Rating by Ranking: An Improved Scale for Judgement-Based Labels

Jack O'Neill
Dublin Institute of Technology
School of Computing
Dublin, Ireland
jack.oneill1@mydit.ie

Sarah Jane Delany
Dublin Institute of Technology
School of Computing
Dublin, Ireland
sarahjane.delany@dit.ie

Brian Mac Namee
University College Dublin
School of Computer Science
Dublin, Ireland
brian.macnamee@ucd.ie

ABSTRACT

Labels representing value judgements are commonly elicited using an interval scale of absolute values. Data collected in such a manner is not always reliable. Psychologists have long recognized a number of biases to which many human raters are prone, and which result in disagreement among raters as to the true *gold standard* rating of any particular object. We hypothesize that the issues arising from rater bias may be mitigated by treating the data received as an ordered set of preferences rather than a collection of absolute values. We experiment on real-world and artificially generated data, finding that treating label ratings as ordinal, rather than interval data results in an increased inter-rater reliability. This finding has the potential to improve the efficiency of data collection for applications such as *Top-N* recommender systems; where we are primarily interested in the ranked order of items, rather than the absolute *scores* which they have been assigned.

ACM Reference format:
Jack O'Neill, Sarah Jane Delany, and Brian Mac Namee. 2016. Rating by Ranking: An Improved Scale for Judgement-Based Labels. In *Proceedings of Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, Como, Italy, August 27, 2017,* 6 pages.
DOI:

1 INTRODUCTION

Value-judgements, personal preferences and attitudes — often supplied in the form of a numerical rating — are an important source of data for training machine learning models, ranging from emotion recognition applications [5] to recommender systems [12]. Typically, oracles providing such data are asked to supply a rating on an N-point scale, representing either the extent to which a particular attribute is judged to be present (*i.e.* how common or rare a particular emotional state is [4]), or how strongly an attitude or judgement is felt (*i.e.* star-ratings of films [16]). Values provided in this way (1 - N on an N-point scale) are known as *absolute* measurements, as they implicitly depend on an absolute, idealised scale [18] of measurement. The data we collect from such ratings may be treated as *interval* data, as all values are expressed in a single, common, numerical scale.

Psychologists have long known of a number of bias errors to which these scales are prone [17]. Researchers found that many respondents can be categorised as displaying one or more of these biases. Borman [3], has conducted a comprehensive study of these biases; including the tendency to rate primarily at the high or low end of the scale, *lenient* or *severe* biases, respectively; the tendency to avoid making distinctions; and rating primarily around the centre of the scale, known as *range restriction*, and its counterpart, the tendency to rate primarily at either end of the scale, which we refer to as a bias of *extremity* . Worryingly, Stockford and Bissel [20] found evidence that the ratings provided can even be influenced by the order of questions on the form, which has been termed *proximity bias*.

Label rankings represent an alternative method of judgement elicitation to absolute ratings. When we elicit ranking data from an oracle, we present a set of objects to be ranked, and the oracle orders this set of objects in terms of how strongly a particular attribute is judged to be present. Values provided in this manner are known as *relative* measurements, as each value is dependent on the other items present in the ranking set. The data collected from such rankings may only be treated as *ordinal* data; it allows us to determine a natural ordering of the data, but gives us no information as to how inherent a particular attribute may be on any absolute, common scale.

Inter-rater reliability (*IRR*) measures the overall level of agreement between raters as to the values they provide for identical data. Although some disagreement among raters is inevitable in situations involving judgements with some degree of subjectivity; we assume that, given enough ratings for a particular item, the average rating will eventually converge on its *true*, gold standard rating (*i.e.* the average rating which would be obtained were we to sample the entire population). A high *IRR* for a label-set indicates that there is little disagreement among raters as to the *true* gold standard for that item; leaving us with more reliable data. Disagreement between raters stems from genuine differences in opinions on the one hand, and factors such as noise and rater bias on the other. An increase in *IRR* is only desirable if the increase is due to reducing the latter.

