Analyzing the Effects of Annotator Gender Across NLP Tasks

Laura Biester,[1] **Vanita Sharma,**[1] **Ashkan Kazemi,**[1]
Naihao Deng,[1] **Steven Wilson,**[2] **Rada Mihalcea**[1]
[1]Computer Science & Engineering, University of Michigan, USA
[2]Computer Science & Engineering, Oakland University, USA
{lbiester,svanita,ashkank,dnaihao}@umich.edu, stevenwilson@oakland.edu, mihalcea@umich.edu

Abstract

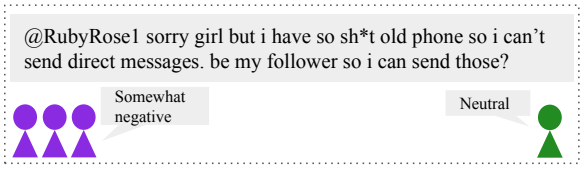
Recent studies have shown that for subjective annotation tasks, the demographics, lived experiences, and identity of annotators can have a large impact on how items are labeled. We expand on this work, hypothesizing that gender may correlate with differences in annotations for a number of NLP benchmarks, including those that are fairly subjective (e.g., affect in text) and those that are typically considered to be objective (e.g., natural language inference). We develop a robust framework to test for differences in annotation across genders for four benchmark datasets. While our results largely show a lack of statistically significant differences in annotation by males and females for these tasks, the framework can be used to analyze differences in annotation between various other demographic groups in future work. Finally, we note that most datasets are collected without annotator demographics and released only in aggregate form; we call on the community to consider annotator demographics as data is collected, and to release dis-aggregated data to allow for further work analyzing variability among annotators.

Keywords: annotator demographics, dataset construction, crowdsourcing

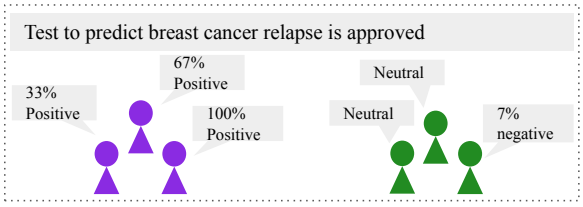
1. Introduction

Natural Language Processing (NLP) has seen a surge in the number of tasks as well as datasets during the last decade (Storks et al., 2019; Li et al., 2022; Nelson et al., 2022). With the success and requirements of deep learning techniques, large scale datasets have been proposed for various NLP tasks (Bojar et al., 2014; Yang et al., 2015; Zhang et al., 2015; Hendrycks et al., 2021). The mainstream formulation of supervised learning tasks across a range of areas trends towards preserving a single ground truth label for each example. However, such a setting ignores the possibility that different annotators may annotate the same example differently (Al Kuwatly et al., 2020). According to Basile et al. (2021), such disagreements between annotators are widespread. Moreover, Geva et al. (2019) showed that the annotator disagreement might significantly affect the performance of a model, indicating that our community may benefit from paying closer attention to annotator disagreement (Davani et al., 2022). Instead of focusing on high agreement scores for subjective datasets, we can be more cognisant of disagreements and build systems that are accommodating of different perspectives and needs, leading to novel insights and reducing harm (Uma et al., 2021; Davani et al., 2022).

In this work, we study how annotator demographics might relate to disagreements across four NLP tasks. Some examples of anecdotal differences in annotation in the datasets we study are shown in Figure 1. We include tasks that are commonly considered to be highly subjective (e.g., affect in text) and tasks that are considered more objective (e.g., natural language inference). In particular, we are interested in determining whether there are systematic, statistically significant differences in annotation that can be attributed to the gender of the annotators.



(a) An example from Sentiment Analysis Dataset



(b) An example from Affective Text Dataset

Figure 1: Examples of annotation difference between female annotators (left, purple) and male annotators (right, green).

First, taking a holistic view of the datasets, we develop a method to test if the overall distributions of annotations differ between male and female annotators. We visualize how the distribution of scores given by male and female annotators differs; for all four tasks (and a number of subtasks), the visualizations appeared to show some differences in the distribution of annotations by male and female annotators. However, after performing permutation testing, we find that for most tasks, we can not reject the null hypothesis that the observed differences could be due to random noise. For one task, sentiment analysis, we found that the male annotators gave more intermediary labels (e.g., somewhat positive/somewhat negative) than female annotators.

