Social Commonsense for Explanation and Cultural Bias Discovery

Lisa Bauer **Hanna Tischer** **Mohit Bansal**
UNC Chapel Hill
{lbauer6, hannalt, mbansal}@cs.unc.edu

Abstract

Social commonsense contains many human biases due to social and cultural influence (Sap et al., 2020; Emelin et al., 2020). We focus on identifying cultural biases in data, specifically causal assumptions and commonsense implications, that strongly influence model decisions for a variety of tasks designed for social impact. This enables us to examine data for bias by making explicit the causal (if-then, inferential) relations in social commonsense knowledge used for decision making, furthering interpretable commonsense reasoning from a dataset perspective. We apply our methods on 2 social tasks: emotion detection and perceived value detection. We identify influential social commonsense knowledge to explain model behavior in the following ways. First, we augment large-scale language models with social knowledge and show improvements for the tasks, indicating the implicit assumptions a model requires to be successful on each dataset. Second, we identify influential events in the datasets by using social knowledge to cluster data and demonstrate the influence that these events have on model behavior via leave-K-out experiments. This allows us to gain a dataset-level understanding of the events and causal commonsense relationships that strongly influence predictions. We then analyze these relationships to detect influential cultural bias in each dataset. Finally, we use our influential event identification for detecting mislabeled examples and improve training and performance through their removal. We support our findings with manual analysis.

1 Introduction

Social commonsense knowledge helps humans maneuver through everyday life, aiding in situations that may require social nuance or cultural knowledge. Commonsense knowledge acquisition has been an important goal in NLP (Levesque et al., 2012; Davis and Marcus, 2015; Talmor et al., 2019) and in the past few years there has been a surge of research focusing specifically on improving social commonsense understanding for neural models (Sap et al., 2019a,b; Hwang et al., 2020; Forbes et al., 2020). However, social commonsense may contain many human biases due to social and cultural influence (Sap et al., 2020; Emelin et al., 2020). We aim to discover influential cultural biases in social applications datasets via the social and causal commonsense knowledge (Roemmele et al., 2011; Luo et al., 2016; Ponti et al., 2020) present in each dataset, to identify bias in cause-effect commonsense relationships. We define cultural bias as a positive or negative cultural attitude toward a social structure (see Section 3.1) and define causal commonsense following Sap et al. (2019a), as if-then, inferential relations. Specifically, we are interested in exploring biases towards the following social structures: *religion, economy, family, government, education* and *technology*. To the best of our knowledge, we are the first to empirically discover influential cultural biases in social tasks, using causal commonsense knowledge about social interactions, emotional reactions, and human needs to explain underlying cultural trends in social applications datasets.

We focus on two social tasks: emotion detection in social media and community value detection from interviews. We define a social task as any task that is intended to have social impact and requires social knowledge to correctly resolve. It is particularly important to identify influential biases in datasets and applications which are designed for social impact. HurricaneEmo (Desai et al., 2020) is an emotion detection task that focuses on perceived emotional reactions to events that occur during natural disasters. For example, events that exist in this dataset like *"thanks god"* may elicit emotions like AWE, and events like *"PersonX sends – to congress"* may elicit emotions like CONTEMPT, showing a positive cultural bias for reli-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3745–3760
May 2-6, 2023

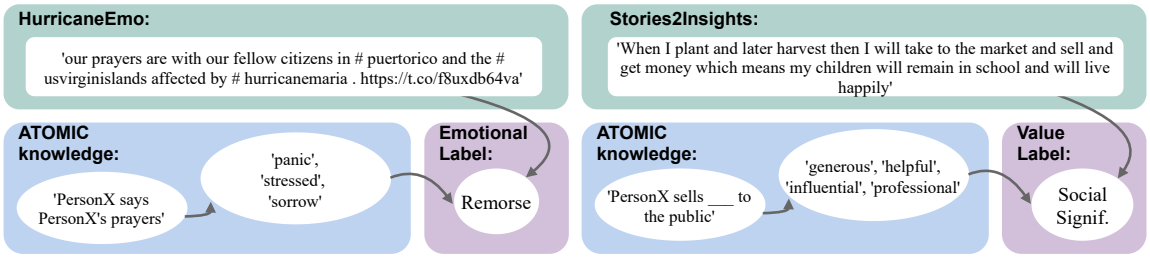

Figure 1: Examples of Social Tasks and Knowledge.

gion and a negative cultural bias for government. Stories2Insights (Conforti et al., 2020) is a value detection task that focuses on interviews conducted to identify community needs in developing countries, aiming to capture perceived values based on certain events. For example, events like *"reading the bible"* indicate INDIGENOUS values (defined by social norms and religion) whereas events like *"providing for children"* indicate INTRINSIC HUMAN needs (defined by health and quality of life). This illustrates a positive culture bias for Christianity and family. We aim to discover causal social knowledge that indicates cultural biases in each dataset, by analyzing model behavior on both tasks to gain a dataset-level understanding of events that influence performance.