By proposing a general framework for recasting queries requiring absolutely valued rating labels as a task needing only a series of relative ratings, we hope to improve the reliability and validity of scale-based data collection. This proposition relies on the intuition that corpora of relatively-valued labels will produce higher levels of inter-rater reliability (*IRR*) than those consisting of absolute values. The intuition, in turn, is based on the assumption that although rater biases affect the numeric label which raters will assign to an item, as bias is constant for each individual, it should not affect the order of preference of items for any given rater.

Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, Como, Italy

DOI:

To take a simple example, consider two raters, rater *S*, a severe rater, and rater *L*, a lenient rater. Both raters are asked to provide labels for a set of items. Rater *S* tends to give lower than average ratings for items in general, whereas rater *L* tends to give higher than average ratings. Table 1 shows the interval ratings provided by both raters; while Table 2 shows these ratings as ordinal (ranked) data. Although there is very little agreement between raters on individual items, there is evident agreement as to the ranking of items.

Table 1: Interval Ratings for Rater *S* and Rater *L*

	A	B	C	D	E	F	G
Rater *S*	4	1	2	6	3	5	7
Rater *L*	7	5	4	8	6	9	10
Difference	3	4	2	2	3	4	3

Table 2: Ordinal Ratings (Rankings) for Rater *S* and Rater *L*

	A	B	C	D	E	F	G
Rater *S*	4	7	6	2	5	3	1
Rater *L*	4	6	7	3	5	2	1
Difference	0	1	1	1	0	1	0

In our current work, we investigate this intuition empirically, hypothesizing that ordinal-valued datasets systematically result in a higher level of inter-rater reliability than their absolute-valued equivalents. Section 2 reviews related research on absolute and relative rating methods motivating the current study. Section 3 describes the methodology behind our experiment to test this hypothesis both on artificially generated and real-world datasets. We relay our findings in Section 4 and discuss the implications for future work in Section 5

2 RELATED WORK

The psychologist, Arthur Blumenthal argues that the human mind is not capable of providing truly absolute ratings; that when we are asked to provide labels on absolute scale, we compare each item to similar items we have seen before. He argues that absolute judgement involves the *"relation between a single stimulus and some information held in short term memory about some former comparison stimuli or about some previously experienced measurement scale using which the observer rates the single stimulus"* [2]. This suggests that raters would be more comfortable with comparison-type, or *relative* measurements, than they would be with the absolute system of ratings so prevalent in data collection for machine learning today. A recent study by Moors *et al.* [15] has shown that both modes of elicitation — rankings and ratings — produce similar within-subjects results. This means it should be possible to use either system without skewing the labels. More importantly, related work has shown empirically that ratings collected using comparative methods can, in certain circumstances, be more reliable between-subjects than data collected using absolute measurement, for example in multi-criteria decision making [18] and collaborative filtering [12].

When gathering absolutely-valued labels, there is no guarantee that all raters share the same understanding of the absolute, idealised scale on which the system is based, and this can result in significantly different behaviours, leading to ultimately unreliable data. On top of this, it has been shown that results obtained from rating exercises can be significantly influenced by the choice of scale itself (5-point scale vs 7-point scale, for example) [7]. Marsh and Ball [14] argue that the effects of these phenomena can be seen in the contemporary peer-review process, where the mean single-rater reliability — the relation between two sets of independent ratings of quality collected for a large number of submissions — of reviewers for journal articles was very low at 0.27.

This possibility is further evidenced by the experience of Devillers *et al.* in collecting emotion-annotation data as part of the HUMAINE project [8]. Raters used the FEELTRACE annotation instrument for recording emotions in videos in real time [6]. FEELTRACE is a video-annotation tool which allows raters to rate the intensity of a given emotion in real time. Raters use a slider to trace the intensity of a target emotion or trait, increasing the value as the target intensifies and decreasing the value as it wanes. The researchers found strong correlations between raters tracing the relative changes in emotions from moment-to-moment; however, they recognised that the absolute values each of the raters chose showed less uniformity. This suggests that although raters disagreed on the (absolute) question of the intensity of the target, there was broad agreement on the (relative) question of whether the intensity right now is greater or less than the intensity in the moment immediately preceding.

