Sensing and Learning Human Annotators Engaged in Narrative Sensemaking

McKenna K. Tornblad,[1] Luke Lapresi,[2] Christopher M. Homan,[2]
Raymond W. Ptucha[3] and Cecilia Ovesdotter Alm[4]
[1]College of Behavioral and Health Sciences, George Fox University
[2]Golisano College of Computing and Information Sciences, Rochester Institute of Technology
[3]Kate Gleason College of Engineering, Rochester Institute of Technology
[4]College of Liberal Arts, Rochester Institute of Technology
mtornblad14@georgefox.edu, lxl6996@rit.edu, cmh@cs.rit.edu,
rwpeec@rit.edu, coagla@rit.edu

Abstract

While labor issues and quality assurance in crowdwork are increasingly studied, how annotators make sense of texts and how they are personally impacted by doing so are not. We study these questions via a narrative-sorting annotation task, where carefully selected (by sequentiality, topic, emotional content, and length) collections of tweets serve as examples of everyday storytelling. As readers process these narratives, we measure their facial expressions, galvanic skin response, and self-reported reactions. From the perspective of annotator well-being, a reassuring outcome was that the sorting task did not cause a measurable stress response, however readers reacted to humor. In terms of sensemaking, readers were more confident when sorting sequential, target-topical, and highly emotional tweets. As crowdsourcing becomes more common, this research sheds light onto the perceptive capabilities and emotional impact of human readers.

1 Introduction

A substantial sector of the gig economy is the use of crowdworkers to annotate data for machine learning and analysis. For instance, storytelling is an essential human activity, especially for information sharing (Bluvshtein et al., 2015), making it the subject of many data annotation tasks. Microblogging sites such as Twitter have reshaped the narrative format, especially through character restrictions and nonstandard language, elements which contribute to a relatively unexplored mode of narrative construction; yet, little is known about reader responses to such narrative content.

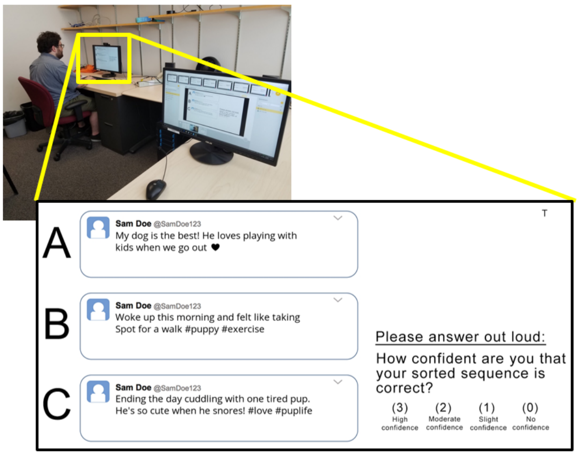
We explore reader reactions to narrative sensemaking using a new sorting task that varies the presumed cognitive complexity of the task and elicits readers' interpretations of a target topic and its emotional tone. We carefully and systematically extracted 60 sets of tweets from a 1M-tweet dataset and presented them to participants via an interface that mimicks the appearance of Twitter (Figure 1). We asked subjects to sort chronologically the tweets in each set, half with true narrative sequence and half without. Each set consisted of 3–4 tweets from a previously collected corpus, where each tweet was labeled using a framework by Liu et al. (2016) as work-related or not. In addition to sequentiality and job-relatedness, sets were evenly distributed across two other variables with two levels each, of interest for understanding narrative sensemaking (Table 1). We recorded readers' spoken responses to four questions (Figure 2) about each set, which involved the sorting task, reader confidence, topical content (job-relatedness), and emotional tone. We used galvanic skin response (GSR) and facial expression analysis to explore potentially quantifiable metrics for stress-based reactions and other aspects of reader-annotator response. Our results add understanding of how annotators react to and process everyday microblog narratives.

This study makes these contributions:
1) Opens a dialogue on annotator well-being;
2) Presents a method to study annotator reactions;
3) Indicates that narrative sorting (task with degrees of complexity) does not cause an increased stress response as task complexity increases;
4) Studies the role of topic saliency and emotional tone in narrative sense-making; and
5) Probes how features model annotator reactions.