We identify these biases via casual social knowledge, which encodes the relationships between events and the triggered reactions. We derived our knowledge from the social knowledge graph ATOMIC (Sap et al., 2019a). Consider the HurricaneEmo example in Fig. 1, in which the ATOMIC knowledge makes explicit the event-emotion causal relationships in the tweet that bring about the perceived emotion REMORSE. Identifying that the event *"says prayers"* causes perceived traits like *sorrow* and *stressed*, allows us to understand which events in the tweet cause the perceived emotion and contribute to the specific cultural biases. Next, consider the Stories2Insights example, in which the knowledge demonstrates the causal relationships that support the perceived value SOCIAL SIGNIFICANCE (defined by identity and status). Identifying that the event *"sells — to the public"* causes perceived traits like *influential* and *professional*, sheds light on how the event's associated cultural biases contribute to the perceived value.

In this paper, we gain a dataset-level understanding of influential causal social commonsense relationships, allowing an exploration of the underlying social and cultural biases that explain model behavior on social tasks. To this end, we first extract ATOMIC knowledge for each datapoint, utilizing a combination of TF-IDF and BERTScore (Zhang et al., 2019). We then append different components of this knowledge to the input, to focus strongly on either the cause, the effect, or the causal relation between the two. We use BERT-AUG, which ingests knowledge augmented data as input to BERT (Devlin et al., 2018), to yield improvements on both tasks. These improvements illustrate that the underlying causal social assumptions made explicit by the integrated knowledge do indeed increase model accuracy on the task and thus augment a model's understanding of the task. To discover underlying cultural biases, we investigate the influence of underlying events in each task. We first identify underlying events in the data by clustering datapoints using k Nearest Neighbors (kNN) around commonsense events, using fine-tuned contextual embeddings. We then identify which of these events are influential by performing leave-K-out experiments, which detect the influence of train clusters on task performance.

Understanding cultural biases in social applications data via the implicit commonsense knowledge present in the data is important for analyzing the limitations of the dataset and the respective tasks. By making explicit the underlying cultural assumptions and causal relationships, we are able to identify the biases that data from a certain source may have, which is paramount when using this data in the development of technology for other applications (Bender and Friedman, 2018).

Finally, we use our methods to identify mislabeling in event clusters. Datasets used for training deep learning models are large and often contain noisy labels, even if the data was collected via crowdsourcing (Frénay and Verleysen, 2013; Rajani et al., 2020). We identify events whose train clusters cause performance improvement when removed in leave-K-out experiments, indicating mis-

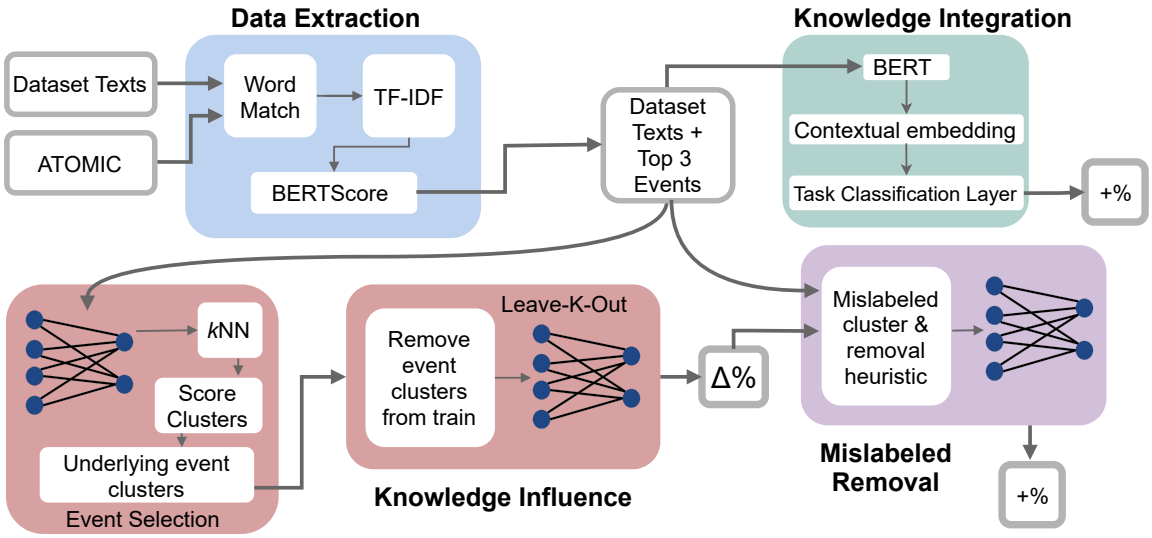

Figure 2: Data Extraction, Knowledge Integration, Knowledge Influence, and Mislabeled Removal.