Next, we expand on an existing method (Prabhakaran et al., 2021) to study the extent to which male and female annotators agree with aggregate labels. In partic-

Dataset	# Male Annotators	# Female Annotators	# Datapoints	Mean Annotations per Datapoint	Annotation Type	Ratings per Datapoint
Affective Text	3	3	1000	6.00	Interval	7; anger, disgust, fear, joy, sadness, surprise, valence
Word Similarity	196	157	498	38.74	Ordinal	2; similarity, relatedness
Sentiment Analysis	736	744	14071	4.21	Ordinal	1
Natural Language Inference	282	211	1200	9.26	Ordinal	1

Table 1: List of linguistic tasks included in this study.

ular, we ask: (1) Is there a difference in the extent to which males and females agree with aggregate labels across the full dataset? and (2) Do female annotators have a higher agreement score with the aggregate of female annotators than with the aggregate of male annotators (and vice-versa for male annotators). For all pairs of agreement distributions that we study, we find no statistically significant difference.

While the results largely reveal no systematic difference in annotation that can be attributed to the gender of annotators and should thus be considered a negative result, our work contributes a robust framework with which to study differences in annotation between two or more groups from multiple angles. The framework is developed for two demographic categories and either ordinal or interval data, but could easily be applied to categorical or binary labels in a one-vs-all setup to work with multiple groups. We hope that this work can instigate further work on demographic differences in annotation, as our negative result cannot be generalized to all NLP tasks and datasets, nor can it be generalized to all demographic groups.

2. Related Work

Annotator Disagreement. As pointed out by Basile et al. (2021), annotator disagreement is ubiquitous, especially in the AI field (Smyth et al., 1994; Poesio and Artstein, 2005; Aroyo and Welty, 2015). People have long proposed that instead of ignoring such a disagreement and having a single groundtruth, we need to preserve annotations from different annotators (Poesio and Artstein, 2005; Recasens et al., 2012).

Reasons for Disagreement. Prior work has detected differences in data annotation with respect to gender in hate speech detection (Gold and Zesch, 2018), POS tagging and dependency parsing (Garimella et al., 2019). This work is often inspired by findings in linguistics, e.g., gender differences in the use of finite adverbial clauses (Mondorf, 2002). Beyond differences related to gender, researchers have studied difference in data annotation with respect to individual annotators (annotator bias) (Ross et al., 2010; Otterbacher, 2018; Larimore et al., 2021) and annotator disagreements (Pavlick and Kwiatkowski, 2019). Furthermore, Geva et al. (2019) reveals that annotators' individual differences affect model performance on natural language understanding tasks, which can lead to problems in model generalization to new users. Most prior work focuses on a single task or a single benchmark to study the data disagreement (bias) introduced by demographic features. In contrast, our paper considers four different NLP datasets, giving a more comprehensive analysis of potential differences across groups of annotators in a range of NLP tasks.

Disagreement Measurement. In order to systematically investigate the bias or disagreements, Geva et al. (2019; Garimella et al. (2019; Wich et al. (2020; Al Kuwatly et al. (2020) trained classifiers on subset of annotators, and use performance difference to demonstrate the existence of bias. Additionally, Wich et al. (2020) used an unsupervised graph method to group annotators and studied the difference between the groups. To measure the agreement between subgroups of annotators, Larimore et al. (2021) used Krippendorf's alpha score (Krippendorff, 2011), Gold and Zesch (2018) used Best-Worst-Scaling by Louviere et al. (2015), and Wich et al. (2020) reported Cohen's kappa (Cohen, 1960) and Krippendorf's alpha score.

3. Data

We study four NLP tasks using datasets that share the following properties: individual annotations are made available along with gender labels for those individuals, and items in the dataset have multiple annotations. We include datasets with interval and ordinal ratings; a summary of our datasets is presented in Table 1.

In the early stages of this study, we surveyed a number of language resource papers describing benchmark datasets to see if they mentioned the demographics of their annotators. To a large extent, we found that they did not; a few authors explicitly stated that no demographic information was collected, while one author stated that they included exclusively annotators located in the United States and Canada, likely to restrict the varieties of English represented by the annotators (Rajpurkar et al., 2016). With respect to user privacy, it is responsible practice not to collect more user-level information than is needed for data processing, and so it is reasonable that many previous studies chose not to store attributes of annotators like gender. However, not collecting these attributes also precludes the possibility of studying whether certain groups are under or over-represented in the dataset, and what effects representation may have on models.

We then emailed authors of twenty-three papers that did not explicitly state that they did not collect annota-

tor metadata, and we received responses from sixteen authors. Most authors stated that they did not collect or consider collecting annotator demographics alongside their annotations. It is therefore worth noting that the tasks we chose to study were largely chosen due to feasibility (access to data) rather than due to our intuitions about the tasks themselves. However, there are some inherent reasons why these tasks are interesting to study. First, affect and sentiment are subjective, but perhaps less clearly linked to identity than hate speech detection, a task for which annotator identity has been shown to correlate with differences in annotation (Gold and Zesch, 2018). Moreover, while it is typically considered to be more objective, systematic disagreement has also been found in natural language inference annotation (Pavlick and Kwiatkowski, 2019).

A limitation of the data we use is that there is little representation of people who do not fit in the gender binary; accordingly, we only study differences between male and female annotators in this work. We hope that larger datasets that indicate annotator characteristics will allow for studying gender differences in annotation beyond the gender binary in the future.

