Annotating and Training for Population Subjective Views

Maria Alexeeva[†] **Caroline Campbell Hyland**[†] **Keith Alcock**[†] **Allegra A. Beal Cohen**[‡]

Hubert Kanyamahanga[◁] **Isaac Kobby Anni**[◁] **Mihai Surdeanu**[†]

[†] University of Arizona, Tucson, AZ, USA
[‡] University of Florida, Gainesville, FL, USA
[◁] International Crops Research Institute for the Semi-Arid Tropics, Dakar, Senegal
{alexeeva, msurdeanu}@arizona.edu

Abstract

In this paper, we present a dataset of subjective views (beliefs and attitudes) held by individuals or groups.[1] We analyze the usefulness of the dataset by training a neural classifier that identifies belief-containing sentences that are relevant for our broader project of interest—scientific modeling of complex systems. We also explore and discuss difficulties related to annotation of subjective views and propose ways of addressing them.

1 Introduction

Collecting annotated data for training natural language processing (NLP) models is a difficult and expensive task, involving selection of data to annotate, preparing guidelines, training annotators, and more. With best prepared annotation efforts, one has to deal with disagreement among annotators, also known as Human Label Variation (Plank, 2022), and find ways to mitigate or embrace it.

The issue is even more prominent when it comes to annotation of tasks that deal with subjectivity—when an annotation assignment is not guaranteed to have one correct answer, but is open to interpretation. An objective task, e.g., determining whether a word is a noun under an annotation schema informed by a certain linguistic framework, would be less complicated than a subjective task of determining whether or not a tweet is sarcastic. Other examples of highly subjective tasks are emotion (e.g., Davani et al., 2022), humour (e.g., Meaney et al., 2021), and, to an extent, fake news detection (e.g., Pomerleau and Rao, 2017, Thorne et al., 2018).

In this paper, we present work on another task that has a high level of subjectivity: identifying subjective views of populations. We describe the task, the associated annotation effort, model-training experiments with the resulting dataset, and initial work on using the outputs of the trained model.

This work and its goals, as well as the definitions of subjective views and related terms, stem from work on using computational models to understand complex systems, e.g., agricultural value chains (AVC), food supply chains, or pandemics. Philosophers such as Heidegger argue that our "being in the world" means that all our decisions are subjective and depend on our current operating context (Dreyfus, 1990). With people being active participants and decision makers in the systems that modeling experts are trying to understand, these systems have to be, to an extent, driven by subjective beliefs of the human participants. Thus, in order to have a complete mechanistic understanding of a complex process, it is crucial for modelers to access subjective views of the populations involved. With the abundance of information available online making it difficult for modelers to identify relevant subjective views, our goal is to identify them automatically.

With this paper, we make the following contributions:

- We release a dataset for identifying subjective views of individuals or groups.
- We train a model for identifying such subjective views in text using the created dataset.
- We discuss ways in which we mitigate issues related to human label variation and provide support for embracing it through error analysis of the model predictions and application of the models trained on the data by intended users.

2 Dataset

2.1 Task Definition

With this project, we aim to help scientific modelers improve their models of complex systems through

[1] The dataset and the code are available at https://github.com/clulab/releases/tree/master/wassa2023-beliefs.

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 416–430

incorporating views of populations, which could potentially impact those systems. With this in mind, we annotate two types of subjective views: *beliefs* and *attitudes*.

We define beliefs as people's views on how the world works, or in other words, their mental models or parts of it. For instance, the following example shows people's understanding of the relation between price and quality, which can impact their purchasing behaviors, and, in turn, impact the food supply chain:

> *Consumers generally recognize that cheaper prices correspond with lower quality and tend to remain loyal to their preferences when prices increase.*

We define attitudes as subjective views that indicate people's feelings towards objects and events. The example below shows how people's attitude (wanting to secure more food) led to their behavior (cultivating crops twice within a season):

> *However, the members cultivated rice twice in 2009/10 [...] because they did not plan to cultivate rice in 2010/11 and wanted to secure a whole year's worth of rice for their own consumption.*

What unites these two types of subjective views is that they both have a potential to impact human behavior, which can in turn impact complex systems that need to be modeled. For simplicity, we refer to both of them as *beliefs* in this paper. For a comparison between our definition of beliefs and that in other datasets and related tasks (e.g., opinion mining and stance detection), see Section 6.

In this effort, sentences are annotated with respect to a trigger word—a word that can potentially indicate a belief, e.g., *think, feel, hope*, and *want*. The list of trigger words used (further referred to as *known triggers*) was created by modeling domain experts and augmented by the authors during initial data analysis. In cases where a trigger or the sentence can have multiple possible meanings, annotators are encouraged to use the paragraph context of the sentence for disambiguation.