3 EXPERIMENTAL METHODOLOGY

We hypothesize that label sets collected on an interval scale (*i.e.* absolute numbers, for example, on a scale of 1 - 9) will exhibit less *IRR* than labels collected on an ordinal scale (*i.e.*, labels provided in terms of ranks within the dataset, from best to worst). Although absolute ratings contain more information than relative ratings (a relative ordering can be deduced from absolute ratings, but not vice versa) —and as such, may naturally increase the *IRR* of the resulting labels —we believe that some of the improvement is not accounted for by this factor alone. In order to investigate our hypothesis we examine 4 datasets introduced in previously published literature. All of the datasets under consideration were rated using interval labels. We then convert these labels into ordinal data using simple intra-rater ranking; and use Krippendorff's α, which adjusts for the inherent *difficulty* of the problem, to compare the inter-rater reliability of both datasets. This section describes the datasets used in the experiment and discusses the suitability of Krippendorff's α as a comparison metric.

3.1 Artificial Datasets

In order to explore the impact of rater bias on *IRR* we generated artificial datasets of ratings for items provided by raters exhibiting one of the four rater biases, discussed in Section 1; *lenient*, *severe*, *restricted* and *extreme*. Our first step is to model a set of items to be rated. These items could represent, for example, movies, where the

goal is to assign a score to each movie representing how good it is; or a joke, where the goal is to rate the joke based on how funny it is. In any case, we are not interested so much in the particular item being rated so much as the gold standard towards which the average rating from a large number of raters would converge. Each item is represented as a randomly selected real number between 1 and 9 representing this gold standard rating. We generated 50 such items to be rated.

We next generate a *base rating* for each rater. This base rating is determined by adding a modifier randomly selected from a normal distribution with μ 0 and σ 1. This value is rounded to the nearest whole number and represents the rating this rater would assign in the absence of rater bias. Each rater's base rating differs slightly, representing the inherent subjectivity of labels based on value-judgements.

Next, we generate a *bias modifier*, representing the extent to which a rater's inherent bias will affect the label provided. The bias modifier is a strictly positive value drawn from a normal distribution with μ 0 and σ 1. We ensure the modifier is positive by taking the absolute value of the number drawn and disregarding the sign. We split the raters into 4 groups of 20; with each group exhibiting one of the rater biases discussed in 1. The manner in which this bias modifier is applied to each rater's base rating is dependent on the type of bias this rater exhibits. These are as follows:

Lenient These raters tend to give ratings above the average for all items. The bias modifier is added to the base rating for raters falling into this category

Severe Raters falling into this category tend to give lower-than-average ratings for all items. The bias modifier is subtracted form the base rating for raters falling into this category

Restricted These raters favour ratings around the mid-point of the scale. If the base rating is below the mid-point, the bias modifier is added to the base rating. If the base rating is above the mid-point, the bias modifier is subtracted from the base rating.

Extreme This final group of raters favours ratings at either end of the scale. If the base rating is below the mid-point of the scale, the bias modifier is subtracted from the base rating. If the base rating is above the mid-point, the bias modifier is added to the base rating.

Figure 1 shows the impact of each of these rating biases on the labels provided. The histograms in grey show the distribution of base ratings for a particular item before bias has been applied. The coloured histograms show the distribution of labels for each of the 4 groups after bias has been applied.

3.2 Empirical Datasets

The **Jester Dataset**, first introduced by Goldberg *et al.* [9], is a corpus containing 4.1 million continuous ratings on a scale of −10.00 to 10.00 of 100 jokes from over 70,000 anonymous users. The dataset in its original format is sparse, with many missing values. In order to simplify the evaluation, we use a small subset of this data containing 50 jokes each rated by the same 10 raters. Ratings are specified to two decimal places.

The **BoredomVideos** dataset is adapted from the corpus introduced by Soleymani *et al.* [19]. A subset of the dataset has been chosen to eliminate missing values; resulting in 31 clips, each rated on a scale of 1-10 by 23 different raters. Ratings in this dataset are provided in integer format.

The **MovieLens dataset** in its original format consists of 10 million ratings provided by 72,000 users across 10,000 different movies. We extracted a subset of 5,720 ratings, provided by 20 users across 286 movies. Ratings were provided on a scale of 0.5 - 5, in steps of 0.5. This subset contained no missing values.

