Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

AnnoBERT: Effectively Representing Multiple Annotators' Label Choices to Improve Hate Speech Detection

Wenjie Yin*[1], Vibhor Agarwal*[2], Aiqi Jiang*[1], Arkaitz Zubiaga[1], Nishanth Sastry[2]
[1]Queen Mary University of London, United Kingdom
[2]University of Surrey, United Kingdom
{w.yin, a.jiang, a.zubiaga}@qmul.ac.uk, {v.agarwal, n.sastry}@surrey.ac.uk

Abstract

Supervised machine learning approaches often rely on a "ground truth" label. However, obtaining one label through majority voting ignores the important subjectivity information in tasks such hate speech detection. Existing neural network models principally regard labels as categorical variables, while ignoring the semantic information in diverse label texts. In this paper, we propose AnnoBERT, a first-of-its-kind architecture integrating annotator characteristics and label text with a transformer-based model to detect hate speech, with unique representations based on each annotator's characteristics via Collaborative Topic Regression (CTR) and integrate label text to enrich textual representations. During training, the model associates annotators with their label choices given a piece of text; during evaluation, when label information is not available, the model predicts the aggregated label given by the participating annotators by utilising the learnt association. The proposed approach displayed an advantage in detecting hate speech, especially in the minority class and edge cases with annotator disagreement. Improvement in the overall performance is the largest when the dataset is more label-imbalanced, suggesting its practical value in identifying real-world hate speech, as the volume of hate speech in-the-wild is extremely small on social media, when compared with normal (non-hate) speech. Through ablation studies, we show the relative contributions of annotator embeddings and label text to the model performance, and tested a range of alternative annotator embeddings and label text combinations.

Introduction

Hate speech is defined as speech that directly attacks or promotes violence towards an individual or a group based on actual or perceived aspects of personal characteristics, such as ethnicity, gender and religion (Waseem and Hovy 2016; Davidson et al. 2017; Fortuna and Nunes 2018; Yin and Zubiaga 2021). Hate speech has become a pressing problem for all major social media platforms and its prevalence has brought about the need to come up with automated solutions (Fortuna and Nunes 2018; Jurgens, Hemphill, and Chandrasekharan 2019).

Most studies tackle the problem of hate speech detection from an algorithmic perspective and state-of-the-art approaches rely more on supervised learning (Badjatiya et al. 2017; Schmidt and Wiegand 2017; Fortuna and Nunes 2018; Arango, Pérez, and Poblete 2020; Jiang and Zubiaga 2021), where the agreement level can impact the quality of data resources, potentially leading to model performance issues. Given its subjective nature, it is difficult to achieve high inter-rater agreement when labelling hate speech (Kocoń et al. 2021). Online hate speech often varies widely across groups, communities and countries (Kocoń et al. 2021). Effective identification of hate speech strongly depends on the annotator's world view (Akhtar, Basile, and Patti 2021), leading to lower inter-rater agreement than more objective classification tasks (Davani, Díaz, and Prabhakaran 2021). Differences among annotators with different demographic and cultural backgrounds can in turn be useful, providing deeper insights into the content (Waseem 2016; Davani, Díaz, and Prabhakaran 2021). However, these differences are often overlooked while aggregating all annotations into a majority label (Beelen, Kanoulas, and van de Velde 2017).

Taking into account annotator perspectives is an emerging direction in subjective tasks, and consequently in hate speech or offensive language detection. Existing works usually first separate annotators into groups according to demographic features (Binns et al. 2017; Wich, Al Kuwatly, and Groh 2020) or personal views (Akhtar, Basile, and Patti 2020, 2021) and then train classifiers on each group, or use annotator metadata as part of model input (Kocoń et al. 2021).

Despite consistent attempts to incorporate annotators' perspectives, most related studies (Akhtar, Basile, and Patti 2021; Kocoń et al. 2021) overlook another attribute inside annotated datasets - explicit information from labels. They only implicitly learn the underlying content of the labels from training data by treating the labels as categorical variables, ignoring the latent linguistic knowledge of the words used for the labels. Since online texts from social media might not be long enough to convey comprehensive information (Guo et al. 2013), label information could be advantageous to improve the performance of neural text classification models by constructing the text-sequence representations, enriching the textual representations and reducing the ambiguity from input texts (Wang et al. 2018; Pappas

[*]These authors contributed equally.

and Henderson 2019). Wang et al. (2018) firstly embed text and label information into the joint space and effectively enhance the text classification performance. Recent studies also demonstrate the capacity of capturing complex label relationships to improve semantic representations of input text in text classification (Pappas and Henderson 2019; Halder et al. 2020; Luo et al. 2021). More recent works effectively capture the conversational context in online conversations for text classification tasks such as polarity prediction (Agarwal et al. 2022a; Young et al. 2022) and hate speech detection (Agarwal et al. 2022b).

Given the potential impact of annotator perspectives and labelling history, this paper presents a novel architecture, AnnoBERT, a first-of-its-kind model that effectively incorporates label choices from different annotators into the hate speech detection task without the need to train separate classifiers or use annotator metadata. In order to obtain text representations containing more comprehensive text semantics from annotators' behaviours, AnnoBERT models annotators' preferences from their annotation history and associates such preferences with the specific hate speech annotation tasks through label texts. Our AnnoBERT architecture yields performance gains over the base transformer model and other baselines for the hate speech detection task on two commonly-used datasets and demonstrates its detection superiority on class-imbalanced datasets. In addition, our ablation study results indicate that, compared to other alternative embeddings, the CTR annotator embedding shows its benefit, and label semantics with high-contrast text also boosts the performance of AnnoBERT for identifying online hate.

Our main contributions are the following:

1. We introduce the first integrated model – AnnoBERT[1] to associate multiple annotator perspectives to text and label representations for hate speech detection;
2. We utilise an intuitive and explainable model to construct unique behaviour representations for each annotator based on the previous annotations;
3. We demonstrate the advantage of AnnoBERT compared to all baselines and its benefit for the minority class in hate speech data;
4. We show the relative contribution to the overall performance of modelling annotator preferences and label texts through ablative experiments.

