The Rating Game: Sentiment Rating Reproducibility from Text

Lasse Borgholt, Peter Simonsen, Dirk Hovy
Center for Language Technology
University of Copenhagen
{fkr838|cbl123}@alumni.ku.dk, dirk.hovy@hum.ku.dk

Abstract

We investigate *(i)* whether human annotators can infer ratings from IMDb movie reviews, *(ii)* how human performance compares to a regression model, and *(iii)* whether model performance is affected by the rating "source" (i.e. author vs. annotator ratings). We collect a data set of IMDb movie reviews with author-provided ratings, and have it re-annotated by crowdsource and expert annotators. Annotators reproduce the original ratings better than a linear regression model, but are off by a large margin in more than 5% of the cases. Models trained on annotator-labeled data outperform those trained on author-labeled data, questioning the usefulness of author-rated reviews serving as labeled data for sentiment analysis.

1 Introduction

Machine learning-based approaches have become the dominant paradigm for sentiment analysis since they were introduced by Pang et al. (2002). While these approaches produce good results, they need a great deal of labeled data in order to be trained. Since human annotation can be both slow and expensive, many studies use data with an inherent subjectivity indicator, such as movie or product reviews with user ratings. While it seems a fair assumption that the *rating* expresses the author's attitude towards the subject, it is less obvious to what extent the review *text* reflects this attitude, and hence what the relation between text and numerical rating is. In this study, we ask

(i) whether human readers are able to infer the author's numerical rating based on the author's review text,

(ii) how well learning algorithms perform on the task compared to human readers, and

(iii) whether model performance is affected by the rating source used for labeling (i.e. how the numerical rating is obtained) .

In order to investigate these questions, we compile a data set of user-generated movie reviews with *author ratings* and collect both *crowdsourced annotator ratings* and *expert annotator ratings*. This setup allows us to evaluate the reproducibility of ratings for both humans and models.

We address *(i)* by comparing author ratings to crowdsourced and expert ratings. Author ratings are supposed to capture the essence of the author's sentiment, although we do not expect readers to perfectly infer author ratings based on text alone. Annotator ratings are solely based on the review text and thus expected to be similar to each other.

We investigate *(ii)* by evaluating a linear regression model on author labeled data. Sentiment analysis models supposedly emulate the cognitive process of text-based rating inference. The gap between human and model performance is interesting, because if human annotators are unable to consistently infer author ratings, we cannot expect learning algorithms to achieve this goal.

Finally, we address *(iii)* by comparing regression models trained on data labeled with crowdsourced and author ratings. Existing work treats different labeling sources as ontologically interchangeable. That is, if text is annotated for sentiment, it does not matter whether the text was labeled by the author in the process of writing said text, or by an annotator who has been paid to label the text a posteriori. This does not seem at all self-evident.

To the best of our knowledge, no previous studies have investigated the assumption that the sentiment of a text can be objectively inferred. Since sentiment analysis is still far from being a solved task, addressing this core bias could help overcome current limitations.

2 Data

We collect 2,000 user-generated IMDb movie reviews and randomly sample 200 authors, each contributing 10 reviews of a length between 800 and 2,000 characters. All reviews are marked with an author rating ranging from 1 to 10. Some authors mention their rating in the review text. This mention is of course an unwanted clue for the annotators, why we remove these reviews.

We pay annotators on CrowdFlower to rate the semantic orientation of reviews on a scale from 1 (negative) to 10 (positive). Each review is labeled by five experienced annotators. We incorporate control items in the annotation task by defining an array of permitted ratings (within two steps of the original author rating). Each annotator starts by completing eight of these test questions. Further test questions are inserted randomly throughout the annotation tasks. If annotators fall below an accuracy of 70%, they are removed from the project. Reviews used as test questions (10% of the initial data) are not included in the data set.

We use three expert annotators to rate a 20% subset of the reviews from our data set: two authors of this study, and a student. All three annotate the full subset. We use stratified sampling to select the subset, considering each rating as a stratum. The distribution of author ratings in our subset thus matches the distribution of author ratings in the full data set. The subset contains 317 reviews, the full data set 1,629 reviews. Notice that only the subset is used to answer *(i)*, whereas the full dataset is used for the regression-based tasks *(ii)* and *(iii)*.

3 Experiments

We want to establish the reproducibility of author ratings by human annotators and statistical models. In order to measure performance of the different methods, we use *mean absolute error (MAE)* and *root mean squared error (RMSE)*. While RMSE is more common, MAE is more directly interpretable, as it does not emphasize outliers. For this reason, we focus on MAE in our analysis.

