"You are an expert annotator": Automatic Best–Worst-Scaling Annotations for Emotion Intensity Modeling

Christopher Bagdon[1,3,4], **Prathamesh Karmalker**[2], **Harsha Gurulingappa**[2], and **Roman Klinger**[4]
[1]Merck Data & AI Organization, Merck Group, Darmstadt, Germany
[2]Merck Data & AI Organization, Merck IT Centre, Merck Group, Bengaluru, India
[3]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[4]Fundamentals of Natural Language Processing, University of Bamberg, Germany
{christopher.bagdon,roman.klinger}@uni-bamberg.de
{prathamesh.karmalkar,harsha.gurulingappa}@merckgroup.com

Abstract

Labeling corpora constitutes a bottleneck to create models for new tasks or domains. Large language models mitigate the issue with automatic corpus labeling methods, particularly for categorical annotations. Some NLP tasks such as emotion intensity prediction, however, require text regression, but there is no work on automating annotations for continuous label assignments. Regression is considered more challenging than classification: The fact that humans perform worse when tasked to choose values from a rating scale lead to comparative annotation methods, including best–worst scaling. This raises the question if large language model-based annotation methods show similar patterns, namely that they perform worse on rating scale annotation tasks than on comparative annotation tasks. To study this, we automate emotion intensity predictions and compare direct rating scale predictions, pairwise comparisons and best–worst scaling. We find that the latter shows the highest reliability. A transformer regressor fine-tuned on these data performs nearly on par with a model trained on the original manual annotations.

1 Introduction

Labeling data with trained experts or via crowdsourcing is a resource-intensive and time-consuming process (Zaidan, 2012; Wang et al., 2021; Bunte et al., 2021). This motivates automated annotation methods, including weak supervison (Ratner et al., 2017), zero-shot predictions (Radford et al., 2019), or, more recently, generative models (Wang et al., 2021; Kasthuriarachchy et al., 2021; Radford et al., 2019). Depending on the downstream task at hand, the labels to be assigned to a textual instance are categorical (e.g., in text classification), structured (for instance in parsing or named entity recognition), or continuous (e.g., emotion intensity, sentiment strength, or personality profiling predictions).

Annotating for continuous value labels comes with its own set of challenges. It can be difficult to obtain consistent labels from humans by asking them to assign a value from a rating scale (Schuman and Presser, 1996). Not only is it difficult for the annotator to rate texts consistently, but it is also difficult for researchers to design rating scales, as there are many design decisions which can bias the annotator, such as scale point descriptions and scale granularity. This lead to comparative annotation setups, in which annotators are tasked to compare multiple instances for the same task, which is easier to accomplish and has fewer design decisions to make. Consider the two example sentences for sentiment strength:

(1) She's quite happy.
(2) He is extremely delighted.

It is difficult to assign a value $v(s_i) \in [-1; 1]$ to these sentences in isolation, or even in context, but it is straight-forward to decide that $v(s_2) > v(s_1)$.

Best–worst scaling (BWS, Finn and Louviere, 1992; Louviere and Woodworth, 1991) is such a comparison-based annotation method that has proven to be more reliable than assigning values on rating scales, when annotating continuous values. The idea is to task an annotator to decide which instance is the one with the highest and the lowest value. The tuple size can be varied but Kiritchenko and Mohammad (2017) observe that quadruples provide a good trade-off between context and numbers of comparative judgements.

In this paper, we question if BWS is also an appropriate approach for large language model-based annotations. On the one hand, one might argue that comparative tasks are also more reliably conducted with language models. On the other hand, one might argue that a large language model (LLM) has more access to other text instances implicitly

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7921–7933

from training data which it can compare a text to, than a human. That would be an argument that when using a LLM for annotations, BWS may not be necessary.

We therefore set up prompts for continuous value assignments for two direct and two comparison-based annotation approaches. We use rating scales (RS, Likert, 1932), in which the model directly labels texts with a numerical value, in two variants: annotating single texts and tuples of four texts per prompt. Our comparison-based approaches are paired comparisons (PC, Thurstone, 1927) in which every text is compared to every other text, and best–worst scaling (BWS, Louviere and Woodworth, 1991), where "best" and "worst" instances are picked from a tuple.

In our evaluation, we focus on comparisons against human annotations for the emotion intensity prediction task (Mohammad and Kiritchenko, 2018). We compare the LLM-based annotations directly against human annotations and further train a transformer-based regressor both on the human-labeled data and the LLM-labeled data. The motivation for this regressor is to avoid the requirement to put together instances in tuples and query a potentially expensive API at inference time. We answer the following research questions:

1. Does the best–worst scaling annotation method perform better than rating scales or paired comparisons when using generative models for labeling text with continuous values? *(Yes, it does.)*
2. How does the performance of a transformer-based regressor compare when trained with automated annotations vs. human annotations? *(The models perform on par.)*

2 Related Work

2.1 Automated Annotation

Annotating texts can be an arduous and costly task (Zaidan, 2012). Crowd-sourced annotation is sometimes cheaper than following a more traditional approach to hire few expert annotators (Snow et al., 2008), but generally, the costs increase with the difficulty of the task, either because more careful training is needed or more annotators need to be involved to obtain reliable aggregated scores.

This situation lead to the development of automatic annotation methods. *Weak supervision* uses noisy automated annotations to train models. The expectation is that they might be less accurate than supervised models, but still better than unsupervised learning (Ratner et al., 2017). Examples include the use of heuristics, keyword searches, or distant supervision from databases (Ratner et al., 2017; Rudra and Anand, 2020; Rao et al., 2021). A more recent approach to automatic data labeling is zero-shot classification (Radford et al., 2019). It relies on the information present in a pretrained language model to solve the task, either via textual entailment (Yin et al., 2019; Obamuyide and Vlachos, 2018; Plaza-del Arco et al., 2022), by mapping verbalizations of classes to outputs of autoregressive models (Shin et al., 2020, i.a.), or with instruction-tuned models (Zhang et al., 2023; Gupta et al., 2022; Ghosh et al., 2023, i.a.).