Detailed descriptions of each task follow.

Affective Text

The affective text dataset is from the SemEval-2017 Task 14 (Strapparava and Mihalcea, 2007). In particular, we use the test dataset, which consists of one thousand headlines, each annotated by six annotators. The original authors provided the gender of the annotators, three of whom were male and three of whom were female. We note that unlike our other datasets, gender was not self-reported by the annotators; rather, it was ascribed by the dataset collector, who was acquainted with the annotators. We are releasing the individual annotations for the SemEval-2007 Task 14 in conjunction with this paper, along with the gender of each annotator.[1]

The text is annotated for six emotions: anger, disgust, fear, joy, sadness, and surprise. The scale used for rating is 0 (the emotion is absent from the headline) to 100 ("maximum emotional load"). Additionally, each headline is annotated for valence on a -100 (highly negative) to 100 (highly positive) scale; 0 is neutral.

Word Similarity

The word similarity dataset was collected using Amazon Mechanical Turk. The annotators self-report a number of their demographic characteristics, including gender, which was reported in a dropdown listing Male, Female, and Other.

The annotators were given pairs of words, and asked to rate them on two five-point Likert scales. In the similarity task, they were asked how similar words were, on a scale from "completely different" to "very similar". In the relatedness task, they were asked how related words were, on a scale from "unrelated" to "very related". A number of examples were given to guide annotators:

Similar words: alligator/crocodile, love/affection

Related words: car/tire, car/crash

Annotations where the annotator incorrectly answered a qualification question were excluded. Approximately 25% of the annotated word pairs were drawn directly from SimLex-999 (Hill et al., 2015); the remaining pairs were inspired by Garimella et al. (2017). Specifically, they were balanced such that approximately 1/4 of the remaining pairs represented common word associations for four demographic groups: males, females, people located in the United States, and people located in India. This sampling strategy suggests that gender differences in the annotations are more likely than they would be in word pairs selected without considering gender.

Sentiment Analysis

We use a sentiment analysis dataset created with the intention of measuring age-related bias in sentiment analysis (Diaz et al., 2018). The training data text is sourced from samples in the Sentiment140 dataset (Go et al., 2009) containing the strings "young" and "old"; the test data text is scraped from blog posts that discuss aging. In collecting this data, the authors also collected a number of the annotator's self-reported demographic attributes, including but not limited to gender, age, and race. Genders reported in the dataset included Male, Female, and Nonbinary (one annotator). More than 1400 annotators rate sentiment on a five-point Likert scale (very negative, negative, neutral, positive, very positive). There are on average 4.21 annotations per datapoint, but we note that not all datapoints have a variety in annotator gender. The dataset is publicly available (Diaz, 2020).

Natural Language Inference

The natural language inference (NLI) dataset we use is CommitmentBank (De Marneffe et al., 2019). The annotators for the dataset were asked to determine the extent of speaker commitment to complements of clause-embedding predicates under an entailment canceling operator (e.g. question, negation, and so on). The authors provided us with the annotator gender and age, which were collected during the original annotation as part of the survey given to annotators. Gender was reported as free-text; we mapped MALE and MALE+ to the male category and FEMALE, WOMAN, FEMAL, and FEMALLE to the female category. We removed one annotator who reported different demographics in different Amazon Mechanical Turk tasks, and a small number of annotators whose reported gender did not fall into the male/female binary due to lack of data. Each datapoint is ranked on a seven-point Likert scale (-3: the annotator believes that the author of the text is certain that the prompt is false, 0: annotator believes that the author of the text is not certain whether the

[1] https://github.com/MichiganNLP/Affective-Text-Individual-Annotations

prompt is true or false, 3: the annotator believes that the author of the text is certain that the prompt is true). For the NLI task, items were labeled based on whether at least 80% of annotations were within three ranges: $[1, 3]$ (entailment), $[0]$ (neutral) or $[-3, -1]$ (contradiction) (Jiang and de Marneffe, 2019). We use the original ratings in the range $[-3, 3]$ in our analysis.

4. Methodology

We use two methods to robustly measure whether there are underlying differences in how male and female annotators annotate each of our four datasets. The first method, described in Section 4.1, directly measures the differences in overall scores given by male and female annotators. This type of analysis is likely to capture shifts in the distribution of scores given by different sets of annotators – for instance, it would capture if male annotators are more likely to label positive sentiments than female annotators. Even a simple linear shift in the distribution of annotations could affect models, especially if ordinal labels are converted to binary, which is a common experimental setting, e.g., in sentiment analysis (Socher et al., 2013). The second method, described in Section 4.2, takes into account aggregate scores to determine to what extent male and female annotators differ from various aggregates. If significant differences were found, this type of analysis would signal the need for multi-perspective modelling.