The **Vera am Mittag** German Audio-Visual Emotional Speech Database **(VAM)**, created by Grimm *et al.* [10] contains emotion recognition ratings gathered from segmented audio clips from a non-acted German-language talk show. Ratings were provided on three dimensions, *activation* (how active or passive the speaker is), *evaluation* (how positive or negative the emotion is), and *dominance* (how dominant the speaker is). For the purposes of this study, we selected only the activation ratings. For this experiment, we used a subset of the data containing 478 speech instances rated by 17 different raters. All ratings were provided on a scale of -1.0 to 1.0, in steps of 0.5.

3.3 Converting Rating Data to Ranking Data

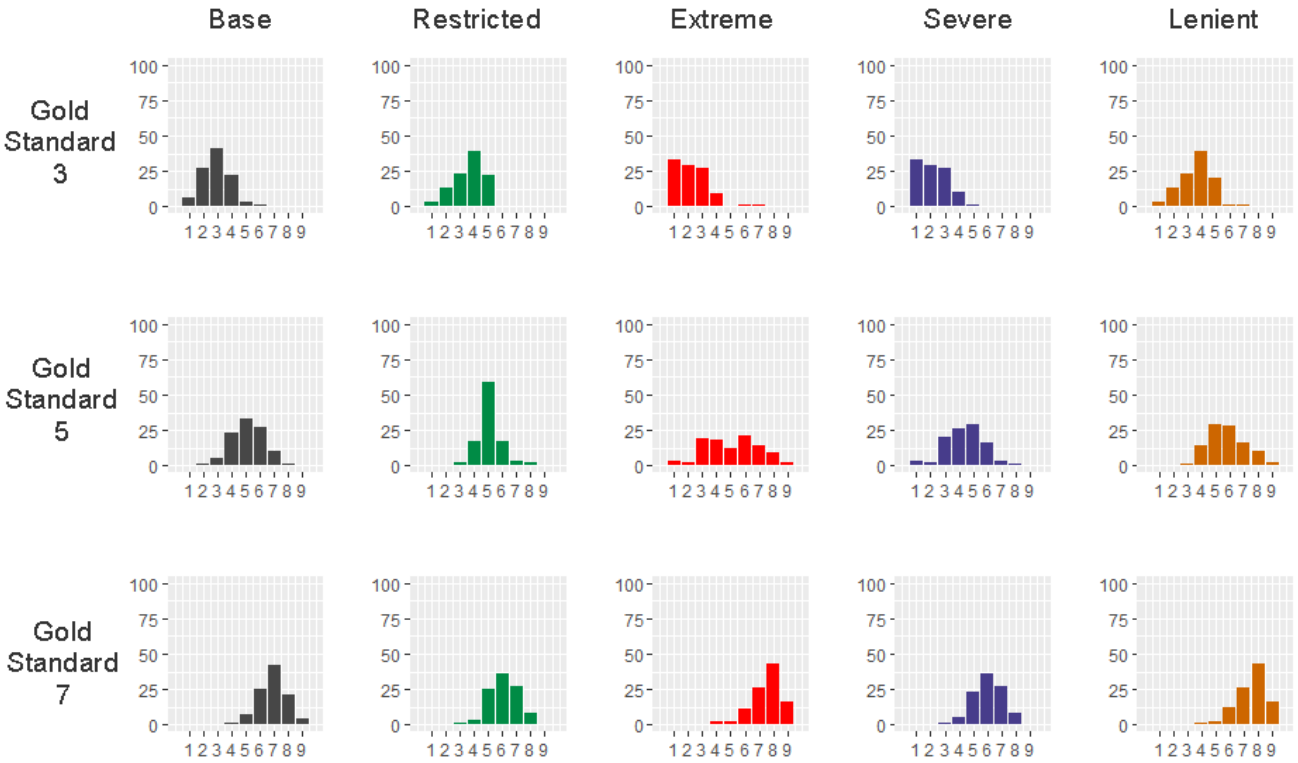
All of the datasets described in Section 3.2 consist of ratings provided on an absolute scale. In order to convert these ratings to rankings we use a simple intra-rater ranking function with average rank used in the case of ties. Given a set of raters R, and a set of items to rate, X, with r_{ij} representing the rating provided for the i^{th} item by the j^{th} rater and R_i representing the set of all ratings provided by rater i, the interval and ordinal values for these ratings are described in Equation 1.