2.2 Annotation Criteria

The guidelines used for the annotation exercise were created in consultation with scientific modeling domain experts based on the needs of the broader modeling project this work is part of.[2]

[2] https://www.darpa.mil/program/habitus

To be usable for the broader project (further, referred to as *modeling project*), sentences we annotate as beliefs have to satisfy a number of criteria, as detailed below.

Beliefs have to actually be held by some individual or population. That is, we annotate existing beliefs, e.g., *Rice production is considered a supplementary, non-commercial activity in the region.* Based on this criterion, we exclude sentences that contain (a) hypothetical beliefs, (b) variables without values (i.e., a type of belief is mentioned, but it is not stated whether or not the belief is held by anyone), (c) statements about research methodology of individual studies, and (d) recommendations:

(a) *If local actors perceive too much initial risk to invest in their own brands, [...]* (**Cf.** beliefs that are true under some conditions: *If these debts are subsequently collected, they are considered to be income subject to tax.*)

(b) *Willingness to discuss experiences of violence may also differ according to the cultural context.* (**Cf.** *Farmers expressed willingness to [...]*)

(c) *For the purposes of this report, the Northwest Territories and Nunavut was considered one jurisdiction.*

(d) *Farmers should believe that land can be rehabiliated.*

Beliefs have to be specific. We annotate complete beliefs that do not leave ambiguity as to the objects of the belief. Under this criterion, we exclude (e) beliefs that require coreference resolution outside of the sentence and (f) beliefs in restrictive clauses, in which the belief is what helps identify an object instead of a belief being held about an otherwise specified object:

(e) *It is considered the most numerous bird worldwide with population numbers totaling about 1.500 million [...]*

(f) *[This] requires subjective judgments about subgroups which are believed to be present in large numbers.* (**Cf.** non-restrictive clauses, which provide additional information about a known object: *[...] plus marked increases in Asian and Hispanic populations, who prefer rice.*)

Beliefs should not be simply reporting on facts. Beliefs are frequently discussed with a pattern <*Believer*> *said* followed by a subjective judgment, e.g.: *They said the chemicals were harmful*. We exclude beliefs that include people merely reporting facts, e.g., (g) in reported speech or (h) as research findings:

(g) *The UN said that 5.2 million people in the northeast remained in urgent need of food assistance.*

(h) *Genesee found that students in early , delayed and late immersion programs displayed no negative effects on the development of their first language*

2.3 Dataset Description

For this dataset, we annotate sentences as containing or not containing beliefs with respect to a given trigger word (when present). We annotate over a collection of scientific publications and reports in PDF format written in English on a number of topics, including education, agriculture, finance, etc. in several countries.

The dataset consists of two partitions. The training partition was annotated via the crowdsourcing platform Mechanical Turk (MTurk)[3] and quality controlled by the authors of this paper (further referred to as *the team*) for adherence to the annotation guidelines. The testing partition was created in collaboration with a modeler domain expert and supplemented with annotations provided by the team.

Each partition contains two main categories of data points: those with and without known belief triggers. In both partitions, the known trigger subset was manually annotated for presence of beliefs in the sentence. In the training partition, the triggerless examples were presumed to not contain beliefs based on the absence of known belief triggers and served as negative examples for the classifier training. In the testing partition, the number of triggerless examples was low enough to be manually annotated as well, so it contains both positive and negative examples. The training partition additionally includes the subset of triggerless examples that was used in our experiments. The statistics on each partition are reported in Table 5 of Appendix A.

At a minimum, each data point in the dataset comes with the sentence annotated, the paragraph and the name of the document that the sentence appeared in, and the annotation field, indicating whether or not the sentence contains a belief. Data points with known triggers also contain a field for the trigger and a separate field for a short text span around the trigger, the latter to specify the location of the trigger within the sentence in case the sentence contains multiple instances of the same trigger. Sentences annotated with MTurk additionally list all annotations that we accepted, i.e., that we did not discard based on annotator-level filtering criteria (see Section 2.4.2).

2.4 Data Collection Procedure

2.4.1 Document Collection and Preprocessing

The documents for annotation were collected in two ways: manually by the modelers involved in the modeling project and by querying the Google API. In the first case, the documents were collected based on their relevance to either the domain of the modeling project (agriculture) or the geographic area of interest (Senegal). This set of documents was used for creating the test set for testing how well the models we train handle the modeling project use case.

In the second case, we extracted documents with information on several countries that contained key terms relevant for the modeling project, e.g., *agriculture* or *rice*; however, since those key words can be mentioned in a number of different contexts, the resulting set of documents ended up being on a variety of topics. We attempted to exclude papers on sensitive topics, e.g., domestic violence, but information like that may still have made it into the dataset if it was present in papers on other topics. The documents collected using Google API were used for creating the training partition.