MAE and RMSE measure the proximity between two sets of observations, but we also need a measure of the *relative* movement between observations. For this purpose, we use Pearson's *r*.

We have two sources of human annotation, namely three expert annotators and five crowdsource annotators per review. In order to obtain our final set of ratings, we average over each of those annotation sources. This result is more robust towards individual biases and misinterpretations. This concept is known as *wisdom of the crowd* and is well documented in the literature, e.g. Steyvers et al. (2009). However, we also wish to investigate how well an individual annotator performs. Therefore, we also compute error and correlation to the author ratings for each individual annotator, and then compute a mean over these individual comparisons.[1] This measure is equivalent to a *macro*-score and captures the average influence of individual annotators. When comparing across the two groups of annotators, we use all possible 3x5 combinations.

We use the same measures as outlined above to compare the different annotators to each other within the two groups. Hence, we compute both MAE, RMSE and correlation calculated between the individual crowdsource and expert annotators, respectively.

3.1 Model

We use a linear least squares model with L2 regularization *(ridge regression)*.[2] This method helps reduce overfitting by imposing a term α, which penalizes the parameters w of the model if they grow too large. Formally parameters w can be calculated by

$$\min_{w} \|Xw - y\|_2{}^2 + \alpha\|w\|_2{}^2$$

The underlying distribution modeled here is Gaussian, but, as we will see, the true author distribution over ratings is a mixture. We thus also experiment with incorporating a prior into the model. We use a beta distribution with shape parameters $(0.8, 0.8)$. This prior models the tendency of authors to use the extremes more than predicted by a Gaussian distribution.

We use 10-fold cross validation to ensure robust results. Furthermore, we use 5-fold cross validation on the training portion of each of the ten folds in order to determine an optimal α.

We use a bag-of-words feature representation, including all unigrams and dependency triples of type *dobj* and *nsubj* appearing more than twice in the training data. We obtain the triples by using the Stanford Parser (Klein and Manning, 2003).

[1] Since the data was not annotated by the same five, but by 200 different annotators, "individual crowdsource annotator" simply denotes one column of ratings.

[2] Experimenting with support vector regression did not yield better results, why we chose this simpler model.

4 Results

We use the mean over the *entire* rating distribution as baseline prediction. Note that the baseline differs across the different tasks, as author ratings and annotator ratings are not distributed equally.

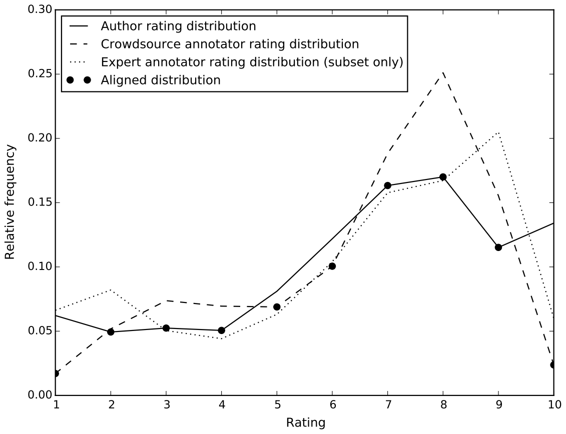

Figure 1. Rating distributions for author, crowdsourced, and expert annotator ratings. Dots indicate aligned distribution.

4.1 Human Rating Inference *(i)*

Figure 1 shows the rating distributions of the three human sources. Note the higher peaks of both annotator distributions as compared to the author distribution. Especially the crowdsource annotators peak around a few ratings. Furthermore, the author distribution is characterized by an increased use of the extreme ratings, while the annotator distributions show no such “flaps”.

Expert annotators are more correlated with one another (0.90) than the crowdsource annotators (0.76). Likewise, we see a lower MAE among expert annotators (0.75) than among crowdsource annotators (1.13), indicating a more diverse set of ratings for the latter.

		aut - cs	aut - exp	exp - cs
Corr.	Mean	0.84	0.86	0.91
	Ind.	0.76	0.83	0.81
MAE	Mean	0.96	0.96	0.71
	Ind.	1.15	1.05	1.07
RMSE	Mean	1.31	1.30	1.01
	Ind.	1.61	1.44	1.48

Table 1. Pairwise relations between author (aut), expert (exp), and crowdsourced (cs) ratings.

Table 1 shows the relation between the different rating sources. We find a higher correlation between the two sets of annotator ratings than between the author ratings and any of the annotator ratings. However, when we look at the individual rating correlations, we see that the correlation between author ratings and expert annotator ratings is now the highest, which underlines the uncertain nature of crowdsource annotators. The same pattern is visible with regard to the MAE. Notice that it is the MAE between the author ratings and the two sets of annotator ratings that will be compared with the performance of a linear regressor in the next section. Conveniently, these numbers are equal to each other, making it unnecessary to distinguish between the two types of annotators.

