A Dataset for Multi-Scale Film Rating Inference from Reviews

Frankie Robertson, Stefano Leone
University of Jyväskylä, Independent
frankie.r.robertson@jyu.fi, stefano.leone.992@gmail.com

Abstract

This resource paper introduces a dataset for multi-scale rating inference of film review scores based upon review summaries. The dataset and task are unique in pairing a text regression problem with ratings given on multiple scales, e.g. the A-F letter scale and the 4-point star scale. It retains entity identifiers such as film and reviewer names. The paper describes the construction of the dataset before exploring potential baseline architectures for the task, and evaluating their performance. Baselines based on classifier-per-scale, affine-per-scale, and ordinal regression models are presented and evaluated with the BERT-base backbone. Additional experiments are used to ground a discussion of the different architectures' merits and drawbacks with regards to explainability and model interpretation.

Keywords: ordinal regression, text regression, rating inference task, explainability, opinion mining, sentiment analysis

1. Introduction

This paper introduces the MS-RottenTomatoes dataset, which consists of enriched data scraped from the Rotten Tomatoes film review aggregator. Critics submit their reviews to Rotten Tomatoes, who aggregate them and then display them on their website, grouped by film, along with their own summarising rating. As an aggregator, Rotten Tomatoes includes reviews from different film critics across different publications, such as online magazines and film critics' own websites. Critics can rate films on a variety of different scales such as letter (A-F) or star scales (e.g. 4 or 5 point). The main aim of the publication of this dataset is as a benchmark for multi-scale text regression on what Pang and Lee (2005) called the *rating inference problem*, that is predicting a rating given by a reviewer from the associated review text.

Prediction of item ratings based on review text has come up previously in the NLP literature, where it has historically been framed as a proxy for the sentiment analysis task with high ratings correlated with positive sentiment. As Pang and Lee (2005) note, film ratings can actually provide additional information to a film review, by providing an overall impression contrary to an otherwise negative or positive seeming review. Thus, it is likely that in some cases, the connection between predictor and outcome variable may be rather poor, making the task somewhat noisy. Additionally, different critics may interpret the same scale differently, for example being more or less generous with awarding the top scale point. The ratings themselves are ordinal data: ordered like nominal data; but with a finite number of outcomes like multi-class data. To the best of our knowledge, MS-RottenTomatoes is the first openly published dataset with sufficient detail for rating scales themselves to be modelled as part of the rating inference task.

The dataset presents a number of challenges. In many cases the review summaries simply do not contain enough information from which to predict the grade, creating a heteroscedastic situation when regressing the review scores based upon review text, in which more vague reviews have a much wider range of possible grades versus more concrete reviews. In addition, some reviewers use ratings to summarise their impressions of the films rather than their reviews. This creates a gap between their review and their rating, resulting in data points which provide random errors, or noise, to the network during training.

In this dataset, critics give their perspective upon films using both a rating and by giving their opinion in the form of text, in doing so, they implicitly express another perspective on how they perceive the rating scale. The presence of multiple rating scales gives rise to a desire to induce a single latent scale, so as to pool common attributes across critics, and in order to work well with gradient-based explainability techniques. On the other hand, since critics have different perspectives upon the rating scales, each *(critic, rating scale)* pair should be modelled individually. Finally, as ordinal data, using specialised techniques such as ordinal regression seems natural given metric methods implicitly assume data lies on an interval scale, where scale points are equal distances from one another. Liddell and Kruschke (2018) give a detailed account of the downsides of using metric methods with ordinal data, and indeed the analysis of Section 6.1 shows that their usage also introduces systematic errors in this setting.

The rest of this paper begins with some background, before describing the construction of the dataset and presenting baseline architectures and evaluating them. The paper then moves onto an analysis of the baseline results and discusses the degree to which different systems lend themselves to interpretability techniques. The paper closes by reviewing some related work and discussing possibilities for future work.

2. Background

Deep ordinal regression techniques can be seen as either modifying classification objectives, e.g. as in Castagnos et al. (2022), or regression objectives, considered further here. Cao et al. (2020) used a ResNet-34 backbone together with an ordinal regression head for age prediction from photos. Their technique, referred to as CORAL, induces a single latent scale which is used to predict an ordinal output Y, by modelling for each label threshold k, $P(Y \geq k+1)$. It is in this sense closely related to the backward cumulative probability family from the EL-MO class of models presented by Wurm et al. (2021), where an Element Link (EL) such as the logit most familiar from machine learning is combined with a Multinomial-Ordinal (MO) family function which reduces the ordinal outcome to a number of boolean outcomes.