For preprocessing, we converted the PDF documents to text using a package that combines the Science Parse[4] converter and a set of methods to refine text, e.g., to eliminate words broken between lines, fix encoding issues, and find appropriate paragraph breaks. We processed text using the `processors` library[5] to break it into sentences. We filtered out strings of text that were erroneously tokenized to be sentences with simple heuristics, e.g., filtering by length and excluding uncapitalized and non-letter-symbol-heavy strings. We then ex-

[3] https://www.mturk.com/

[4] https://github.com/allenai/science-parse

[5] https://github.com/clulab/processors

tracted potential belief-containing sentences using a set of string-match-based rules that capture sentences containing known belief triggers and stored the sentence and the trigger. We extracted trigger-less examples, which we presume to be negative (non-belief-containing) data points, in a similar way: using a rule, we only selected sentences that did not contain known belief triggers. For instance, the following sentence is such a negative example: *Agriculture occupies 44% of the workforce and accounts for 25% of the GDP.*

We attempted to make the training partition thematically varied. To achieve that, we sampled sentences on several topics (education, technology, agriculture, traditions, etc). For every topic, we ranked all available belief trigger sentences by their similarity to the topic and took the top N sentences, with N depending on the sample size needed. The similarity of sentences to topics was calculated using the `SentenceTransformers package`[6] (Reimers and Gurevych, 2019) with the *all-mpnet-base-v2* model.[7] We used the model to encode potential beliefs and the topic names (e.g., education) and calculated the similarity between the belief and the topic embeddings as a dot product.

2.4.2 Annotation

For annotation, we wanted to follow a realistic annotation protocol where crowd sourcing is used to generate the training data and the test data are generated in a controlled environment by domain experts.

Team Annotation. We started the annotation process by annotating a set of sentences with guidance from a domain expert, which allowed us to decide on the initial guidelines and create the modeling project domain test set. The test set was later supplemented with additional annotations from team members. Before working on the task, the team annotators were asked to complete a series of qualification tasks of 20 sentences each followed by feedback, until their Cohen's kappa annotator agreement score calculated against the answer key reached the higher bound of moderate agreement or higher. Moderate agreement, considered to be in the 0.41– 0.60 range, was deemed sufficient for this task given its complexity.

Annotators were encouraged to provide comments along with annotations, which aligns with field recommendations (see, for instance, Plank 2022).

Mechanical Turk Annotation. The training partition of the dataset was annotated using the crowdsourcing platform Mechanical Turk with additional quality control by the team members. For every data point, we collected annotations from three MTurk workers. Before starting the task, workers had to read through the guidelines and pass a short qualification task, which covered points of potential confusion, with a score of 90% or higher, which allowed one incorrect answer. To qualify, workers also had to be over 18 years old, located in the U.S., and had completed at least 100 assignments with at least 97% assignment acceptance rate. The workers were compensated at 5 cents per data point (i.e., for evaluating one sentence as containing or not containing a belief).

With the task being highly subjective, we do not have many ways to eliminate possible bad faith annotators other than the qualification task and the annotator statistics filters. However, we removed a small number of annotations that were provided by workers that marked every sentence as containing a belief. We additionally removed the annotations provided by workers who annotated fewer than 10 sentences as they may not have had enough exposure to the task.

2.5 Annotation Issues

2.5.1 Mechanical Turk Quality Control

We evaluated the MTurk annotations by asking two team members to provide their feedback on subsets of annotations. With about 50% of annotations requiring correction in order to align with the guidelines, we chose to proceed with manual quality control of crowd-sourced data.

During quality control, a team member read the sentence and, when needed for disambiguation, the paragraph, and marked their agreement with the MTurk annotation. In complicated cases, a team member provided feedback and had the option to request a second opinion from another team member. In cases of disagreement, a third team member was available as well. Overall, 47% of the labels we assigned during quality control did not match those assigned based on majority vote on MTurk annotations.

Even with the need for quality control, we still collected the data through MTurk for several rea-

[6] https://www.sbert.net/

[7] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

sons. First, at a rate of about two data points a minute, quality control took less time for team members than providing annotations from scratch. Second, we found quality control task to be less mentally taxing than annotating from scratch. Finally, we believe that annotations from MTurk workers, although frequently misaligned with the guidelines, provide some useful signal that helps the quality controller to make a faster decision regarding the label to assign. For experimental support for quality control, see Section 4.1.

2.5.2 Task-specific Difficulties

As mentioned above, we believe the task of belief annotation is complicated because it is inherently subjective. Both triggers and sentences can have multiple meanings and be open to interpretation. In the example below, a positive connection between use of a fertilizer and the health of a plant is discussed, but this case can be interpreted as either the farmers believing it or stating facts. In cases like this, we err on the side of over-annotating beliefs:

[...] and some farmers apply urea (called 'salt'), saying that leaf color becomes healthy.