4.1. Distribution Analysis

We split annotations into those provided by male and female annotators, then visualize the scores given by those annotators; for the affective text dataset, we use a kernel density estimation plot because the annotations are on an interval. For ordinal data, we use a barplot. A key advantage of this type of analysis is that it produces clearly interpretable results; the plots allow us to directly see how the male and female annotators differ. To ensure the significance of our findings, we employ permutation tests; our null hypothesis is that gender does not affect the distributions of annotations. We define two test statistics, which we will refer to as t_{obs}.

For interval data, we begin by computing the cumulative sum of % of annotations for each gender with each possible rating from min (the minimum score in the range) to max (the maximum score in the range), which represents the empirical distribution function. Our test statistic is the area between the curves of the two empirical distribution functions. With cumulative sum vectors M and F, this area can be computed as $t_{obs} = \sum_{i=min}^{max} |M_i - F_i|$.

For ordinal data, we formulate our alternative hypothesis for each task by observation of how the two groups differ in the bar charts. We compute the difference in percentage of annotations with scores that meet the conditions of the alternative hypothesis. Specifically, given the total number of annotations and choices given to the annotators within the relevant condition C for the task, we compute $t_{obs} = \sum_{c \in C} |P(c|f) - P(c|m)|$, which represents the extent to which the distributions across labels differ for the two annotator groups.

We then randomly assign annotators to groups a (size = # of male annotators) and b (size = # of female annotators) and recompute the test statistic 10,000 times[2] with those groups instead of m (all male annotators) and f (all female annotators), creating an array of test statistics T_{perm}. Finally, we compute our p-value as the percentage of values in T_{perm} that are greater than t_{obs} (e.g., have a larger difference in the distribution).

4.2. Agreement Analysis

We expand upon the methodology from (Prabhakaran et al., 2021). They compute agreement using Cohen's kappa between each in-demographic annotator and the *majority vote* of the overall annotator pool. We use the same sentiment analysis dataset they study, but do not condense the labels to a binary scale. This means that we change our agreement metric to Krippendorff's alpha, due to its ability to compute agreement of ordinal and interval data between any number of annotators. We then compute the agreement of each in-demographic annotator with the *aggregate* of the overall annotator pool (labeled F-ALL, M-ALL). We also add two other measurements: the agreement of each in-demographic annotator and other in-demographic annotators (labeled F-ALLF, M-ALLM) and the agreement of each in-demographic annotator with all out-of-demographic annotators (labeled F-ALLM, M-ALLF). In all computations, the in-demographic annotator who is being compared to the aggregate is excluded from the aggregation.

We aggregate labels using the mean for interval data and a median for ordinal data; if the median is not an integer, we take the mean of two agreement scores for each annotator: one with the ceiling of the medians and one with the floor. The algorithm is formalized in Algorithm 1.

To measure the significance of our results, we performed t-tests for three metrics of interest across all of our datasets:

F-ALL vs. M-ALL This two-sided t-test determines if there is a statistical difference between the extent to which male and female annotators agree with the aggregate of all annotators. A difference here would show that the aggregate is more representative of one gender.

F-ALLF vs. F-ALLM This one-sided t-test determines if female annotators agree with other female annotators more than they agree with male annotators.

M-ALLM vs. M-ALLF This one-sided t-test determines if male annotators agree with other male annotators more than they agree with female annotators.

[2]Or fewer, if every possible permutation is covered with fewer tests

Algorithm 1 Agreement Comparison Algorithm

The algorithm takes as input A, a matrix of annotations where annotators are rows and datapoints are columns, G, a list of genders of annotators in A, and i, an individual annotator index.

Our aggregation function, `agg`, is median for interval data and mean for interval data. We use the `krippendorff` function for agreement.

$krip_{ALL}$ is used for F-ALL and M-ALL, $krip_{EQ}$ is used for F-ALLF, M-ALLM, $krip_{OTH}$ is used for F-ALLM, M-ALLF. The scores for each annotator i are used in the visualization.

procedure FILTER(A, G, i, all, eq)
  $A^G \leftarrow []$
  for $k \leftarrow 1, |A|$ **do**
    ▷ exclude target annotator from aggregate
    if `i != k` **then**
      if `all` **then**
        Append A_k **to** A^G
      else if `eq &&` G_i `==` G_k **then**
        Append A_k **to** A^G
      else if `!eq &&` G_i `!=` G_k **then**
        Append A_k **to** A^G
      end if
    end if
  end for
  ▷ Return other annotators depending on `all`/`eq`
  return A^G
end procedure

procedure ANN_AGREEMENT(A, G, i)
  ▷ aggregate and filter set of annotators
  $agg_{ALL} \leftarrow$ `agg`(FILTER($A, G, i, true, true$))
  $agg_{EQ} \leftarrow$ `agg`(FILTER($A, G, i, false, true$))
  $agg_{OTH} \leftarrow$ `agg`(FILTER($A, G, i, false, false$))