There is no discernible difference between the two annotator groups in terms of error margins. 80% of mean annotator rating, regardless of source, are one step off, or the rating is correctly inferred. Slightly more than 5% are more than two steps off. If we look at individual annotator ratings instead of mean ratings, however, some crowdsource annotators are a full nine steps off, and in a single case, one expert annotator was eight steps off.

	Ridge	+Prior	Aligned
Aut	1.66 / 2.14	1.70 / 2.21	1.52 / 1.95
Base	2.15 / 2.62	2.15 / 2.62	1.85 / 2.24
Ann	1.31 / 1.69	-	1.34 / 1.72
Base	1.85 / 2.23	-	1.85 / 2.24
Ann/Aut	1.60 / 2.05	-	-
Base	2.15 / 2.62	-	-

Table 2. MAEs/RMSEs for baselines and regressors trained and tested on different sources.

4.2 Human vs. Machine *(ii)*

Table 2 shows the regression results. Since we want to compare the ability of people and regressors to infer *author ratings* from text, we only look at the *Author* row. Both the full rating distributions and the aligned distribution(s) are presented in Figure 1. All settings easily outperform the baseline.

The regression model achieves a MAE of 1.66, whereas both sets of human annotators achieve a MAE of 0.96 (see Table 1). This is an absolute difference of 0.70 in favor of the annotators. Or, in relative terms: the MAE of the learning algorithm is 72.6% larger than the human MAE.

4.3 Author vs. Annotator Labels *(iii)*

In order to test whether the model is influenced by the label source, we compare the results in the *Author* and *Annotator* rows of Table 2. The regressor performs noticeably better on the annotator ratings than on the author ratings when the full data set is utilized.As in the case of the human annotators, the regressor under-utilizes the extreme ratings. Alleviating this problem by incorporating the prior does not increase performance.

As mentioned, the annotator distribution lack the "flaps" in the extremes. In order to control for the influence of this bias, we *align* the two distributions. The number of reviews per rating is determined by the distribution with fewer reviews for the given rating. Thus, we end up with two data sets containing the same number of reviews per rating, and a total of 1,319 reviews.

The performance difference between the models trained on the aligned distributions is smaller, but still notable. This is an important result, indicating that the model's performance drop when trained on authors is not solely due to the underlying distribution, but to the *quality* of the ratings.

Even if the goal is to predict author ratings, it could still be advantageous to train on annotator-labeled data as indicated by the Ann/Aut row.

5 Related Work

Since the seminal study by Pang et al. (2002) using author-labeled IMDb user reviews, author-labeled data has been used for a wide range of domains, like user-generated product reviews (Dave et al., 2003), restaurant reviews with several aspect ratings (Snyder and Barzilay, 2007), movie reviews written by experienced film critics (Pang and Lee, 2005) and many more.

Pang and Lee (2005) also argue that it is unreasonable to expect a learning algorithm to predict ratings on a fine-grained scale if humans are not able to do so. They performed a minor study on human rating-inference to establish a suitable classification regime. They presented pairs of movie reviews from a single author rated on a 10-point Likert scale to two subjects (the authors themselves). Subjects had to decide whether one review was more, less, or equally positive than the other. They find that subjects correctly discern reviews separated by more than three steps, but accuracy drops when relative difference decreases. Pang and Lee (2005) also identify three obstacles for humans to accurately infer author ratings, namely *lack of calibration*, *author inconsistency* and *textually unsupported ratings*.

While suitable for their purposes, the preliminary study does not answer our research questions. First of all, the experiment is rather small (178 instances), which limits general validity and reliability. Second, Pang and Lee (2005) studied the human ability to discern *relative* and not *absolute differences*. If two reviews rated 7 and 8 are judged a 3 and a 4, the *relative* difference will be correctly identified, even though the guess is far off in absolute terms. Furthermore, single-author reviews dilute the effects of the three aforementioned obstacles. Inconsistencies within a single author will undoubtedly be smaller than inconsistencies between multiple authors. Single-author use also affects lack of calibration, since subjects can get a better feel for the writing style of one rather than several authors. Finally, we expect experienced authors to be less prone to producing reviews that do not support their ratings.