Additionally, sentences can be seen as containing or not containing beliefs based on the context in which they appear. The sentence below in bold, without the context, can be interpreted as a belief held by an organization. However, the broader context indicates that this is not a belief held by any population, but a study-specific definition introduced by the researcher:

I therefore consider that the global rice VC is part of the context, and I do not make it the focus of the research. ***Nevertheless, since importers are involved through government intervention in trading the rice produced in Senegal, they are considered as part of the domestic VC.***

2.5.3 Human Factor

Team annotation. In many cases, especially in the more complicated ones, team annotators provided comments on their annotations, both during initial annotation and quality control. These comments helped us pinpoint a few issues that may arise during annotation exercises.

The guidelines that were provided to the team annotators were quite extensive. From the comments, we learn that different annotators focused on different aspects of the guidelines. This can be illustrated by the following example:

Because of their precarious employment conditions, they are considered to be in "vulnerable" types of employment.

During the quality control, two annotators discussed via comments the meaning of the sentence while deciding on whether or not it should be annotated as containing a belief, disregarding the guideline to exclude sentences where belief is not complete (we do not know who *they* refers to).

Similarly, while explaining why they did not annotate some sentences as beliefs, some annotators kept listing the same criterion (e.g., completeness or clarity on the believer) as a reason for multiple, unrelated cases.

Mechanical Turk. During several rounds of team annotations, the proportion of sentences annotated as beliefs mainly remained in the 30–50% range. However, we observed high level of variation in the proportion of belief annotations between MTurk workers (59% mean with a standard deviation of 22%, a minimum of 11%, and a maximum of 94%). This could be an indicator of either the difficulty of the task, inadequacy of the guidelines (e.g., not informative enough or overly detailed and, therefore, not read in full), or bad-faith annotation. Another indicator pointing to possible bad-faith annotation is marking beliefs in sentences with belief triggers used in the meaning clearly not stating a belief, e.g., the known belief trigger *think* in the collocation *think tank* or the trigger *trust* in *partnerships, joint ventures, and trusts*.

3 Belief Identification

3.1 The Model

We use our dataset to fine-tune a model mimicking the task performed by the annotators: the model is intended to *provide a binary label indicating whether or not a given sentence contains a belief*. We start with the pretrained BERT model (Devlin et al., 2018) and fine-tune it for the task using the MTurk-annotated examples with known triggers from the training partition combined with a sample of triggerless examples four times the number of sentences annotated as beliefs. We run fine-tuning for 20 epochs, with a batch size of 16 and weight decay of 0.01. We do not do any hyper-parameter tuning.

3.2 Evaluation

We evaluate the performance of the model in two ways. We use cross validation ($k = 5$) to evalu-

Model	P	R	F1
In-domain	$0.68_{\pm 0.05}$	$0.73_{\pm 0.07}$	$0.7_{\pm 0.02}$
Out-of-domain	$0.77_{\pm 0.03}$	$0.80_{\pm 0.03}$	$0.78_{\pm 0.02}$

Table 1: Performance (means and standard deviations) of the belief identifier on cross-validation (in-domain) and the test set aligned with the goals of the modeling project (out-of-domain).

ate model performance in-domain, that is, to evaluate its performance on the data from the same distribution it is trained on—the quality-controlled MTurk data. We further test the model on the out-of-domain test set—the set annotated by the team in collaboration with a modeling domain expert—after training a model on all the training data available. In both cases for all experiments, we report means and standard deviations; for cross-validation, they are calculated over k folds, and for the evaluation of the full model on the test set, they are calculated using bootstrap resampling.

We note that here we use the term *domain* loosely since there may be a thematic overlap between the two sets. What the two sets differ in is that the training set is expected to have more thematic variety and was annotated in a different way. The results are reported in Table 1 as precision, recall and F1 score for the positive label.

3.3 Error Analysis

For error analysis, we manually analyze the sentences that were marked as incorrect during cross-validation evaluation in one of the five cross-validation partitions (folds). We also use the `lime` package[8] (Ribeiro et al., 2016) to analyze how the model assigns weights to features.

False Positives. By using `lime`, we learn that the model is learning to pay attention to the words related to our set of known belief triggers (Figure 1, Appendix B). However, it does not always successfully disambiguate multiple meanings of the triggers. In Example 3 (Figure 1), the model successfully learns a previously unknown (i.e., not used during training as a known trigger) trigger *enjoy*, but fails to pick up on its less frequent meaning *have* as in *enjoying a competitive advantage* and falsely predicts the sentence as containing a belief.

Some false positive predictions turn out to not be false, but result from the fact that during training we make an assumption that examples with no known belief triggers do not contain beliefs. Instead of being incorrect, these examples demonstrate that the model is able to generalize and find new belief triggers (e.g., *aspiring* in the example below), which can later be used for belief extraction:

When asked about the type of job they would like, more than 80 percent of those currently employed in agriculture, indicate to be aspiring a job outside agriculture.