Gradient-based explainability techniques such as integrated gradients (Sundararajan et al., 2017) can attribute perturbations in a deep neural network's outputs to specific areas of the input. Latent variable models give a natural choice of output perturbation to answer questions of interest about the input, e.g. *Which parts of this film review are more associated with the lower and higher ends of the rating scale?*

3. Dataset creation

This section describes how the Rotten Tomatoes data was scraped and enriched into three derived datasets, summarised in Table 1.

Table 1: Summary of the released datasets

Name	RT-critics
Repository	Kaggle[1]
# Movies	17 712
# Critics	11 109
# Ratings	1 130 017
Name	RT-normalized
Repository	Huggingface Hub[2]
# Movies	17 619
# Critics	6 832
# Ratings	617 819
Task	Text regression
Split	Random
Train/Val/Test	60% / 20% / 20%
Name	RT-critics 500
Repository	Huggingface Hub[3]
# Movies	16 251
# Critics	266
# Ratings	315 802
Task	Multi-scale rating inference
Split	Stratified (critic, rating scale)
Train/Val/Test	60% / 20% / 20%

3.1. Scraping

The film ratings were scraped from Rotten Tomatoes using a scraper written in Python using the `requests` library. The process was split into the following steps:

1. Collect film URLs from the film index;
2. Scrape film information from each film;
3. Scrape the critic reviews section of each film.

After scraping, films without at least one critic review, and critic reviews without film information are dropped. At this point, we have the non-task specific RT-critics dataset from Table 1.

3.2. Normalisation

In order to normalise the dataset for usage in multiscale prediction tasks, the type of scale being used in each rating needs to be determined. Following this, ratings can be converted into a whole number numerical, ordinal scale, ready to be treated as either a classification or ordinal regression task.

The scales are first divided into either letter scales or number scales, which include e.g. star ratings as well as percentage and out-of-10 scales. Within the letter scales there are short and long scales, with short scales ranging from F up to A and long scales, which include plus and minus grades, ranging from F- up to A+.

Number scales are broken down by three factors: the maximum score; whether the scale includes 0; and by granularity. A scale's granularity is whether it contains only whole numbers, or whether fractional ratings such as 0.5 or 0.25 are included, and if so, what is the minimum division between them. Since Rotten Tomatoes allows free text entry by reviewers submitting rat-

[1] https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

[2] https://huggingface.co/datasets/frankier/processed_multiscale_rt_critics

[3] https://huggingface.co/datasets/frankier/multiscale_rt_critics_subsets

ings, it is possible to have ratings not in the above categories. These are either misentered data or extremely rare, unusual forms of ratings and are dropped altogether.

Grades are matched against either the list of letter grades or matched as a fraction: two numbers separated by '/'. After this, grades are considered to have a preliminary *grade type*.

For numerical grades, the granularity is found by considering the Least Common Multiple (LCM). This is done group-wise, keyed on the *(publisher, grade type)* pair. Next, the grade is divided by the granularity to obtain a normalised integer grade. A single non-conforming grade will change the LCM. Rather than rounding grades, rare grades, defined here as those with less than 50 reviews across the whole dataset, together with 8 manually chosen entries with unusual grade values are dropped. These grades include for example, 2.4/5 on a grading scale which otherwise has 0.5 as the granularity. Since these do not fit with their grading scale, they may be the result of a typo, and including them would change the LCM, affecting all other entries in the same group. This cleaned dataset is released as RT-normalized (see Table 1).

3.3. Schema and tasks

The columns of the dataset can be broken down into: *Entity identifiers useful for linking within or beyond the dataset*: the movie's title, the review publisher's name, and the critic's name; *Textual content* consisting of the review text itself; *Numerical data* including the review score as originally presented, along with a normalised integer 0-based label, and an accompanying number of scale points thought to exist within the scale; and finally *information about how the grade was normalised*, such as the detected grade scale granularity, which can be used to convert back and forth between the normalised and unnormalised form of the grade.