2 Related Work

That annotation tasks cause fatigue is widely recognized (Medero et al., 2006), and crowdsourcing

Proceedings of NAACL-HLT 2018: Student Research Workshop, pages 136–143
New Orleans, Louisiana, June 2 - 4, 2018. ©2017 Association for Computational Linguistics



Sam Doe @SamDoe123
My dog is the best! He loves playing with kids when we go out ♥

Sam Doe @SamDoe123
Woke up this morning and felt like taking Spot for a walk #puppy #exercise

Sam Doe @SamDoe123
Ending the day cuddling with one tired pup. He's so cute when he snores! #love #puplife

Please answer out loud:

How confident are you that your sorted sequence is correct?

(3) High confidence (2) Moderate confidence (1) Slight confidence (0) No confidence

Figure 1: The experimental setup showing the experimenter's screen, a view of the participant on a second monitor, and a closeup view of the participant's screen. A fictional tweet set is shown to maintain privacy.

1. What is the correct chronological order of these tweets?
2. How confident are you that your sorted sequence is correct?
3. Are these tweets about work? [Target topic] If so, are they about starting or leaving a job?
4. What is the dominant emotion conveyed in this tweet set?

Figure 2: The four study questions that participants answered for each tweet set in both trials.

annotations has been critiqued for labor market issues, etc. (Fort et al., 2011). Another important concern is how stressful or cognitively demanding annotation tasks may adversely impact annotators themselves, a subject matter that has not yet been prominently discussed, despite its obvious ethical implications for the NLP community and machine learning research.

Microblogging sites raise unique issues around sharing information and interacting socially. Compared to traditional published writing, Twitter users must choose language that conveys meaning while adhering to brevity requirements (Barnard, 2016). This leads to new narrative practices, such as more abbreviations and nonstandard phraseology, which convey meaning and emotion differently (Mohammad, 2012). Tweets are also published in real time and the majority of users make their tweets publicly available, even when their subject matter is personal (Marwick and Boyd, 2011). Stories are shared among—and everyday narrative topics are explored by—communities, providing a window into the health and well-being of both online and offline networks at community levels.

We selected job-relatedness as our *Target Topic* variable. Employment plays a prominent role in the daily lives and well-being of adults and is a popular theme in the stories microbloggers share of their lives. One study reported that slightly more than a third of a sample of nearly 40,000 tweets by employed people were work-related (van Zoonen et al., 2015). Employees can use social media to communicate with coworkers or independently to share their experiences (Madsen, 2016; van Zoonen et al., 2014), providing many interesting job-related narratives. Because it is so common in on- and off-line storytelling, annotators can relate to narratives about work. Work is also a less stressful topic, compared to other ones (Calderwood et al., 2017). Tweets with high or low emotional content in general may also affect readers in ways that change how they understand texts (Bohn-Gettler and Rapp, 2011). We used the sentiment functionality of TextBlob[1] to distinguish high- and low-emotion tweets in the study.

Variable	Lvl 1	Abbr.	Lvl 2	Abbr.
Narrative Seq.	yes	seq+	no	seq-
Target Topic	yes	job+	no	job-
Emo. Content	high	emo+	low	emo-
Num. Tweets	3		4	

Table 1: Overview of the four study variables that characterize each tweet set. Lvl 1 indicates the first condition for a variable, and Lvl 2 indicates the second. The primary distinction is between sets in Trial 1 (all seq+) and 2 (all seq-).

3 Study Design

The study involved 60 total tweet sets across two trials, reflecting a presumed increase in the cognitive complexity of the narrative sorting task. Trial 1 consisted of sets with a true narrative order (seq+) and Trial 2 of sets without (seq-); with the assumption that the latter is more cognitively demanding. The tweets were evenly distributed across the other three study variables (Table 1).