Recently, (Wadhwa et al., 2023) successfully used rating scales with LLM's to improve crowd-sourced annotations. Their goal was, however, not automatic annotations but improving existing annotations.

2.2 Annotating Continuous Values

Some NLP tasks require the prediction of continuous values, for instance rating emotion intensity (Mohammad and Bravo-Marquez, 2017). A typical operationalization is to ask annotators to choose a position on a rating scale, such as Likert scales (Likert, 1932). The exact position that humans chose does, however, depend on various subjective aspects, including preferences for particular intervals (Kiritchenko and Mohammad, 2017). This leads to inconsistencies between annotators. Annotations can also be inconsistent from a single annotator – after seeing more examples they might adjust their interpretation of a specific value range.

An alternative is to rate data-points via comparison such as Paired Comparisons (PC, Thurstone, 1927). PC ranks texts by comparing every text to every other text. This approach comes with the major drawback of a quadratic number of required annotator's decisions in the number of instances.

2.3 Best–Worst Scaling

Best–worst Scaling (BWS) addresses the issues of rating scales and paired comparisons (Louviere and Woodworth, 1991; Finn and Louviere, 1992). Annotators are provided an n-tuple (typically $n = 4$) of data-points and asked to choose the *best* and *worst* of the tuple, given the scale they are annotating. 4-tuples are efficient because by giving the *best* and *worst* ratings the annotator has effectively done five out of six possible pair-wise

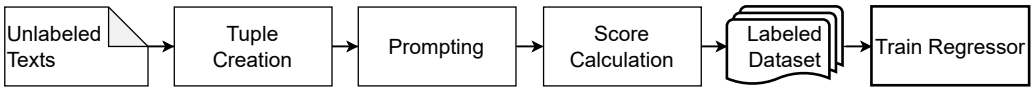

Figure 1: Workflow diagram of our method to use automated annotations for training a regression model.

comparisons. Kiritchenko and Mohammad (2017) showed that $1.5N$ to $2N$ annotations, where N is the number of instances, lead to reliable results. The final score $s(i)$ of an instance i is computed by $s(i) = \frac{\#\text{best}(i) - \#\text{worst}(i)}{\#\text{overall}(i)}$.

Kiritchenko and Mohammad (2017) found BWS produces more consistent annotations than RS, while also being efficient in the number of annotations required. They compared annotations from RS and BWS using split-half reliability (SHR). As qualitative values cannot be precisely known, one measure of their accuracy is reproducibility across multiple annotators. SHR evaluates reproducibility by splitting all annotations for each data-point randomly into two bins. Each bin is used to calculate the target label separately, and the two sets of scores are compared. This is repeated for multiple iterations and the correlation scores are averaged.

2.4 Emotion Intensity

Emotion analysis consists of various subtasks, including emotion categorization (Calvo and Mac Kim, 2013; Mohammad et al., 2018, i.a.), in which emotion labels from a predefined set are assigned. The labels commonly stem from psychological models of basic emotions (Plutchik, 2001). Another popular task in emotion analysis is affect prediction, where continuous values of valency and arousal are assigned (Posner et al., 2005).

A combination of these approaches in which intensities for a given emotion are to be predicted was first proposed with a shared task in 2017 (Mohammad and Bravo-Marquez, 2017). The Affect-in-Tweets Dataset is an extension of the original data (AIT, Mohammad and Kiritchenko, 2018).[1] They have been created via best–worst scaling with a setup in which annotators have been asked to select the most intense and least intense instance from a quadruple of tweets for a given emotion. The data set is partitioned into the four emotions joy, fear, anger, and sadness. An individual instance can appear in multiple of the data sets.

[1]More information, including terms and conditions, can be found at https://competitions.codalab.org/competitions/17751.

3 Methods

We now provide an overview of our automatic annotation method (§3.1), describe our prompting strategies (§3.2), and finally explain how the automatically created corpora are used to estimate a transformer-based regressor (§3.3).

3.1 Overview

Our method automates the annotations of a training set which we use to estimate a regressor. Figure 1 shows the workflow, discussed in the following.

Tuple Creation. In this step of the pipeline, we create the instances for annotation. Rating scales (RS) require single instances. For rating scales tuples (RS-T) we use 4-tuples to give the model the same context per prompt as BWS, where we also use 4-tuples. Paired comparisons (PC) require each text to be paired with every other text once. In BWS, nearly no pair appears in more than one tuple and all texts appear in nearly the same number of tuples. We follow the recommendation by Kiritchenko and Mohammad (2017) to use $1.5N$ to $2N$ tuples.

Prompting. The prompt tasks the LLM to output annotations for the given text tuple. Independent of the concrete prompting approach (RS, RS-T, PC, BWS), each prompt contains a role description, a task description, the texts, and formatting instructions. As the back-end, we mainly use GPT-3.5-turbo[2] and compare the best performing setup to GPT-3 and Llama2 (Radford et al., 2019; Touvron et al., 2023). We discuss the prompts in Section 3.2.

Score Calculation. For RS and RS-T, we directly use the output value as the score. For PC and BWS, we use the counting method (§2.3). All results are linearly normalized to $[0; 1]$.