This data is rich enough to be viewed from a number of perspectives, which are summarised in Table 2. Namely, the fact that the authors are named means this dataset could be used for author identification. Additionally, the ability to form a matrix of films and critics means that item-response theory[4] could be applied in order to analyse aspects of the critics and films themselves, e.g. film quality and critic fussiness estimated on a common scale. However, it is the multi-scale rating inference problem, where film ratings are regressed based upon the review text that is considered for the rest of this paper.

3.4. Splits and subsets

We further process the dataset for the multi-scale rating inference task. In order to model the behaviour of each critic on each scale, we consider each group keyed by a *(critic, rating scale)* pair as a task within a multi-task learning setup. We drop all groups with less than 500 items. This is done so as to create a dataset where each task contains a sufficient number of samples to model it independently. It has the additional benefit of reducing the total dataset size, leading to a more energy efficient dataset to use for benchmarking.

Next, we create training, test and validation splits using stratified sampling grouped by task. The rest of this paper considers only this RT-critics 500 dataset (see Table 1).

3.5. Analysis

Different critics treat each rating scale differently. Figure 1 shows the marginal distributions of four different critics across an out-of-5 scale and a long letter scale. The critic in the bottom left panel appears to avoid fractional grades, while the critic in the top left panel uses them in proportiona with the other grades, but is more cautious about giving 5/5. On the right, both critics avoid giving C+ grades, but the top critic gives plus and minus grades at the top end of the scale in proportion to bare letter grades, whereas the bottom critic gives relatively less plus and minus grades. These marginal distributions show different behaviours, and suggest that critics should be at least partially modelled independently. In addition, they fairly clearly show the ordinal nature of the data. Different grade points are not equally spaced, but rather it is a matter of determining some reasonable thresholds on a latent scale.

4. Systems

In order to demonstrate the dataset, some baselines are presented here[5]. In particular, we look at the performance of fine tuning foundational language models on the dataset. In all cases, the idea is to train a single backbone model for all *(critic, rating scale)* pairs. The baseline systems consist of fairly typical ways of approaching the problem of predicting film ratings based upon their text.

A block-diagram level overview of the different baselines is given in Figure 2. The backbone used in all experiments is BERT-base (Devlin et al.,

[4] Item-response theory is used here as an umbrella term for methods utilising latent variable models which estimate parameters for some general type of "items" and "respondents" on a common scale based on a cross tabulated response matrix.

[5] Code to reproduce the baselines and experiments is made available at `https://github.com/frankier/ms_text_regress`.

Table 2: A list of possible tasks relating to the MS-RottenTomatoes dataset

Task	Target	Regressors	Techniques
Author identification	Critic	Review text	Text classification
Film/critic analytics	Film quality and Critic “difficulty”	Film × Critic matrix Entries: Review text, Rating	Item-response theory
Rating inference	Film rating	Review text	Multi-scale text regression
Opinion generation	Review text	Rating, Film, Critic	Text generation