[1] https://textblob.readthedocs.io

For selecting job+ sets in Trial 1, queries of the corpus combined with keywords (such as *coworker*, *interview*, *fired* and other employment-related terms) identified job-related seed tweets while additional querying expanded the set for the same narrator. For example, a 3-item narrative could involve interviewing for a job, being hired, and then starting the job. For the job- sets in Trial 1, queries for seed tweets focused on keywords for other life events that had the potential to contain narratives (such as *birthday*, *driver's license*, and *child*), continuing with the same method used for job+ sets. For each Trial 2 set, we conducted similar keyword queries, except that we chose the seed tweet without regard to its role in any larger narrative. We selected the rest of the tweets in the set to match and be congruent with the same user and job+/job- and emo+/emo- classes as the seed tweet. The final selection of 60 tweet sets was based on careful manual inspection.

4 Methods

Participants: Participants were nineteen individuals (21% women) in the Northeastern U.S. who ranged in age from 18 to 49 years ($M = 25.3$). 58% were native English speakers, and active Twitter users made up 42% of the sample.

Measures: *Galvanic Skin Response (GSR):* We used a Shimmer3 GSR sensor with a sampling rate of 128 Hz to measure participants' skin conductance in microsiemens (μS). More cognitively difficult tasks may induce more sweating, and higher μS, corresponding to a decrease in resistance and indicating cognitive load (Shi et al., 2007). We used GSR peaks, as measured by iMotions software (iMotions, 2016), to estimate periods of increased cognitive arousal.

Facial Expression: To also differentiate between positive and negative emotional arousal, we captured and analyzed readers' expressed facial emotions while interacting with the task using Affectiva's facial expression analysis in iMotions (McDuff et al., 2016; iMotions, 2016).

Procedure: Participants sat in front of a monitor with a Logitech C920 HD Pro webcam and were fitted with a Shimmer3 GSR sensor worn on their non-dominant hand. They then completed a demographic pre-survey while the webcam and GSR sensor were calibrated within the iMotions software. Next, participants completed a practice trial with an unsorted tweet set on the left-hand side of the screen and questions on the right (Figure 1).

Participants answered each question aloud, minimizing movement, and continued to the next question by pressing a key. We provided a visual of Plutchik's wheel of emotions (Plutchik, 2001) if the participant needed assistance in deciding on an emotion for Question 4 (Figure 2), although they were free to say any emotion word. After the practice trial, the participant completed Trial 1 (seq+), followed by a short break, then Trial 2 (seq-). Finally, we gave each participant a post-study survey on their self-reported levels of comfort with the equipment and of cognitive fatigue during each trial.

One experimenter was seated behind the participant in the experiment room with a separate video screen to monitor the sensor data and answer any questions. A second experimenter was in another room recording the participants' answers as they were spoken via remote voice and screen sharing.

Classifier: Pairing each tweet set with each participant's response data as a single item yields 1140 data items (19 subjects * 60 sets). We used the LIBSVM (Chang and Lin, 2011) support vector machine (SVM) classifier library, with a radial basis function kernel and 10-fold cross-validation, to predict variables of interest in narrative sensemaking. Table 2 lists observed and predicted features. When a feature served as a prediction target it was not used as an observed one. The mean value of the feature was used in cases of missing data points.

Study	Self-Report	Sensing
Narrative Seq.	Kendall τ Dist.	Facial Expression
Emo. Content	Confidence	Average GSR
Target Topic	Topic Judgment	Num. GSR Peaks
Num. Tweets .	Dom. Emotion	Time
	Twitter User	
	Soc. Media Use	

Table 2: Features used for a SVM classifier were put into three groups: 1) Study variables; 2) Subject self-reported measures from participants' answers to study questions and pre-survey responses related to social media use; and 3) Sensing attributes relating to collected sensor and time data.

5 Results and Discussion

Trial Question 1: Tweet Sorting To quantify how close participants' sorted orders were to the correct narrative sequence for each tweet set in Trial 1, we used the Kendall τ rank distance between the two, or the total number of pairwise swaps of adjacent items needed to transform one sequence into the other. An ANOVA revealed that participants were significantly closer to the correct narrative order when tweets were job-related (job+), compared to not job-related (job-), regardless of emotional content, $F(1, 566) = 13.30, p < .001$ (Figure 3a). This result indicates that the target topic was useful for temporally organizing parts of stories. Without this topic's frame to start from, readers were less accurate in the sorting task.

