Corpus Considerations for Annotator Modeling and Scaling

Anonymous NAACL submission

Abstract

Recent trends in natural language processing research and annotation tasks affirm a paradigm shift from the traditional reliance on a single ground truth to a focus on individual perspectives, particularly in subjective tasks. In scenarios where annotation tasks are meant to encompass diversity, models that solely rely on the majority class labels may inadvertently disregard valuable minority perspectives. This oversight could result in the omission of crucial information and, in a broader context, risk disrupting the balance within larger ecosystems. As the landscape of annotator modeling unfolds with diverse representation techniques, it becomes imperative to investigate their effectiveness with the fine-grained features of the datasets in view. This study systematically explores various annotator modeling techniques and compares their performance across seven corpora. From our findings, we show that the commonly used user token model consistently outperforms more complex models. We introduce a composite embedding approach and show distinct differences in which model performs best as a function of the agreement with a given dataset. Our findings shed light on the relationship between corpus statistics and annotator modeling performance, which informs future work on corpus construction and perspectivist NLP.

1 Introduction

An integral aspect of dataset creation is obtaining multiple annotations from annotators on the same instances of data. More often than not, these are aggregated through a majority vote to arrive at a single ground truth label. However, an annotator's background influences the label they assign. This divergence is often evident in the judgments and perceptions of annotators of subjective tasks. In the social media domain, where people's reactions are influenced by their personal experiences and vested interests, relying solely on a majority label to determine or reach a consensus based proves challenging. Hence, it becomes crucial to improve annotator modeling frameworks for robust user representations that capture the diverse views inherent in our datasets, while preserving individual perspectives.

Interest in data perspectivism has been growing and with it, approaches for annotator modeling. The approaches for annotator modeling are built from corpora with unaggregated labels widely ranging in the number of annotators, data type and volume available per annotator, the type of task, and the magnitude of disagreement. Though a few recent works have mentioned the impact of the number of annotators or the level of agreement on annotator modeling methods, they have not been systematically explored.

In this paper, we perform the first systematic study of the scalability of annotator modeling methods and the relationship between annotator modeling methods and corpus statistics. We implement annotator modeling and personalization techniques from recent work and implement our own novel composite embedding approach. This work sheds light on the effectiveness of annotator modeling methods under various real-world scenarios for subjective tasks. We provide recommendations for which methods to use based on the available data.

We find that when agreement is high, our novel composite embedding performs best, while when agreement is lower, the user token approach common in previous work performs best. We find that the user token approach often outperforms other more complex methods developed in previous work, including the averaged SBERT, authorship attribution embeddings, and multi-tasking. We investigate the scalability of these methods with respect to the amount of overall data, the number of annotations per annotator, and the number of annotators in the dataset across seven different datasets

and approximately 3k subsamples representing artificial datasets with controllable properties. We find that the number of annotations per annotator is the most important factor for annotator modeling, though the number of instances and annotators in the corpus both had weak but significant correlations with performance. We will release our code and the statistics of all trials in our experiments to support future work that examines the relationship between corpus statistics and annotator modeling performance.

2 Related Work

Disagreement in the data labeling task is an integral aspect of the annotation process. Even though achieving a higher Inter-Annotator Agreement (IAA) is generally pursued as a goal in the annotation task, the acknowledgment of disagreement will contribute significantly to the annotation task, especially when it is subjective. For a general discussion and overview of this data perspectivism viewpoint, see Cabitza et al. (2023).

The rising interest in perspectivism also becomes apparent in competitions. The recent SemEval shared task on learning with disagreements (Leonardelli et al., 2023) invited teams to develop annotator modeling methods using four textual datasets from different domains, ranging from social media posts to structured conversations. Besides predicting a single "hard" label, the goal of the task was to predict a soft label, i.e., label distribution. Consequently, most participants included a soft loss in their model, directly optimizing for collective disagreement captured by the soft label. Still, some teams modeled individual annotators' perspectives, aggregating predictions afterwards.

Shahriar and Solorio (2023) trained individual models per annotator, observing strong performance gains in soft label prediction. This radical approach does not scale well due to computational complexity. Kohli and Tiwari (2023) reduced this complexity, by first identifying annotator groups with hierarchical clustering on annotation patterns and then training individual models per cluster. However, this approach improved the soft label prediction only slightly over a majority baseline (hard label).