3.2 Annotation

The prompt differs for each of the four annotation methods. We show all of them side-by-side in Table 1. Each prompt contains up to six parts: The *role* informs the model how we expect it to behave. For GPT-3.5, this is applied as a system level prompt separate from the main prompt. With models which do not utilize a system prompt, it is

[2]https://platform.openai.com/docs/models/gpt-3-5

Section	Rating Scales	Rating Scales Tuples	Paired Comparison	Best–Worst Scaling
Role	You are an expert annotator specializing in emotion recognition.		You are an expert annotator specializing in emotion recognition.	You are an expert annotator specializing in emotion recognition.
Task Descr.	Please rate the following text from social media for how intense the authors feels {emo}.		Which of the two speakers is likely to be the MOST {emo} and which of the two speakers is likely to be the LEAST {emo}?	Which of the four speakers is likely to be the MOST {emo} and which of the four speakers is likely to be the LEAST {emo}?
Scale	Use the following scale [Round to the fourth decimal.]: 4: extremely intense {emo} 3: very intense {emo} 2: moderately intense {emo} 1: slightly intense {emo} 0: Not {emo} at all			
Format	Only reply with the numerical rating.		Only give the Speaker number. Do not repeat the text content.	Only give the Speaker number. Do not repeat the text content.
Texts	Text: {text}	Text 1: {text1} Text 2: {text2} Text 3: {text3} Text 4: {text4}	Speaker 1: {text1} Speaker 2: {text2}	Speaker 1: {text1} Speaker 2: {text2} Speaker 3: {text3} Speaker 4: {text4}
Format Example	Format your response as: {emo} intensity:		Format your response as: Most {emo} Speaker: Least {emo} Speaker:	Format your response as: Most {emo} Speaker: Least {emo} Speaker:

Table 1: Prompts for Rating Scales, Rating Scales Tuples, Paired Comparisons, and Best–Worst Scaling. Variables are typeset in {curly brackets}. Unique text blocks are in the same color across a row. Rounding is only requested for decimal scales. Rating Scales and Rating Scales Tuples are identical except for the Texts section.

the first line of the prompt. The *task description* is similar to the instructions for humans. The *scale* is only used for RS and RS-T, and the scale in Table 1 is only an example. It is updated according to the actual rating scale that is used. The *format* explains the expected output based on the *texts* that are labeled so that the model can refer to them. The final element is an *example* for the expected output.

Rating Scales. For rating scales, which directly annotate individual texts, we do not need tuples. The scale, which we consider a parameter of this method, is included as part of the task description. When the scale includes decimals we instruct the model to round to the fourth digit.

Rating Scales Tuples. Comparison-based approaches have the advantage of showing the model more examples of text from the dataset per prompt. To mitigate this advantage we have the model rate four texts per prompt. Otherwise, this approach is identical to our rating scale approach.

Paired Comparisons. Paired Comparisons compare every text with every other text. We instruct the model to choose both the speaker with the *most* emotion and the *least* emotion. We use the terminology "speaker" to follow AIT's task description (Mohammad and Kiritchenko, 2018). We accept an output if it includes two distinct predictions.

Best–worst Scaling. Our setup of the BWS annotation process follows Kiritchenko and Mohammad (2017); Mohammad and Kiritchenko (2018). We use tuples of four texts, with no pair of texts appearing in more than one tuple. The tuples are annotated one at a time, though a single prompt is used to annotate both *most* and *least*. We accept an output if it includes two distinct predictions. We request an output for the same tuple multiple times until we receive an acceptable answer.[3]

3.3 Regressor

In principle, the output of a LLM can directly be used to label unseen instances at application time. Nevertheless, we consider it reasonable to train a regressor on top of the annotations for three reasons. Firstly, it makes the use of the annotations comparable to human annotations which are also the input to a model training. Secondly, it allows the smaller regressor to run locally and does not require a potentially expensive API to be called, that might also change in behaviour. Thirdly, for PC and BWS, annotating in a zero-shot learning setting at inference time would require combining the instance of interest in tuples. We fine-tune roberta-base (Liu et al., 2019) with a regression head with default parameters for 5 epochs.[4]

[3] In our experiments with GPT3.5 (200,000 tuple requests), few non-acceptable answers have been returned (388). The first repetition typically lead to an acceptable answer.

[4] https://huggingface.co/transformers/v3.0.2/model_doc/auto.html, we tested other epoch counts,

Emotion	Training	Dev	Test
Joy	1,616	290	1,105
Anger	1,701	388	1,002
Fear	2,252	389	986
Sadness	1,533	397	975
Total	7,102	1,464	4,068

Table 2: Details of the AIT dataset, of which the training set is made up of the EmoInt dataset.

4 Experiments

We perform experiments to compare the annotation methods on the emotion intensity prediction task, firstly to understand if BWS, PC, RS, or RS-T perform differently and secondly to understand their performance in comparison to human annotators. Finally, we perform additional experiments to understand the role of the tuple count, which can be flexibly adapted in automated annotations.

4.1 General Setup

Dataset. We use the Affect-in-Tweets Dataset (AIT). AIT has been manually annotated for the 2018 shared task *SemEval-2018 Task 1: Affect in Tweets* (Mohammad et al., 2018). It consists of tweets, manually annotated using BWS for emotion intensity scores for joy, sadness, anger, and fear. AIT is an extension of the Emotion-Intensity dataset (EmoInt) Mohammad and Bravo-Marquez (2017); its training set is composed of the entirety of EmoInt and its development and test sets are newly added. Table 2 shows the statistics.

In the BWS annotation, each tuple consists of four tweets. The target emotion for choosing the most and least emotion intensity is predefined. The total number of tuples created per emotion is $2N$ where N is the number of tweets to be annotated. Each tuple was annotated by 3–4 independent workers on the crowd-sourcing platform Crowdflower. The final scores are calculated as described in Section 2.3 and linearly scaled to $[0; 1]$.

