Intelligent Calibration for Bias Reduction in Sentiment Corpora Annotation Process

Idan Toker[1], David Sarne[1], Jonathan Schler[2]
[1]Bar-Ilan University
[2]Holon Institute of Technology
idanokim@gmail.com, david.sarne@gmail.com, schler@gmail.com

Abstract

This paper focuses in the inherent anchoring bias present in sequential reviews-sentiment corpora annotation processes. It proposes employing a limited subset of meticulously chosen reviews at the outset of the process, as a means of calibration, effectively mitigating the phenomenon. Through extensive experimentation we validate the phenomenon of sentiment bias in the annotation process and show that its magnitude can be influenced by pre-calibration. Furthermore, we show that the choice of the calibration set matters, hence the need for effective guidelines for choosing the reviews to be included in it. A comparison of annotators performance with the proposed calibration to annotation processes that do not use calibration or use a randomly-picked calibration set, reveals that indeed the calibration set picked is highly effective—it manages to substantially reduce the average absolute error compared to the other cases. Furthermore, the proposed selection guidelines are found to be highly robust in picking an effective calibration set also for domains different than the one based on which these rules were extracted.

Introduction

Customer feedbacks are becoming increasingly crucial for the success of firms as they increase their understanding of strengths and weaknesses, as well as the general attractiveness of the products and services they offer (Dey, Haque, and Raj 2010; Milner and Furnham 2017). Feedbacks are also useful for prospective customers, serving as a signal for quality, usefulness and cost-effectiveness. As such, many companies are now offering a public feedback mechanism on their websites (Chen et al. 2017; Dellarocas 2003; Assimakopoulos et al. 2014). One important analysis tool for customer feedback is the automatic understanding of their sentiment (Moghaddam 2015; Gamon 2004). Still, despite rapid progress in the automation of data analysis, and recent advances in the field of sentiment analysis, automatic sentiment extraction still requires much user involvement, primarily in terms of labeling of training data (Mitra 2020).

The labeling (or "annotation") process, either carried out in house or with the use of crowdworkers (e.g., with Amazon Mechanical Turk, Prolific or Toloka), is inherently sequential. The annotator is presented with a textual feedback (termed "review" onwards) and returns a numeric sentiment score, repeatedly. This may result in a score bias, particularly when the annotator runs into a reviews sequence of rather extreme reviews, creating a biased reference point (anchor) which influences following scores.

Methods for dealing with anchoring bias are human-labor intensive. These include, among others, taking the majority sentiment label assigned by multiple crowdworkers for a given text (Sheng, Provost, and Ipeirotis 2008), incorporating gold standard questions or attention check questions in the task to identify and exclude unreliable or biased workers (Snow et al. 2008), and labeling a small subset of the data by experts in sentiment analysis to serve as a benchmark for the crowdworkers to compare their own labels against (Pang, Lee et al. 2008). As an alternative to the above, we propose an approach aiming at directly mitigating the anchoring effect. The method is based on providing a small set of reviews to be labeled by the annotators before they begin annotating the (substantially) larger batch of reviews assigned to them. The purpose of these calibration reviews is to reduce the average annotation error of the other reviews, meaning that the effort exerted in labeling of these reviews is overhead. Alas, the saving on manpower engaged in redundant labeling is often well worth this additional overhead.

Simply adding random reviews at the beginning of the process is not enough. We suggest that the selection of the calibration reviews be carried out intelligently, based on learning. In such case the selection of the calibration reviews should be made from an already labeled reliable set of reviews. We emphasize that even if such repository of labeled reviews should be constructed from scratch, the additional overhead is significantly smaller compared to the substantial overhead associated with the above reviewed methods.

In this study, we introduce an innovative review annotation model that employs a calibration set to meticulously quantify anchoring bias, utilizing the metric of mean absolute error (MAE). Through a comprehensive series of 17 experiments encompassing 1640 annotators on established platforms such as Prolific or Amazon Mechanical Turk, we delineate the trajectory of sentiment bias as annotators progress in their tasks. A significant contribution of our research is the elucidation of the profound impact of initial review exposure on annotator bias and the consequent emphasis on the judicious selection of calibration sets. Building

on this insight, we used a set of machine-learning models, enriched with behavioral features (Noti et al. 2016; Plonsky, Hazan, and Tennenholtz 2016). This approach provides a nuanced mechanism to predict MAE outcomes based on initial review interactions. Using feature importance analysis, we developed guidelines for the strategic selection of calibration sets. Our empirical comparisons reveal that our methodology yields a marked reduction in MAE, outperforming traditional methods such as random set selection or the absence of calibration. Further validation in alternative domains underscores the robustness and broad applicability of our research findings.