The most prevalent paradigm to model individual annotators was multi-task learning as introduced by Davani et al. (2022) and for instance applied by Vitsakis et al. (2023) without further modifications. Instead of individual models per annotator, this approach extends a shared common model with individual outputs per annotator. While the multi-task approach reduces computational complexity, it still does not scale well and Sullivan et al. (2023) observed that "some systems struggle to account for large or variable numbers of annotators." Grötzinger et al. (2023) found that multi-tasking with respect to datasets instead of annotators performed slightly better, but the annotator multi-tasking approach "could be a worthwhile endeavor to explore further, as it seems to be a natural way to model the task." Wang and Plank (2023) applied the multi-task model to active learning, where the goal is not to predict annotations, but to actively select annotators and instances to be labeled in order to reduce annotation costs.

Maity et al. (2023) integrated annotator representations (embeddings) via attention-based mixing, but scored outside the top-20 of the SemEval task. In addition to learnable annotator representations, Deng et al. (2023) also trained representations for annotations. Further, they trained compatibility matrices between the input text embedding and annotator/annotation embedding to control the influence of the latter. The authors evaluated their approach on eight subjective tasks, observing performance improvements in particular for tasks with high disagreement and found that single dimension demographic features are not enough to model disagreement across annotators.

On the social norms dataset, a corpus of Reddit data (see section 3 for details), Plepi et al. (2022) explored various personalization methods on annotator information sourced from previous posts or network information, finding that averaging embeddings of previous posts to represent annotators "provided a strong and relatively simple approach." A simple approach to integrate annotators' perspectives into a predictive model is to concatenate annotator information (e.g., an identifier) to the instance input. This approach was used as a comparative baseline by Plepi et al. (2022) and Deng et al. (2023).

In summary, we identified four principled paradigms in related work: concatenation of annotator information to the input, multi-task learning, integration of annotator representations (embedding), and individual models per annotator. We exclude the last paradigm from our analysis, due to its computational complexity.

	#A	#I	N	A/I	K-α
GoEmotions	82	58,012	2,576± 2,292	3.64± 0.94	0.27
Gab Hate Speech	18	27,665	4,807± 3,185	3.13± 0.39	0.25
Social Chemistry	2,500	18,431	46.5± 45.9	6.3± 13.3	0.58
MD-Agreement	819	10,753	65.65± 143.77	5.00± 0.00	0.36
HS-Brexit	6	1,120	1,120.00± 0.00	6.00± 0.00	0.35
ConvAbuse	8	4,050	1,521.00± 206.91	3.00± 0.88	0.65
ArMIS	3	943	943.00± 0.00	3.00± 0.00	0.52

Table 1: Dataset statistics including the number of annotators (A), the number of total instances (I), the number of annotations per annotator (N), annotations per instance (A/I), computed as the percentage of labels on an instance that falls into the majority class for that instance, and the agreement as measured by Krippendorff's alpha.

3 Datasets

We use seven datasets from recent work on annotator modeling. All datasets use binary labels for classification. We include four datasets from the recent SemEval-2023 task on learning with disagreements (Leonardelli et al., 2023), two datasets used by Davani et al. (2022), and the Social Chemistry dataset (Forbes et al., 2020) that was adapted for personalization and annotator modeling by Plepi et al. (2022). Dataset statistics are presented in Table 1.

Gab Hate Speech Corpus The Gab Hate Corpus (GHC) (Brendan Kennedy and Dehghani, 2022) comprises 27,665 social media posts from Gab.com annotated by a minimum of three annotators. The GHC features an extensive coding framework that encompasses hierarchical labels denoting dehumanizing and violent speech, markers indicating targeted groups, and rhetorical framing. As in previous annotator modeling work (Davani et al., 2022), we use labels indicating the presence or absence of hate speech.