Background and Related Work

Multiple Annotator Perspectives

To perform the annotation aggregation discussed above, several well-known methods, such as probabilistic models of annotation (Plank et al. 2014; Paun et al. 2018) and neural-based aggregation methods (Yin et al. 2017; Rodrigues and Pereira 2018; Li, Takamura, and Ananiadou 2020), have been applied to address this problem.

Due to imprecise or vague annotation schemes, however, disagreements among annotators have universally appeared (Paun, Artstein, and Poesio 2022), especially in highly subjective annotation projects, such as sentiment analysis (Kenyon-Dean et al. 2018), abusive language detection (Akhtar, Basile, and Patti 2020) and hate speech detection (Agarwal et al. 2022b). Then, instead of only aggregating annotations based on a majority vote (hard labels), some studies start to directly learn from a distribution of annotation labels without aggregation (soft labels) (Peterson et al. 2019; Rodrigues and Pereira 2018; Chou and Lee 2019; Uma et al. 2020; Fornaciari et al. 2021; Chu, Ma, and Wang 2021), which can be modelled jointly with hard labels (Chou and Lee 2019), incorporated into the loss function (Plank, Hovy, and Søgaard 2014; Plank, Søgaard, and Goldberg 2016; Uma et al. 2020), learnt to generate representations of annotators (Rodrigues and Pereira 2018), integrated as an auxiliary task (Fornaciari et al. 2021), or learnt with the context (Uma et al. 2020). Chu, Ma, and Wang (2021) propose a similar end-to-end approach, with two types of annotation noise layers, modelling annotator confusion in terms of expertise and instance difficulty.

In the increasing body of research exploring annotator agreement in hate speech, some have focused on exploring how to aggregate annotations based on demographic characteristics (Binns et al. 2017; Al Kuwatly, Wich, and Groh 2020; Wich, Al Kuwatly, and Groh 2020) or annotation history (Akhtar, Basile, and Patti 2020, 2021), resulting in classifiers tailored to each group. Other studies have tended to directly adopt the distribution of annotations into a multi-task (Davani, Díaz, and Prabhakaran 2021) or a multi-modal (Kocoń et al. 2021) architecture.

Demographic data is usually unavailable and could be incomplete or imprecise even when available, as a result of a crowdsourcing task where identity isn't verified. Dividing annotators into groups also has its shortcoming — the number of groups can be arbitrary; training a classifier for each group is inefficient. Hence, we produce annotator representations based on their annotation history, which, without grouping, are directly exploited as part of the input for the hate speech detection model.

Label Information

A growing number of works related to the text classification task has recently realised the value of label information in improving textual representations (Wang et al. 2018; Pappas and Henderson 2019; Halder et al. 2020; Luo et al. 2021). Work on modelling label information can be divided into two types.

On one hand, labels can be modelled with label embeddings, as another input independent of text embeddings. An attentive architecture with joint label embeddings is used to fuse texts and labels into a space and learn label-attentive representations by assigning attention scores to the texts with homologous labels (Wang et al. 2018). Pappas and Henderson (2019) also learns joint input-label embeddings by encoding label descriptions and texts separately, and projecting them into a joint input-label multiplicative space for classification.

On the other hand, instead of being modelled separately, label information can be integrated with the input text. Prompt-based methods (Schick and Schütze 2021) encode

[1] The AnnoBERT code is available at https://github.com/socsys/AnnoBERT.

label and task information in the form of questions as part of language model input. Luo et al. (2021) simplified this by using one single label word instead of manually constructing questions as model input, and Halder et al. (2020) also use label words and text pairs as the input to pre-trained language models to learn a task-aware representation of sentences. We adapt Luo et al. (2021)'s approach to our model.

Annotation bias exists for certain sub-types of hate speech (Waseem 2016; Davidson, Bhattacharya, and Weber 2019; Davani, Díaz, and Prabhakaran 2021), thus the task label (e.g. "misogyny", "hateful") carry significant information for the hate speech detection task. However, existing hate speech detection models primarily adopt a binary prediction task, mapping a fixed set of categorical task labels into k-hot vectors. To some extent, they disregard semantic information in the label text, as well as the influence of the behaviours of multiple annotators on the single label of each text. Therefore, we propose an intuitively motivated hate speech detection model, the first model utilising task label information to enrich input text representations and annotator perspective representations in order to boost the detection performance.

Methodology

Here we describe our proposed model – AnnoBERT.

Collaborative Topic Regression for Annotator Representation

The Collaborative Topic Regression (CTR) model (Wang and Blei 2011) is a probabilistic model originally used in recommendation systems. It discovers a set of topics from the corpus and models user preference with regards to the topics with their ratings of items in the corpus, enabling prediction of user ratings of unseen items based on the topics present.

We chose to adapt CTR to hate speech data based on the analogy that hate speech was also generated by a set of topics, and that annotators are expected to have different levels of sensitivity to each topic. In the fitted model, each user is represented by a fixed-length vector, where each element represents the user's preference for a topic, despite not having any explicit information about the topic itself. This user vector is used as part of the AnnoBERT input.

Integrating Label Text

Besides the annotator embeddings, we also leverage the annotators' chosen label text to associate relevant annotators to their opinions on a particular instance. For each entry in the dataset, we create two separate instances: one containing the annotators who labelled it as hate and the other containing the annotators who labelled it as non-hate, unless all annotators agree. During training, we then append the annotated class label text used by the specific dataset at the end for each of these instances, separated by the [SEP] token. This class label is not appended in the development and test set. The intuition is that during training, the model associates annotators with their label choices given a piece of text; so that during evaluation, when label information is not available, the model predicts the aggregated label given by the participating annotators by utilising the association learnt during training. The default model uses "misogynistic" or "hate" for the positive class, domain-specific to the original datasets as described in the "Datasets" section, and the label "not" for the negative class, assuming that non-hateful instances across datasets are not qualitatively different.