Related Work

Although sentiment analysis is a well-researched area, the majority of research to date has focused on algorithms and techniques for automatically extracting sentiment from text (Salinca 2015; Shandilya and Jain 2009). For example, (AL-Smadi et al. 2016) use classifiers to predict the sentiment polarity of reviews. Although simple, Naïve Bayes classifiers were often found relatively effective at performing tasks of sentiment analysis on different kinds of data. (Daniela, Hinde, and Stone 2009) compare different methods for automatic analysis and classification of training course web pages, finding the Naïve Bayes classifier to achieve the best performance in terms of F–measure score in that domain. Similar attempts for using Naïve Bayes classifiers for sentiment analysis on English Tweets and micro-blogging resulted in accuracy of 63% and 84%, respectively (Gamallo and Garcia 2014; Ou, Cao, and Mu 2015).

All in all, despite the great interest in automatic sentiment analysis and the encouraging results obtained, many of the above works acknowledge the fact that the task is highly challenging, for reasons such as incomplete information in the reviews analyzed, and biased and diverse information. Most importantly, many of the methods suggested, require rich annotated corpora for training, which is labor intensive and often suffer from annotators' bias.

Bias is a fundamental human trait that can significantly impact various aspects of life, including data annotation. In recent years, there has been a growing interest in understanding the bias encapsulated in annotations carried out by crowdworkers, which is often attributed to their diverse personal backgrounds. These annotator biases can adversely impact the overall accuracy of machine learning models. Here, bias can manifest in various forms, such as exhibiting preferences or prejudices towards specific demographics (e.g., gender, race, or age) in one's labeling (Ding et al. 2022) and having one's own gender influence his labeling of data (Lucy et al. 2020). Mitigating the influence of annotator biases greatly helps in extracting reliable labels, facilitating the refinement of machine learning models and the enhancement of their predictive capabilities.

Annotation bias can also be the product of the ordering by which the reviews that need to be presented to the annotator are presented, as this has a great impact over their perception and interpretation of the content. This is often referred to as "anchoring effect", which was shown by many researchers. For example, (Wilson et al. 1997) refer to several innovative studies showing an anchoring bias in the absence of explicit instructions regarding a task. (Brewer and Chapman 2002) carry out several experiments proving how simple it is to bias numeric answers in surveys. (Wall et al. 2019) suggest that in visual data analysis, people leverage cognitive and perceptual systems to think about data by analyzing the views created.

Our solution to the problem aims to reduce this effect by generating a more balanced (or neutral) anchor to begin with. Our use of this initial calibration set mitigate the initial bias that would otherwise occur based on the original first reviews of the dataset. This, as we show experimentally, results in reducing the potential for bias-induced inaccuracies in the final dataset.

Model

The labeling process. We consider a set of N product or service reviews that need to be labeled for their author's sentiment as part of the corpus labeling process, such that it can then be used to train a machine learning model. The labeling is being carried out by people (typically crowdworkers), in a sequential manner, i.e., an annotator receives the reviews one by one, and upon reading a review assigns a sentiment score to it (typically between 1-5, where 1 is most negative and 5 most positive).[1] The goal is to minimize the average labeling error, cross the different reviews, where the labeling error per review is the absolute difference between the sentiment score received from the annotator for that specific review and some ground truth according to its definition in the following paragraph. Formally, let x_i and y_i denote the annotator's label value and the ground truth sentiment value of the ith review in the annotator's sequence, respectively. The average labeling error of the annotator is the mean absolute error (MAE), defined as:

$$MAE = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N} \tag{1}$$

Ground Truth. The ground truth for the sentiment of a given review is the average of sentiment scores assigned to the review by the entire population (Borgholt, Simonsen, and Hovy 2015). In reality, the extraction of the ground truth based on this latter definition is infeasible,[2] hence we take the ground truth to be the average of a sufficiently-large set of sentiment scores elicited for that review from different annotators (Chung et al. 2019).

MAE Lower Bound. We emphasize that while the annotator is using a discrete scale for providing his sentiment score, the ground truth, by its definition above, uses a continuous scale. This means that even a theoretically perfect

[1] While traditionally the sentiment classification is of a binary nature (i.e., positive, negative and neutral), the use of a scale is far more informative and can be trivially mapped to the first.

[2] Let alone the fact that if one has the option to calculate the ground truth then the labeling process is not needed in the first place.

labeling process will result in some unavoidable MAE, measured as:

$$MAE = \frac{\sum_{i=1}^{N} |\lfloor y_i \rceil - y_i|}{N} \quad (2)$$

where $\lfloor y_i \rceil$ denotes the closest integer to y_i. The above is actually a theoretical lower bound to the average labeling error, clearly unachievable as people are heterogeneous in their beliefs and individual sentiment scores typically vary. Still, it can be used as a baseline for reasoning about and evaluating the improvement achieved with the proposed calibration method in comparison to other methods.