Some types of errors stem from the decisions that we made for the annotation exercise that may need to be reconsidered. For instance, the model predicts sentences requiring out-of-sentence coreference resolution as beliefs. This tells us that imposing artificial constraints on annotations with the desire to simplify the task may not be feasible:

The youth of today understand this–think about courageous young people like Greta Thunberg and others like her. (**Note:** the pronoun *this* is unresolved, i.e., we do not know what the youth understand).

False Negatives. Based on the analysis of false negatives, we believe that the model learned several incorrect heuristics for belief identification:

- possible anti-modal verb bias (Figure 2 of Appendix B), which could be explained by the fact that we avoid hypothetical beliefs, but applies even when modal verbs are not modifying a belief;
- possible anti-long sentence bias (Figure 2 of Appendix B), with long sentences potentially providing more opportunity for certain non-belief terms to appear, skewing the prediction;
- possible anti-first person bias (Figure 3 of Appendix B)—since we mainly focus on reported beliefs, the model may learn that the word *we* is an indicator of non-beliefs.

General Observations. In both false positive and false negative cases, we find that some examples were possibly mislabeled by annotator, most likely because of either possibly conflicting interpretations of the guidelines (e.g., Example 3 in Figure 1, which could be interpreted as either an attitude or reporting facts) or because of the complicated structure of the sentence:

In December, 44 people arbitrarily detained for what local NGOs considered to have been

[8] https://github.com/marcotcr/lime

Amnesty International Report 2017/18 395 politically motivated reasons were released [...] (**Note:** the use of pronoun *what* can lead to a false conclusion that the sentence requires out-of-sentence coreference resolution.)

Importantly, from the `lime` analysis, we observe that the model does learn new, previously unknown potential belief triggers, e.g., *likely*, *enjoy*, and *problematic*.

3.4 Discussion

The model performs better on the out-of-domain (i.e., team annotated data developed together with the modeling project domain expert) test set (Table 1). This could be explained by an existing thematic overlap between the train and test data, the lower number of topics in the test set, and the fact that in the test set, both known and unknown trigger examples are annotated, which means there can be no false positive predictions based on unknown triggers.

From the error analysis, we learn that in some cases, e.g., when annotations show that multiple interpretations are possible for a sentence, human label variation is to be expected and should be embraced as it can help guide the development of annotation guidelines. Human label variation should also be taken into account when evaluating systems: as discussed in literature (e.g., Plank, 2022) and seen from our manual error analysis, doing evaluation only on hard labels may not be informative.

We can add that providing rationale during quality control also helped with error analysis since it made it possible to determine the meaning of the sentence and the issues that could arise without rereading the whole sentence and paragraph.

4 Additional Experiments

4.1 MTurk Annotation Threshold

Along with mitigating annotation quality issues with manual quality control, we explored the possibility of automatically cleaning the original MTurk annotations. For every data point, after filtering out suspected bad faith annotators, we had between two and three MTurk worker annotations. From our evaluation of one of the MTurk trial runs, we observed that about 65% of sentences annotated as beliefs by three annotators were judged by the quality controller to indeed be beliefs, while it was about 25% for sentences annotated as beliefs by only one or two annotators.

Partition	Setting	**P**	**R**	**F1**
CV	MTurk0.5	$0.72_{\pm 0.06}$	$0.82_{\pm 0.08}$	$0.76_{\pm 0.02}$
	MTurk1.0	$0.41_{\pm 0.07}$	$0.49_{\pm 0.09}$	$0.44_{\pm 0.04}$
	MTurkQC	$0.68_{\pm 0.05}$	$0.73_{\pm 0.07}$	$0.7_{\pm 0.02}$
Test	MTurk0.5	$0.54_{\pm 0.03}$	$0.87_{\pm 0.02}$	$^{*}0.67_{\pm 0.02}$
	MTurk1.0	$0.54_{\pm 0.04}$	$0.42_{\pm 0.03}$	$0.47_{\pm 0.03}$
	MTurkQC	$0.77_{\pm 0.03}$	$0.8_{\pm 0.03}$	$^{*}0.78_{\pm 0.02}$

Table 2: Performance of the models trained on different versions of the MTurk data (CV: cross validation, in-domain performance. Test: test partition, out-of-domain performance). On the team-annotated test set, the quality controlled data model (MTurkQC) significantly(*) outperforms the next best model, which used the original MTurk data with a belief label majority vote threshold of 50% (MTurk0.5). The MTurk0.5 model outperforms the other two models on cross-validation evaluation.

With that in mind, we conducted an experiment to evaluate which belief-annotation-proportion threshold results in best performance of the model and whether using the original MTurk data can compete with the quality controlled version. We try two thresholds: 1.0 (100%), with all the available annotators agreeing that a sentence contains a belief, and 0.5 (50%), with at least half the annotators making that judgment. Same as with the belief identifier model trained on the quality-controlled MTurk data, we evaluate the models trained with original MTurk data in-domain (using cross validation) and out-of-domain (using the test partition for evaluating the model trained on all the training data available). The results appear in Table 2.