GoEmotions The GoEmotions (GE) dataset (Demszky et al., 2020) has fine-grained emotions comprising 58k English Reddit comments and labeled for 27 emotion categories and a neutral label for no emotion. We focused on the six Ekman emotions (Ekman et al., 1999) from the experiments of (Davani et al., 2022) *anger, disgust, fear, joy, sadness, and surprise*. Each post received annotations from three to five of a total of 82 annotators. The agreement varies across the Ekman emotions, with Krippendorff's α highest at 0.35 for fear, followed by 0.29 for sadness, 0.28 for surprise, 0.27 for anger, 0.26 for joy, and 0.21 for disgust.

HS-Brexit The Hate Speech Brexit (HSB) dataset (Akhtar et al., 2021) is a dataset on Abusive language detection consisting of 1,120 tweets related to Brexit and Immigration. These were annotated for hate speech, aggressiveness, and offensiveness by two distinct groups of three annotators consisting of a target group of three Muslim immigrants in the UK and a control group of three neutral individuals. In contrast to the MD-Agreement dataset which enlisted over 800 crowd-sourced annotators, the peculiarity of the HS-Brexit dataset lies in its utilization of only six annotators, distributing annotations across a smaller group with each annotator assessing many instances.

Multi-Domain The MultiDomain Agreement (MD) dataset (Leonardelli et al., 2021) comprises 10,753 tweets from three domains: Black Lives Matter (BLM), 2020 USA Presidential Elections, and COVID-19. Each tweet was annotated for offensiveness by a group of five annotators.

ArMIS The Arabic Misogyny and Sexism (ArMIS) dataset contains over 900 Arabic tweets annotated specifically for the detection of misogyny and sexism (Almanea and Poesio, 2022). This annotation was based on identifying bias in the assessment of sexism, with a focus on the varying perspectives of the annotators concerning liberality. Three distinct individuals, self-identifying as a moderate female, liberal female, and conservative male, engaged in the annotation task. The structure of the annotation task is similar to the HS-Brexit dataset.

ConvAbuse The Conversational Abuse (CVA) dataset contains approximately 4k English dialogues between users and two conversational agents (Cercas Curry et al., 2021). Users' conversations were annotated by at least three experts in gender studies using a hierarchical labeling scheme categorized into abuse presence, abuse severity, and directness.

Social Chemistry The Social Norm (SoC) dataset (Welch et al., 2022b) is sourced from Reddit, an online platform with various communities known as subreddits. This dataset specifically utilizes posts and comments from the `/r/amitheasshole` (AITA) subreddit where users share social experiences and seek community opinions on the appropriateness of their behavior and that of others involved. AITA members express their views on whether the original poster is at fault in a scenario, using labels YTA (you are the asshole) and NTA (not the asshole), often providing additional reasoning. The commentors are treated as annotators and personalization methods can be applied to their comments on other parts of the subreddit or other subreddits to compute annotator representations that are not possible with most datasets that contain no more than annotator IDs (and occasionally demographic information). The dataset comprises 21k posts and 327k verdicts (229k NTA, 98k YTA) from 86k different authors. Due to the large number of annotators, we downsampled to only the top 2,500 annotators with the highest annotation counts, resulting in 18,431 instances.

In Table 1, we report the aggregated statistics for our datasets. These include the number of annotators (A), the number of total instances (I), the number of annotations per annotator (N), annotations per instance (A/I), and the agreement level across annotators quantified using Krippendorff's alpha.

4 Methodology

In our study, we investigated five distinct methods for capturing personalized attributes within a collective context, encompassing various levels of complexity.Our models take the annotator ID and text that was annotated as input and are designed to distinguish between unique perspectives. With the exception of Social Chemistry, the annotator tokens are the sole explicit distinguishing attribute of the annotators within the dataset. For several of our methods, annotator embeddings were obtained using special tokens and concatenated with the input strings for each text instance to link each text instance to its corresponding annotator. For the multi-tasking model, the annotator ID instead determines which layer of a multi-task model is responsible for predicting the label.

Additionally, we experimented with personalization techniques for the Social Chemistry dataset. This dataset contains additional writing for each annotator that can be used to derive annotator embeddings. We implemented the authorship attribution approach (Welch et al., 2022a) and averaged SBERT embeddings (Plepi et al., 2022).