Calibration. We augment the above model by adding the option to include $k << N$ additional "calibration reviews" to be labeled by the annotator prior to annotating the N original reviews. The k reviews can be selected from a set of $k' > k$ reviews for which the ground truth is known.[3] These calibration reviews are added solely for the purpose of minimizing the average annotation error of the other N reviews, hence we have no use or interest in the labeling provided by the annotators for them. The value k is therefore exogenously set and represents the amount of overhead the requester (i.e., the one benefiting from the labeling process) is willing to incur in order to improve the quality of the N labels to be received.

The goal of this research is to come up with an effective method for selecting the subset of calibration reviews to be used in the above archetypal setting.

Calibration-Set Extraction Method

For selecting an effective calibration set we rely on learning, as detailed in the following paragraphs.

Dataset. The training data to be used consists of annotated review sequences. Each sequence holds the labeling resulting from a sequential annotation process of $k + N$ reviews by a single annotator and the ground truth sentiment of each review in the sequence.

For each sequence we exclude the labeling of the first k reviews, as these serve as calibration for the set, and compute the MAE, according to (1), over the remaining N labels. Although one might be inclined to select the specific k-reviews set that results in the minimum MAE as the calibration set, it is important to recognize that the minimum MAE score achieved within sets pertains to the particular individual who labeled that specific subset of $k + N$ reviews. Hence, our focus is on the k-sized subset that produces low MAE according to the prediction model developed in following steps.

Learning process and methods. We construct multiple machine learning models to predict high versus low MAE of the provided N reviews, contingent upon the initial k calibration reviews. To accomplish this, it is necessary to define a set of features that effectively characterize those initial k reviews and can be employed to predict high or low MAE of the labeling of the subsequent N reviews following the initial k reviews.[4] The considered features include:[5]

- The MAE and standard deviation of the sentiment scores used for calculating the ground truth of the k reviews in the subset.
- The proximity of the ground truth (continuous) value of each review to the nearest integer.[6]
- The distance between the ground truth of each of the k reviews and the extreme sentiment scores 1 and 5.
- The length (number of words) of the review, which can be used to estimate its complexity.
- The number of sentences in the review.

Each of the aforementioned review features should be calculated with respect to its position in the sequence of the k reviews.

Ultimately, we identify the most suitable prediction model among those constructed. Suitability can be determined based on various measures, e.g., ROC AUC and logloss, depending on the nature of the methods used. Utilizing the selected model, we have two options for extracting the calibration set. The first approach involves extracting all possible k-sized subsets of reviews out of the k'-size corpus and use the model to predict the MAE of a considerably large sample of N-sized different review sequences with each. From these, we can select the calibration subset associated with the lowest average predicted MAE. The second approach entails examining the significance of the various features comprising the prediction model, utilizing them as guidelines for the calibration set selection. The two methods are likely to produce similar calibration reviews, as naturally the reviews that best comply with the model's extracted rules will also generate low expected MAE. The advantage of the first approach is that it directly correlates the selected calibration set with its MAE-performance across an extensive set of examples, hence it can be more accurate. The second approach, on the other hand, offers selection rules rather than a specific reviews subset, hence provides further insights regarding what kind of review combinations will effectively influence annotators. Furthermore, a set of selection rules is more intuitive for use in a new domain. Hence we opt for using the second approach to compute the calibration sets in our experiments.

[3] Either a publicly available set or one constructed from scratch by eliciting sentiment from a large group of annotators.

[4] The selection of models depends on the number of available observations. Since data collection can be relatively expensive, we are likely to rely on a small set of observations, thus utilizing methods such as Linear Regression and Tree-based models. Possessing a larger number of observations would facilitate the use of Deep Neural Networks and Gradient Boosting approaches.

[5] While features pertaining to the reviews content can also be considered for model construction, we choose to exclude them primarily due to their complexity (requires NLP capabilities). Furthermore, their inclusion precludes the transfer of the model to a new domain (which our approach enables). Similarly, including annotators-related features can sure improve the models, though in most crowdsourcing platform annotators background is not disclosed to the requester.

[6] See a related discussion in the former section.

Experimental Infrastructure

The proposed approach was tested experimentally. As an infrastructure for our experiments, we developed a web-based system that enables manual annotation of reviews for their sentiment. The system receives a set of reviews and presents them to the annotator either in a pre-defined or random order, one at a time. The annotator then needs to label each such review with an integer sentiment score, ranging from 1 ("the reviewer is strongly negative towards the reviewed service or product") to 5 ("the reviewer is strongly positive towards the reviewed service or product"). Once assigning a sentiment score to a review, the annotator moves on to the next one, with no option to go back and change any of the sentiment scores assigned to previous reviews.