labeling. Overall, our contributions are:

- We improve HurricaneEmo & Stories2Insights performance when we augment models with ATOMIC social knowledge, indicating the implicit assumptions a model requires to be successful on each dataset.
- We find representative events in each of the datasets via ATOMIC and *kNN*, and use leave-K-out experiments to discover causal social commonsense relationships that strongly influence model behavior.
- We analyze the influential cultural biases for different social structures and strongly suggest a cultural bias analysis for train data.
- We demonstrate that our methods can be used to identify mislabeled examples in the dataset.

2 Related Work

2.1 Social Applications

We explore tasks for social applications, specifically perceived emotion detection and perceived value detection. Emotion prediction has been studied for many different domains (Strapparava and Mihalcea, 2007; Katz et al., 2007; Ezhilarasi and Minu, 2012; Chen et al., 2018; Mohammadi et al., 2019), and has been extensively applied to social media posts (Mohammad, 2012; Wang et al., 2012; Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017), particularly in the social good domain. Sharifirad et al. (2019) performed emotion classification on sexist tweets and Sanders et al. (2021) analyzed sentiment in tweets during the early COVID-19 pandemic. Similar to our disaster-related application, Lin et al. (2018) used semantic matching to discover disaster recovery trends in large text corpora. In this paper, we focus on implicit perceived emotion prediction, which requires models to capture context and perform reasoning about perceived emotions (Desai et al., 2020), rather than intended emotions. Desai et al. (2020) released an emotion prediction dataset, focusing on tweets related to Hurricanes Irma, Harvey, and Maria. Hirmer and Guthrie (2016) focused on User-Perceived Values (UPVs), particularly concentrating on the needs and values of project beneficiaries in developing countries. More recently, Conforti et al. (2020) analyzed emotion as user perceived values in statements made by Ugandan rural individuals.

2.2 Social Causal Commonsense Reasoning

Commonsense reasoning has been a long-standing challenge in NLP (Levesque et al., 2012; Davis and Marcus, 2015; Talmor et al., 2019) and more recently, social commonsense reasoning has gained popularity (Rashkin et al., 2018; Nematzadeh et al., 2018; Talmor et al., 2019; Sap et al., 2019b; Hwang et al., 2020; Forbes et al., 2020; Sap et al., 2020; Emelin et al., 2020), with a strong focus on social, moral, and cultural understanding and norms. We further this work by extracting causal relations for social and cultural norms and integrating this knowledge to understand biases and model explanation in social tasks. Similar to our integration methods, Chang et al. (2020) implicitly and explicitly incorporates social knowledge from ATOMIC

(Sap et al., 2019a) and ConceptNet (Speer et al., 2017) into a social commonsense reasoning task (in contrast to event-driven tasks with real-world impacts), yielding performance gains. We, instead, focus on a causal dimension of ATOMIC and use knowledge to both improve and explain both perceived emotion and value detection. Causal commonsense reasoning has also been explored for a variety of tasks, focusing on identifying event causality in social media (Sil et al., 2010; Riaz and Girju, 2013; Kayesh et al., 2019, 2020b,a), commonsense causal QA (Roemmele et al., 2011; Luo et al., 2016; Hassanzadeh et al., 2019; Ponti et al., 2020), and conversational emotion recognition (Ghosal et al., 2020). We, instead, focus on social tasks and use knowledge to gain a dataset-level understanding of influential events.

2.3 NLP Interpretability

Our analysis methods are related to influence functions (Koh and Liang, 2017), which have been recently extended to neural text classifiers in NLP (Han et al., 2020). Our methods share particular similarity with group influence functions, which identify an influential group of training examples in a particular test prediction (Basu et al., 2020). We largely differ by using a heuristic for clustering events and verifying the influence of each cluster's train data on the cluster's dev data, allowing a dataset-level analysis of influential social relationships for text classification. Similar to our work, Rajani et al. (2020) proposed using a *k*NN framework to gain a dataset-level understanding of model behavior by identifying training examples responsible for NLI predictions. In contrast, we identify influential causal social commonsense relationships.