We also introduce a new approach, the *composite embedding*, where all instances of annotations for each annotator were aggregated for each class. We calculated the average embeddings of the positive class and negative class and concatenated them to derive a unified representation for each annotator. Note that the averaged SBERT embeddings are averaging the additional text provided by annotators (which is only available in SoC) rather than the embeddings of the text that was assigned an annotation.

4.1 User Token Annotator Embedding

Consider an annotated dataset defined by $D = (X, A, Y)$ where $X = \{x_1, x_2, ...x_n\}$ is the set of text instances, $A = \{a_1, a_2, ...a_k\}$ is the set of annotators and Y is the annotation matrix, where $y_{ij} \in \{0, 1\}$ represents the annotation given to $x_i \in X$ by $a_j \in A$. Since it is typical for each annotator to assign labels to only part of the dataset instances X, therefore Y may contain many missing values. We represent all annotations available for an instance x_i as $\bar{y_i}$ and all annotations by an annotator a_j are represented as $\bar{y_j}$.

Suppose we have D defined with additional annotator information such that $D = (X, A, Y, T)$ where $T = \{u_1, \ldots, u_k\}$ denotes additional information, e.g. annotator IDs. T can be used to augment the input to our classification model. When learning the user token, as is commonly done in previous work, we append a learnable special token, u_j to the text to be encoded by the model, $R_i = \{w_1, \ldots, w_{|x_i|}\}$, resulting in $R_{ij} = R_i || u_j$.

4.2 Composite Embedding

We define the average of all the positive embeddings of all the positive instances of x_i that is $\{x_i\}_{i=1}^n$ by an annotator a_j where $y_i = 1$ as E_p. E_p is derived as $E_p = \sum_{\{i|y_i=1\}} x_{i_{embed}}$. E_n denotes the average of all the instances x_i where $y_i = 0$ and is similarly defined as $E_n = \sum_{\{i|y_i=0\}} x_{i_{embed}}$. Given E_p and E_n, We calculate the composite embedding for an annotator E_c as $E_c = [E_p || E_n]$ where || denotes concatenation.

This embedding is used to initialize a special token representing the annotator, whereas the user token approach is a random initialization. Intu-

Method	GE	GHC	SoC	MD	HSB	CVA	ArMIS
SBERT	68.6	68.5	53.3	73.0	68.6	85.9	61.7
User Token	**70.2**	**76.5**	58.5	**77.7**	**77.6**	88.5	62.1
Composite Embed (Ours)	68.2	68.2	58.6	73.1	67.6	85.8	61.4
Composite+User Token (Ours)	70.0	76.4	**60.4**	77.5	77.3	**88.6**	**62.5**
Multi-task	68.3	70.5	53.5	75.7	71.7	82.3	56.6

Table 2: Full dataset result F1 scores on the **individual annotator** labels for each annotator representation method and dataset.

Method	Anger	Disgust	Fear	Joy	Sadness	Surprise
SBERT	67.9	64.3	71.1	69.2	69.7	70.2
User Token	**69.6**	**66.9**	**73.1**	**69.9**	**70.6**	71.1
Composite Embed (Ours)	66.3	63.6	71.5	69.0	68.9	69.8
Composite+User Token (Ours)	69.1	66.2	72.3	69.8	70.5	**71.6**
Multi-task	67.8	64.0	70.4	68.9	68.8	69.6

Table 3: GoEmotions Ekman emotion F1 scores on the **individual annotator** labels for each method.

itively, an initialization computed from all training data for a given annotator should provide a better starting point for the model.

4.3 Multi-task

The multi-tasking model is implemented on top of a BERT-base model as in previous work. One linear prediction layer is added for each annotator. The loss is summed over all annotators for a given instance. We tuned the learning rate of the multi-tasking model on the validation sets and used $1e^{-}5$ for subsequent experiments. This model has more parameters dedicated to each annotator, so we hypothesized that it would outperform the other annotator methods. We expect to see a trade-off between the cost in time and model complexity versus the improvement in annotator modeling.

4.4 Personalization Methods

We implement the authorship attribution and averaged SBERT embeddings used in previous work for the Social Chemistry dataset. The averaged SBERT embeddings are computed for a given annotator by taking the set of texts that annotator has written independently of their annotations, encoding them with SBERT, and subsequently averaging the representations.