AnnoBERT Architecture

We input the annotators' class labels and annotator embeddings into the AnnoBERT model (see Figure 1). The sentence and the corresponding annotators' class label, separated by the [SEP] token, are input into a BERT encoder. The contextual embeddings are extracted from the final layer, and then concatenated with the aggregated annotator embeddings from those who annotated a particular instance with the same class label. These concatenated embeddings are then input into six BERT feature extraction layers, following Zhang et al. (2021) who found six feature extraction layers to be the best. Finally, the resultant embeddings are fed into the classification layer.We used the most widely acknowledged Bidirectional Encoder Representations from Transformers (`BERT-base-uncased`) (Devlin et al. 2019) as the base model in our experiments, but the choice of the base model is completely flexible and can be applied to any other pretrained BERT-like model.

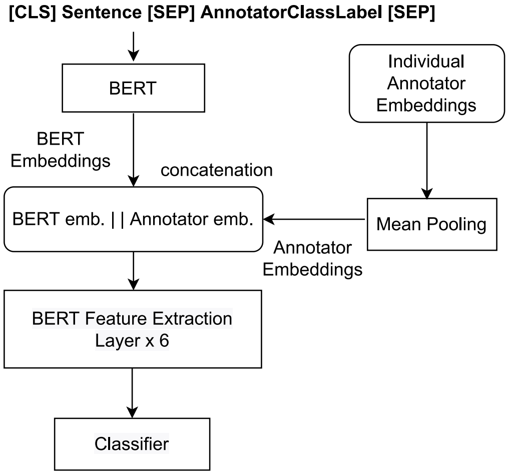

[CLS] Sentence [SEP] AnnotatorClassLabel [SEP]
BERT
Individual Annotator Embeddings
BERT Embeddings
concatenation
BERT emb. | | Annotator emb.
Mean Pooling
Annotator Embeddings
BERT Feature Extraction Layer x 6
Classifier

Figure 1: AnnoBERT Architecture. | | = concatenation.

Experiment Setup

Datasets

Due to the limited availability of annotators' meta-information and labels, in our work, we choose two contrasting datasets (see Table 1), with the number of instances of each class and other features displayed.

- *Guest* (Guest et al. 2021): Expert annotated, sourced from Reddit. Non-hate instances are the overwhelming majority. The domain and the official positive label text

		Guest	Dynabench
Number of Instances	total	6124	41094
	non-hate	5612	**22132**
	hate	**512**	18962
Label		misogynistic, nonmisogynistic	hate, nothate

Table 1: Dataset statistics, with minority class bolded.

		Guest	Dynabench
Number of Annotators		6*	20
Annotators per Instance		3	1
Gender	Female	4	13
	Male	1	7
Education	HS	2	2
	UG	2	4
	PG	1	9
	PhD	0	5
Nationality	British	5	12
	Other	0	8

Table 2: Annotator statistics. Education background is listed, namely high school (HS), undergraduate (UG), postgraduate (PG), and research degree (PhD). *Five of the six annotators in *Guest* give permission to share their basic information.

is "misogynistic". Each instance is annotated by 3 annotators on average. 6 unique annotators in total.

- *Dynabench* (Vidgen et al. 2020): Crowd generated and annotated, deliberate synthetic data designed to fool classifiers. No specific hate type, hence hate label is "hate". Slightly more data is labelled as hate speech than as non-hate. Each instance is only associated with one annotator who constructed the instance originally, but in some cases the class label could have been changed by an expert. This dataset has 20 unique annotators in total.

These two datasets were, at the time when we started the experiments, the only datasets with sufficient information about the decisions of each annotator. Table 2 summarises annotators' details in the two datasets. All annotators regularly use social media at least once per day, and more than half of them reported to have been targeted personally by online abuse before. We use the official training and testing splits from the original dataset release, and further randomly sample 10% from the training split to form the validation set.

Text preprocessing was minimal: All usernames and urls were replaced with special tokens ("<user>", "<url>"). Non-ascii letters were considered non-English and thus removed for being out of scope. Additional whitespaces were removed.

Baselines

We compare AnnoBERT with three baseline models:

- **BERT** (Devlin et al. 2019): Bidirectional Encoder Representations from Transformers (BERT) is a Transformer network (Vaswani et al. 2017) pre-trained with a language modelling objective and a vast amount of raw text. Having been acknowledged as a top performing model on many hate speech datasets, it was chosen as a representative of the BERT-like transformers.
- **CrowdLayer** (Rodrigues and Pereira 2018): It is a neural network layer, which enables models to be trained end-to-end directly from the distribution of annotations. Such a layer is added on top of a BERT classifier to capture biases of multiple annotators and classification results can be obtained from the BERT classifier after the training stage.
- **LEAM** (Wang et al. 2018): Label-Embedding Attentive Model (LEAM) is an attention-based architecture with joint label embeddings, which embeds the text and label information into the same space to learn label-attentive text representations.

Subset	Total	Hate ratio (%)
Guest, disagreement	170	60.00
Guest, agreement	1,061	0.00
Guest, overall	1,231	8.29

Table 3: Subsets investigated with total instance counts and hate class ratio in the testing data.

Imbalanced Sample (Dynabench)

Guest is much smaller and imbalanced than *Dynabench*. This is a major but not the only difference between the two datasets. To isolate the effect of class imbalance on the relative performances, we randomly sampled from *Dynabench* 6124 instances (512 hate, 5612 non-hate) to form a sample with the same class imbalance as *Guest*.