Evaluation. AIT uses split half reliability (SHR) to evaluate its annotations for reproducibility. This approach is problematic for evaluating our method; we are not taking measurements from multiple people, we are taking multiple measurements from the same language model. Evaluating reproducibility is more a test of the model's consistency than it is a measure of distance from truth. Hence we compare the model's output to human-annotated scores.

We use a direct and an indirect, downstream evaluation: (1) in the direct evaluation we compare the AIT's training data gold annotations to the annotations from the LLM on the training data. This tells us how well the generative model is able to replicate manual annotations. (2) The indirect downstream evaluation is based on the trained regressor model that produces scores in a second step. Therefore, it is applied to the official test data set. We consider this evaluation to be more important as it replicates the actual use-case of such automatic annotations. We use Pearson's correlation.

4.2 RQ 1: Does the best–worst scaling annotation method perform better than rating scale-based annotations?

4.2.1 Experiment Settings Details

Generative Model. We use GPT-3.5-turbo via the AzureOpenAI API. The models are no different from those made available directly by OpenAI. However, AzureOpenAI has additional content filters which prevent prompts containing violence, bigotry, self-harm, or sexual content.[5] In the case that such filter prevents an output, we rerun the prompt-tuple in OpenAI's API. Overall, we paid 125.12 Euro for our experiments.

Best–worst Scaling. We use the exact same tuple sets as used in the original manual annotation of AIT, kindly provided to us by the authors.

Rating Scales. We vary the scale in the prompting method by granularity and description. We refer to an instruction that does not contain descriptions as "Bare" (B), instructions that only contain descriptions at the maximal and minimal level as "Outlined" (OL), and with complete descriptions as "Descriptive" (D)[6]. We experiment with intervals of 0.0–1.0 (X-1), 0–10 (X-10) and 0–4 (X-4). The latter corresponds to the rating scale used by (Kiritchenko and Mohammad, 2017) to compare BWS and rating scale annotations done by humans, clipped to positive values.

Rating Scales Tuples. For each prompt we ask the model to rate four texts. Each text appears in only one prompt. Otherwise, the approach is identical to our rating scales approach.

Paired Comparisons. In the PC setup, we are not able to create the entirety of N^2 comparisons for reasons of resource constraints. We use a subset of

including early stopping, in preliminary experiments but did not find the results to differ substantially.

[5] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter

[6] The exact descriptions can be seen in Table 1.

	Rating Scales						Rating Scales Tuples						BWS	PC
Emo.	B-1	OL-1	B-10	OL-10	D-4	D-10	B-1	OL-1	B-10	OL-10	D-4	D-10	$2N$	200P
Joy	5.3	5.5	49.1	49.2	48.8	58.2	67.1	66.9	40.7	65.2	65.3	64.6	81.0	81.2
Ang.	15.0	16.6	39.7	42.4	52.7	57.3	68.2	68.0	70.8	67.3	69.1	69.0	74.5	72.9
Fear	16.4	15.9	54.0	55.9	63.4	65.0	52.3	50.5	65.2	51.7	60.7	63.0	76.2	75.1
Sad.	17.9	15.6	62.6	62.1	60.1	67.0	67.8	68.4	71.4	68.4	69.5	68.6	80.3	76.8
Mean	14.9	13.7	52.0	53.1	57.3	62.4	63.6	63.1	63.7	62.8	66.0	66.3	78.1	76.8

Table 3: Direct comparison via Pearson's correlation (*100) between original AIT annotations and automated annotations from various annotation approaches: Rating scales, Rating scales tuples, Best–worst Scaling (BWS), and Paired Comparisons (PC).

	Original	RS	RS-T	BWS	PC
Emotion	AIT	D-10	D-10	$2N$	220P
Joy	78.8	64.2	67.9	76.9	60.7
Anger	78.9	68.5	71.0	71.5	60.3
Fear	78.7	65.4	64.3	70.9	53.3
Sadness	75.1	62.1	63.9	72.1	55.7
Mean	78.3	65.5	67.1	73.5	58.1

Table 4: Indirect downstream comparison via training a RoBERTa model on the annotated data from various annotation approaches: Rating scales (RS), Rating scales tuples RS-T), Best–worst Scaling (BWS), and Paired Comparisons (PC).

200 randomly selected texts per emotion to test if vastly increasing the number of comparisons per text improves annotation quality. In this case, direct annotation comparison only consider these 200 texts. When training a regressor, we also only use these 200 annotated texts. In that latter case, the evaluation is performed, as in all other regressor evaluations, on the independent test corpus.

4.2.2 Results

Table 3 shows direct evaluation results and Table 4 shows indirect downstream results.[7]

Direct Comparison. For the direct comparison of the annotated data, we consider Table 3. We see results for different setups of the RS, the RS-T, the BWS results and the PC. The results vary dramatically between different rating scales. The best performance is achieved with the D-10 model, which could be considered unsurprising as it provides the most detail to the LLM. Removing part or all of the descriptions leads to a performance drop; the values alone appear to not be interpretable. Interestingly, changing only the scale and keeping the descriptions (B-1/OL-1 vs. B-10/OL-10) also leads to a substantial performance decrease. Apparently, floating numbers are less informative to the model than natural numbers.

The strong difference between rating scales does not hold with RS-T. D-10 still performs the best. B-10 shows the best results for anger, fear and sadness, but its poor performance on joy drags its mean score below D-10.

On all four emotions, the BWS scores are higher than any RS or RS-T annotation. Joy and sadness show the most similar scores to the original annotations. Similar to performance differences of human annotators, anger appears to be most challenging.

The PC performance scores are on a similar level as BWS, but lower. This comes, however, at a substantially higher cost: PC uses roughly six times the annotations as BWS for only 200 texts. Note that this result is achieved on a different data subset.