  ▷ find agreements with krippendorff's alpha
  $krip_{ALL}$ = `krippendorff`(A_i, agg_{ALL})
  $krip_{EQ}$ = `krippendorff`(A_i, agg_{EQ})
  $krip_{OTH}$ = `krippendorff`(A_i, agg_{OTH})

  return $krip_{ALL}, krip_{EQ}, krip_{OTH}$
end procedure

For both types of analysis, we use the Benjamini-Hochberg (Benjamini and Hochberg, 1995) False Discovery Rate correction to account for performing multiple statistical tests.[3]

5. Results

5.1. Distribution Analysis

Our plots of the affective text distributions (Figure 2) revealed an interesting pattern: the male annotators more commonly gave a rating close to zero, indicating the text was absent of an emotion. A similar pattern is observed for the valence task, for which annotations ranged from -100 to 100; the male annotators more frequently used 0, which was the "neutral" label.

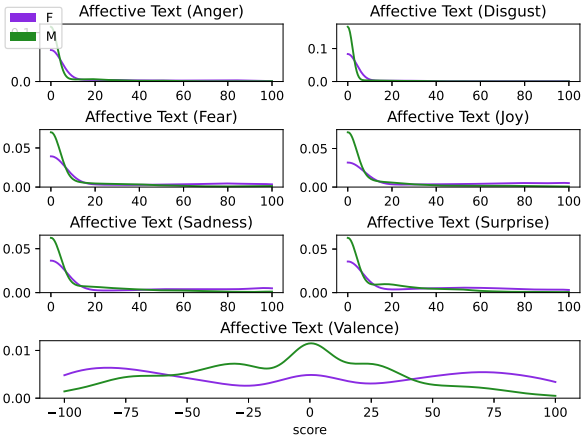


Figure 2: Kernel density estimation plots of affective text annotations.

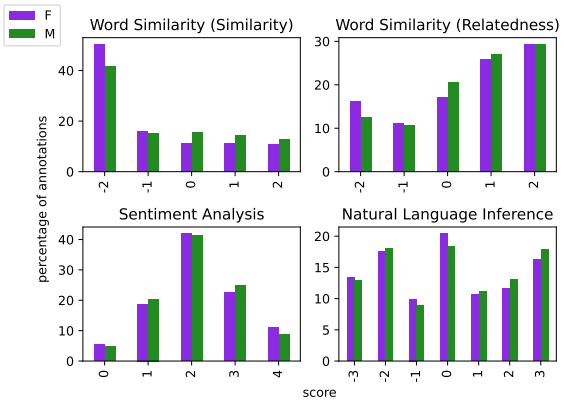


Figure 3: Bar plots of word similarity, sentiment analysis, and NLI annotations.

The plots for word similarity (Figure 3) do not reveal such stark differences; however, we do see that male annotators appear to generally give higher scores, while female annotators more commonly chose -2; while the differences are not as clear, a similar pattern can be observed for word relatedness and NLI. On the sentiment analysis dataset, female annotators appear to more commonly give scores neutral, very positive, or very negative ratings, while male annotators give more intermediary ratings of somewhat positive/somewhat negative.

These observations form the basis for the metrics used in our permutation tests. For the word similarity task, we compute the difference in percentage of scores greater than or equal to 0. For NLI, we compute the difference in percentage of scores greater than or equal

[3] $\alpha = 0.05$.