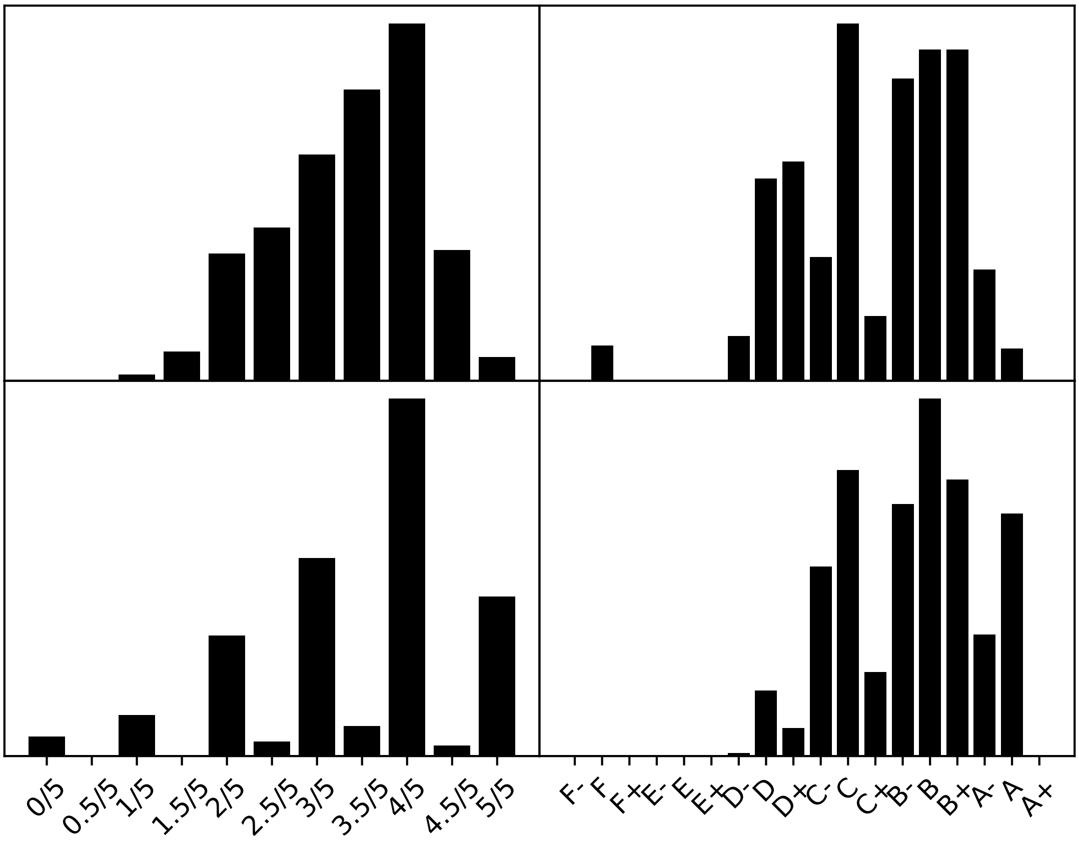


Figure 1: Plots of label distributions for different critics, illustrating varying behaviours of critics across an out-of-5 scale with half-star granularity (left) and a long letter scale (right)

2019). Although this backbone is no longer state-of-the-art, the main aim of these experiments is to establish baselines for this task. Since this dataset is derived from a publicly available website, it is of particular importance that this model has only been trained on Wikipedia and BookCorpus, and not on large scale web text, which would run the risk that the backbone has been exposed to the MS-RottenTomatoes test set during pretraining.

4.1. Classifier-Per-Scale

The Classifier-Per-Scale (CPS) system consists of a single linear predictor followed by another linear layer per *(critic, rating scale)* pair. Following softmax, the multinomial vector can be aggregated to get a single prediction. In preliminary experiments, the mode provided the best accuracy, while the median gave the lowest MAE, so both are presented in Table 3, as CPS_{mode} and $\text{CPS}_{\text{median}}$, respectively.

4.2. Affine-Per-Scale

Affine-Per-Scale (APS) learns an affine transformation of a common linear scale per *(critic, rating scale)* grouping. In order to initialise the latent scale and heads, 8 pilot batches of the training set are run, and the pre-latent linear initialised so its output has a mean of 0 and a standard deviation of 1. The heads are then initialised based on the mean and standard deviation of their critics’ labels on their rating scale according to the training dataset. So a fair comparison can be made