BWS creates the annotations with the most similar performance scores to the original annotations. However, one might argue that this is not surprising given the alignment of the annotation method with the original corpus creation approach. Therefore, it is important to consider the indirect comparison.

Indirect Downstream Comparison. The indirect evaluation results, through training regressors, are shown in Table 4. The results labeled with "Original" stem from the model trained on the original human data annotations.

The RoBERTA models trained on human annotations perform well. The results are en par with the winning team's approach in the AIT-2018 shared task (Duppada et al., 2018). BWS outperforms RS, RS-T, and PC, but not the models trained on the original data. The drop in performance to training on the original data is lower for joy and sadness than for anger and fear, with the latter performing the worst. These observations are in line with Mohammad et al. (2018).

The RS and RS-T performances are similar to their direct evaluation counterparts, though the gap

[7]The annotations can be found at https://www.uni-bamberg.de/en/nlproc/resources/autobws/

in performance between emotions is smaller. Anger is not the most challenging emotion and nearly performs the same as BWS. The paired comparisons perform worse, due to the smaller training set.

Summary. BWS performs the best for both evaluation setups. It does use twice the annotations as rating scales ($2N$ vs. N), however the increase in performance is worth the cost, given that all annotations are automated. Paired comparisons are too inefficient to be considered a viable alternative.

4.3 RQ 2: Can automated annotations be as good as human annotations?

In the previous experiment, we kept the annotation setup close to the original setup of BWS annotations to understand the impact of the LLM use. In this section, we will exploit the advantages of automatic annotations to see if scaling it up can improve the predictions to be closer to human performance.

4.3.1 Experiment Settings Details

For the AIT annotations each tuple was annotated by three annotators. However, running tuple sets through GPT three times is not equivalent to annotating with three annotators; while the temperature could be raised to increase randomness in the model's output, we do not equate this to the new perspectives, experiences, intuitions, and opinions an additional annotator would provide. Hence we need another method of increasing our total number of annotations. The quality of BWS annotations can be increased by either increasing the number of annotators or the number of tuples (Kiritchenko and Mohammad, 2017). Therefore we increase the number of tuples annotated. We start with $6N$ tuples to match the number of annotations done by AIT, which uses $2N$ unique tuples and each tuple is annotated by 3 annotators, giving a total $6N$ annotations. Then we explore half and double that number. We only run each prompt-tuple once, but we increase the number of unique tuples:

- $3N$: 50% more unique tuples, but half the total number of annotations.
- $6N$: 200% more unique tuples with the same number of total annotations.
- $12N$: 600% more unique tuples and twice the number of total annotations.

The sets of tuples used for this experiment are randomly created using the same design as AIT, but they do not purposefully contain the original tuples.

	Direct Eval.				Ind. Downstream Eval.				
	GPT-3.5				GPT-3.5				Orig.
E	$2N$	$3N$	$6N$	$12N$	$2N$	$3N$	$6N$	$12N$	$2N$
J	81.0	78.7	80.5	81.3	76.9	76.0	75.7	77.6	78.8
A	74.5	74.2	74.9	76.1	71.5	71.7	72.2	73.2	78.9
F	76.2	74.3	76.3	77.3	70.9	72.6	73.5	71.7	78.7
S	80.3	77.4	79.7	80.8	72.1	74.5	74.4	74.5	75.1
∅	78.1	76.2	78.0	78.9	73.5	74.2	74.4	74.8	78.3

Table 5: Evaluation with higher tuple counts.

4.3.2 Results

We show all results in Table 5, for both a direct comparison of the annotation output and the performance of a trained RoBERTA-based regressor. We see that increasing the tuple counts does lead to an improvement for both evaluations, with 12N performing best for every emotion except fear. This follows the findings by Kiritchenko and Mohammad (2017) that increasing total annotation counts improve annotation quality. Regarding the indirect evaluation with a trained model, we also see an increase of performance for higher tuple counts. With 2N, we see an average performance of 73.5 in contrast to 78.3 for humans. These gaps shrink with larger tuple counts, but not too dramatically – the best result is achieved with $12N$ tuples, leading to 74.8 correlation.

5 Further Analyses

5.1 Generative Model Comparison

We performed all experiments with GPT3.5, but the results might not carry over to other models, and they might not remain replicable if the API, the model, or the licenses change. We therefore compare the results to GPT-3-davinci (Radford et al., 2019) and Llama2 (Touvron et al., 2023) with 13B parameters. We use the same evaluation setup as above, with the original $2N$ tuples. Table 6 shows the results. GPT3.5 outperforms GPT3 by 14pp and Llama2 by 27pp. There is a notable drop in performance for fear (24pp,36pp). GPT-3 performs decently on joy and sadness (7pp/8pp drop).

These results translate closely to the indirect comparison. GPT-3's regressions' performance lines up with its direct comparison results. Llama2 does however perform better than in the direct comparison. Joy and fear all close the gap between their Llama2 and GPT-3 performances.

Llama2 does not perform well in our experiments. The results for joy and sadness are accept-

	Direct Eval.			Ind. Downstream Eval.		
Emo.	GPT3.5	GPT3	Llama2	GPT3.5	GPT3	Llama2
J	81.0	73.9	52.2	76.9	71.7	68.0
A	74.5	61.5	52.1	71.5	65.9	49.4
F	76.2	51.9	39.5	70.9	54.7	49.4
S	80.3	71.9	64.1	72.1	64.7	62.7
Avg.	78.1	64.6	51.5	73.5	65.3	58.6

Table 6: Evaluation across models.