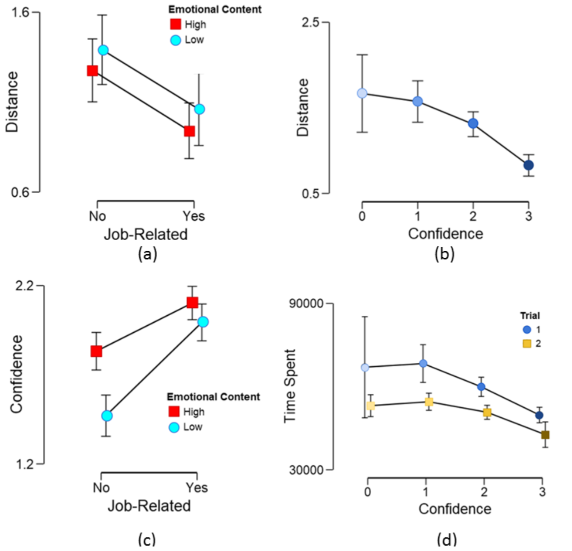


Figure 3: The four panels show: (a) Average Kendall τ rank distance between participants' orders and the correct order by job-relatedness (target topic) and emotional content in Trial 1 only. Participants were significantly more accurate with sets about the target topic, regardless of emotional content. (b) Average Kendall τ rank distance between participants' sorted orders and the correct order by confidence level in Trial 1. $0 =$ *No Confidence* and $3 =$ *High Confidence*. Participants tended to be closer to the correct order as their confidence increased. (c) Average confidence ratings across both trials by job-relatedness and emotional content. The only non-significant difference was between Job-Related (job+) x High Emotional Content (emo+), and Job-Related (job+) x Low Emotional Content (emo-) groups (see Table 3), indicating an interaction between target topic and emotional content. (d) Average time spent (ms) per tweet set by confidence and trial. Participants spent overall less time on sets in Trial 2 than Trial 1, and less time for sets in both trials as confidence increased.

Trial Question 2: Confidence An ANOVA showed that as participants became more confident in Trial 1, their sorted tweet orders were closer to the correct order, $F(3, 566) = 14.72, p < .001$. This demonstrated that readers are able to accurately estimate their correctness in the sorting task when tweets had a true narrative sequence (Figure 3b). This is an interesting result suggesting that participants were not under- or overconfident but self-aware of their ability to complete the sorting task. An ANOVA also showed that participants were significantly more confident in Trial 1 ($M = 2.20, SD = 0.79$) than Trial 2 ($M = 1.50, SD = 0.92$), $F(1, 1138) = 188.00, p < .001$, regardless of other study variables (Figure 4). This indicates that it was more difficult for participants to assign a sorted order when the tweets did not have one.

Grp 1	Grp 2	t	p	SE	d
Y × H	Y × L	1.646	.101	0.066	0.097
Y × H	N × H	4.378	<.001*	0.062	0.259
Y × H	N × L	10.040	<.001*	0.063	0.595
Y × L	N × H	2.541	.012†	0.064	0.151
Y × L	N × L	8.030	<.001*	0.065	0.476
N × H	N × L	5.750	<.001*	0.063	0.341

Table 3: Student's t-test indicates differences in confidence by job-relatedness and emotional content (visualized in Figure 3c) with df=284 for all conditions. Groups are named with the following convention: Job-Relatedness x Emotional Content. Y = Yes (job+), N = No (job-), H = High (emo+), and L = Low (emo-). (* $p < .001$ † $p < .05$).

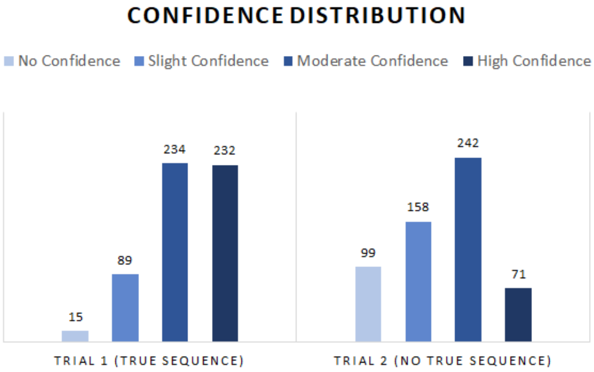


Figure 4: Total instances of each confidence level by trial (0 = *No Confidence* and 3 = *High Confidence*).