Experimental Design

Application Domains and Review Sets

We set the number of reviews that need to be labeled, N, to 15 and the number of calibration reviews, k, to 3.[7] That requires at least 18 reviews for which the ground truth is known in a given domain in order to pick the calibration reviews and calculate the MAE in retrospect. The primary set of reviews used, k', included 20 *Food reviews* (denoted $FOOD_A$) taken from a public Kaggle dataset.[8] For validation purposes we also used an alternative set of 15 different food reviews (denoted $FOOD_B$) taken from the same repository and a set of 18 *Hotel reviews* (denoted $HOTELS$) also taken from a public Kaggle dataset[9]. For comparing the proposed calibration-set selection method to a random set of calibration reviews we used a set of 10 additional food reviews to choose from (denoted $FOOD_C$). While sentiment labeling is primarily required whenever the star based rating (termed "star-score" onwards) of the different reviews is not available, having access to this parameter turned useful in guiding the process of sampling the reviews to be included in the sets $FOOD_{A/B/C}$ and $HOTELS$. Specifically, we picked at least three reviews for each star-score value in order to properly cover the entire spectrum of sentiments, with some additional reviews of star-scores 2-4 which are typically more controversial in terms of sentiment.

Subject Recruitment

We relied on subjects from online labor markets, as this is where one would normally go when recruiting annotators for sentiment analysis (Anderson et al. 2019). Specifically, we used Prolific, which is now being used to wide effect for scientific inquiry and was found to produce high quality data with a reliable ability to replicate existing results (Peer et al. 2017; Palan and Schitter 2017). Several treatments, as detailed in the following paragraphs, were replicated and executed also in Amazon Mechanical Turk (AMT) in order to validate that population diversity cross-platforms does not change our results, qualitatively. Subjects were qualified to participate in our experiments if they had an approval rate greater than 96% (98% in AMT) and are coherent in English. IRB approval was received from the institutional committee.

Treatments

Overall, we had 17 experimental treatments. These can be divided into four categories, each aiming to support a different aspect of the proposed calibration method: (1) providing ground truth for the different datasets; (2) validating that indeed the anchoring bias emerges in sentiment annotation with these datasets; (3) providing a comprehensive performance evaluation of the proposed calibration-set selection method; and (4) evaluating the robustness of the proposed approach, using transfer learning to a new domain. To eliminate any carryover effects, we used a between-subject design, i.e., each of our subjects participated in only one of the 17 experiments. All experiments, except those explicitly marked as "AMT" were carried out using Prolific.

Ground Truth Treatments. Here we had three treatments: $FOOD_A20Truth$, $FOOD_B15Truth$, and $HOTEL15Truth$, using the datasets $FOOD_A$, $FOOD_B$, and $HOTELS$, respectively. Each subject in any of the treatments received a random ordering of the reviews within the relevant set and was asked to provide their sentiment score. The ground truth of each review was calculated as the average of the scores it received.

Anchoring-Bias Validation Treatments. To validate the presence of anchoring bias, we conducted 7 treatments by fixing the first reviews as follows:

- $FOOD_AAnchorNegPro$, $FOOD_AAnchorNegAMT$: The 2 most negative $FOOD_A$ reviews presented first (once on prolific and once on AMT).
- $FOOD_AAnchorPosPro$, $FOOD_AAnchorPosAMT$: The 3 most positive $FOOD_A$ reviews presented first (once on prolific and once on AMT).
- $FOOD_AAnchorRandAMT1/2/3$ 1/2/3: 3 random $FOOD_A$ reviews first (with 3 ordering options of positive, negative and moderate reviews)

Evaluating the performance of the proposed calibration-set selection method. Here we relied on the primary $FOOD_A$ set for selecting the three calibration reviews (see more details below). These were then used for evaluating the performance of the calibration set when labeling the $FOOD_B$ reviews and the remaining $FOOD_A$ reviews:

- Treatments using $FOOD_A15$ - upon removing five calibration reviews from $FOOD_A$, we used 15 reviews from those remaining. These were used in three treatments:
  - $FOOD_A15NonCalibrated$ - reviews of $FOOD_A15$ presented in random order, each subject receiving a different ordering.[10] This treatment did not use the calibration reviews, hence can be used as a baseline for comparison.

[7] This choice aligns with the order of magnitude of reviews used in prior work, e.g., (Yano, Resnik, and Smith 2010) use 10.

[8] www.kaggle.com/snap/amazon-fine-food-reviews

[9] www.kaggle.com/datasets/datafiniti/hotel-reviews

[10] This cannot be extracted from the $FOOD_A20Truth$ experiment, as it uses 20 reviews.