Ex. 1: Not giving a fuck is better than revenge. AIT: .63 GPT3.5: .06			
		M	L
Tuple 1	Yay bmth canceled Melbourne show fanfuckingtastic just lost a days pay and hotel fees not happy atm #sad #angry	1 2 3 G	
	Just saw lil homie @NICKMERCS rage on cam. Weren't roids a thing in the late 90's or has it come back? I'm lost...		2
	Not giving a fuck is better than revenge.		G
	ESPN just assumed I wanted their free magazines		1 3
Tuple 2	@MMASOCCERFAN @outmagazine No offense but the only way this makes sense is if you work for the magazine. Otherwise,who are you apologizing		1
	She's foaming at the lips the one between her hips @realobietrice, one of many great lyrics	1 2 G	3
	the bee sting still suck i feel sick		
	Not giving a fuck is better than revenge.	3	2 G

Table 7: Example with the highest difference between manual and automatic annotation, and the associated tuples. The 1/2/3 refer to the AIT annotator IDs and G to the annotator GPT3.5.

able, but the results for anger and fear are even less than the paired comparisons trained on only 200 texts. The score for joy of 68.0 is especially surprising given the low score in the direct evaluation. This highlights that lower correlation to the original annotations does not guarantee worse performance as a training set. The exact cause of this is worth further investigations. It is noteworthy that Llama2 provided further challenges, in addition to its lower performance. Nearly 10% of all prompts sent to the model returned non-acceptable answers.

5.2 Error Analysis

To provide an intuition why annotations differ, we manually inspect the top 10 instances per emotion that have the largest absolute difference in emotion intensity annotation between the human and the automatic annotation. We show these instances in the Appendix, Table 9.

Out of the 40 texts, 21 explicitly mention the target emotion, while 19 either only refer implicitly to the emotion and could be considered neutral, or do not describe the emotion at all. In the cases in which the emotion is explicitly mentioned, GPT-3.5 tends to assign a higher score than the human annotators – in such cases, it does more often rate the instance as the most intense (14/21 cases). In cases in which one might argue that the text is in fact comparably neutral, GPT assigns lower values than humans (13/19 cases). In summary, our error analysis shows that GPT has a tendency to make consistent decisions for explicitly mentioned emotions, but humans might have a tendency to interpret the text, unsurprisingly, more carefully regarding implicit information.

For reasons of space constraints in this paper, we cannot show all tuples for all these instances. We do, however, believe that a more in-depth error analysis requires such analysis. We resort to showing the one example that has the highest difference in prediction for the emotion anger (Table 7). This instance ("Not giving a fuck is better than revenge") contains a strong metaphorical negative expression. One might however argue, that it does not in fact express anger – it offers some freedom for interpretation. In the first tuple all human annotators agree regarding a different instance exhibiting most anger. GPT3.5 assigns it the lowest anger. The picture is less clear in the second tuple. Annotators are more distributed across instances. This analysis suggests that more combinations with more varied tuples can lead to more reliable results. In this instance, the error decreases from .56 to .47 when increasing the tuple count to $12N$.

5.3 Task Validation on Another Corpus

Setup. To validate our method's performance on emotion intensity prediction we apply our method to a second dataset, *SemEval-2007 Task 14: Affective Text* dataset (Strapparava and Mihalcea, 2007), which is comprised of 1250 news headlines annotated for six emotions: joy, anger, fear, sadness, disgust, and surprise. The annotations were done by six annotators using rating scales of 0–100, with each text annotated for all emotions at once. The inter-annotator agreement score was found by taking average Pearson's correlation scores between annotators (shown in Table 8 on the right).

We test two RS-T scales to compare with BWS

$2N$: D-10, our previously best performing scale, and OL-100, adjusted to from OL-10 to match the dataset's label range. The experiment setup and prompts are the same as our previous experiments. We refer to this setup as *Basic Approach*. To further test if an annotation setup that is closer to the original annotation environment further improves the result; we have each prompt annotate all six emotions at once in an adapted setup (*Adapted Appr.*).

Results. Table 8 shows the results. D-10 outperforms OL-100 on every emotion and performs better than BWS on anger, fear, and disgust in the basic approach. BWS does better for joy, sadness, and surprise, giving BWS a better overall performance. RS-T and BWS annotations score higher than the average human annotator for all emotions except surprise. In the adapted approach, BWS scores do not change substantially, but RS-T results improve. D-10 performs better than BWS overall and OL-100 performs en par with BWS.

Interpretation. These results validate the method's ability to do emotion intensity prediction, however they challenge our initial finding that BWS is the better approach. Our interpretation is that the similarity between the automatic annotation setup and the original setup matter – the label distributions substantially differ: Where rating scales allow for all annotations to be skewed towards specific areas of the scale, BWS's comparative nature forces scores towards a normal distribution. While this can be a benefit when annotating fresh data, it limits BWS's ability to replicate rating scale annotations. SemEval-2007's original annotations are all skewed towards the low end of the scale. Our RS-T annotations are also skewed, but the BWS annotations are normally distributed.

We take this as an indicator that BWS is a better choice for annotating novel corpora from scratch automatically. Kiritchenko and Mohammad (2017) show that BWS produces more reliable annotations than rating scales. So if our method can replicate both BWS and rating scale annotations to a similar degree, then it follows that we should choose the approach which performs better overall. Furthermore, when simulating existing data, the gap in performance between BWS and RS-T is much larger for the BWS-native dataset than for the RS-native dataset.

	Basic Approach			Adapted Approach			Original
	RS-T		BWS	RS-T		BWS	RS
Emo.	B-100	D-10	$2N$	B-100	D-10	$2N$	B-100
Joy	47.7	66.8	70.4	71.2	76.2	70.1	59.9
Ang.	44.3	60.1	59.6	57.1	61.0	60.2	49.6
Fear	65.1	68.4	65.8	66.2	72.9	67.5	63.8
Sad.	66.5	71.5	74.2	73.4	77.3	73.0	68.2
Dis.	33.1	47.8	47.5	51.4	52.7	49.4	44.5
Sur.	18.5	15.6	32.3	21.5	23.8	24.6	36.1
Avg.	57.2	67.5	68.4	68.0	72.9	68.5	53.7

Table 8: Results for applying our method to SemEval-2007 Task 14 dataset. Basic approach uses the same prompts as previous experiments. Adapted approach rates all emotions in a single prompt. Original shows the average inter-annotator agreement score of the 6 human annotators.