$$\begin{aligned} L_{Interval} &= r_{ij} \\ L_{Ordinal} &= rank_{R_i}(r_{ij}) \\ x_i \in X\ &,\ r_i \in R \end{aligned} \tag{1}$$

3.4 Evaluation Metrics

Inter-rater reliability can be computed using a wide variety of measurements, the choice of which is often dependent on the properties of the data being investigated [13]. An important factor in determining the most suitable *IRR* metric is the scale of the data under examination. For example, Kendall's τ assumes that the labels provided are on an ordinal scale, whereas Pearson's ρ is stricter in that it requires labels to be on an interval scale. Further complicating the comparison is the question of commensurability of results. Interval data is more fine-grained than ordinal data, providing both an absolute value measurement and a relative ordering of items on the scale. For raters to agree on an interval scale, they must agree on both the ordering and the abolute value of each item rated. On an ordinal scale, raters are required to agree only on the relative ordering of each item. This suggests that inter-rater agreement on ordinal data is fundamentally easier to achieve, and we want to ensure that any improvement in *IRR* for ordinal data is not due simply to the higher standard of agreement required by interval data.

Figure 1: Showing example distribution of ratings before and after bias modifiers are applied



To overcome these difficulties, we use Krippendorff's α [11] to compare the IRR between the ordinal and interval label sets. Krippendorff's α is a generalized reliability measurement which expresses the ratio of observed agreement over agreement expected due to chance. In its most general form, Krippendorff's α is defined as

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2}$$

where D_o is the observed disagreement, and D_e is the disagreement that can be expected when chance prevails.

Krippendorff's α can be applied to multiple raters simultaneously, and allows comparisons to be made between values obtained on data using differing scales of measurement. The α value ranges from -1 (perfect disagreement) to 1 (perfect agreement). Having calculated the overall α for each dataset, we decompose the results by calculating the α for each pair of raters individually. The α value obtained for each pair of raters shows the level of agreement for that pair of raters independent of all others. We plot the individual α values on a heatmap, demonstrating that the improved accuracy results from a consistent increase in the pair-wise accuracy across all raters.

4 RESULTS

4.1 Artificial Datasets

We compared the *IRR* of the simulated datasets using Krippendorff's α, treating the data first as interval and then as ordinal data. After 30 repetitions, the α for ordinal rankings was consistently higher than that of the interval data, and a paired Wilcoxon Signed-Rank test rejected the alternative hypothesis, that the shift in mean α was 0, with a p value of ¡ 0.001. Figure 2 depicts the Krippendorff's α for both interval and ordinal data over 30 iterations as a box plot. This experiment demonstrates that ordinal data is more reliable than interval data, under the assumptions we made when modelling our artificial labels. However, it remains to be seen whether this improvement persists when working with real-world datasets.

4.2 Empirical Datasets

Table 3 summarises the α values for each dataset when treated both as interval and as ordinal data. Although the underlying agreement in each of the datasets was low —as can be expected with datasets containing fundamentally subjective labels —we obtained consistently higher *IRR* by treating the ratings as ordinal data. The MovieLens dataset, in particular, showed a considerable improvement when treated as ordinal data. Although the improvement in the Jester dataset and the BoredomVideos dataset was smaller, the improvement relative to the original inter-rater agreement was quite high. The relative improvement of the VAM dataset was lower

Figure 2: Krippendorff's α on artificial rating datasets using both ordinal and interval measurements, showing a relatively consistent improvement in *IRR* for ordinal data