Disagreement Subset (Guest)

One might expect that taking into account annotator characteristics helps clarify edge cases which annotators disagree on. To verify this hypothesis, we provide a comparison of the original AnnoBERT model and the strongest baseline, BERT, on instances where disagreements between annotators are present in *Guest*[2], compared to the overall data. Possibly unsurprisingly, most instances with disagreement were predominantly deemed hateful; all instances without disagreement were consistently labelled as pertaining to the negative class. Thus, we did not include the subset without disagreement as the metrics would be obsolete when one class is absent. Note that the subset distinction only happens at the evaluation stage, using the models trained on the entirety of data. We present results on the entire dataset as well as on the disagreement subset only, for more detailed analysis of edge instances vs all instances.

Their sizes and the ratios of the positive class of subsets are shown in Table 3.

[2] The same experiment was not available for *Dynabench*, in which each instance was only associated with one annotator.

Model Ablation: Annotator Embeddings

We use AnnoBERT variants with two other types of annotator embeddings to test the effectiveness of the CTR-generated annotator embeddings. These model variants have the same architecture as the original AnnoBERT model (Figure 1), with only the individual annotator embeddings replaced.

- **History**. Instead of the CTR annotator embeddings, we insert raw annotation history as annotator vectors. In line with Basile (2021), each annotator vector is made of their annotations across all instances in the training data: -1 for non-hate, 1 for hate, and 0 when the annotator didn't label that instance. The dimensionality of the vector is then linearly reduced to the same as BERT hidden size. Models with the embedding layer weights frozen and unfrozen were experimented.
- **Learnt**. To test whether annotation preference offers sufficiently more useful information than annotator identifiers, we use a model which randomly initialises an embedding layer for each annotator and learns the annotator embeddings end-to-end. The embedding size is the same as BERT hidden size.

Apart from the classification performance comparison of different kinds of annotator embeddings, we also conduct a detailed analysis of CTR-generated and history embeddings. Note that we do not use Learnt embeddings in this analysis since they are learnt end-to-end by the model during training. We compute the correlation between inter-annotator agreement (cohen's kappa) and cosine distance between every possible pair of annotator embeddings. Intuitively, an efficient annotator embedding should be similar for annotators with high inter-annotator agreement or, in other words, correlation should be strongly negative since cosine distance would be smaller between annotators with the higher agreement.

Model Ablation: Label Text

We also test the same models on different label texts. In addition to our experiments evaluating the impact of using label texts over not using them, the objective of testing different label texts is to assess the impact of using more or less relevant label texts. On both datasets, five conditions were tested:

1. **"hate", "not"**. The default setting on *Dynabench*; inaccurate description condition on *Guest*.
2. **"misogynistic", "not"**. The default setting on *Guest*; inaccurate description condition on *Dynabench*.
3. **"hate", "not hate"**. A variation of the first condition, to test whether a neutral label or an antonym label is better as the negative class for the domain-specific conditions.
4. **"misogynistic", "nonmisogynistic"**. A variation of the first condition, to test whether a neutral label or an antonym label is better as the negative class for the domain-specific conditions.
5. **"yes", "no"**. Non-domain-specific label text, to test the importance of domain-specificity in the label text for the model.

In summary, the comparison between 1, 2, 3, 4 and 5 tests whether domain-specific labels are needed for the model; the comparison between 1, 3 and 2, 4 tests the effect of accurate domain specific class description; the comparison between 1, 2 and 3, 4 tests whether negative instances in different datasets can be represented in the same way.

Training Settings

We fitted CTR models using an open-source implementation[3] on the training set of each dataset for the default setting of 100 expectation-maximisation iterations, and used the user latent vectors as annotator embeddings. The non-parallelised and iterative implementation required 2 minutes per iteration on the *Guest* dataset and 11 minutes on the *Dynabench* dataset respectively on a single CPU. We experimented with a range of dimensions (5 to 15, as 10 is the default setting) and aggregation strategies (mean-pooling, max-pooling, summation, and concatenation), in terms of macro F1 on the validation set. We found 10-D latent vectors and mean-pooling to be the best performing option with moderate advantage over other options (1.8 to 9 macro F1); these settings are used in the final AnnoBERT model for all downstream tasks.

We implemented AnnoBERT and the BERT baseline using the Huggingface transformers library (Wolf et al. 2020). Models are trained for 4 epochs (except LEAM for 15 epochs), with a learning rate of $1e^{-5}$, cross-entropy loss, and a batch-size of 32. *Adam* (Kingma and Ba 2014) is used as the optimiser, with *Tanh* activation function. LEAM model uses FastText embeddings of dimension 300 (Grave et al. 2018) as the encoder, and softmax function in the output layer. We report average results for 10 separate runs of each model.

Evaluation Metrics

Given the imbalanced nature of the hate speech classification task, we use the following metrics to evaluate our model and other baselines:

Macro F1: Macro F1 is the arithmetic mean of all the per-class F1 scores.

Sensitivity of the positive class: also known as true positive rate or recall of the positive class, is the probability of a positive test, given a positive truth.

Specificity of the positive class: also known as true negative rate or the recall of the negative class, is the probability of a negative test, given a negative truth.

We use macro F1 as the high-level representation of the overall model performance, but contrast sensitivity and specificity to reflect a model's bias towards either the positive or the negative class.

Results and Discussions

Basic model performances are summarised in Table 4.

The main weakness shared by baselines is a bias towards the majority class, mediated by class imbalance severity.

[3] https://github.com/dongwookim-ml/python-topic-model

Model	Macro F1	Sensitivity	Specificity
Dynabench			
BERT	76.04±0.17	79.43±0.82	72.58±0.66
CrowdLayer	42.77±1.87	68.52±10.14	31.65±10.12
LEAM	39.37±0.94	**94.64±1.42**	5.93±1.49
AnnoBERT	**76.19±0.20**	75.64±1.24	**77.17±1.16**
Guest			
BERT	66.26±1.77	30.20±3.88	**97.53±0.41**
CrowdLayer	48.91±0.49	9.12±2.33	90.02±2.52
LEAM	50.62±0.66	8.82±2.72	94.40±1.91
AnnoBERT	**69.31±0.57**	**37.26±1.90**	96.81±0.42

Table 4: Performance of AnnoBERT model and baselines on two datasets. Macro F1, sensitivity, and specificity scores, mean from 10 runs each, with standard errors on mean. Best performance bolded, second best performance underlined.