  - $FOOD_A15Calibrated$ - same as previous one, except that subjects were presented with 3 calibration reviews.
  - $FOOD_A15RandCalibration$ - each subject was presented with 3 random reviews from $FOOD_C$, as the calibration set, followed by reviews of $FOOD_A15$. The idea was to check the effectiveness of the specific calibration set used in $FOOD_A15Calibrated$ compared to a random selection set.

- Treatments using $FOOD_B$:
  - $FOOD_B15NonCalibrated$ - similar to the treatment above on $FOOD_A$, except that using the $FOOD_B$ dataset as a baseline for comparison.
  - $FOOD_B15Calibrated$ - same as $FOOD_B15NonCalibrated$, but presenting the 3 calibration reviews first (as in $FOOD_A15Calibrated$).

Evaluating robustness using transfer learning to a new domain. To evaluate robustness, we tested transferring the calibration rules learned from the food reviews domain to a new hotels domain using the $HOTELS$ dataset.

- $HOTEL15NonCalibrated$ - To be used as baseline. Similar to the $FOOD_A15NonCalibrated$ but using the $HOTELS15$ dataset. The justification for this treatment, given that we already had the $HOTEL15Truth$ treatment, is the same as the one given for $FOOD_A15NonCalibrated$.
- $HOTEL15Calibrated$ - Using 3 calibration reviews (similar to $HOTEL15NonCalibrated$).

Experimental Flow

Each subject first received a brief explanation of the task and expressed his consent to participate. Then, demographic details were collected (age and gender). Next, the subject received thorough instructions regarding the task, including the definition of the term "review's sentiment", the sentiment's numeric scale, the main elements of the user interface and the way he is supposed to provide the sentiment scores of the reviews to be presented. After a short quiz that checked the subject's understanding of the above, he was sequentially presented with the reviews assigned to him and was asked to provide the proper labels.

Subject Filtering

Two means were used to eliminate the presence of noisy labeling. The first is the qualifying quiz reported above. The second included two attention checks based on two highly polarized reviews, one with extremely negative and one with extremely positive sentiment. Annotators who did not adhere to the correct response pattern of providing the two lowest or highest corresponding scores were filtered out.

	Treatment	Subjects
Ground Truth	$FOOD_A20Truth$	180
	$FOOD_B15Truth$	123
	$HOTEL15Truth$	123
Anchoring Bias Validation	$FOOD_AAnchorNegPro$	107
	$FOOD_AAnchorNegAMT$	51
	$FOOD_AAnchorPosPro$	103
	$FOOD_AAnchorPosAMT$	47
	$FOOD_AAnchorRandAMT1/2/3$	40/45/50
Performance Evaluation	$FOOD_A15NonCalibrated$	155
	$FOOD_A15Calibrated$	125
	$FOOD_A15RandCalibration$	110
	$FOOD_B15NonCalibrated$	60
	$FOOD_B15Calibrated$	60
Robustness Evaluation	$HOTEL15NonCalibrated$	106
	$HOTEL15Calibrated$	155

Table 1: Experiments

Results

Overall we had 1640 annotators taking part in our experiments (excluding those filtered according to the above guidelines), according to the following breakdown to treatments:[11]

The elevated number of participants in the ground truth experiments derives from the need to provide as accurate baselines as possible, as this is the basis for evaluation. In the other treatments the varying number of subjects is the result of the need to reach statistical difference, i.e., we always started with an initial batch, and added as necessary for statistical validness. The percentage of men and women within subjects is 59% and 41%, respectively, and ages averaged at 30.3 and ranged between 18-88 cross treatments.

MAE lower bounds, according to equation 2, calculated based on the data collected in the $FOOD_A20Truth$, $FOOD_B15Truth$ and $HOTEL15Truth$ treatments are 0.19, 0.25 and 0.25, respectively. The comparative analysis that follows uses these (theoretically unachievable) lower bounds as a baseline, focusing on the added MAE (i.e., beyond these lower bounds) resulting from the use of (or lack of) calibration. In the following paragraphs, unless otherwise stated, we use a t-test for checking the statistical significance of any comparison between the difference in the MAE of different treatments.

Anchoring Effect Validation

Calibration aims to rectify significant absolute errors in early sentiment scores, often due to biases from insufficient comparison baselines and the anchoring effect of initial reviews (Callegaro, Manfreda, and Vehovar 2015). Figure 1, based on treatments $FOOD_A15NonCalibrated$ and $FOOD_B15Truth$, underscores that the error, after initially increasing due to anchoring bias, decreases as the annotator gains exposure to a broader review range. The error for the initial five reviews exceeds that of the last five by 19%, and intriguingly, doesn't stabilize even after 15 reviews, indicating the effect's persistence.

[11] All data and results can be downloaded from https://tinyurl.com/ftd83pjb.

