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

Sentiment analysis models often use ratings as labels, assuming that these ratings reflect the sentiment of the accompanying text. We investigate *(i)* whether human readers can infer ratings from review text, *(ii)* how human performance compares to a regression model, and *(iii)* whether model performance is affected by the rating "source" (i.e. original author vs. annotator). We collect IMDb movie reviews with author-provided ratings, and have them re-annotated by crowdsourced and trained annotators. Annotators reproduce the original ratings better than a model, but are still far off in more than 5% of the cases. Models trained on annotator-labels outperform those trained on author-labels, questioning the usefulness of author-rated reviews as training data for sentiment analysis.

1 Introduction

Machine learning approaches have become the dominant paradigm for sentiment analysis since introduced by Pang et al. (2002). While these approaches produce good results, they need to be trained on sufficiently large labeled data sets. 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 (Dave et al., 2003; Pang and Lee, 2005; Snyder and Barzilay, 2007; Elming et al., 2014, i.a.). While it is 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 rating is. In this study, we ask

(i) whether 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 *trained 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 trained annotator ratings. Author ratings supposedly capture the essence of the author's sentiment, but we do not expect annotators to perfectly reproduce these ratings based on text alone.

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 both labeling sources as ontologically interchangeable. That is, it does not matter whether a 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 is not at all self-evident.

To the best of our knowledge, no previous study has investigated the assumption that the sentiment of a text can be objectively inferred. Since sentiment analysis is still far from being solved, investigating this core bias can help address current limitations.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2527–2532,
Lisbon, Portugal, 17-21 September 2015.

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 rated on a 10-point scale. 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. Overall, 171 annotators participated in the task. We incorporate control items in the annotation task, and each annotator starts by completing eight of these test questions. Further test questions are inserted randomly throughout the annotation tasks. We define a range of permitted ratings (within two steps of the original author rating). 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 part of the final data set.

We use three trained annotators to rate a 20% subset of the reviews: 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 data set is used for the regression-based tasks *(ii)* and *(iii)*.

3 Experiments

We want to establish the reproducibility of author ratings from text 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 mainly Spearman's ρ, but also report Krippendorff's α and Cohen's κ. The latter is a standard agreement measure, but does not work as well for ordinal ratings such as these, since it assumes a uniform distribution to compute chance agreement.

We have two sources of human annotations, namely three trained annotators and five crowdsource annotators per review. In order to obtain our final ratings, we average over each of those annotation sources.[1] This result is more robust towards individual biases and misinterpretations. This effect is known as *wisdom of the crowd* and well-documented in the literature, e.g. Steyvers et al. (2009). However, we also wish to investigate how well individual annotators perform. Therefore, we also compute error and pairwise correlation for each individual annotator with the authors or other annotators, and then average over the pairwise comparisons for each annotator type.

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 trained annotators, respectively.

In order to control for different levels variance in the rating distributions, we *align* the crowdsource annotator and author distribution by sub-sampling. The number of reviews per rating is determined by the distribution with fewer reviews for the given rating. The resulting two data sets contain the same number of reviews per rating, and a total of 1,319 reviews. The main implication of aligning the distributions, is that variance for both distributions will be identical, thus making the comparison more appropriate.

3.1 Model

We use a linear least-squares model with L_2 regularization *(ridge regression)* to reduce overfitting.[2] L_2 imposes a term α, which penalizes the parameters w of the model if they grow too large. Formally, w can be calculated by

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

We also experiment with incorporating a prior,

[1] Aggregating with an item-response model like MACE (Hovy et al., 2013) results in worse estimates, since it requires nominal data.

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

to model the tendency of authors to use the extremes more than predicted by a Gaussian distribution. We use a beta distribution with shape parameters $(0.8, 0.8)$.

We use 10-fold cross validation for robust results, and 5-fold cross validation on each of the then training folds in order to determine the optimal α.

We use bag-of-words features, including all unigrams appearing more than twice in the training data. [3]

4 Results

As baselines, we use the average rating over each of the *entire* rating distributions. Since the distributions differ between author and annotator ratings, the baseline differs from task to task.

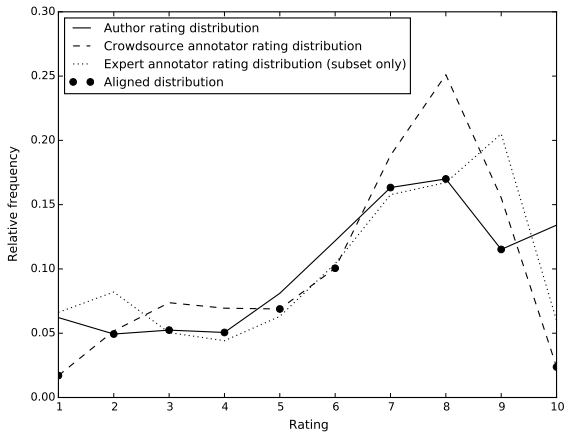

Figure 1. Rating distributions for authors, crowdsourced, and trained annotator ratings. Dots indicate aligned distribution.

Human Rating Inference *(i)* Figure 1 shows the rating distributions of the three human sources. Note the more peaked distributions of both annotator types, as compared to the author distribution. Especially crowdsource annotators have a smaller variance. Furthermore, the author distribution includes more extreme ratings, while the annotator distributions show no such "flaps".

[3]To rule out that the lack of any syntactic information (which human annotators use) disadvantages the model, we also experimented with including dependency triples (*dobj* and *nsubj*, the most frequent dependencies) using the Stanford Parser (Klein and Manning, 2003). However, performance did not improve, so due to limited space, we did not further explore this option.