On the out-of-domain test set (*Test* in Table 2), neither of the two threshold conditions result in performance surpassing that of the model trained on quality controlled MTurk data, with the second best model (threshold of 0.5) still performing significantly[9] worse than the best model ($p < 0.001$).

We note that the 0.5 threshold model performs better than the other two during cross-validation (*CV* in Table 2). However, given its performance on the testing partition, we believe that the high cross-validation performance could be an indicator of consistent noise present in the training data. One way to address this, other than with manual quality control, is to work on improving the guidelines provided to MTurk workers before collecting additional data. In the meantime, we believe these results support the need for quality control.

[9]Statistical significance is calculated using bootstrap resampling with 10000 samples.

Model	P	R	F1
Unmarked trigger	$0.68_{\pm 0.05}$	$0.73_{\pm 0.07}$	$0.7_{\pm 0.02}$
Marked trigger	$0.72_{\pm 0.06}$	$0.72_{\pm 0.05}$	$0.72_{\pm 0.05}$

Table 3: Performance of the belief identifier during cross-validation ($k = 5$) over the training dataset—quality-controlled MTurk with two trigger marking conditions.

Model	P	R	F1
Unmarked trigger	$0.77_{\pm 0.03}$	$0.8_{\pm 0.03}$	$0.78_{\pm 0.02}$
Marked trigger	$0.81_{\pm 0.03}$	$0.74_{\pm 0.03}$	$0.77_{\pm 0.02}$

Table 4: Performance of the belief identifier on the team-annotated modeling-project-based partition with two trigger marking conditions. No statistical significance between the two configurations was observed.

4.2 Marked Trigger Experiment

After testing the efficacy of the model at predicting beliefs given a sentence, we tested whether using another piece of information available—the trigger—would improve the model performance. We test that by marking the trigger with special tokens **<t>** at the beginning of the trigger and **<\t>** at the end of the trigger (e.g., "... he **<t>**believes**<\t>** that…"). The data was formatted the same way in the marked trigger experiment as it was in the unmarked trigger experiment, with the exception of the special tokens marking the trigger.

The results of the experiment are in Tables 3 (cross-validation performance during training) and 4 (test set performance). In cross-validation, the marked trigger model demonstrates higher performance than the unmarked trigger model, with the difference especially prominent in terms of precision. This could mean that marking known triggers while training the belief identifier can be beneficial. However, the unmarked trigger model does slightly better on the test set, although the difference is not statistically significant.

We also experimented with using predicted instead of extracted triggers; however, the performance of the trigger classifier has not yet been high enough to test it in the belief identifier. See Appendix C for details of the experiment.

5 Application

We are in early stages of using the the belief identifier for the modeling project. We rank beliefs based on similarity to topics and provide them to modeling experts. While automatically identified beliefs, expressed in natural text, cannot be directly fed into models, they can inform modelers' decisions on what parameters to include in models and how to weigh them. So far, the work on belief identification has been met with enthusiasm since, at a minimum, we can save modelers time by surfacing the information that they would normally need to manually search for. We are working on ways to improve the quality of the belief identifier as well as to make the information regarding population beliefs that we provide more systematic.

6 Related Work

Handling noisy data. Noise in annotated data is a common issue discussed in literature, with recent work focusing on embracing it during modeling and evaluation (Davani et al., 2022, Fornaciari et al., 2021, Plank, 2022). Chen et al. (2022) describe a different approach—data cleaning, or targeted relabeling,—in which they use a large portion of the annotation budget to build a model and preserve the remaining budget to relabel the examples that the model gets wrong because those are more likely to be incorrect. For a comprehensive overview and recommendations on handling annotator disagreement, see Plank (2022).

Our approach of using quality control is more similar to that of Chen et al.: with a rather limited number of data points available for training (about a thousand) and between two and three annotations per data point, modeling uncertainty did not seem feasible. Additionally, we had reasons to believe that some annotation variation came from annotator- and guidelines-related issues (see Section 2.5.3) and not from the inherent subjectivity of the task, in which case uncertainty would need to be modeled.

Belief annotation. We are not aware of any datasets that handle beliefs the way we do; however, there exist datasets that focus on beliefs, but define and annotate them from a different perspective. Most recently, Tracey et al. (2022) released *BeSt*, the corpus of beliefs and sentiment, which is concerned with capturing agents' cognitive states. The authors equate belief with factuality and annotate data in terms of whether or not the author believes the described events to be true.

Tracey et al. (2022) provide a detailed summary on related datasets. Since their corpus shares many properties with other related datasets, we will use it as a point of comparison with our work.

While Tracey et al. focus on the authors evaluation of truthfulness of described events and distinguishes between committed and not committed beliefs (the author believes the events are true vs. the author thinks they are true, but is not certain), we target both committed and non-committed beliefs—the level of certainty of the agents does not impact how we annotate or use beliefs.