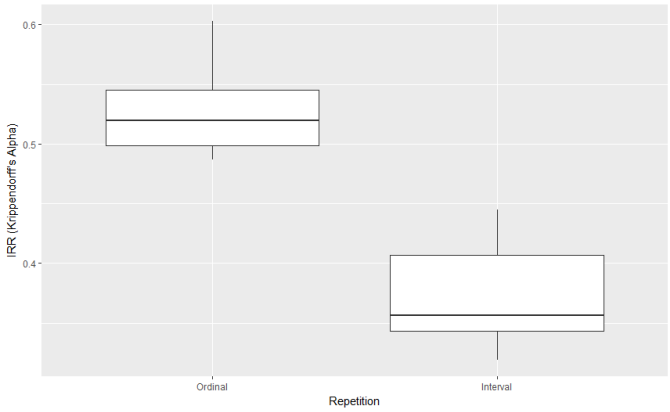


Table 3: Krippendorff's α for all datasets

Dataset	Interval	Ordinal
Jester	0.0066	0.0608
BoredomVideos	-0.0149	0.0139
MovieLens	0.2119	0.3046
VAM	0.1980	0.2260
Artificial	0.3692	0.5253

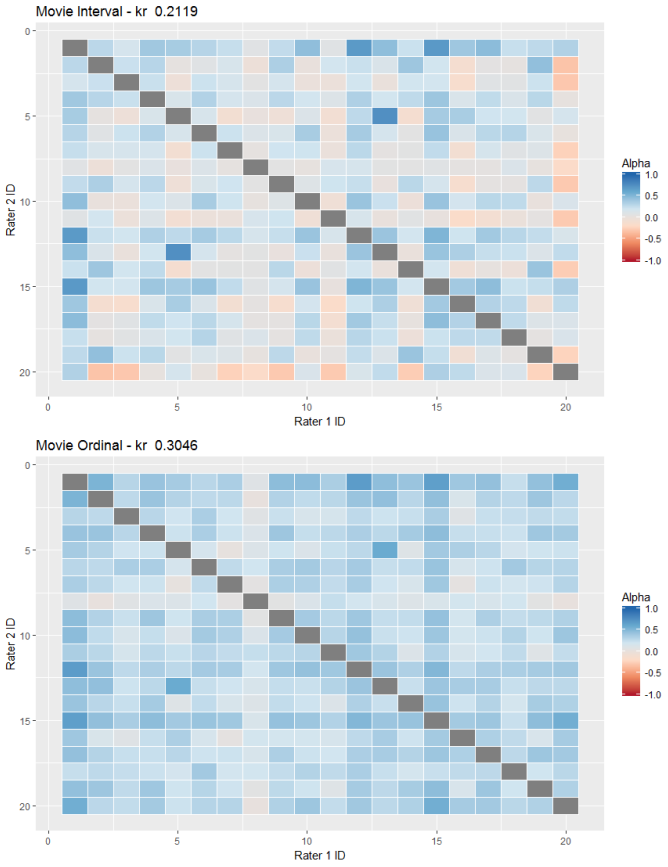
than the others. We noted in Section 4.1 that rater bias is a particular feature of subjective ratings. The lower performance of the VAM dataset may be due in part to the fact that identifying emotions in others is less subjective than making personal value judgements and so the inherent rater bias in this dataset is less pronounced than the genuine differences in opinions between raters.

Figure 3 visualises the *IRR* improvement of ordinal treatment over interval on the BoredomVideo dataset. Each rater is shown on both the X axis and the Y axis, with each cell representing the Krippendorff's α of *IRR* between the raters. Krippendorff's α scores above 0, indicating higher agreement than that expected by chance are shown as blue, while Krippendorff's α scores below 0, indicating lower agreement than that expected by chance are represented as red. This visualisation shows that the improvement in *IRR* was consistent across all pairs of raters reinforcing the findings in Section 4.1 that ranking data yields signficantly higher *IRR* than rating data.

5 CONCLUSION

Our experiments suggest that treating label sets as ordinal, rather than interval in scale tends to suppress the effects of label bias in generating inter-rater disagreement, and consequently leads to more reliable datasets. As this increased reliability results from a reduction of noise (*i.e.* rater bias), we believe this finding could be utilised to generate reliable labels with fewer oracle queries,

Figure 3: Pairwise Krippendorff's α per rater in BoredomVideo dataset, interval and ordinal data



reducing the overall cost of data collection. This finding, however, comes with a number of caveats.