All baselines have higher sensitivity than specificity on *Dynabench* and higher specificity than sensitivity on *Guest*. On the much more imbalanced *Guest* (Table 1), a much lower sensitivity is observed in all baselines. This contributed greatly to overall lower macro F1 scores. BERT is the most robust baseline on this front, yet still achieving only 30.20 sensitivity on *Guest*and 72.58 specificity on *Dynabench* respectively. LEAM and CrowdLayer are consistently worse than BERT. These highlight the challenge of class imbalance in hate speech detection, and that transformer models are strong baselines even compared to alternative models with additional annotator and label features.

AnnoBERT improves performance through addressing the challenge of class imbalance.

As shown in Table 4, on both datasets, AnnoBERT achieves the highest recall of the minority class – specificity on *Dynabench* and sensitivity on *Guest*, showing better robustness against class imbalance. Although with a small trade-off on the complementing metric, the advantage with the minority class is reflected on the overall performance as a higher macro F1 on both datasets, with the advantage much larger on the highly imbalanced, realistically sampled *Guest* than the balanced, synthetically constructed *Dynabench*.

On the imbalanced subset of *Dynabench* (Table 5), model relative performances follow the same pattern as those on *Guest*, with AnnoBERT having a considerable advantage on sensitivity and in turn macro F1. This clarifies that the performance difference between *Guest* and *Dynabench* is indeed mainly caused by different class distributions.

These results altogether show that AnnoBERT is less biased against the majority class and benefits hate speech detection the most when the data is highly imbalanced. Specifically, it is better at picking up the characteristics of the minority class through limited instances during training, resulting in a higher recall of that class.

AnnoBERT has the most advantage on edge cases which annotators disagree on.

Model	Macro F1	Sensitivity	Specificity
BERT	54.17±1.08	9.12±1.78	**98.10±0.44**
AnnoBERT	**59.69±0.65**	**25.00±2.22**	93.93±1.00

Table 5: Performance of AnnoBERT and BERT on imbalanced sample of *Dynabench*. Macro F1, sensitivity, and specificity scores, mean from 10 runs each, with standard errors on mean. Better performance bolded.

The comparison between AnnoBERT and BERT on all instances and only instances with disagreement in *Guest* is shown in Table 6.

Model	Macro F1	Sensitivity	Specificity
Guest, disagreement			
BERT	53.33±2.71	30.20±3.88	**93.97±1.43**
AnnoBERT	**58.44±1.18**	**37.26±1.90**	92.79±1.21
Guest, overall			
BERT	66.26±1.77	30.20±3.88	**97.53±0.41**
AnnoBERT	**69.31±0.57**	**37.26±1.90**	96.81±0.42

Table 6: Performance of AnnoBERT and BERT on instances with disagreement and all instances from *Guest*. Macro F1, sensitivity, and specificity scores, mean from 10 runs each, with standard errors on mean. Better performance bolded.

Both models had lower specificity on the subset with disagreement, which is within expectation as these are instances that also confused human annotators.

Model relative performance is similar to disagreement in a sense that AnnoBERT marginally trades off specificity for a large improvement on sensitivity and macro F1. However, the advantage of AnnoBERT over BERT is noticeably larger on the disagreement instances (+5.11 macro F1) than on the whole dataset (+3.15 macro F1). This highlights the remarkable benefit of using AnnoBERT's annotator and label information to improve hate speech detection in edge cases.

Model Ablation

Theoretically grounded CTR embeddings benefit the classification task more than end-to-end trained embeddings and raw annotation history.

Models with different annotator embeddings are compared in Table 7.

Overall, the original AnnoBERT model with CTR embeddings consistently outperforms the ablation models with alternative embeddings, with the end-to-end learnt embeddings better than using annotation history.

Annotator embeddings trained end-to-end are not affected by annotator frequencies, but are inferior to a theoretically driven model in the overall medium classification performance. Only the CTR-based models consistently outperform the BERT baselines, meaning that with a strong baseline, additional annotator information is only beneficial with high quality annotator embeddings.

Annotator embedding	Macro F1	Sensitivity	Specificity
Dynabench			
CTR, uf	76.05±0.24	73.46±1.31	**79.40±1.19**
CTR, f	**76.19±0.20**	**75.64±1.24**	77.17±1.16
History, uf	38.52±2.60	41.91±13.69	58.81±13.63
History, f	39.22±2.76	22.60±9.55	78.24±8.93
Learnt, uf	64.83±1.03	55.60±2.58	76.57±1.87
Guest			
CTR, uf	69.05±0.58	**37.45±2.10**	96.62±0.37
CTR, f	**69.31±0.57**	37.26±1.90	**96.81±0.42**
History, uf	65.41±0.70	35.59±2.38	94.52±1.06
History, f	65.73±0.89	30.49±2.60	96.78±0.58
Learnt, uf	66.36±0.52	31.86±2.57	96.69±0.69

Table 7: Performance of AnnoBERT model variants with different annotator embeddings. Macro F1, sensitivity, and specificity scores, mean from 10 runs each, with standard errors on mean. f/uf: annotator embeddings frozen/unfrozen during task-specific training. Best performance bolded, second best performance underlined.

Frozen annotator embeddings tend to achieve higher mean macro F1 scores, but the difference is marginal considering the standard errors. Thus, the model does not forget much about the annotators during end-to-end training of the model.