Tracey et al. are interested in author beliefs, while the main focus of our work is what Tracey et al. and Prabhakaran et al. (2015) refer to as *reported beliefs*—the beliefs reported by the author of the text but held by someone else. This type of beliefs is most likely to identify beliefs of some population—which is what we are interested in capturing—while author beliefs could be idiosyncratic and not representative of beliefs of a population. However, in certain cases we annotate author beliefs as well if the author identifies themself as being affiliated with some population:

We here in Germany think that we may have risked too much [...]

Tracey et al. annotate full text, while we aim to locate reported beliefs. Due to sparsity of reported beliefs, annotating full text is not likely to result in the highest number of annotations of the type we are mainly interested in.

Our work is also related to work on opinion mining, or sentiment analysis (Wankhade et al., 2022), and stance detection (Mohammad et al., 2016). Both overlap with our work in how they target people's subjective views (*opinions* and *stance*, closely related to what we refer to as *attitudes*). However, while opinion mining and stance detection focus on subjective view gradation (*positive, negative, or neutral* in opinion mining and *against, neutral*, or *in favor* in stance detection), we are interested in the presence or absence of a subjective view in a given sentence without evaluating properties of the view, with what the view is being much more open and not forced into a Likert scale. Additionally, opinion mining and stance detection work on author views, such as, student feedback (Shaik et al., 2023), consumer product reviews (Kumar et al., 2016), and tweets (Glandt et al., 2021, Mohammad et al., 2016). We, on the other hand, are mainly interested in reported (third person) views with only occasional cases of first person narration included in the dataset.

7 Limitations and Future Work

While our dataset shows promise based on the models we train with it, at about 1000 annotated examples in the training partition, it is relatively small. Before working on increasing the size of the dataset, we need to work on improving the guidelines provided to Mechanical Turk workers and finding more robust ways of excluding bad faith annotators.

The dataset also currently misses some information that could be useful, e.g., polarity, beliefs involving out-of-sentence coreference resolution, as well as believer and belief span annotations. We plan to address all of these in future work.

8 Conclusion

In this paper, we create a dataset of subjective views of populations and test it by training and experimenting with a belief identifier model. We discuss the issues related to annotation and human label variation that we encountered during the annotation exercise such as the difficulty of creating guidelines for a subjective task and lack of certainty in annotators adhering to the annotation guidelines for various reasons (e.g., annotators focusing on different aspects of the guidelines or not annotating in good faith). We also compare two ways of managing human label variation—annotation quality control vs. majority voting with different thresholds—by evaluating a model performance under the two conditions. Finally, we provide support for the idea that human language variation should be embraced by doing an error analysis of the model predictions, which shows how language ambiguity as well as human factor and guidelines-related issues make it impossible to rely strictly on majority voting without qualitative analysis while evaluating systems working on subjective tasks.

9 Acknowledgments

The authors thank the anonymous reviewers for helpful discussion. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program. Maria Alexeeva and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies. The annotation work was partially supported through Research and Project (ReaP) Grant from the University of Arizona Graduate and Professional Student Council (GPSC).

References

Derek Chen, Zhou Yu, and Samuel R. Bowman. 2022. Clean or annotate: How to spend a limited data collection budget. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 152–168, Hybrid. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hubert L Dreyfus. 1990. *Being-in-the-world: A commentary on Heidegger's being in time, division I*. Mit Press.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 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*, pages 2591–2597, Online. Association for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

K L Santhosh Kumar, Jayanti Desai, and Jharna Majumdar. 2016. Opinion mining and sentiment analysis on online customer review. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–4.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge. *Exploring how artificial intelligence technologies could be leveraged to combat fake news. url: https://www. fakenewschallenge. org/(visited on 03/13/2020)*.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, and Linda Galligan. 2023. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Tracey, Owen Rambow, Claire Cardie, Adam Dalton, Hoa Trang Dang, Mona Diab, Bonnie Dorr, Louise Guthrie, Magdalena Markowska, Smaranda Muresan, Vinodkumar Prabhakaran, Samira Shaikh, and Tomek Strzalkowski. 2022. BeSt: The belief and sentiment corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2460–2467, Marseille, France. European Language Resources Association.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

A Dataset Statistics

In Table 5, we report the details on the dataset composition, including the number of data points, documents, known triggers, etc.

B Error Analysis with `lime`

Figures 1–3 illustrate some common errors that were discovered during error analysis using the `lime` package (see Section 3.3).