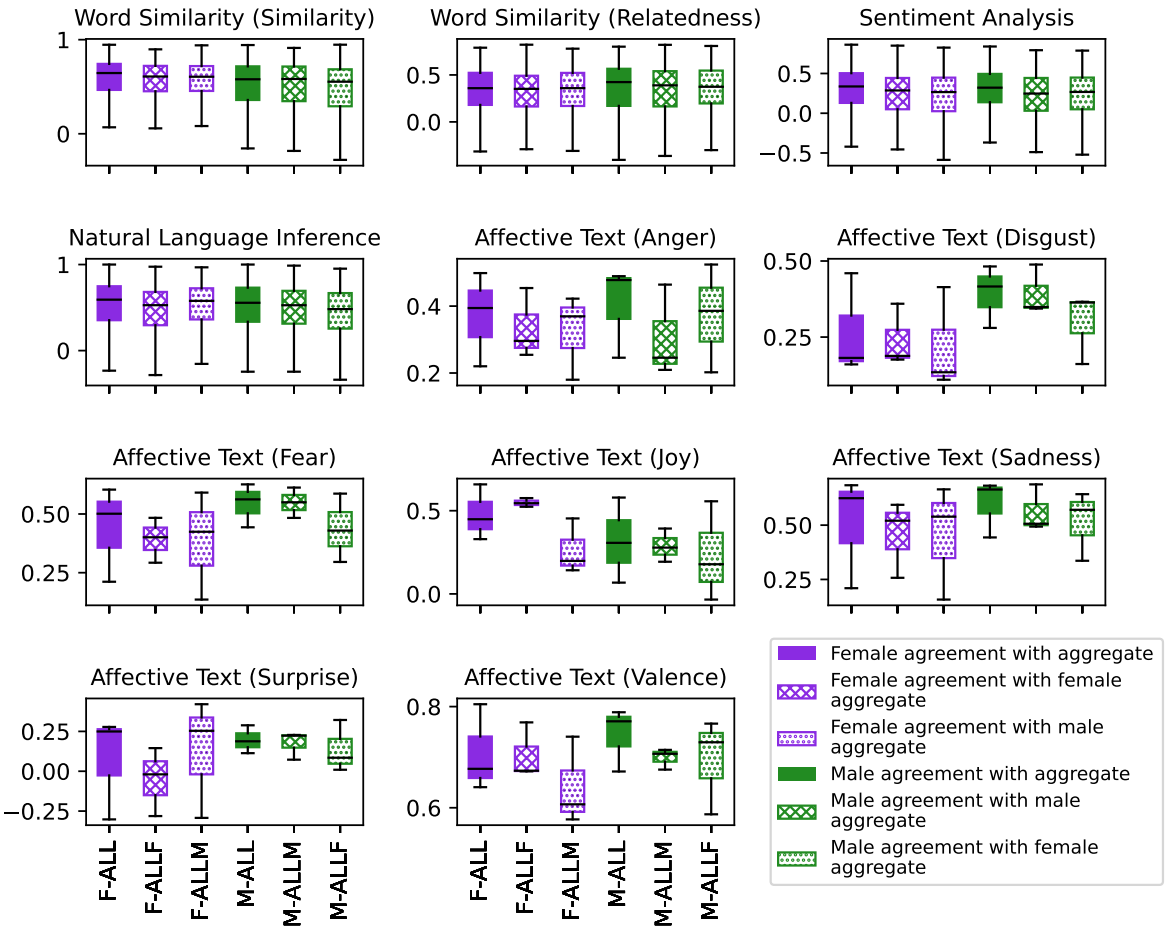


Figure 4: Boxplots representing the results of our agreement analysis.

to 1. For the sentiment task, we compute the difference in percentage of scores of 1 or 3.

The permutation tests (Table 2) reveal a significant difference ($p < 0.05$) in sentiment analysis annotations. While the plots appear to reveal consistent differences in affective text annotations, the permutation tests show that this result may be attributed to only one or two annotators with extreme behavior, and looks meaningful due to the small number of annotators overall. The word similarity and NLI tasks also did not produce significant results; however, the p-value for word similarity was very close to our significance level, indicating that studying gender differences in annotation of other similar datasets with different word pairs may be worthwhile in future work.

Task	p-value
Word Similarity (Similarity)	0.0528
Word Similarity (Relatedness)	0.7910
Sentiment Analysis	**0.0209**
Natural Language Inference	0.7592
Affective Text (Anger)	0.5500
Affective Text (Disgust)	0.3143
Affective Text (Fear)	0.3143
Affective Text (Joy)	0.2750
Affective Text (Sadness)	0.3143
Affective Text (Surprise)	0.6111
Affective Text (Valence)	0.2750

Table 2: Results of permutation tests. Results significant at the level $\alpha = 0.05$ are demarcated in **bold**. The false discovery rate correction is performed for results across the table.

5.2. Agreement Analysis

The plots representing distributions of agreements between different genders and aggregations are presented in Figure 4. Among the four ordinal tasks, we find that male and female annotators tend to have similar levels of agreement with the aggregate scores of all other annotators, as was observed by (Prabhakaran et al., 2021) on the sentiment dataset. Furthermore, we find that for both genders, the agreement with the overall aggregate (*F-ALL, M-ALL*) tends to exceed the agreement for the in-demographic aggregate (*F-ALLM, M-ALLF*). This

	F-ALL vs. M-ALL		F-ALLF vs. F-ALLM		M-ALLM vs. M-ALLF	
	tval	pval	tval	pval	tval	pval
Word Similarity (Similarity)	3.08	0.07	-0.56	0.81	0.96	0.77
Word Similarity (Relatedness)	-0.44	0.81	-0.32	0.81	-0.40	0.81
Sentiment Analysis	-0.11	0.94	0.78	0.77	0.09	0.77
Natural Language Inference	1.10	0.77	-1.83	0.97	1.15	0.77
Affective Text (Anger)	-0.29	0.83	0.11	0.77	-0.52	0.81
Affective Text (Disgust)	-1.11	0.77	0.19	0.77	1.17	0.77
Affective Text (Fear)	-0.81	0.77	0.06	0.77	1.21	0.77
Affective Text (Joy)	0.91	0.77	2.92	0.36	0.30	0.77
Affective Text (Sadness)	-0.54	0.81	0.02	0.77	0.41	0.77
Affective Text (Surprise)	-0.62	0.81	-0.72	0.82	0.34	0.77
Affective Text (Valence)	-0.59	0.81	1.06	0.77	0.08	0.77

Table 3: Results of t-tests. No results are significant at the level $\alpha = 0.05$ after the false discovery rate correction was performed for results across the table.

suggests that the statistical effects of having more annotations in the aggregate has a larger effect on agreement than the demographics of the annotators who are included in that aggregate.