CTR embeddings reflect annotator agreement better than embeddings generated with annotator history.
Figures 2 and 3 visualise the annotator embeddings, dimension-reduced through Principal Component Analysis. For both datasets, variance explained by the first two principal components (in brackets) is higher for annotator embeddings using the CTR model than using the annotators' history, as shown in Figures 2a and 3a. Comparing *Dynabench* to *Guest*, there are a lot more unique annotators (20 as opposed to the Guest dataset which only has 6 annotators), and each instance is only associated with one annotator. As a result, the relative locations of the annotators in the embedding space computed through raw history, as shown in Figure 3b, is largely determined by the volume of annotations each annotator made. For example, 3 annotators annotated a large portion (almost 35%) of the instances in Dynabench and hence these annotators are separated far apart in the embedding space. All the remaining 17 annotators, who produced fewer annotations, are close together in the embedding space because they have lots of missing annotations (0s) in the raw history embeddings. In comparison, the CTR embeddings are robust to this factor and do not just reflect the volume of annotations or the missing annotations (0s), but also the annotators' label choices via topic modelling. This likely underlies the classification difference across different annotator embeddings, which is much larger in *Dynabench* than *Guest*.

Table 8 shows the correlation of inter-annotator agreement (cohen's kappa) with cosine distance between all the possible pairs of annotators. CTR embeddings better reflect agreement since their correlation is strongly negative as compared to history embeddings. Correlation is expected to be negative since annotators with higher agreement are more similar and hence, their cosine distance will be smaller.

Embedding type	Pearson's r
History	-0.2956
CTR	-0.4031

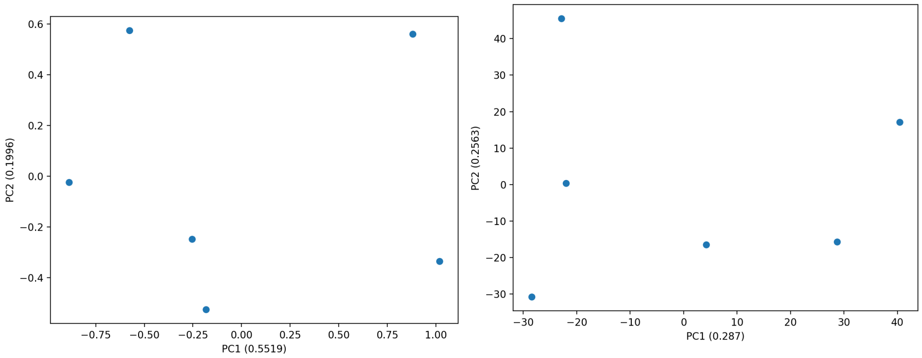
Table 8: Correlation, measured by Pearson's r, of inter-annotator agreement with cosine distance between possible pairs of annotators for *Guest* dataset.

The exact label text contributes less to the model performance than associating annotator embeddings with the label choices.
Performance with alternative label texts is shown in Table 9.

Label text	Macro F1	Sensitivity	Specificity
Dynabench			
hate, not hate	76.05±0.26	74.25±1.69	**78.55±1.75**
hate, not	76.19±0.20	75.64±1.24	77.17±1.16
misogynistic, nonmisogynistic	75.29±0.32	75.38±2.12	75.68±2.05
misogynistic, not	76.03±0.24	74.52±1.42	78.13±1.23
yes, no	**76.36±0.21**	**76.84±1.55**	76.24±1.71
Guest			
hate, not hate	68.97±0.64	41.08±2.03	95.44±0.56
hate, not	68.38±0.41	35.39±2.35	**96.83±0.55**
misogynistic, nonmisogynistic	67.58±0.49	38.23±1.45	89.11±6.00
misogynistic, not	69.31±0.57	37.26±1.90	96.81±0.42
yes, no	**70.17±0.52**	**44.41±2.77**	95.33±0.77

Table 9: Performance of AnnoBERT model and baselines on two datasets. Macro F1, sensitivity, and specificity scores, mean from 10 runs each, with standard errors on mean. Best performance bolded, second best performance underlined.

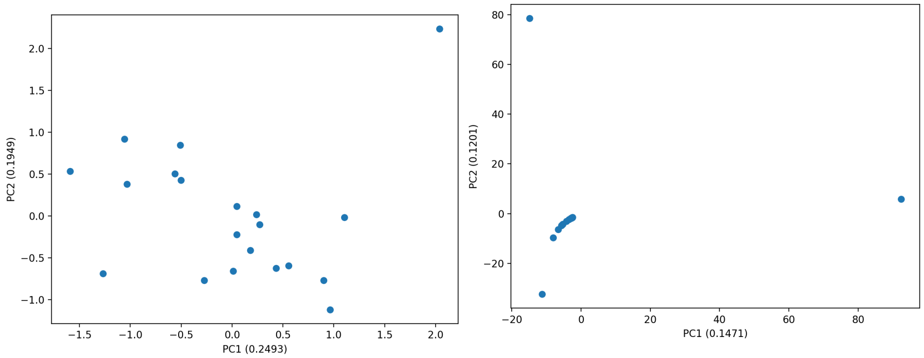
Label text has a very different effect than annotator embeddings. For one thing, classification performance does not vary as much – the gap between the best and worst models is less than 3% in macro F1 across label texts, in contrast to over 30% between annotator embeddings. For another, the more varying dataset is *Guest* here as opposed to *Dynabench*. Our interpretation is that a dataset with a narrow domain (*Guest*, which specialises in misogyny) is more sensitive to label texts.



(a) Annotator embeddings using CTR model.

(b) Annotator embeddings using annotators' history.

Figure 2: Visualisation through PCA dimension reduction of annotator embeddings on *Guest* dataset. Variance explained by each principal component is shown within brackets.



(a) Annotator embeddings using CTR model.

(b) Annotator embeddings using annotators' history.

Figure 3: Visualisation through PCA dimension reduction of annotator embeddings on *Dynabench* dataset. Variance explained by each principle component is shown within brackets.