The authorship attribution method is implemented by training an authorship attribution classifier over the text of all annotators. We first embed the text with SBERT and forward these encodings to a two-layer feed-forward network. The output of the last linear layer provides a distribution over all annotators for predicting the author of the text. For each annotator, we use all of their texts from the training set (each label is accompanied with text) and pass them to the classifier. The predictions of the classifier provide a distribution over all annotators. Intuitively, an annotator that is more often confused with another annotator should be more similar to that annotator.

5 Experimental Setup

First, we implemented our text-only baseline using SBERT (Reimers and Gurevych, 2019), a consistently high-performing BERT-based model trained to encode sentences for a variety of downstream tasks. Our annotator modeling methods are based on the SBERT method as in Plepi et al. (2022), with two exceptions. For the multi-task model, we used the same setup as Davani et al. (2022) who used a BERT model with separate output layers for each annotator (Devlin et al., 2019). Lastly, since all of our datasets are in English except ArMIS, which is in Arabic, we used the Arabic BERT model from Safaya et al. (2020), which was trained on a combination of data from Wikipedia and Common Crawl.

To evaluate, we used the macro F1 score across individual annotator's labels. We trained our models for 10 epochs, employing early stopping based on the validation set performance. The Adam op-

timizer was used, with an initial learning rate set at $2e-5$. For our base SBERT model, we chose DistilRoBERTa (Sanh et al., 2019), which features a 768-dimensional representation and a maximum sequence length of 512 tokens. We split the data into 80% train and 10% for each of validation and test with the same annotators in all splits. Our experiments were conducted on a single NVIDIA A100 40GB GPU. The average running time for both training and inference phases combined was around 15 minutes per model.

6 Results

Individual Annotators In Tables 2 and 3 we report the F1 score for all our models across different datasets. Our analysis reveals a consistent trend across our personalized models: performance generally improves with a smaller number of annotators and a high agreement rate. Our models are performing the worst in the Social Chemistry dataset, which contains the highest number of annotators (triple that of the dataset with the next highest count). In contrast, in the GoEmotions dataset, our models underperform compared to the MD-Agreement model, despite having fewer annotators. This can be attributed to two key differences between both datasets: a significant variance in the number of annotations per annotator and a lower overall agreement rate among annotators. Our models achieve the best performance in the ConvAbuse dataset, characterized by a smaller pool of annotators and a higher rate of agreement. This pattern suggests that both the number of annotators and the level of agreement among them are pivotal factors influencing model performance. This correlation is further observed in the GoEmotions dataset, as detailed in Table 3. For instance, our models perform poorly in the *disgust* category, which coincides with the lowest annotator agreement.

For Social Chemistry, we also calculated the authorship attribution and averaged embeddings scores. The F1 for authorship attribution was 56.7 and for averaged embeddings was 57.1, both underperforming our composite embedding plus user token. However, these results are comparable with the 56 F1 reported by Plepi et al. (2022) for the situation split (the same splitting method we use), even though we are using 10k fewer annotators.

Majority Label We presented results using the F1 scores of the individual annotator labels as our primary concern is with annotator modeling. However, it is interesting to note that the methods that perform best on annotator modeling are not the same as those that perform best when aggregating the results of individual predictions to the majority label. To obtain these results we take the individual annotator labels predicted by the model for a given instance and take the most frequent label as the predicted majority label. This is compared to the gold majority label and the resulting F1 scores are shown for the full datasets in Table 4 and for each Ekman emotion of GoEmotions in Table 5.

For the majority labels, we find that the composite embedding combined with the user token improves performance on GHC as well as Disgust. The performance on ArMIS still outperforms other annotator modeling methods, but the SBERT baseline appears to perform better for ArMIS, as well as HS-Brexit and Fear. For two emotions, Anger and Joy, we see the multi-task model outperform other methods, while it is never the best at predicting individual labels. For Social Chemistry, the authorship attribution method achieved an F1 of 57 and the averaged SBERT embeddings scored 56.

