes.

Applications

The dataset and resulting models can be used to infer readers' induced emotions. Such capability can be used to help online platforms detect and filter out content that can be emotionally harmful, or emphasize and highlight texts that induce positive emotions with the potential to improve users' well-being. For example, when a person is in grief or distress, platforms can give preference to responses which will induce a feeling of caring, gratitude, love, or optimism. Moreover, such technology can be of beneficial use in assistive computing applications. For example, people with emotional disabilities can find it difficult to understand the emotional affect in stories or other narratives, or decipher emotional responses by third parties. By computing the emotional properties of texts, such applications can provide hints or instructions and provide for smoother and richer communication. However, this technology also has substantial risks and peril. Inducing users' affective response can also be used by digital platforms in order to stir users into specific action or thoughts, from product purchase and ad clicking to propaganda and opinion forming. Deployers must ensure that users understand and agree to the use of such systems, and consider if the benefit created by such systems outweigh the potential harm that users may incur.

References

Saeideh Bakhshi, David A. Shamma, Lyndon Kennedy, Yale Song, Paloma de Juan, and Joseph 'Jofish' Kaye. 2016. Fast, cheap, and good: Why animated GIFs engage us. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 575–586, New York, NY, USA. Association for Computing Machinery.

Luca Belli, Sofia Ira Ktena, Alykhan Tejani, Alexandre Lung-Yut-Fong, Frank Portman, Xiao Zhu, Yuanpu Xie, Akshay Gupta, Michael M. Bronstein, Amra Delić, Gabriele Sottocornola, Vito Walter Anelli,

Nazareno Andrade, Jessie Smith, and Wenzhe Shi. 2020. Privacy-preserving recommender systems challenge on Twitter's home timeline. *CoRR*, abs/2004.13715.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

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

Alf Gabrielsson. 2001. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(Special Issue: Current Trends in the Study of Music and Emotion):123–147.

Chao-Chun Hsu and Lun-Wei Ku. 2018. SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 27–31, Melbourne, Australia. Association for Computational Linguistics.

Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting viewer perceived emotions in animated GIFs. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 213–216, New York, NY, USA. Association for Computing Machinery.

Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, Los Alamitos, CA, USA. IEEE Computer Society.

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

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan. The COLING 2016 Organizing Committee.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, page 1556–1560, New York, NY, USA. Association for Computing Machinery.

Leimin Tian, Michal Muszynski, Catherine Lai, Johanna D. Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. 2017. Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same? In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 28–35.

Garreth W. Tigwell and David R. Flatla. 2016. Oh that's what you meant!Reducing emoji misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '16, page 859–866, New York, NY, USA. Association for Computing Machinery.

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

Twitter. 2020. Developer Agreement and Policy. `https://developer.twitter.com/developer-terms/agreement-and-policy`. (Accessed on 02/01/2021).

Record ID	Tweet	GIF Response	Reaction Category	Sentiment	Emotions
13241	“so...I have a job now 😊”		dance	positive	Amusement, Excitement, Joy
1320	“dyed my hair..... Pics soon”		applause	positive	Admiration Approval Excitement Gratitude Surprise
17	“Don't forget to Hydrate!”		yawn	negative	Disappointment Disapproval
808	“Folks, I have a BIG BIG announcement coming tomorrow night at 9 PM EST”		scared	negative	Confusion Fear Nervousness Surprise

Figure 4: ReactionGIF samples.

A Dataset Samples

Figure 4 includes four samples from the dataset. For each sample, we show the record ID within the dataset, the text of the tweet, a thumbnail of the reaction GIF, the reaction category of the GIF, and the two augmented labels: the sentiment and the emotions.