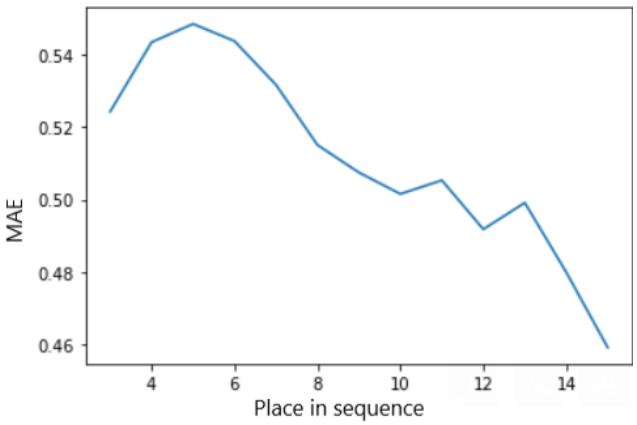

Figure 1: Average MAE over number of reviews already processed (moving average over 5 reviews, calculated on a minimum of 3 reviews).

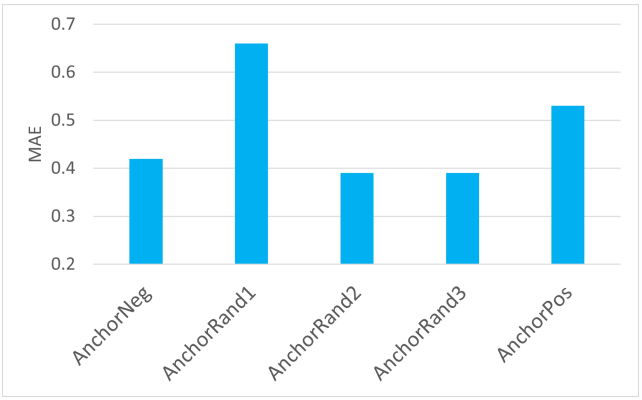

Figure 2: Anchoring-bias validation treatments - the MAE obtained with different calibration subsets using $FOOD_A$.

Our findings suggest augmenting the review set enhances sentiment accuracy. The start of the process should have strategically chosen instances, as indicated in Figure 2. Different introductory reviews significantly alter the MAE and overall sentiment. Positive reviews lead to lower scores overall, whereas negative ones yield higher averages. However, MAE's magnitude doesn't strictly align with average scores, and even random sets might produce higher MAEs than extreme review sets. Statistically, these MAE variations across treatments are significant ($p - value < 0.01$).

To summarize, the initial review set plays a pivotal role in anchoring bias during sentiment labeling. By judiciously selecting a calibration set, this bias, and hence the MAE, can be markedly reduced.

Calibration-Set Extraction Rules

We utilized a logistic regression classifier to discern between high and low MAE, setting the MAE classification threshold at 8.5 (which was the median value per annotator). The L2 penalty was employed for regularization (Qin and Lou 2019), with the model iterating 100 times. Using an 80:20 train-test split, we achieved a model accuracy of 61%, recall of 77%, and precision of 60% on the test set. Two feature selection methods were applied: SHAP values (Lundberg and Lee 2017) prioritized vital features, and Pearson correlations (Nasir et al. 2020) eliminated highly inter-correlated features. These techniques refined the feature set, reducing overfitting risk and enhancing model generalization.

We determined five main binary features guiding calibration set selection:

1. Avoid extremely negative first reviews (i.e. lowest 20 percentile).
2. Eschew extremely positive initial reviews (i.e. highest 20 percentile).
3. The second review should be exceptionally negative (i.e. in lowest 20 percentile)..
4. The third review's ground truth should be around 3.
5. Maintain a low standard deviation for the second and third reviews (i.e. within the 50th percentile of the lowest standard deviations).

The model's ROC AUC (Bradley 1997) was 0.637, confirming its efficacy and stability in differentiating between high and low MAE. Guided by these insights, we chose calibration reviews for the $FOOD_A$, $FOOD_B$, and $HOTEL$ datasets.

Calibration Performance

In assessing the efficacy of our calibration set, we compared the MAE to results from treatments with random or no calibration. The calibrated approach achieved an MAE of 0.47, outperforming non-calibrated (0.58) and randomly calibrated (0.54) approaches. These differences were statistically significant ($p - value \leq 0.0001$), with MAE reductions of 19% and 13% against the non-calibrated and random calibrations, respectively. Figure 3 showcases the MAE trends across the review sequence for three treatments. Contrary to the typical pattern where MAE decreases progressively, calibration (in its both forms) delivers immediate and consistent MAE reductions. Specifically, the calibrated approach achieves an MAE of 0.46 for the initial five reviews, contrasting with 0.61 without calibration. Notably, this calibrated MAE is akin to the value reached by the non-calibrated method after processing 15 reviews. Furthermore, our intelligent calibration lowered the standard error by 8% when compared to other methods.