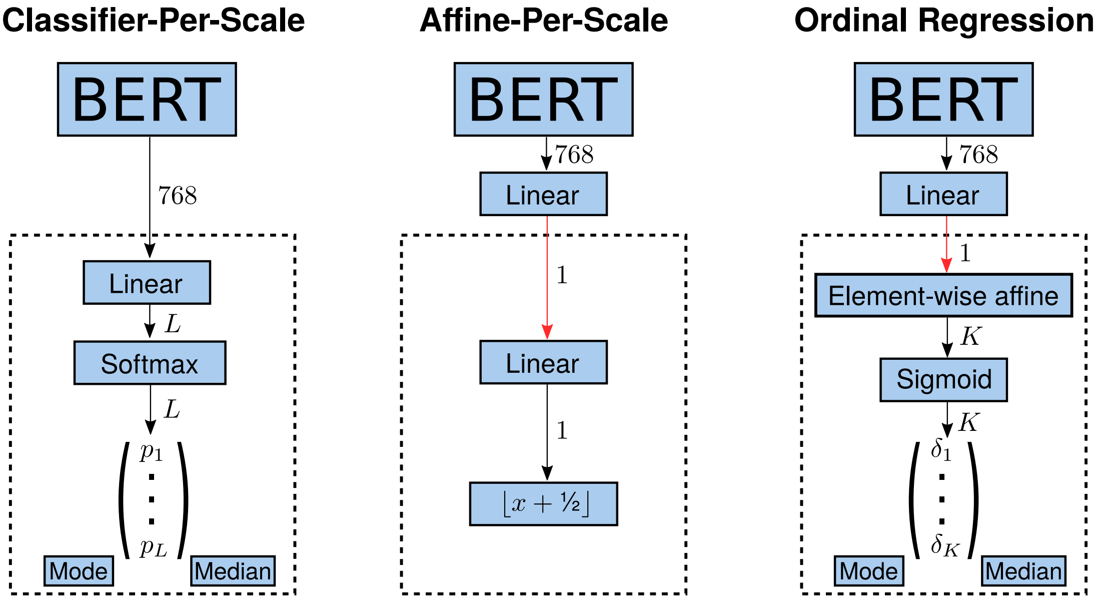


Figure 2: Schematic diagrams of the architectures of the baseline systems considered here. Dotted areas are referred to as *heads* and are repeated for each scale. Red arrows indicate the shared latent scale. L is the number of labels while $K = L - 1$ is the number of class boundaries. $\delta_k = \mathrm{P}(Y \geq k + 1)$

between regression and classification approaches, the output is rounded to the nearest integer with the function:

$$\text{round}(x) = \lfloor x + \frac{1}{2} \rfloor$$

4.3. Ordinal regression

Following Cao et al. (2020), we apply an affine transformation composed with a sigmoid transformation element-wise to get a vector modelling $P(Y \geq k+1)$. This vector can be processed to obtain a multinomial vector and the mode or median applied to obtain a single prediction, denoted as Ord.$_{\text{mode}}$ and Ord.$_{\text{median}}$, respectively, in Table 3.

5. Results

The main metric used here is a multi-scale variant of Mean Absolute Error (MAE), normalised to the range of the relevant scale:

$$\text{MAE}_{\text{MS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{\text{scale-max}(i) - \text{scale-min}(i)}$$

This metric attempts to give even weight to errors from different scales. It is invariant to the scale transformation of the normalisation procedure of Section 3.2. MAE is the chosen base metric here since it has a straightforward interpretation and lends itself easily to this multi-scale adaption.

Since the objective for the latent models is to learn a good latent scale, evaluation metrics are always taken after refitting the heads. This consists of first performing a full evaluation-mode pass over the training set, followed by refitting the heads task-at-a-time using convex optimisation procedures. The Affine-Per-Scale system is fitted using ordinary least squares as implemented in `scikit-learn` (Pedregosa et al., 2011) and the ordinal regression system is fitted using the `VGAM` (Yee, 2010) R package, which uses iteratively reweighted least squares, wrapped using `rpy2`. In both cases, no regularisation is used. No refitting is performed for the classification model.

All experiments were run using HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). The hyperparameters, which are common to all baselines are summarised in Table 4.

All experiments used early stopping with a patience of 3 validation cycles, with a validation cycle run every 1000 steps.

The results in Table 3 show that the ordinal regression system performs quite poorly. It exhibited noisy validation metrics and loss curves during training. It was the only system for which early stopping was applied, after 4000/17000 steps. We speculate that a specialised training procedure may be needed to fit these kinds of models in this setting.

Table 3: Evaluation results for the three baseline architectures, alongside a Most Frequent Class (MFC) baseline. Lower MAE is better. Higher acc. = accuracy is better. MAE$_{\text{MS}}$ and accuracy are given in %.

	Validation					
	Micro			Macro		
	MAE$_{\text{MS}}$	MAE	Acc.	MAE$_{\text{MS}}$	MAE	Acc.
MFC	16.5	2.11	27.3	16.7	2.16	27.5
CPS$_{\text{median}}$	10.5	1.38	37.4	11.2	1.50	35.7
CPS$_{\text{mode}}$	10.7	1.41	**39.3**	11.4	1.55	**37.9**
APS	**9.6**	**1.23**	34.6	**9.7**	**1.28**	34.4
Ord.$_{\text{median}}$	18.7	2.36	16.3	19.1	2.44	16.1
Ord.$_{\text{mode}}$	30.0	3.99	7.8	30.6	4.15	7.9