Because each participant's confidence scores for tweet sets in Trial 1 were not related to scores in Trial 2, we used an independent samples t-test and found that confidence was significantly higher for job+ tweets ($M = 2.05, SD = 0.84$) than job- tweets ($M = 1.65, SD = 0.97$), $t(1138) = -7.38, p < .001$. By including emotional content in an ANOVA, we also found that participants were significantly more confident about emo+ tweets compared to emo-, but only when the topic was not job-related (job-), $F(1, 1136) = 5.65, p = .018$. Comparisons among groups these groups can be seen in Table 3 and Figure 3c.

These findings suggest that the target topic, having a known narrative frame, was the most useful piece of information for participants in ordering tweets. When there was no job narrative to guide interpretation, emotional context was instead used as the primary sorting cue. This result agrees with previous work (Liu et al., 2016) by suggesting that the job life cycle involves a well-known narrative sequence that is used as a reference point for sorting tweets chronologically.

Confidence level could indicate cognitive load, which is supported by the ANOVA result that participants spent less time for each tweet set as their confidence increased, regardless of trial, $F(3, 1103) = 16.71, p < .001$ (Figure 3d). This suggests that participants spent less time on sets that appeared to be easier to sort and more time on sets that were perceived as more difficult. This could be a promising factor to predict how straightforward sorting different texts will be based on readers' confidence and time spent.

However, participants also spent significantly more time for each tweet set in Trial 1 than Trial 2 regardless of confidence level, $F(1, 1103) = 29.53, p < .001$ (Figure 3d), even though Trial 2 was expected to be more difficult and thus time-consuming. This contrary result could be for several reasons. First, participants may have simply got faster through practice. Second, they may have been fatigued from the task and sped up to finish. Lastly, participants could have given up on the task more quickly because it was more difficult, an indicator of cognitive fatigue.

Trial Question 3: Target Topic Participants correctly inferred the topic as job-related 97.5% of the time (true positive rate) and not job-related 96.3% (true negative rate) of the time. For more ambiguous sets, we observed that participants added qualifiers to their answers more often, such as "*these tweets could be about starting a job*," but were still able to determine the topic. These observations indicate that readers perform well when determining the topical context of tweets despite the format peculiarities that accompany the text.

Trial Question 4: Dominant Emotion If participants gave more than one word to answer this question, we used only the first to capture the participant's initial reaction to the text. We categorized words according to Plutchick's wheel of emotions (2001), and used the fundamental emotion in each section of the wheel (such as *joy* and *sadness*) as the category label. Emotion words in between each section were also used as labels and a neutral category was added, resulting in 17 categories explored in classification (below).

Sensor Data Analysis We averaged GSR readings across all four study questions for each tweet set. Because participants spent minutes answering questions 1 and 4 compared to mere seconds on 2 and 3, we chose to examine overall readings for each set instead of individual questions. An extremely short timeframe yields fewer data points and more impact in the event of poor measurement. We were also more interested in differences between the study variables of the tweet sets (Table 1) rather than differences by question, although this could be a direction for future data analysis and research.

Interestingly, an ANOVA indicated no differences in overall normalized GSR levels or number of peaks across any of the four study variables, which leads to important insights. First, it suggests that annotators may not feel more stressed when trying to sort texts that have no true narrative sequence. This could be because piecing together narratives is so fundamental and natural in human interaction and cognition. Second, when the focus is the sorting task, it appears that engaging with tweet sets with high emotional content—whether about or not about the target topic—did not elicit a greater stress response in readers. This adds more understanding to previous research (Calderwood et al., 2017) on emotional and physiological responses to texts.