Additionally, the benefits of our calibration approach were further demonstrated on the $FOOD_B$ dataset. Using calibration, an MAE of 0.34 was achieved, contrasting with 0.44 without calibration, yielding a statistically significant reduction of 22.7% ($p - value < 0.0001$).

Robustness

Finally, we test the applicability of transferring the calibration rules learned for the domain used (food reviews) to a new domain. This last experiment measures the extent to which the rules can be generalized, in order to save the learn-

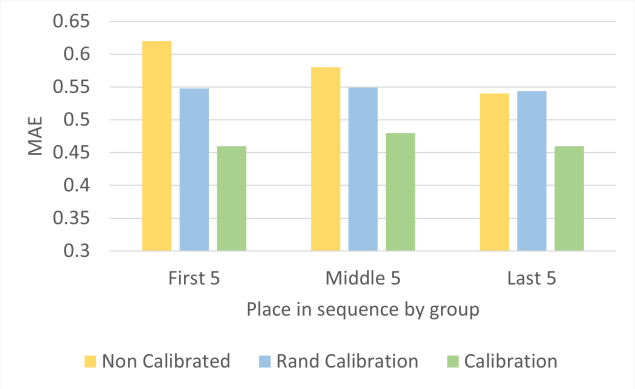

Figure 3: MAE comparison of $Food_A$ experiments, in division into the first, middle and last five reviews in the sequence.

ing process required when labeling sentiments for a new domain. For this purpose we used the $HOTELS$ data set.

We compare the MAE results obtained with calibration ($HOTEL15Calibrated$ treatment) to those obtained with no calibration ($HOTEL15NonCalibrated$ treatment).[12] The MAE with the calibration set is 0.16, compared to 0.21 with no calibration (a 24% improvement). The difference is statistically significant ($p - value \leq 0.0002$), meaning that the three-reviews calibration set picked based on the above guidelines was highly effective, even though the guidelines were extracted for a different domain.

Cross-Platform Validation

Comparing the results obtained with AMT and Prolific for similar settings of highly positive and highly negative initial sets of reviews (treatments $FOOD_AAnchorNegPro/AMT$ and $FOOD_AAnchorPosPro/AMT$), we find that the mean scores are almost identical.

Discussion and Conclusions

The encouraging results presented in the former sections suggest that the proposed calibration-based method for reducing annotation biases resulting from first-reviews anchoring is viable and effective. In particular, the guidelines for picking the calibration set were found useful not only in constructing an effective calibration set for the domain on which training took place ("foods"), but also for a completely different domain ("hotels"), providing a strong signal to its robustness. The importance of these results is leveraged by the importance of the application domain and the growing need in developing annotated corpora of reviews and their sentiment.

Indeed, the proposed method is associated with some overhead, primarily in the form of the unnecessary manual labeling of the reviews in the calibration set. Additional overhead may be experienced in cases there is no labeled corpus from which the calibration set's reviews are chosen. These overheads are substantially smaller than those associated with existing methods for dealing with labeling bias—as they all require much redundancy in the annotation process, either directly (when using several annotators for each review and picking the majority sentiment label assigned) or indirectly, for replacing a poor annotator (when including gold standard questions and attention check questions in the task or when labeling a small subset of the data by experts in sentiment analysis to compare crowdworkers own labels against). Our calibration-based method, on the other hand, directly focuses on improving the quality of the annotator. The overhead it induces is fixed and does not depend on the number of reviews in the annotators batch.

One key parameter that was arbitrarily set in our implementation of the proposed method is the size of the calibration set. Obviously, this parameter has a huge effect over performance hence it is advised to extract the learning-based guidelines using different set sizes and chose the best-performing one. Alas, the size with which the maximum improvement in average absolute error is achieved is not necessarily the optimal one, as one needs to take into consideration the cost/overhead of having the crowdworkers annotate this set. We leave the study of this tradeoff for future research. Another interesting direction for future work is the idea of using calibration sets along the process—not just at the beginning—as it is possible that anchoring may arise throughout. This, of course, would require a different logic for selecting calibration reviews.

[12]The ground truth was calculated in treatment $HOTEL15Truth$.

References

AL-Smadi, M.; Qwasmeh, O.; Talafha, B.; Al-Ayyoub, M.; Jararweh, Y.; and Benkhelifa, E. 2016. An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study. In *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, 98–103.

Anderson, C. A.; Allen, J. J.; Plante, C. N.; Quigley-McBride, A.; Lovett, A.; and Rokkum, J. 2019. The MTurkification of Social and Personality Psychology. *Personality and Social Psychology Bulletin*, 45: 842 – 850.