Firstly, ordinal data cannot be directly transposed back to an interval scale. This approach will work best when we are more interested in the relative ordering of items rather than their absolute values, for example, Top-N recommendation. If absolute label values are required, a further step, (and further information) will be needed to infer absolute values from our ranked data. We believe that investigating possible approaches to making such an inference may expand the applicability of our proposed method of data collection.

Secondly, our experiments were conducted on real-world data gathered by researchers using an absolutely-valued measurement scale. One of the reasons absolute values are more commonly used in label collection is that it they are easier to elicit from oracles. Researchers have long been aware of the difficulty in getting raters to rank large collections of items [1]. A key challenge in collecting ranking data lies in building a method to efficiently break large collections of data into smaller subsets which can be easily ranked by human labellers; and then to recombine these sets of partial orderings into a reliable super-ordering over all labels in the set.

Developing an algorithm to present items for rating in such a manner as to maximise the efficiency of this process is an essential next-step in making our proposed approach production-ready

Thirdly, our experiments do not necessarily reflect the potential results which would be seen on a label set collected using rankings. However, there is reason to be optimistic that labels collected using ranking — rather than rating — methods would result in even higher reliability improvements. Related work, as outlined in Section 2, suggests that human oracles are more reliable in ranking small sets of items than they are at providing absolute ratings. Once a method for effectively collecting ranking labels over a large set of data has been determined, conducting this experiment on data collected using ranking would allow us to ascertain the full benefits of this approach to efficient label collection.

REFERENCES

[1] Alwin, D. F., and Krosnick, J. A. The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly 49*, 4 (1985), 535–552.

[2] Blumenthal, A. L. *The process of cognition. Experimental psychology series.*

[3] Borman, W. C. Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance 20*, 2 (1977), 238–252.

[4] Cowie, R., Apolloni, B., Taylor, J., Romano, A., and Fellenz, W. What a neural net needs to know about emotion words. *Computational intelligence and applications 404* (1999), 5311–5316.

[5] Cowie, R., and Cornelius, R. R. Describing the emotional states that are expressed in speech Describing the emotional states that are expressed in speech. *Speech Communication 40*, April (2003), 5–32.

[6] Cowie, R., Douglas-Cowie, E., Savvidou, S., Mcmahon, E., Sawey, M., and Schröder, M. fiFeeltracefi: An instrument for recording perceived emotion in real time. *ISCA Workshop on Speech & Emotion* (2000), 19–24.

[7] Dawes, J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research 50*, 1 (2008), 61–77.

[8] Devillers, L., Cowie, R., Martin, J.-c., Abrilian, S., and Mcrorie, M. Real life emotions in French and English TV video clips : an integrated annotation protocol combining continuous and discrete approaches. In *Language Resources and Evaluation* (2006), pp. 1105–1110.

[9] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval 4*, 2 (2001), 133–151.

[10] Grimm, M., Kroschel, K., and Narayanan, S. The vera am mittag german audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on* (2008), IEEE, pp. 865–868.

[11] Hayes, A. F., and Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Communication methods and measures 1*, 1 (2007), 77–89.

[12] Kamishima, T. Nantonac Collaborative Filtering fi!? Recommendation Based on Order Responces. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (2003), vol. 90, pp. 4–11.

[13] Krippendorff, K. *Content analysis: An introduction to its methodology.* Sage, 2004.

[14] Marsh, H. W., Ball, S., and Taylor, P. The Peer Review Process Used to Evaluate Manuscripts Submitted to Academic Journals : Interjudgmental Reliability. 151–169.

[15] Moors, G., Vriens, I., Gelissen, J. P., and Vermunt, J. K. Two of a kind. similarities between ranking and rating data in measuring values. In *Survey Research Methods* (2016), vol. 10, pp. 15–33.

[16] Pang, B., and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics 3*, 1 (2005), 115–124.

[17] Saal, F. E., Downey, R. G., and Lahey, M. a. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin 88*, 2 (1980), 413–428.

[18] Saaty, T. L. Rank from comparisons and from ratings in the analytic hierarchy/network processes. *European Journal of Operational Research 168*, 2 SPEC. ISS. (2006), 557–570.

[19] Soleymani, M., and Larson, M. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus.

[20] Stockford, L., and Bissel, H. W. Factors involved in establishing a merit-rating scale. *Personnel* (1949).