Surprisingly, in both datasets, "yes" and "no" texts appear to be the most effective, albeit the difference being marginal. The reason behind this could be that this pair is the most semantically contrasting, for both language models and audiences from different backgrounds. "Misogynistic" vs "nonmisogynistic" is the worst performing. However, pairing "misogynistic" with a neutral negative class text ("not") noticeably improves the classification performance, leading to a second highest mean macro F1 on *Guest*. This is likely because the BERT embeddings with "nonmisogynistic", a much rarer word than any other label texts, is of lower quality, as rare words are a major problem in contextualised embeddings (Schick and Schütze 2020). Nonetheless, all label text variants outperform the baselines.

Overall, the two model ablation studies show that the advantage of AnnoBERT lies within the ability to pair effectively encoded annotator attributes with contrasting classes. The exact label text does not matter as much as annotator embedding quality, but it is likely to be beneficial to have text combinations that are less domain-specific, semantically contrasting, and have higher word frequency for the base transformer model.

Implications on the role of annotators in hate speech annotation and classification. It was not until recently that the annotators' role was brought into attention in abusive language detection (Basile 2021) and contrasting paradigms for subjective annotation were defined (Röttger et al. 2021)[4]: while a descriptive paradigm aims to reflect individuals' viewpoints by providing less instruction, a prescriptive paradigm aims to reflect one predefined taxonomy by providing extensive instruction. Both *Dynabench* and *Guest* datasets follow a prescriptive paradigm with clear class definitions.

On the one hand, our experiments highlight the significant role that annotators play in the creation of subjective data: even with prescriptively constructed datasets, where annotators are expected to disagree much less, annotator information can considerably benefit classification performance, especially on edge cases where disagreement inevitably happens. This calls for the field to include and analyse annotator behaviour as well as to utilise such information when

[4] These definitions were not available when we carried out the experiments.

approaching classification tasks, particularly tasks of a subjective nature. Admittedly, privacy is an important concern when it comes to publicly releasing annotator information, which is why our study focuses on using only annotators' judgements rather than their personal information in the classification model.

On the other hand, our study also points towards a promising direction for identifying and mitigating annotator biases. One example use case would be that, with annotator representations that better represent the level of agreement between them, annotators who perform differently from the others can be more easily identified through visualisation.

Conclusions and Future Work

Our work is the first to integrate annotator characteristics, text and label representations in one hate speech detection model. Our proposed model, AnnoBERT, improves performance in hate speech detection on top of a transformer-based baseline. AnnoBERT works especially well on imbalanced data, which is characteristic of nearly all extant hate speech datasets. The scarcer the data of the minority class, the larger its advantage on the recall of the minority class and the more visible the improvement in overall classification metric. We clarified this effect of class distribution using an imbalanced sample of an originally balanced dataset. Through further analysis, we show that our proposed approach benefits especially instances on which annotators disagree more. Through ablation studies, we showed the relative contributions of annotator embeddings and label text to the model performance, and tested a range of alternative annotator embeddings and label text combinations.

Although we used BERT as the base model for our proposed AnnoBERT, we expect the benefit of our approach to translate to other transformer-based models; although we used CTR user latent vectors as annotator embeddings, this approach shows potential with other user models. Thus, the most immediate future steps would be using more advanced transformer models and user models, especially user models that directly link text features to user ratings, which will produce inherently explainable hate speech detection models.

Additionally, our approach carries significant practical value, with its considerable advantage on learning from limited minority class samples – in the real world, abusive content amounts to around 0.1% to 3% of social media (Founta et al. 2018). Our results encourage future hate speech detection research to consider annotator preference and label information to address this common challenge.

More importantly, our approach enables customised hate speech detection on an aggregated level, by enabling prediction of different groups of annotators through a single language model and only replacing annotator embeddings. Neither does our approach require the collection of annotator demographics as it solely relies on their annotation histories.

Our ablation studies and further discussions also shed light on perspectives and the use of gold standards in such subjective tasks as hate speech detection. Possible future directions include utilising user models that directly link text features to user ratings, which will produce explainable hate speech detection models, or scaling up our approach to, for example, predicting moderation choices of communities with different moderation strategies.

Ethics Statement

Hate speech online is a sensitive issue, and there are some ethical issues in the controversy on free speech. To further improve the fairness and reliability of our work, we claim the following ethical considerations:

- **Confidentiality:** Access to data is critical to the effectiveness of our work. Since the data is already publicly available, we have replaced all personal data with a special token <user> to ensure user anonymity. The information of all the annotators corresponding to the dataset is hidden, and only the ID number of the annotator and its annotation content is displayed.
- **Potential for harm:** We do not intend to harm vulnerable groups that are already discriminated against based on specific characteristics. Our work serves only the benign purpose of detecting and mitigating abusive expressions.
- **Results communication:** We guarantee that there is no plagiarism or academic misconduct, but we acknowledge that there may be limitations and potential misrepresentations when analysing social media data, particularly for those abusive data that does not clearly represent the target of the attack.

It can be debatable whether all annotator perspectives should be treated equally, especially when they lack knowledge or relevance to the speech at question. If annotators are random samples from the population, it is likely that they diverge from the opinions of the minority groups targeted by online hate speech. In an extreme scenario, for example, treating a white supremacist annotator and an oppressed individual with the same weights can further suppress the voice of the latter group. Future work from a social science perspective is needed.

Moreover, adversaries, who actually spread hate speech online, can use this kind of research for malicious purposes such as to learn how to prevent being detected.

References

Agarwal, V.; Joglekar, S.; Young, A. P.; and Sastry, N. 2022a. GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. In *Proceedings of the ACM Web Conference 2022*, 2729–2737.

Agarwal, V.; Young, A. P.; Joglekar, S.; and Sastry, N. 2022b. A Graph-Based Context-Aware Model to Understand Online Conversations. *arXiv preprint arXiv:2211.09207*.

Akhtar, S.; Basile, V.; and Patti, V. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 151–154.