Annotator-labels are typically used when dealing with phrase-level semantics (Wilson et al., 2005; Wiebe et al., 2005; Socher et al., 2013). Alternatively, labels can be induced from salient sentiment-related features like emoticons (Pak and Paroubek, 2010; Go et al., 2009) or hashtags (Kouloumpis et al., 2011). In any case, the source of labeling tends to be a matter of convenience, rather than theoretical reflection. We did not find any considerations regarding potential differences between author and annotator labeling, implying that these are generally perceived as ontologically equivalent. We do not believe this to be the case.

6 Discussion

6.1 Human rating inference *(i)*

When inspecting the three rating distributions, we observe some interesting differences. First of all, the flaps (i.e. the upward going trend toward the extreme ratings) in the author rating distribution were not present in the annotator rating distributions. This phenomenon might very well be explained by the notion that *"the propensity to post online reviews is higher for movies that are perceived by consumers to be exceptionally good or exceptionally bad"* (Dellarocas and Narayan, 2006). However, this tendency does not explain why the same flaps are not present in the annotator distributions. One possible explanation is *risk aversion*. An annotator might estimate a review to be within the range of 6 to 10. They might also recognize that 10 is the most likely rating and 6 the least. However, in order to minimize the margin of error, picking 8 is a better option than 6 or 10, since it will ensure the annotator is within two steps of the author's rating. This behavior might be especially prevalent with crowdsourced annotators, who have a monetary incentive to minimize their error. This circumstance could explain the lack of flaps in their distribution. The expert annotators show some evidence of the flaps, but are less extreme than the authors (i.e. peaks are one step closer to the center of the distribution).

We would also like to underline the role of *wisdom of the crowd*. We saw that individual annotators were performing worse (with regard to both correlation and MAE) than the mean over all annotators. This was the case for both annotator types. Human ability to infer author ratings should thus be seen in light of these results. No individual annotator performed better than the mean of all annotators. The wisdom of the crowd effect might also explain crowdsource annotators performing as well as expert annotators. Using five crowdsource (vs. three experts) provides more robust estimates to counter sloppy annotators.

We might expect a simple yes or no answer to our initial research question whether humans are able to infer author ratings. Of course, this is not the case. Most annotator ratings were within a margin of two steps of the original author rating, which intuitively seems acceptable. Only slightly more than 5% were more than two steps off. These results indicate that humans in *most* cases are able to infer the original author rating with decent accuracy, if allowed to "work together".

6.2 Human vs. Machine *(ii)*

We feel it is safe to say that learning algorithms have *not* caught up with humans yet, when it comes to detection of semantic orientation in natural language text. This difference holds even though humans, too, fail in a considerable number of cases. Overall, our results provide an upper bound for the performance we can expect from learning algorithms.

6.3 Author vs. Annotator Labels *(iii)*

As hypothesized, using annotator labels resulted in a lower MAE than using author labels. Naturally, it is difficult to establish the exact reason for this observation, but we still believe that annotator labels follow a more regular, and thus predictable pattern than author labels. This effect is a result of the annotator labels being generated by the reader's interaction with the text.

The results obtained using the aligned rating distributions support this result. Aligning the distributions removes the flaps from the author distribution. The prediction error of the model trained on this data improves. For annotator labels, on the other hand, the aligned distribution gathers fewer reviews around the main peak of the distribution. Training the model on this data have a worse prediction error. These two results indicate that the regressor is biased towards predicting ratings around the peak of the distribution.

However, aligning the distributions also creates some problems. First, the reviews contained in the author and annotator data sets are not the same: 18.6 % of the reviews differ, although this should not be of significant advantage to any of the data sets. Second, the aligned distributions are not evaluating the natural rating distributions. However, results follow the same trend as when using the unmodified distributions (and hence the exact same reviews): annotator labels outperform the author labels. Overall, though, the data supports our hypothesis that annotator labels are more aligned with the text than author labels.

7 Conclusion

We find that readers are able to infer author ratings from the review text fairly accurately (on average less than one step away from the author rating on a 10-point scale). However, in more than 5% of the cases, the annotators were off by at least a three-step margin.

Human annotators outperform a linear regression model, even when adding a prior. We believe that no trivial adjustments can bridge this gap. However, the model achieves better results using annotator rather than author ratings, even when controlling for distributional shape as a confounding factor. Thus, it is questionable if author ratings are optimal as data labels for sentiment analysis.

References

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

Chrysanthos Dellarocas and Ritu Narayan. 2006. What motivates consumers to review a product online? a study of the product-specific antecedents of online movie reviews. In *WISE*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326.

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 on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Benjamin Snyder and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL*, pages 300–307.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Mark Steyvers, Brent Miller, Pernille Hemmer, and Michael D Lee. 2009. The wisdom of crowds in the recollection of order information. In *Advances in neural information processing systems*, pages 1785–1793.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.