For the majority baseline, it is interesting to note that the text-only baseline is sometimes the best model. However, we know from Tables 2 and 3 that the baseline is more often getting the individual annotator model labels incorrect. This finding supports similar findings indicating that the best model of the majority class often marginalizes the voice of minority annotators (Sap et al., 2019; Fleisig et al., 2023), which can be particularly harmful, for instance, in cases such as when racial bias impacts the perception of hate speech (Sap et al., 2022).

7 Scaling Up

Subsequently, to further test the impact of dataset statistics reported in Table 1, we created subsets of our datasets by scaling the number of annotations per annotator, and the number of annotators.

When scaling the number of annotators, the three smallest datasets were scaled in increments of one annotator. For the other four, we scaled in increments of two annotators for the range of 6 to 18 annotators; the upper limit for GHC. Then for GoEmotions and Social Chemistry we scaled from 18 to 82 annotators in increments of 4. Lastly, we scaled Social Chemistry from 100 to 2.5k in increments of 100. This is repeated for each method and Ekman emotion across five runs. We tested each method on each subset, which yields a total of 1,670 trials

Method	GE	GHC	SoC	MD	HSB	CVA	ArRMIS
SBERT	67.0	67.5	51.6	79.8	**72.7**	87.6	**65.3**
User Token	**67.4**	69.7	58.7	**80.6**	72.2	88.8	62.1
Composite Embed (Ours)	65.5	69.2	56.7	80.1	71.9	87.5	63.0
Composite+User Token (Ours)	66.6	**69.8**	**59.9**	80.5	70.4	**89.3**	63.6
Multi-task	66.7	66.7	49.5	76.4	70.2	83.1	55.3

Table 4: Full dataset result F1 scores for each annotator representation method and dataset using the F1 score of predicting the **majority class** label rather than the individual annotator labels.

Method	Anger	Disgust	Fear	Joy	Sadness	Surprise
SBERT	67.5	62.6	**73.2**	63.3	67.5	67.6
User Token	67.5	63.3	72.5	64.8	**68.0**	**68.4**
Composite Embed (Ours)	69.4	61.0	64.9	63.4	66.9	67.2
Composite+User Token (Ours)	66.8	**63.5**	72.2	64.0	66.2	66.7
Multi-task	**70.3**	62.0	70.9	**67.8**	62.6	66.7

Table 5: GoEmotions Ekman emotion F1 scores for each annotator representation method using the F1 score of predicting the **majority class** label rather than the individual annotator labels.

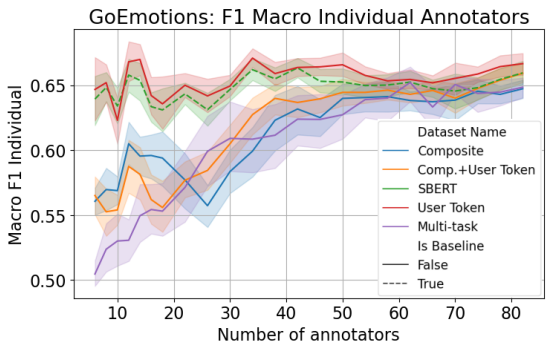

Figure 1: GoEmotions mean performance across emotions when scaling the number of annotators. The SBERT baseline is indicated by the dashed line. Shaded regions correspond to 95% confidence intervals.

using artificially constructed datasets.

The trend across datasets was similar. An example is shown for GoEmotions performance in Figure 1. It shows each method for the dataset when scaling the number of annotators from 6 to 82. The baseline SBERT method is marked by a dashed line. User token performed best overall and even performs strongly when the number of annotators and amount of data is low. The composite embedding and multi-tasking methods perform poorly in this setting but approach the performance of other methods as the amount of data increases.

To identify trends across these three variables we looked at the correlation with performance across all trials. We split the data into those with greater than 18 annotators (the median value across our corpora) and those with 18 or fewer. For each model, we calculate the percentage of relative improvement in the F1 measure over the baseline and measure the correlation using Pearson's coefficient. Since we scaled each dataset based on their available annotations and annotators to examine corpora specific patterns, this resulted in an unequal distribution of trials across datasets. For this analysis, to avoid imbalance of dataset influence, we sampled 60 trials from each of the seven corpora.