Measure	Train			Test	
	known triggers	unk. triggers	unk. in training	known triggers	unk. triggers
N documents	59	65	65	50	43
N data points	1044	9769	1440	400	193
N positive class	360	0*	0*	202	12
% positive class	34%	0*	0*	50.5%	6%
Unique triggers	95	N/A	N/A	72	12

Table 5: Dataset statistics. For the training partition unknown trigger subset, we release all available data points as well as the subsample used for the experiments. Asterisk (*) indicates values assumed based on absence of known belief triggers in the sentence.

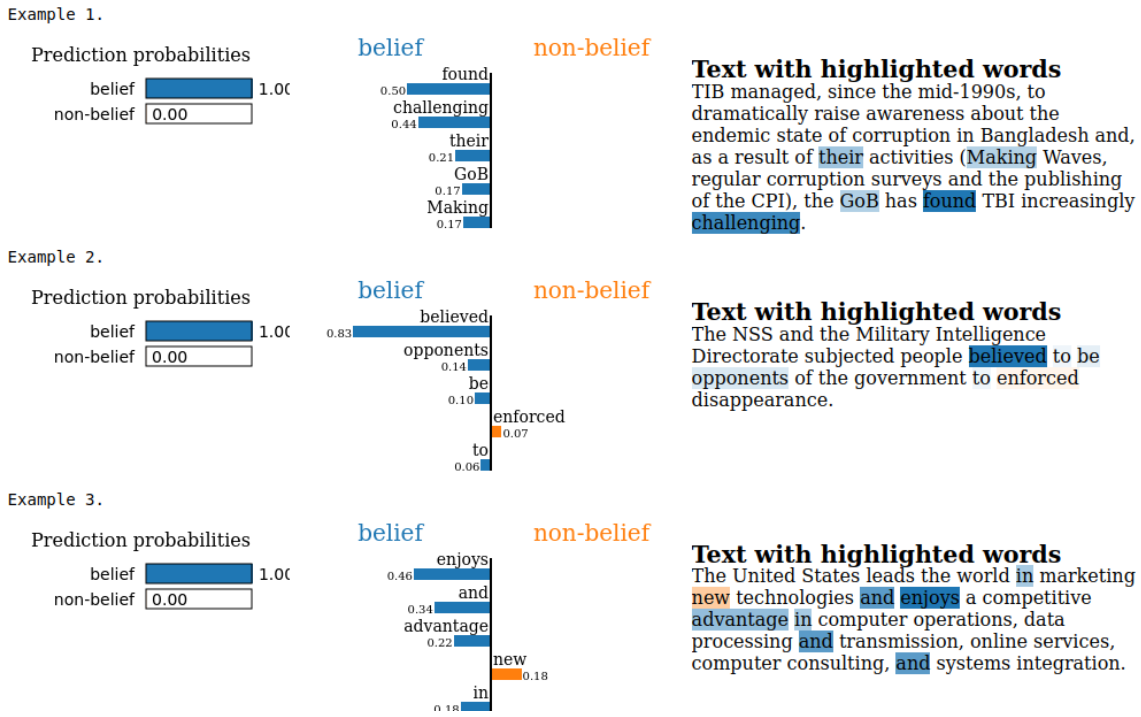

Figure 1: `lime` analysis of false positive examples from the belief identifier model trained on quality controlled MTurk data. The model learns and makes decisions on words that appear to be good quality belief triggers, but also includes some noise—the words that could occur in both beliefs and not beliefs, e.g., *their*, *and*, and *advantage*.

Figure 2: `lime` analysis of a false negative example in which the model incorrectly judges a sentence with a large number of instances of the modal verb *may* as non-belief. The possible reason is that we aim to avoid hypothetical beliefs, which eliminates a lot of belief triggers accompanied by modal verbs. The sentence can also illustrate the anti-long sentence bias, where the model tends to not annotate long sentences as beliefs.

Figure 3: `lime` analysis of a false negative example in which the model incorrectly judges a sentence with the first person pronoun *we* as non-belief. This could be happening because the dataset the model was trained on (quality controlled MTurk) focuses on reported (non-author) beliefs, so the word *we* does not get associated with beliefs.

C Extracted Triggers vs. Predicted Triggers

We wanted to test if a classifier could be trained to predict the trigger words as opposed to the current approach, which searches text for a pre-selected list of trigger words before running the sentences through the belief identifier. The classifier could be helpful in multiple ways: it could help identify new triggers, avoid the need to extract triggers before running the marked belief version of the belief identifier, and potentially improve performance of the belief identifier by marking previously unknown triggers in sentences that do not have any known ones.

We trained a classifier to label each word in a sentence as either "n" for "not a trigger", "tb" ("trigger beginning") for the first token of the trigger, or "tc" ("trigger continued") for subsequent tokens in a multi-word trigger phrase. The predicted triggers were to be added to the dataframe used for training the belief identifier model. Based on the initial experiments, we judged the performance of the trigger classifier, with only about 25% of triggers correctly identified, not to be high enough for us to proceed with the predicted trigger belief classifier experiment. We will continue the work on improving the trigger prediction model.