6 Conclusion & Future Work

We proposed to automate annotations of text data with continuous labels with BWS, which outperforms rating scales and paired comparisons in the case of emotion intensity predictions, when the original annotations were also annotated using BWS. The predictions from a regression model, fine-tuned on automated annotations, perform nearly en par with the models fine-tuned on the original human annotations. We showed that we can improve the annotation quality by increasing the total number of tuples. In general, we conclude that BWS is the better approach to annotate novel data sets for emotion intensity regression.

The results of our experiments are encouraging for emotion intensity regression. We presume that these findings carry over to other regression tasks, but this still needs to be validated. Candidates for other tasks would be the BWS-labeled toxicity data set Ruddit (Hada et al., 2021) or the Affective Norms for English Words dataset (Bradley and Lang, 1999). Word similarity assessment tasks could be an interesting case for evaluation as well (Antoniak and Mimno, 2018).

Finally, it is important to study more open-source generative models, instead of relying on pay-locked and black-box models.

Acknowledgements

We thank Saif Mohammad for helping us and providing us the original tuples of the data set we used. This project is partially supported by the project ITEM (User's Choice of Images and Text to Express Emotions in Twitter and Reddit, funded by

the German Research Foundation, KL 2869/11-1). We thank the reviewers and the action editor at ACL Rolling Review for their helpful feedback.

7 Limitations

Before our method can be a reliable alternative to manual annotations it must be tested on more NLP tasks. While the results are promising on predicting emotion intensity, several possible shortcomings come to mind. Without manually annotated data to compare to, it is difficult to judge the quality of the automated annotations. Especially for novel tasks, it is hard to judge if mediocre performance of the regression model is evidence of poor quality annotations or of the task being difficult to model.

The results we present on the task of predicting emotion intensity rely on the assumption that none of the models used were pretrained on the AIT dataset. The poor results from prompts using rating scales leads us to believe the model does not have prior knowledge of the dataset. This highlights the problem with using GPT-3 as it is a black-box system which inhibits interpretability.

At the time of our experimentation, the token limit of generative large language models posed a challenge due to the need for four texts to be apart of each prompt. With the recent (and fast) development of improved models, this constitutes a limitation of our research. Future research may look into annotating longer texts.

Our method relies on the information embedded in an LLM to accurately compare target texts. If the task is too specialized than the LLM might require fine-tuning to perform adequately. This defeats the purpose of the method as it is intended to circumvent the need for training data. Furthermore, the method is not as accessible as working with non-generative LLMs such as RoBERTa. The success found in this paper relied on GPT models which are not open source, making them less transparent and more unreliable.

8 Ethical Considerations

Our method does not contribute a new data set or introduce a novel task. Therefore, it does not add any additional risks from these perspectives to the already existing research landscape. All data that we use, we use for their originally intended case, namely the creation of emotion intensity prediction models.

However, it is noteworthy that previous research showed that emotion analysis systems are biased for various reasons (Kiritchenko and Mohammad, 2018). The use of language models for automatic creation might lead to different biases and this requires further research.

The core idea behind our method is to reduce the need for manual annotations. Ideally, researchers would only need manual annotations for their development and test sets, reducing manual annotations by roughly 80%. While this is great for research projects, this can reduce the amount of work available for people who depend on annotation-based jobs.

Though, the reduction could be attractive for tasks which require annotators to read emotionally or mentally damaging texts, such as hate speech or toxicity. Abusive content detection systems are needed more in online spaces but the creation of such systems requires annotators to spend time interacting with content which can cause undue emotional trauma (Vidgen et al., 2019; Kiritchenko and Nejadgholi, 2020).

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology. University of Florida.

Andreas Bunte, Frank Richter, and Rosanna Diovisalvi. 2021. Why it is hard to find ai in smes: A survey from the practice and how to promote it. In *International Conference on Agents and Artificial Intelligence*.

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

Venkatesh Duppada, Royal Jain, and Sushant Hiray. 2018. SeerNet at SemEval-2018 task 1: Domain adaptation for affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 18–23, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Finn and Jordan J. Louviere. 1992. Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy & Marketing*, 11(2):12–25.

Sayan Ghosh, Rakesh R. Menon, and Shashank Srivastava. 2023. LaSQuE: Improved zero-shot classification from explanations through quantifier modeling

and curriculum learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7403–7419, Toronto, Canada. Association for Computational Linguistics.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for English Reddit comments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2700–2717, Online. Association for Computational Linguistics.

Buddhika Kasthuriarachchy, Madhu Chetty, Adrian Shatte, and Darren Walls. 2021. Cost effective annotation framework using zero-shot text classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Svetlana Kiritchenko and Isar Nejadgholi. 2020. Towards ethics by design in online abusive content detection. *CoRR*, abs/2010.14952.

Rensis. Likert. 1932. *A Technique for the Measurement of Attitudes*. Number Nr. 136-165 in A Technique for the Measurement of Attitudes. Archives of Psychology.

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. *ArXiv*, abs/1907.11692.

Jordan J. Louviere and George G. Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nikitha Rao, Chetan Bansal, and Joe Guan. 2021. Search4code: Code search intent classification using weak supervision. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, pages 575–579.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings VLDB Endowment*, pages 269–282.

Koustav Rudra and Avishek Anand. 2020. Distant supervision in bert-based adhoc document retrieval. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2197–2200, New York, NY, USA. Association for Computing Machinery.