Akhtar, S.; Basile, V.; and Patti, V. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

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*, 184–190.

Arango, A.; Pérez, J.; and Poblete, B. 2020. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 101584.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 759–760. International World Wide Web Conferences Steering Committee.

Basile, V. 2021. It's the End of the Gold Standard as We Know It: Leveraging Non-aggregated Data for Better Evaluation and Explanation of Subjective Tasks. In Baldoni, M.; and Bandini, S., eds., *AIxIA 2020 – Advances in Artificial Intelligence*, volume 12414, 441–453. Cham: Springer International Publishing. ISBN 978-3-030-77090-7 978-3-030-77091-4. Series Title: Lecture Notes in Computer Science.

Beelen, K.; Kanoulas, E.; and van de Velde, B. 2017. Detecting Controversies in Online News Media. SIGIR '17, 1069–1072. New York, NY, USA: Association for Computing Machinery.

Binns, R.; Veale, M.; Van Kleek, M.; and Shadbolt, N. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, 405–415. Springer.

Chou, H.-C.; and Lee, C.-C. 2019. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5886–5890. IEEE.

Chu, Z.; Ma, J.; and Wang, H. 2021. Learning from Crowds by Modeling Common Confusions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7): 5832–5840.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. Florence, Italy: Association for Computational Linguistics.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Fornaciari, T.; Uma, A.; Paun, S.; Plank, B.; Hovy, D.; and Poesio, M. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2591–2597. Online: Association for Computational Linguistics.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4): 85.

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; and Margetts, H. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1336–1350. Online: Association for Computational Linguistics.

Guo, W.; Li, H.; Ji, H.; and Diab, M. 2013. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 239–249. Sofia, Bulgaria: Association for Computational Linguistics.

Halder, K.; Akbik, A.; Krapac, J.; and Vollgraf, R. 2020. Task-Aware Representation of Sentences for Generic Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3202–3213. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Jiang, A.; and Zubiaga, A. 2021. Cross-lingual Capsule Network for Hate Speech Detection in Social Media. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 217–223.

Jurgens, D.; Hemphill, L.; and Chandrasekharan, E. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3658–3666. Florence, Italy: Association for Computational Linguistics.

Kenyon-Dean, K.; Ahmed, E.; Fujimoto, S.; Georges-Filteau, J.; Glasz, C.; Kaur, B.; Lalande, A.; Bhanderi, S.; Belfer, R.; Kanagasabai, N.; Sarrazingendron, R.; Verma, R.; and Ruths, D. 2018. Sentiment Analysis: It's Complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1886–1895. New Orleans, Louisiana: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocoń, J.; Figas, A.; Gruza, M.; Puchalska, D.; Kajdanowicz, T.; and Kazienko, P. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5): 102643.

Li, M.; Takamura, H.; and Ananiadou, S. 2020. A Neural Model for Aggregating Coreference Annotation in Crowdsourcing. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5760–5773. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Luo, Q.; Liu, L.; Lin, Y.; and Zhang, W. 2021. Don't Miss the Labels: Label-semantic Augmented Meta-Learner for Few-Shot Text Classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2773–2782. Online: Association for Computational Linguistics.

Pappas, N.; and Henderson, J. 2019. GILE: A Generalized Input-Label Embedding for Text Classification. *Transactions of the Association for Computational Linguistics*, 7: 139–155.

Paun, S.; Artstein, R.; and Poesio, M. 2022. *Statistical Methods for Annotation Analysis*. Morgan & Claypool Publishers.

Paun, S.; Chamberlain, J.; Kruschwitz, U.; Yu, J.; and Poesio, M. 2018. A Probabilistic Annotation Model for Crowdsourcing Coreference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1926–1937. Brussels, Belgium: Association for Computational Linguistics.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

Plank, B.; Hovy, D.; McDonald, R.; and Søgaard, A. 2014. Adapting taggers to Twitter with not-so-distant supervision. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1783–1792. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Plank, B.; Hovy, D.; and Søgaard, A. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 742–751. Gothenburg, Sweden: Association for Computational Linguistics.

Plank, B.; Søgaard, A.; and Goldberg, Y. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 412–418. Berlin, Germany: Association for Computational Linguistics.

Rodrigues, F.; and Pereira, F. 2018. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Röttger, P.; Vidgen, B.; Hovy, D.; and Pierrehumbert, J. B. 2021. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. *CoRR*, abs/2112.07475.

Schick, T.; and Schütze, H. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8766–8774.

Schick, T.; and Schütze, H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269.

Schmidt, A.; and Wiegand, M. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. Valencia, Spain: Association for Computational Linguistics.

Uma, A.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2020. A Case for Soft Loss Functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1): 173–177.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, u.; and Polosukhin, I. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Vidgen, B.; Thrush, T.; Waseem, Z.; and Kiela, D. 2020. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. *arXiv:2012.15761 [cs]*. ArXiv: 2012.15761.

Wang, C.; and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, 448. San Diego, California, USA: ACM Press. ISBN 978-1-4503-0813-7.

Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; and Carin, L. 2018. Joint Embedding of Words and Labels for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2321–2331. Melbourne, Australia: Association for Computational Linguistics.

Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

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*, 191–199.

Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S.;

et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Yin, L.; Han, J.; Zhang, W.; and Yu, Y. 2017. Aggregating Crowd Wisdoms with Label-aware Autoencoders. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1325–1331.

Yin, W.; and Zubiaga, A. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7: e598.

Young, A. P.; Joglekar, S.; Agarwal, V.; and Sastry, N. 2022. Modelling online debates with argumentation theory. *ACM SIGWEB Newsletter*, (Spring): 1–9.

Zhang, Y.; Wang, J.; Yu, L.-C.; and Zhang, X. 2021. MA-BERT: Learning Representation by Incorporating Multi-Attribute Knowledge in Transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2338–2343. Online: Association for Computational Linguistics.