	Test					
	Micro			Macro		
	MAE$_{\text{MS}}$	MAE	Acc.	MAE$_{\text{MS}}$	MAE	Acc.
MFC	16.5	2.12	27.1	16.6	2.16	27.5
CPS$_{\text{median}}$	10.5	1.39	37.3	11.2	1.50	35.8
CPS$_{\text{mode}}$	10.7	1.42	**39.2**	11.3	1.54	**38.0**
APS	**9.6**	**1.25**	34.6	**9.8**	**1.28**	34.5
Ord.$_{\text{median}}$	17.8	2.27	17.2	18.1	2.34	17.0
Ord.$_{\text{mode}}$	30.2	4.05	7.9	30.8	4.22	7.9

Table 4: Hyperparameters used for the baselines

Backbone	BERT-base
Optimisation routine	AdamW
Batch size	32
Learning rate	$1\text{e}{-5}$
Schedule	Linear
Warmup	10%
Training time	$17\,000$ steps ($= 2.87$ epochs)
Validation metric	MAE$_{\text{MS}}$
Validation	Every 1000 steps
Early stopping patience	3 validations

Table 5: Counts and proportions of refit APS heads with significant x^2 parameters according to different p-values

p-value	Correction	count	%
0.05	None	101	38
0.01	None	77	29
0.05	Bonferroni	33	12
0.01	Bonferroni	24	9

6. Experiments

This section demonstrates the need for alternatives to APS, before motivating further work in the direction of latent variable models by demonstrating their potential with regards to explainability and model interpretation using the APS model.

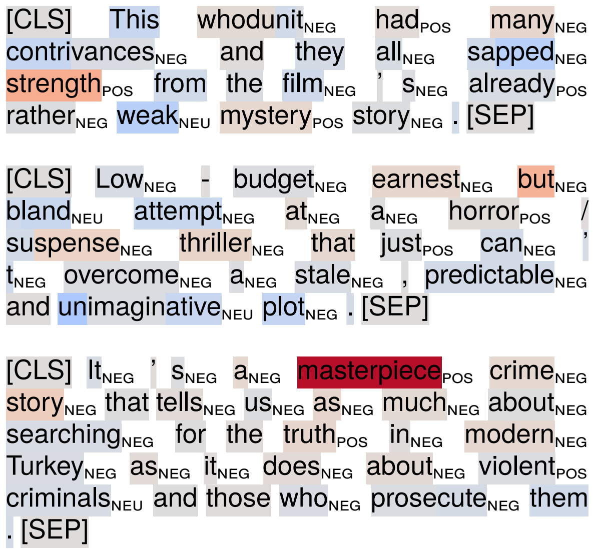
6.1. Model fit

One of the aims of releasing this dataset is to help spur interest in combining non-standard regression methods with NLP. In order to show the inadequacy of treating ratings as real numbers, we diagnose the model fit of the heads of the APS final model by running a pass over the training data and refitting a linear model of the form $x^2 + x + c$ where x is the latent scale. The fit was performed with `statsmodels` (Seabold and Perktold, 2010) with the default settings: nonrobust regression and two-tailed parameter significance testing based on the Student's t-distribution. The results in Table 5 give strong evidence that at least 9% of the heads are not fitted well by a linear relationship.

This poor fit of the heads suggests the outcomes modelled by some heads do not have a linear relationship with the latent variable, while others perhaps do. This poor fit will result in the backbone systematically receiving poor gradients during training.

6.2. Gradient-based attribution

Gradient-based attribution methods are an explainable machine learning technique which work by looking at how gradients at different network inputs change according to perturbations in the outputs of the network. For each review in the validation set, we use layer integrated gradi-



[CLS] This whodunit$_{\text{NEG}}$ had$_{\text{POS}}$ many$_{\text{NEG}}$ contrivances$_{\text{NEG}}$ and they all$_{\text{NEG}}$ sapped$_{\text{NEG}}$ strength$_{\text{POS}}$ from the film$_{\text{NEG}}$ ' s$_{\text{NEG}}$ already$_{\text{POS}}$ rather$_{\text{NEG}}$ weak$_{\text{NEU}}$ mystery$_{\text{POS}}$ story$_{\text{NEG}}$. [SEP]