Howard Schuman and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Louis Leon Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34:273–286.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Manya Wadhwa, Jifan Chen, Junyi Jessy Li, and Greg Durrett. 2023. Using natural language explanations to rescale human judgments. *ArXiv*, abs/2305.14770.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Omar F. Zaidan. 2012. *Crowdsourcing annotation for machine learning in natural language processing tasks*. Ph.D. thesis, Johns Hopkins University.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada. Association for Computational Linguistics.

Appendix: Text Examples with the Largest Manual and Automatic Annotation Difference

Emo.	Text	AIT	GPT	Δ
Joy	#LethalWeapon A suicidal Vet with PTSD... so FUCKING FUNNY.... let the hilarity begin...	.64	.06	-.58
	It was very hard to stifle my laughter after I overheard this comment. It really is amazing in the worst ways.	.65	.12	-.52
	@ardit_haliti I'm so gutted. I loved her cheery disposition.	.50	.0	-.50
	Rojo is so bad it's hilarious.	.60	.13	-.48
	I fear that if United bottle this my heart would actually collapse from laughter.	.69	.25	-.44
	I wish there were unlimited glee episodes:(so I could watch them forever. #gleegoodbye	.31	.75	.44
	@OrbsOfJoy plan a date... like a date u find pleasing or smth. fuckign\n\n10/10. because the child will grow to be a ten out of ten	.31	.75	.44
	Lea doing a mini set tour of glee my heart just cried tears of happiness and sadness	.49	.94	.45
	#RIP30 Heaven is rejoicing because they've gained an angel, the Keifer family are in my prayers	.40	.88	.48
	Headed to Montalvo w/@jaxster3—bring on the #mirth, bitches!\nd(-_-)b\n@Nick_Offerman\n@MeganOMullally\n#SummerOf69Tour2016	.44	.94	.50
Anger	Not giving a fuck is better than revenge.	.63	.06	-.56
	@FluDino Event started! everyone is getting ready to travel to the lake of rage, where everything glows	.52	.0	-.52
	could never be a angry drunk lol yall weirdos just enjoy your time	.52	.0	-.52
	@Iucifaer you can go on what you usually do its just their own personal reason and not mean to offend anyone :(	.63	.13	-.50
	Inner conflict happens when we are at odds with ourselves. Honor your values and priorities. #innerconflict #conflict #values	.50	.0	-.50
	The war is right outside your door #rage #USAToday	.50	.94	.44
	Marcus Rojo is the worst player i have ever seen. Useless toasting burning bastard	.56	1.0	.44
	@Bell @Bell_Support Cancelling home Fibe, Internet and TV this afternoon - as soon as I can arrange alternate Internet. 2/2 #angry #fedup	.48	.94	.45
	You boys dint know the game am I the game... life after death... better chose and know who side you on before my wrath does come upon us	.52	1.0	.48
	And Republicans, you, namely Graham, Flake, Sasse and others are not safe from my wrath, hence that Hillary Hiney-Kissing ad I saw about you	.35	.88	.52
Fear	@RJAH_NHS @ChrisHudson76 @mbrandreth #course day # potential Leadership #excited #nervous # proud	.65	.13	-.52
	I was literally shaking getting the EKG done lol	.88	.38	-.50
	MSM stoking #fear. Please remember the beautiful prayerful protests in Dallas and Atlanta. Smile at a stranger. We make each other strong.	.56	.06	-.50
	@ChrissyCostanza and have social anxiety. There is many awkward things wrong with me.	.77	.31	-.46
	Avoiding #fears only makes them scarier. Whatever your #fear, if you face it, it should start to fade. #courage	.71	.25	-.46
	Staff on @ryainair FR1005. Asked for info and told to look online. You get what you pay for. #Ryanair @STN_Airport #Compensation #awful	.27	.75	.48
	I'm mad at the injustice, so I'm going to smash my neighbours windows'. Makes perfect sense. #CharlotteProtest #terrible	.46	.94	.48
	O you who have believed, fear Allah and believe in His Messenger; He will [then] give you a double portion of His mercy...' (Quran 57:28)	.33	.88	.54
	Don't think I'll hesitate to run you over. Last time I checked, I still had 'Accident Forgiveness' on my insurance policy...	.39	.94	.55
	@dc_mma @ChampionsFight think shes afraid to fight Holly. One can only imagine what goes through her head when she thinks of Cyborg #terror	.40	1.0	.60
Sadness	It feel like we lost a family member	.71	.19	-.52
	@chelseafc let them know it's the #blues	.52	.06	-.46
	It's a gloomy ass day	.89	.44	-.45
	@Theresa_Talbot @fleurrbie Haha...sorry about the dreadful puns... I need to get out more....I've been cooped up lately...	.56	.13	-.44
	@Beakmoo hmmmm...you may have a point... I thought Twitter had got dull. LAMINATION	.54	.13	-.42
	So unbelievably discouraged with music as of late. Incredibly behind on Completing my album. Not digging this at all.	.60	1.0	.40
	Nothing else could possibly put a damper on my day other than doing X-rays on someone with kickinnnnn ass breath	.40	.81	.41
	@LBardugo Start w/ the 3 songs in Blue Neighborhood\n1) Wild\n2)Fools\n3)Talk Me down for #Wesper\nAlso,\n4)Too Good. #serious kaz/inej feelz	.38	.81	.43
	@ticcikasie1 With a frown, she let's out a distraught 'Gardevoir' saying that she wishes she had a trainer	.48	.94	.46
	@courtneymee I'm 3 days sober don't wanna ruin it	.33	.81	.48

Table 9: Top ten texts with largest errors per emotion for GPT-3.5 annotations compared to original AIT annotations.