The results are mixed for the affective text tasks. This could be in part due to the small number of annotators, but there are a few notable results. We see that for one emotion (joy), female agreement with other females has very little variance, and is much higher than female agreement with other males. However, after controlling for multiple comparisons, this result is not statistically significant. Furthermore, the results differ across the six emotions and valence; we see that for some, there appears to be more agreement between people of the same gender, or between females with the overall aggregate than males. It would be interesting to do these comparisons on a larger dataset with more annotators to determine whether or not there is a difference in how people of different genders annotate each of these emotions, a hypothesis that was supported by some individual examples in the dataset (see Figure 1).

T-tests detailed in Table 3 reveal that there were no statistically significant differences in the pairs of distributions that we compared (see Section 4.2).

6. Discussion

Our results indicate that there is no strong evidence that there are statistical differences between how male and female annotators annotate the four tasks that we studied. The only statistically significant difference we found was for the sentiment analysis dataset; male annotators gave more "intermediary" scores of 1 (somewhat negative) and 3 (somewhat positive) than females when annotating this task. We had initially hypothesized that demographic characteristics of annotators (including gender) may affect annotations and therefore the models trained on various NLP datasets. We were particularly surprised to not find differences in the word similarity dataset, which intentionally included word pairs that represented differing word associations of demographic groups. These differences in word associations were revealed by Garimella et al. (2017), and differences in word associations based on age have also been observed by psychologists (Tresselt and Mayzner, 1964).

This result conflicts with some previous studies, which found a difference in annotations (Al Kuwatly et al., 2020; Larimore et al., 2021; Shen and Rose, 2021; Excell and Al Moubayed, 2021) based on annotator demographics and identities. While our results differ from prior work, it is worth nothing that much of this work focuses on annotation tasks that are more directly related to the identities that were proven to correspond with differing annotations. These works frequently focus on racism, hate speech, and toxicity, which are often targeted at people with certain identities. Hate speech in particular is commonly defined as offensive or degrading language towards a person based on a specific group identify, such as race, ethnicity, gender, or sexual orientation (Parekh, 2006), increasing the likelihood that it will perceived differently by people depending on whether or not they are part of the targeted group(s). The same is true for the labeling of text as corresponding to political ideologies, where the ideologies of the annotators differ (Shen and Rose, 2021).

It is worthwhile to continue studying this problem, as this paper only shows that there are not differences in annotation that can be attributed to *one demographic attribute* (gender) across *four datasets*. We have not proven that there is no difference across the space of all NLP datasets, and we have not proven that there is no difference for other demographic attributes like race or nationality.

A major contribution of our work, therefore, is robust methodology that can expose statistical differences in annotation across groups. By performing permutation tests, we are able to compare the differences we see between male and female annotators to differences that might appear by chance in our annotation pool. Unlike prior work (Prabhakaran et al., 2021), we take this

a step further, formalizing metrics for comparing if annotators agree with the set of annotators who share their gender to a larger extent than they agree with annotators who have a different gender. While these methods are used with interval and ordinal data in our work, they could easily be adapted to use with binary or categorical data.

These methods provide multiple ways in which researchers could study whether annotator demographics result in differences in annotation, and we hope that they will be adopted in future work. In order to ease adoption of our methods, our code is publicly available.[4] To aid this important work, we would recommend that dataset curators consider collecting annotator characteristics and releasing dis-aggregated datasets to the extent possible while preserving the privacy of annotators.

7. Limitations and Future Work

The scope of our study is limited to investigating the effects of annotator gender on NLP benchmark datasets. In collecting data for this project, we learned that nearly all widely used NLP benchmarks have not recorded annotator characteristics their construction process. With the scarcity of annotator demographics associated with NLP benchmarks, several challenges arise. First in such data scarcity, studying annotation differences among non-binary crowdworkers is a challenging but important area of future work. Second, our results do not reveal statistically meaningful discrepancies in data annotation among different genders, but we remain cautious of over-generalization as studying gender effects among a handful of annotators and datasets poses challenges to drawing broader conclusions. Third, while it is helpful to include annotator characteristics in constructing new NLP benchmarks, crowdworker privacy should also be considered. We identify privacy preserving approaches for collection and distribution of annotator demographic data as an important area for future work. Additionally, inclusive practices should be followed when asking crowdworkers to identify their gender (Spiel et al., 2019; Larson, 2017).