[CLS] Low$_{\text{NEG}}$ - budget$_{\text{NEG}}$ earnest$_{\text{NEG}}$ but$_{\text{NEG}}$ bland$_{\text{NEU}}$ attempt$_{\text{NEG}}$ at$_{\text{NEG}}$ a$_{\text{NEG}}$ horror$_{\text{POS}}$ / suspense$_{\text{NEG}}$ thriller$_{\text{NEG}}$ that just$_{\text{POS}}$ can$_{\text{NEG}}$ ' t$_{\text{NEG}}$ overcome$_{\text{NEG}}$ a$_{\text{NEG}}$ stale$_{\text{NEG}}$, predictable$_{\text{NEG}}$ and unimaginative$_{\text{NEU}}$ plot$_{\text{NEG}}$. [SEP]

[CLS] It$_{\text{NEG}}$ ' s$_{\text{NEG}}$ a$_{\text{NEG}}$ masterpiece$_{\text{POS}}$ crime$_{\text{NEG}}$ story$_{\text{NEG}}$ that tells$_{\text{NEG}}$ us$_{\text{NEG}}$ as$_{\text{NEG}}$ much$_{\text{NEG}}$ about$_{\text{NEG}}$ searching$_{\text{NEG}}$ for the truth$_{\text{POS}}$ in$_{\text{NEG}}$ modern$_{\text{NEG}}$ Turkey$_{\text{NEG}}$ as$_{\text{NEG}}$ it$_{\text{NEG}}$ does$_{\text{NEG}}$ about$_{\text{NEG}}$ violent$_{\text{POS}}$ criminals$_{\text{NEU}}$ and those who$_{\text{NEG}}$ prosecute$_{\text{NEG}}$ them . [SEP]

Figure 3: The results of integrated gradients as applied to three example review texts. Warmer colours are associated with increases in the rating, while colder colours are associated with decreases in the rating. Words found in SentiWordNet are annotated with their corresponding classes.

ents (Sundararajan et al., 2017) as implemented in Captum (Kokhlikyan et al., 2020) with 50 steps to find subtokens associated with higher and lower ratings. For each token, we reconstruct the word it is part of and look up the first sense in SentiWordNet (Esuli and Sebastiani, 2006), classifying it as positive if the positive score is greater than the negative score, neutral if they are the same and negative otherwise. Both sources of information are overlaid onto example review summaries in Figure 3, while the resulting cross-tabulation table is shown in Table 6.

We can now calculate two association measures: mutual information and the G statistic. We calculate these between the most negatively and most positively associated token according to integrated gradients and their classification according to the negative/positive classes of SentiWordNet. The resulting mutual information calculated using the plug-in estimator gives a moderate value of 0.21 bits. Calculating the G-statistic results in a value of 16796 (rounded) which results in a P-value for association equal to 0 when calculated using double floating point precision, i.e. there is close to zero chance we would see these results without an association between the two.

6.3. Interpretation of model fit

Given a text regression model, we can apply linear modelling diagnostics to each head's linear model on the training set to answer questions about the level to which the heads are able to model the relation between the true rating and the latent variable derived from the deep learning prediction model. Critics with poor fit may be "film impression summarisers" who use a rating to give an overall impression of the film, sometimes contrary to the review text, while critics with a good fit may be "review summarisers", whose review rating agrees with the text. If we look at the MAE_{MS} for all critics, we see that *Philip Martin* is the biggest review summariser with an MAE_{MS} of 2%, and *Walter Chaw* is the biggest film impression summariser with an MAE_{MS} of 17%. However, the poor model fit of some linear heads shown in Section 6.1 means it is not entirely clear whether these metrics truly reflect critic behaviour, or whether they result from a non-linear relationship between the latent rating prediction and this rater's scores.

Table 6: Cross tabulation of words with the strongest high (hi) and low (lo) token rating attribution against negative, positive, neutral, or unknown sentiment in SentiWordNet. The top half gives raw counts and the bottom half gives % of reviews in the validation set.

		SentiWordNet			
		Neg.	Pos.	Neu.	Unk.
Cnt.	Lo	19541	6978	2964	6998
	Hi	6485	24441	28514	3717
%	Lo	31	11	47	11
	Hi	10	39	45	6

7. Related work

A number of NLP datasets deriving from review data have been published openly and used in the literature. Maas et al. (2011, § 4.3.2) introduced a fairly popular example. Their dataset consists of 50 000 user reviews from IMDB processed to create, as is often the case, a binary positive/negative review dataset by thresholding a negative class from scores ≤ 4/10, and a positive class from scores ≥ 7/10.