Assimakopoulos, C.; Eugenia, P.; C., S.; and Georgiadis, C. 2014. Online reviews as a feedback mechanism for hotel CRM systems. *Anatolia*, 26: 1–16.

Borgholt, L.; Simonsen, P.; and Hovy, D. 2015. The rating game: Sentiment rating reproducibility from text. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2527–2532.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7): 1145–1159.

Brewer, N.; and Chapman, G. 2002. The Fragile Basic Anchoring Effect. *Journal of Behavioral Decision Making*, 15: 65 – 77.

Callegaro, M.; Manfreda, K.; and Vehovar, V. 2015. *Web Survey Methodology*. ISBN 9780857028617.

Chen, L.; Jiang, T.; Li, W.; Geng, S.; and Hussain, S. 2017. Who should pay for online reviews? Design of an online user feedback mechanism. *Electronic Commerce Research and Applications*, 23.

Chung, J.; Song, J.; Kutty, S.; Hong, R.; Kim, J.; and Lasecki, W. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. *Proceedings of the ACM on Human-Computer Interaction*, 3: 1–25.

Daniela, X.; Hinde, C.; and Stone, R. 2009. Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science Issues*, 4.

Dellarocas, C. 2003. Goodwill Hunting: An Economically Efficient Online Feedback Mechanism for Environments with Variable Product Quality. volume 2531.

Dey, L.; Haque, S. M.; and Raj, N. 2010. Mining Customer Feedbacks for Actionable Intelligence. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, 239–242.

Ding, Y.; You, J.; Machulla, T.-K.; Jacobs, J.; Sen, P.; and Höllerer, T. 2022. Impact of Annotator Demographics on Sentiment Dataset Labeling. 6(CSCW2).

Gamallo, P.; and Garcia, M. 2014. Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. In *International Workshop on Semantic Evaluation*.

Gamon, M. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *COLING 2004: Proceedings of the 20th international conference on computational linguistics*, 841–847.

Lucy, L.; Demszky, D.; Bromley, P.; and Jurafsky, D. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open*, 6(3): 2332858420940312.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Milner, R.; and Furnham, A. 2017. Measuring Customer Feedback, Response and Satisfaction. *Psychology*, 08: 350–362.

Mitra, A. 2020. Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset). *Journal of Ubiquitous Computing and Communication Technologies*, 2: 145–152.

Moghaddam, S. 2015. Beyond Sentiment Analysis: Mining Defects and Improvements from Customer Feedback. In Hanbury, A.; Kazai, G.; Rauber, A.; and Fuhr, N., eds., *Advances in Information Retrieval*, 400–410. Cham. ISBN 978-3-319-16354-3.

Nasir, I. M.; Khan, M. A.; Yasmin, M.; Shah, J. H.; Gabryel, M.; Scherer, R.; and Damaševičius, R. 2020. Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training. *Sensors*, 20(23).

Noti, G.; Levi, E.; Kolumbus, Y.; and Daniely, A. 2016. Behavior-based machine-learning: A hybrid approach for predicting human decision making. *arXiv preprint arXiv:1611.10228*.

Ou, X.; Cao, Y.; and Mu, X. 2015. Classification of Micro-blog Sentiment Based on Naive Bayesian Classifier. In Zhang, R.; Zhang, Z.; Liu, K.; and Zhang, J., eds., *LISS 2013*, 585–589. Berlin, Heidelberg. ISBN 978-3-642-40660-7.

Palan, S.; and Schitter, C. 2017. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17.

Pang, B.; Lee, L.; et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2): 1–135.

Peer, E.; Brandimarte, L.; Samat, S.; and Acquisti, A. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70: 153–163.

Plonsky, O.; Hazan, T.; and Tennenholtz, M. 2016. Psychological Forest: Predicting Human Behavior. *SSRN Electronic Journal*.

Qin, J.; and Lou, Y. 2019. L1-2 Regularized Logistic Regression. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 779–783.

Salinca, A. 2015. Business Reviews Classification Using Sentiment Analysis. In *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 247–250.

Shandilya, S. K.; and Jain, S. 2009. Automatic Opinion Extraction from Web Documents. *Computer and Automation Engineering, International Conference on*, 0: 351–355.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263.

Wall, E.; Blaha, L.; Paul, C.; and Endert, A. 2019. A formative study of interactive bias metrics in visual analytics using anchoring bias. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part II 17*, 555–575. Springer.

Wilson, T.; Houston, C.; Etling, K.; and Brekke, N. 1997. A New Look at Anchoring Effects: Basic Anchoring and Its Antecedents. *Journal of experimental psychology. General*, 125: 387–402.

Yano, T.; Resnik, P.; and Smith, N. A. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 152–158.