The evaluation framework used in this study only considers the discrepancies correlated with a single annotator characteristic. We consider generalized additive models (GAMs) with pairwise interactions (Lou et al., 2013) as a potential avenue for modeling intersectionality of annotator demographics (e.g. gender, race, socioeconomic background) in future work. While language generation tasks are an exciting area in NLP, grounded observations about the discrepancies caused by crowdworker gender are difficult to make, as our methodology is mainly applicable to interval, ordinal and categorical benchmarks.

[4] https://github.com/MichiganNLP/Analyzing-the-Effects-of-Annotator-Gender-Across-NLP-Tasks

8. Conclusion

In this paper we studied the effects of annotator gender on four NLP benchmarks and developed a robust evaluation framework for studying annotator demographic effects on datasets. Our results reveal that there are not statistical differences in how male and female annotators annotated the four benchmark datasets we studied. However, we focused on a small number of datasets and one demographic attribute (gender). We chose the datasets included in our study in large part because they were the ones that were available; most existing NLP benchmarks have been collected without annotator demographics.

We strongly advocate that the community should consider collecting demographics of annotators as part of the data annotation process. This data can be used to perform analyses such as those presented in this paper and to ensure that there is no large demographic imbalance in the annotator pool, relative to the population, as such an imbalance could lead to ineffective models if the annotations differ based on demographics.

9. Acknowledgements

We would like to thank Simin Fan and Andrew Lee for their assistance in early stages of this work.

10. Bibliographical References

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online, November. Association for Computational Linguistics.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A., et al. (2021). We need to consider disagreement in evaluation. In *1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., et al. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the

majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D., (2018). *Addressing Age-Related Bias in Sentiment Analysis*, page 1–14. Association for Computing Machinery, New York, NY, USA.

Excell, E. and Al Moubayed, N. (2021). Towards equal gender representation in the annotations of toxic language detection. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 55–65, Online, August. Association for Computational Linguistics.

Garimella, A., Banea, C., and Mihalcea, R. (2017). Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark, September. Association for Computational Linguistics.

Garimella, A., Banea, C., Hovy, D., and Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498.

Geva, M., Goldberg, Y., and Berant, J. (2019). Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, November. Association for Computational Linguistics.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Gold, M. W. T. H. D. and Zesch, T. (2018). Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 110–120, Vienna, Austria.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December.

Jiang, N. and de Marneffe, M.-C. (2019). Evaluating BERT for natural language inference: A case study on the CommitmentBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6086–6091, Hong Kong, China, November. Association for Computational Linguistics.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Larimore, S., Kennedy, I., Haskett, B., and Arseniev-Koehler, A. (2021). Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online, June. Association for Computational Linguistics.

Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April. Association for Computational Linguistics.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2), apr.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631.

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Mondorf, B. (2002). Gender differences in english syntax. *Journal of English Linguistics*, 30(2):158–180.

Nelson, P., Urs, N. V., and Kasicheyanula, T. R. (2022). Progress in natural language processing and language understanding. In *Bridging Human Intelligence and Artificial Intelligence*, pages 83–103. Springer.

Otterbacher, J. (2018). Social cues, social biases: stereotypes in annotations on people images. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Parekh, B. (2006). Hate speech. *Public policy research*, 12(4):213–223.

Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky*, pages 76–83.

Prabhakaran, V., Mostafazadeh Davani, A., and Diaz, M. (2021). On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Re-

public, November. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Recasens, M., Martí, M. A., and Orăsan, C. (2012). Annotating near-identity from coreference disagreements. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 165–172.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872.

Shen, Q. and Rose, C. (2021). What sounds "right" to me? experiential factors in the perception of political ideology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1762–1771, Online, April. Association for Computational Linguistics.

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. (1994). Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, 7.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Spiel, K., Haimson, O. L., and Lottridge, D. (2019). How to do better with gender on surveys: a guide for hci researchers. *Interactions*, 26(4):62–65.

Storks, S., Gao, Q., and Chai, J. Y. (2019). Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Tresselt, M. E. and Mayzner, M. S. (1964). The kent-rosanoff word association: Word association norms as a function of age. *Psychonomic Science*, 1(1):65–66.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., and Poesio, M. (2021). Semeval-2021 task 12: Learning with disagreements. Association for Computational Linguistics.

Wich, M., Al Kuwatly, H., and Groh, G. (2020). Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199.

Yang, Y., Yih, W.-t., and Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

11. Language Resource References

De Marneffe, Marie-Catherine and Simons, Mandy and Tonhauser, Judith. (2019). *The commitmentbank: Investigating projection in naturally occurring discourse*.

Diaz, Mark. (2020). *Age Bias Training and Testing Data*. Harvard Dataverse.

Strapparava, Carlo and Mihalcea, Rada. (2007). *SemEval-2007 Task 14: Affective Text*. Association for Computational Linguistics.