Pang and Lee (2005) released a dataset of 5006 reviews from 4 authors, with ratings normalised to a single 4-point scale. This results in quantisation error, and discards scale information. Although they noted the subproblem of using a single model for multiple authors would require some degree of calibration, this was not explored further.

It is worth noting that while early work in sentiment analysis treated review ratings as a proxy for sentiment as a pragmatic way to create a dataset quickly, both sentiment and review analysis has developed quite significantly since. On the one hand, sentiment analysis has given way to datasets which model emotions expressed in text, such as that of Öhman et al. (2020) who take a multi-label setting and tag subtitles with zero or more emotions from the eight emotions in Plutchik's Wheel of Emotions. On the other hand, the desire to produce more detailed analyses of review texts has led to aspect-based sentiment analysis tasks, such as that of Pontiki et al. (2014), where certain aspects of opinions expressed in reviews are to be extracted from the text, and then individually given sentiment tags.

8. Conclusion

We have presented a dataset for multi-scale film rating inference based on reviews. The dataset preserves data useful for downstream tasks, such as the original rating scales and entity identifiers. In this sense it has fewer data quality problems when compared to comparable review rating datasets. Critics and grading scale types are included in the dataset, allowing for critic behaviour against particular scales to be modelled.

Baseline experiments show that the Affine-Per-Scale (APS) and Classifier-Per-Scale (CPS) systems were able to fit the dataset. The ordinal regression -based approach outlined here, on the other hand, did not manage to beat a Most Frequent Class (MFC) baseline. While the model is a largely analogous to the model of Cao et al. (2020), we speculate that it failed to converge here due to the multi-scale setting and noisy dataset. As we saw in Table 2, different critics show different behaviours in response to different rating scales. Since ordinal regression directly models the relationship between rating scale thresholds and the latent scale, it appears to be the correct tool to handle such non interval scale data.

Furthermore, since ordinal regression induces a latent scale, as APS does, it lends itself to the interpretation experiments of Section 6.2 & 6.3, however, it would mitigate the problems of the APS system. In particular, 1. systematic error in the gradients to the backbone due to poor model fit of linear heads as outlined in Section 6.1, and 2. this same systemic error making interpretation of the values of these heads difficult, as outlined in Section 6.3 . Thus, a clear future direction for this work is to adapt the training procedure of the ordinal regression system so that it is able to make stable progress when fine-tuning a language model to tackle a noisy multi-scale task like the one posed by this paper.

Another line of future work is to consider different perspectives and tasks related to the dataset as outlined in Table 2. Particularly promising is the possibility of applying item-response theory in order to better understand different styles of using grading scales. After all, perspectives are expressed via ratings given on a variety of grading scales across a variety of domains including Education and Psychometrics. Thus this dataset holds promise for deepening understanding of rating based preferences when combined with text into these fields also.

9. Ethical considerations

Critics are named in this dataset, however, we do not believe that releasing this dataset including the names constitutes a violation of privacy. The reviews have been created by professional critics as part of their public persona. In terms of regulations, the EU General Data Protection Regulation makes such an exemption in *Article 9(2)(e)* for cases where data has been made "manifestly public" by the data subject. Furthermore, from the less strictly legalistic perspective of seeking to avoid harm to those named, since the reviews have been submitted for aggregation, there is no reasonable expectation of relative privacy (as can be the case with social media) and no reason to believe that aggregation as part of this dataset will cause the critics to come under disproportionate public scrutiny.

Rights and regulations regarding text mining vary widely between jurisdictions. While the EU has specific exceptions for text mining, the US has wider reaching fair and transformative usage exceptions. Thus, users of the dataset are advised that they have the same rights to it as if they had created it themselves. Depending on jurisdiction, this may vary according to whether the usage is commercial or done in the service of the public interest as in the case of publicly disseminated research.

10. References

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A simple log-based loss function for ordinal text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Torrin M. Liddell and John K. Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation*.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Michael J Wurm, Paul J Rathouz, and Bret M Hanlon. 2021. Regularized ordinal regression and the ordinalnet r package. *Journal of Statistical Software*, 99(6).

Thomas W. Yee. 2010. The vgam package for categorical data analysis. *Journal of Statistical Software*, 32:1–34.