When there are more than 18 annotators, there are no significant correlations. However, when we have 18 or fewer, we find that there is a significant correlation with the dataset size ($R = 0.18$, $p < 0.0005$), number of annotators ($R = 0.16$, $p < 0.001$), and most significantly with the number of annotations ($R = 0.42$, $p < 0.0001$).

Subsequently, we decided to look at the impact of the number of annotations per annotator on performance. We ran experiments on all datasets when fixing the number of annotators to the maximum number for ArMIS, ConvAbuse, and HS-Brexit. For the others, we fixed the number of annotators to 14 and lastly, for GoEmotions, MD Agreement, and Social Chemistry we also ran trials with 50 annotators. We varied the number of annotations per annotator in increments of 10% up to the mini-

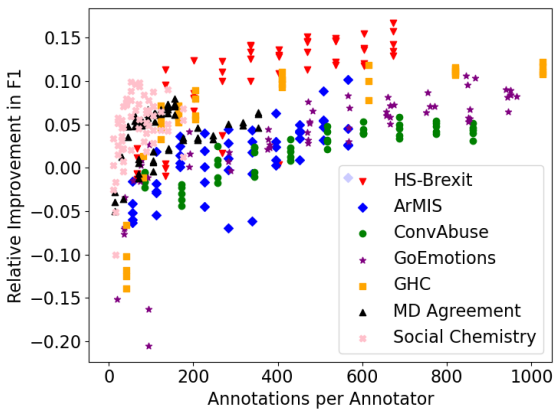

Figure 2: Relative performance increase in F1 as a function of the number of annotations per annotator.

mum number of annotations per annotator in each set of annotators for a given dataset. This resulted in an additional 1,260 trials, which we sampled as in the previous analysis. When examining the relationship between performance and the number of annotations per annotator, our correlation coefficient was $R = 0.47$ ($p < 0.0001$). The samples shown in Figure 2 are from the best performing method on each dataset according to Table 2We see that the number of annotations per annotator is important, but on most datasets the improvement levels off after a couple hundred annotations.

8 Discussion

When examining the dataset statistics and performance, we notice that the user token method performs well when the Krippendorff alpha is relatively low (K-$\alpha < 0.4$). For the Social Chemistry, ConvAbuse, and ArMIS datasets, the agreement is higher (K-$\alpha > 0.5$) and we find that in these cases the composite embedding boosts performance, becoming the best model. This may be because when agreement is higher, the composite embedding is more informative as an initialization of the annotator representation. When annotators agree more with each other, the aggregate of their positive and negative labels is more generalizable, as opposed to the random initialization of the user token.

The largest boost provided by the composite embedding occurred in the Social Chemistry dataset, where we have the lowest number of annotations per annotator. This led us to check the correlation between the performance of the user token and composite embedding methods with the number of annotations per annotator when examining the scaling experiments. We found a slightly stronger correlation for the user token ($R = 0.40$) than for the composite embedding ($R = 0.35$) suggesting that the low number of annotations per annotator is more detrimental to the user token method, but further work is needed to understand this relationship.

Interestingly, while the multi-task model performed the best in previous work on GoEmotions and GHC (Davani et al., 2022), we found that it performed worse in our experiments and was more expensive to train. This was the case even when using the full datasets, while our reimplementation of the approach outperforms results reported in their paper. It tends to perform worse than the baseline for the datasets that had higher agreement. Our hypothesis when we began our study was that multi-tasking would outperform other methods, as it allocates a larger number of dedicated parameters to learning representations of each annotator. We expected to see a trade-off in the model complexity and annotator modeling performance, with multi-tasking being the highest on both.

9 Conclusions

As research on subjective tasks in natural language processing has grown, the importance of modeling annotators and minority opinions has become more apparent. With the recent growth in work on annotator modeling and data perspectivism, it is becoming important to understand how the properties of our datasets and methods impact the effectiveness of annotator modeling methods. We examined seven corpora to better understand recent methods and introduced our method, the composite embeddings. We found that when annotator agreement on a dataset is low (K-$\alpha < 0.4$), the user token embedding was most effective. When the annotator agreement was higher (K-$\alpha > 0.5$), our new composite embedding gave the best performance. Surprisingly, the user token and composite embeddings, which are simple and efficient to implement, outperformed the multi-tasking model that was to be the best model in prior work. Importantly, we also note that the number of annotations per annotator is correlated much more strongly with performance improvements than other corpus statistics, suggesting that this should be an area of focus for those constructing new datasets and collecting annotations. We plan to release our code and collection of over 3k trial experiments statistics to support further work in this area.