We analyzed facial expressions using Affectiva (McDuff et al., 2016), obtaining the total number of instances per minute of each facial expression (anger, contempt, disgust, fear, joy, sadness, and surprise) for each tweet set by participant. We recorded the emotion that was displayed most often by a participant for a given tweet set (*neutral* if no emotion registered over Affectiva's threshold of 50). Disgust and contempt appeared to be the primary emotions displayed across all narratives regardless of topic or emotional content.We observed that participants' neutral resting faces tended to be falsely registered as disgust and contempt. Other factors such as skin tone and glasses influenced facial marker tracking.

By observation it was clear that participants had reactions (Figure 5) to tweet sets that were funny, including complex forms of humor such as irony and sarcasm. This is an important finding because annotating humor is a task that human readers are very good at compared to machines. The facial response to humor when reading texts could be used in a machine learning-based model to identify and classify humorous content.

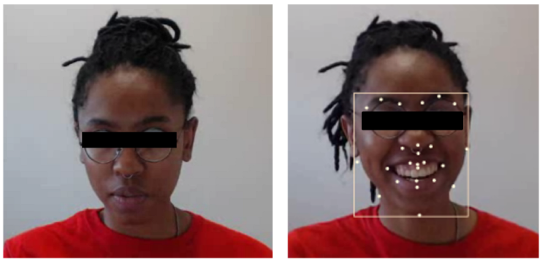


Figure 5: Participant displaying a neutral facial expression followed by a joy reaction to a humorous tweet.

Classification Results To further study the collected human data and explore patterns of interest in the data, we used SVMs to model a tweet set's *Emotional Content* (emo+/-) and *Narrative Sequence* (seq+/-); and a participant's *Confidence* and self-reported *Dominant Emotion* (Dom. Emo.) (see Table 4). The label set differed for each classification problem, ranging from binary (emo+/- and seq+/-) to multiple, less balanced classes (4 for confidence and 17 for dominant emotion). When the label being predicted was also part of the observed feature group used to build the classifier, it was excluded from the feature set.

Table 4 displays the accuracy for these classifiers, using various combinations of the features sets from Table 2. It also shows, for each variable modeled, results from a *Good-3* selection of features. This approach uses top-performing features via exhaustive search over all valid combinations of three. The same Good-3 features (Table 5) were used for all trials for a classification problem in Table 4.

Feat.	Conf.			Emo+/-			Dom. Emo.			Seq+/-
	T1	T2	C	T1	T2	C	T1	T2	C	C
Leave 1-subject out cross-validation										
Study	**49**	41	**45**	53	53	53	32	**34**	33	53
Rep.	48	35	43	65	62	59	28	28	26	67
Sens.	43	39	37	54	56	52	26	22	25	49
All	46	38	42	62	**64**	62	**35**	32	**34**	68
Good-3	**49**	**42**	**45**	**71**	62	**66**	34	30	32	**70**
Leave 1-question out cross-validation										
Study	46	41	44	27	27	53	23	19	21	53
Rep.	47	40	42	**63**	**58**	43	21	22	20	64
Sens.	45	**43**	39	45	52	44	24	21	25	51
All	47	41	45	39	44	42	26	19	23	63
Good-3	**50**	41	**46**	55	31	**57**	**34**	**25**	**30**	**67**
Leave 1-subject-and-question out cross-validation										
Study	49	**45**	**47**	27	27	27	**35**	**36**	**34**	27
Rep.	46	41	43	64	57	56	24	27	25	68
Sens.	46	42	39	52	55	50	26	24	25	59
All	50	42	**47**	57	61	60	**35**	31	**34**	**74**
Good-3	**51**	42	46	**71**	**62**	**66**	34	33	33	70

Table 4: Classification accuracies rounded to nearest percent for Trial 1 (T1), Trial 2 (T2) and both trials combined (C). Bold values indicate the most accurate feature set's prediction percentage per trial or combined. Because trials 1 and 2 differed in having a true narrative sequence, no seq+/- prediction is reported by trial.