Trained annotators are more correlated with one another ($\rho = 0.90, \alpha = 0.67, \kappa = 0.34$) than the crowdsource annotators ($\rho = 0.75, \alpha = 0.52, \kappa = 0.09$). Likewise, we see a lower MAE among trained annotators (0.75) than among crowdsource annotators (1.13), indicating a more diverse set of ratings for the latter.

		aut - cs	aut - tr	tr - cs
Corr. (ρ)	Mean	0.84	0.85	0.94
	Ind.	0.71	0.83	0.80
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 comparisons between author (aut), trained (tr), and crowdsource (cs) ratings.

Table 1 compares the different rating sources. We find a higher correlation and lower error between the two sets of annotator ratings than between the author ratings and any of the annotator ratings. However, when comparing the individual rating correlations, author ratings are highly correlated with trained, but not with crowdsource annnotators, showing the uncertain nature of crowdsource annotators.

There is no discernible difference between the two annotator groups in terms of error margins. 80% of mean annotator ratings, regardless of source, are correctly inferred or one step off. Slightly more than 5% of ratings are more than two steps off. However, comparing *individual* annotator ratings instead of mean ratings, some crowdsource annotators are a full nine steps off, and in a single case, even one of the trained annotators 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 Aut=authors; Ann=crowdsource annotators; Ann/Aut=trained on annotators and tested on authors.

In the next section, we compare the MAE be-

tween author ratings and the two sets of annotator ratings with the performance of the linear model. Conveniently, the numbers for the two types of annotators are equal (0.96), making it unnecessary to distinguish between them.

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 an 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.

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 using the full data set. Just as human annotators, the regressor under-estimates the extreme ratings (i.e., the "flaps"). Even incorporating a prior to address this shortcoming does not increase performance.

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

The Ann/Aut row indicates that even if the goal is to predict author ratings, it could still be advantageous to train on annotator-labeled data.

5 Related Work

Since Pang et al. (2002) used author-labeled IMDb user reviews in their seminal study, 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 from experienced film critics (Pang and Lee, 2005), business reviews (Hardt and Wulff, 2012; Elming et al., 2014; Hovy, 2015), 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. To test this, 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. Subjects correctly discerned reviews separated by more than three steps, but accuracy dropped 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 study does not answer our research questions. First of all, the experiment is rather small (178 instances), which limits general validity and reliability. Second, the study tests the human ability to discern *relative*, 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 are undoubtedly smaller than inconsistencies between multiple authors. Single-author use also affects lack of calibration, since subjects can adjust to the writing style of one author better than that of several. Finally, we expect experienced authors to be less prone to producing reviews that do not support their ratings.

Annotator labels are typically used for 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; Tang et al., 2014) or hashtags (Kouloumpis et al., 2011). Often, the label source tends to be a matter of convenience, rather than theoretical reflection. The lack of considerations regarding potential differences between author and annotator labels implies that these are often perceived as ontologically equivalent. We do not believe this to be the case.

6 Discussion

Human rating inference ***(i)*** We observe some interesting differences between the three rating distributions. First, the "flaps" in the extreme

ratings in the author ratings are not present in the annotator rating distributions. This phenomenon might be explained by the observation 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 flaps are not present in the annotator distributions. One possible explanation is *risk aversion*. An annotator might estimate a review to be between 6 and 10. She might also estimate 10 to be 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 is especially prevalent with crowdsourced annotators, who have a monetary incentive to minimize their error, which could explain the lack of flaps in *their* distribution. Indeed, the trained annotators show *some* evidence of flaps, but are still less extreme than the authors (i.e. their mean is closer to the center of the scale).

We also want to stress the role of *wisdom of the crowd*. Individual annotators perform worse (with regard to both correlation and MAE) than the mean over all annotators. This holds 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 why crowdsource annotators perform as well as trained annotators: using five crowdsourced (vs. three trained) annotators provides more robust estimates to counter sloppy annotators.

We might expect a simple 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 two steps of the original author rating. Only slightly more than 5% were further 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".

Human vs. Machine ***(ii)*** Based on our results, learning algorithms are *still* worse than humans in detecting semantic orientation of 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.

Author vs. Annotator Labels ***(iii)*** As hypothesized, using annotator labels lowered the MAE more than using author labels. Presumably, annotator labels follow a more regular, and thus predictable, pattern than author labels, since the former are generated by the reader's interaction with the text.

The aligned-distribution results support this theory. Aligning the distributions controls for different levels of rating variation in the distributions, thus ruling it out as confounder for the MAE difference. The aligned-distribution results also indicate that the model is biased towards mean ratings: MAE improves for author labels, since the relatively high variation is eliminated, but worsens for the annotator labels, as variance increases.

However, alignment also creates problems. First, the reviews contained in the author and annotator data sets differ in 18.6 % of the cases, although this should not be of significant advantage to either set. Second, aligned distributions do not evaluate the natural rating distributions. However, results follow the same trend as when using unmodified distributions (and hence the exact same reviews): annotator labels outperform author labels. All this suggests that annotator labels are more aligned with the text than author labels.

7 Conclusion

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

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 rating variance as a confounding factor. This suggests that author ratings are *not* optimal data labels for text-based sentiment analysis models.

Acknowledgements

The authors would like to thank the anonymous reviewers and the members of the CoAStAL group for their helpful comments. This research was funded in part by the ERC Starting Grant LOWLANDS No. 313695.

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*.

Jakob Elming, Barbara Plank, and Dirk Hovy. 2014. Robust cross-domain sentiment analysis for low-resource languages. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7, Baltimore, Maryland, June. Association for Computational Linguistics.

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

Daniel Hardt and Julie Wulff. 2012. What is the meaning of 5*s? an investigation of the expression and rating of sentiment. In *Proceedings of EMNLP*, pages 319–326.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *NAACL*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of ACL*.

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.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

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.