Limitations

We examined seven different corpora for our experiments and covered the tasks of emotion recognition, hate and offensive speech detection, and judgement of social norms and morality. However, there are many more types of subjective tasks we have not considered in our work, including those that were previously thought to be more objective in nature (Pavlick and Kwiatkowski, 2019). We do not know how our results will generalize to unseen tasks that are significantly different than we we have examined in this paper. It would be interesting the extend the analysis by the type of task to measure the task influence on annotator modeling performance.

The scaling experiments we performed were initially designed to target the number of annotators and number of annotations per annotator and for experiments with one, we held fixed the value of the other to serve as a control. While we ran a large number of trials, many of the annotation scaling experiments use only a handful of different quantities of annotators. Future work should diversify the sampling for such trials to confirm these results, but our work serves as a first exploratory study that exemplifies the relationship with regards to the correlation and thresholds for sufficient numbers of annotations per annotator.

Our datasets did not all include demographic information. Previous work by Deng et al. (2023) included a detailed analysis and breakdown of annotator identity groups and their relation to annotator modeling performance. Since we do not know which annotators are represented in four of our seven datasets, we cannot say that our results are robust across demographic groups. Future work in this area should include more corpora with annotator information, and future data collection should strive to contextualize collected annotations with annotator meta-data.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *ArXiv preprint*, abs/2106.15896.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France.

Aida Mostafazadeh Davani Leigh Yeh Ali Omrani Yehsong Kim Kris Koombs Shreya Havaldar Gwenyth Portillo-Wightman Elaine Gonzalez Joe Hoover Aida Azatian Alyzeh Hussain Austin Lara Gabriel Olmos Adam Omary Christina Park Clarisa Wang Xin Wang Yong Zhang Brendan Kennedy, Mohammad Atari and Morteza Dehghani. 2022. The gab hate corpus. https://doi.org/10.17605/OSF.IO/EDUA3. Open Science Framework.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online.

Naihao Deng, Siyang Liu, Xinliang Frederick Zhang, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *ArXiv preprint*, abs/2305.14663.

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

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online.

Dennis Grötzinger, Simon Heuschkel, and Matthias Drews. 2023. CICL_DMS at SemEval-2023 task 11: Learning with disagreements (le-wi-di). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1030–1036, Toronto, Canada.

Guneet Singh Kohli and Vinayak Tiwari. 2023. Arguably at SemEval-2023 task 11: Learning the disagreements using unsupervised behavioral clustering and language models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2137–2142, Toronto, Canada.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic.

Ankita Maity, Pavan Kandru, Bhavyajeet Singh, Kancharla Aditya Hari, and Vasudeva Varma. 2023. IREL at SemEval-2023 task 11: User conditioned modelling for toxicity detection in subjective tasks. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2133–2136, Toronto, Canada.

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

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint*, abs/1910.01108.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States.

Sadat Shahriar and Thamar Solorio. 2023. SafeWebUH at SemEval-2023 task 11: Learning annotator disagreement in derogatory text: Comparison of direct training vs aggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 94–100, Toronto, Canada.

Michael Sullivan, Mohammed Yasin, and Cassandra L. Jacobs. 2023. University at buffalo at SemEval-2023 task 11: MASDA–modelling annotator sensibilities through DisAggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985, Toronto, Canada.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. iLab at SemEval-2023 task 11 le-wi-di: Modelling disagreement or modelling perspectives? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669, Toronto, Canada.

Xinpeng Wang and Barbara Plank. 2023. Actor: Active learning with annotator-specific classification heads to embrace human label variation. *ArXiv preprint*, abs/2310.14979.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022a. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland.

Charles Welch, Joan Plepi, Béla Neuendorf, and Lucie Flek. 2022b. Understanding interpersonal conflict types and their impact on perception classification. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 79–88, Abu Dhabi, UAE.