Often, either All or Good-3 sets result in higher performance. Confidence and emo+/- classification improves performance with Trial 1 classification; however, since the dataset is modest in size, it is difficult to make a judgment as to whether or not presenting users with chronologically ordered tweets yield better classifiers for narrative content. As expected, regardless of which set of features is being used, simpler boolean problems outperform the more difficult multiclass ones.

6 Conclusion and Future Work

This study adds understanding of how annotators make sense of microblog narratives and on the importance of considering how readers may be impacted by engaging with text annotation tasks.

The narrative sorting task—and the self-evaluated confidence rating—appears useful for understanding how a reader may frame and interpret a microblog narrative. Confidence displayed a strong relationship with several factors, including target topic, time spent, Kendall τ distance, and cognitive complexity between trials. This points to the importance of considering confidence in an-

Confidence	Emo+/-	Dom. Emo.	Seq+/-
Seq+/-	# of Tweets	# of Tweets	Job+/-
Kendall τ dist.	Job+/-	Emo+/-	K. τ dist.
Soc. Med. use	Dom. Emo.	Seq+/-	Confid.

Table 5: Good-3 features used for each SVM classifier.

notation. Confidence ratings can also help identify outlier narratives that are more challenging to process and interpret. The increase in cognitive complexity in Trial 2 did not appear to cause a potentially unhealthy stress response in annotators.

Despite generating interesting results, this study had limitations. For example, the sample size was modest and trial order was not randomized. Additionally, the topics of tweets were not overly stressful, and we avoided including tweets we thought could trigger discomfort. As an exploratory study, the quantitative results presented represent preliminary findings. More nuanced and advanced statistical analysis is left for future work.

Future work could benefit from developing classifiers for predicting whether a microblog post is part of a narrative; a useful filtering task completed by careful manual inspection in this study. Additional development of classifiers will focus on further aspects related to how readers are likely to interpret and annotate microblog narratives.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. IIS-1559889. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Josie Barnard. 2016. Tweets as microfiction: On Twitter's live nature and 140-character limit as tools for developing storytelling skills. *New Writing: The International Journal for the Practice and Theory of Creative Writing*, 13(1):3–16.

Marina Bluvshtein, Melody Kruzic, and Victor Massaglia. 2015. From netthinking to networking to netfeeling: Using social media to help people in job transitions. *The Journal of Individual Psychology*, 71(2):143–154.

Catherine M Bohn-Gettler and David N Rapp. 2011. Depending on my mood: Mood-driven influences on text comprehension. *Journal of Educational Psychology*, 103(3):562.

Alexander Calderwood, Elizabeth A Pruett, Raymond Ptucha, Christopher Homan, and Cecilia O Alm. 2017. Understanding the semantics of narratives of interpersonal violence through reader annotations and physiological reactions. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 1–9.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Karën Fort, Gilles Adda, and K Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

iMotions. 2016. iMotions Biometric Research Platform 6.0.

Tong Liu, Christopher Homan, Cecilia Ovesdotter Alm, Megan C Lytle, Ann Marie White, and Henry A Kautz. 2016. Understanding discourse on work and job-related well-being in public social media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1044–1053.

Vibeke Thøis Madsen. 2016. Constructing organizational identity on internal social media: A case study of coworker communication in Jyske Bank. *International Journal of Business Communication*, 53(2):200–223.

Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133.

Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 3723–3726, New York, NY, USA. ACM.

Julie Medero, Kazuaki Maeda, Stephanie Strassel, and Christopher Walker. 2006. An efficient approach to gold-standard annotation: Decision points for complex tasks. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, volume 6, pages 2463–2466.

Saif M Mohammad. 2012. #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 246–255. Association for Computational Linguistics.

Robert Plutchik. 2001. The Nature of Emotions. *American Scientist*, 89:344.

Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *Proceedings of CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, pages 2651–2656, New York, NY, USA. ACM.

Ward van Zoonen, Toni GLA van der Meer, and Joost WM Verhoeven. 2014. Employees work-related social-media use: His master's voice. *Public Relations Review*, 40(5):850–852.

Ward van Zoonen, Joost WM Verhoeven, and Rens Vliegenthart. 2015. How employees use Twitter to talk about work: A typology of work-related tweets. *Computers in Human Behavior*, 55:329–339.