3 Tasks & Datasets

3.1 Definition of Cultural Bias

We define cultural bias as a positive or negative cultural attitude toward a certain social structure. Thus, positive and negative bias refer to the sentiment of the particular commonsense relation toward a social structure in the text. This is encoded differently for the two datasets that we use throughout this work. HurricaneEmo has labels for positive or negative perception (e.g., contempt = negative, love = positive). However, the target of the cultural bias depends on the subject of the text and the intended target of the emotion, thus we stress the importance of analyzing the automatically retrieved instances manually (see list of influential events in Table 4 and datapoints in Table 5). We also consider the respective commonsense relation in order to understand whether the sentiment is positive or negative. On the other hand, Stories2Insights does not use sentiment-based labels, but instead focuses on different categories of values, indicating a positive-only cultural attitude/bias toward the topic of the text. Therefore, we consider all Stories2Insights labels as positive with respect to cultural attitude.

It is crucial to note that it is not the intention of this work to draw conclusions about various cultures that these datasets may derive from. The cultural attitudes we discover in a single dataset are not an accurate reflection of the entire culture that this dataset may derive from and we would find this conclusion to be particularly harmful. We instead aim to explore biases with respect to social structures in a particular dataset and are only able to discover and examine cultural attitudes that are particular and limited to the target dataset and are additionally limited by our knowledge recall. See Section 8 for more details.

3.2 HurricaneEmo

For emotion classification, we use HurricaneEmo (Desai et al., 2020). This dataset was constructed from 15,000 English tweets about Hurricanes Irma, Harvey, and Maria. Through crowd-sourcing, each tweet was classified based on the 24 Plutchik emotions (Plutchik, 2001), and then summarized into eight emotions: AGGRESSIVENESS, AWE, CONTEMPT, DISAPPROVAL, LOVE, OPTIMISM, REMORSE, and SUBMISSION. These were split into eight binary classification tasks. For example, this tweet in the LOVE binary classification task, *"to my friends offering that support during the hurricane, i thank you. we are safe and sound. https://t.co/yl8wdbhi4"* is labeled positive. See Section A.1 in the appendix for dataset construction and size.

3.3 Stories2Insights

For Automatic User-Perceived Value classification, we obtain the Stories2Insights (Conforti et al., 2020) corpus, consisting of labeled (English) interviews from villages in Uganda. Each statement in the dataset is labeled with User-Perceived Values (UPVs), and the data is divided into six different labels: EMOTIONAL, EPIST, FUNCTION, INDIGENOUS, INTRINSIC HUMAN, and SOCIAL SIGNIFICANCE (Conforti et al., 2020). We used these

labels to create binary datasets from the original data. For example, *"Also my children and my husband will get entertained and be happy."* is labeled as EMOTIONAL. See Section A.2 in the appendix for dataset construction and size.

3.4 ATOMIC

We utilize the causal social commonsense knowledge in the knowledge graph ATOMIC (Sap et al., 2019a), a graph for if-then reasoning, connecting events through one of nine different relation edges. In this paper, we focus on the "xAttr" edge, which describes the perceived attributes of an event's subject. This edge allows us to extract perceived social knowledge, as it often covers the perceived emotion or value we are interested in for HurricaneEmo and Stories2Insights. For example, the ATOMIC event *"PersonX helps people"* uses the "xAttr" edge to show the perceived attributes of PersonX based on this event, i.e., PersonX is seen as *"kindhearted; incredible; pleasant; kind"*.

4 Models & Knowledge Augmentation

We aim to 1) extract knowledge that makes explicit the underlying causal social commonsense relationships in each datapoint and 2) propose a simple integration method to show that this knowledge is able to increase model accuracy, demonstrating the role of these relationships in improving a model's understanding of the task.

4.1 Baseline

We finetune BERT base (Devlin et al., 2018) as a baseline for both datasets. We input the text (tweet or statement) to BERT to obtain contextual embeddings, which we then project with a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times 2}$.

4.2 Knowledge Integration

We augment the BERT baseline with knowledge, which we call BERT-AUG. We illustrate this process in the *Knowledge Integration* section of Fig. 2. Each input consists of text from the task and k (where k=3) extracted ATOMIC events with attributes (the extraction process is described in Section 4.3). We directly append knowledge as text to the input text via 5 different methods. To illustrate these methods, consider the following top 2 events that may be extracted for a given tweet:

- "PersonX gets a warning" -> "unhappy", "behaved", "negative", "mischievous", "badly"
- "PersonX ignores the warnings" -> "reckless", "unworried", "confused", "dangerous", "unsafe", "careless"

Dataset	%Event-Text	%Attribute-Label
HEmo	58	54
S2I	77	65

Table 1: Data Extraction Manual Analysis. Event-Text is the match between the ATOMIC event and datapoints. Attribute-Label is the match between the respective ATOMIC attributes and the dataset label. Higher % indicates more matches.

Each integration method uses different components of the knowledge as follows, which are then appended to the input text:

- Method 1: all attributes of first event
  Example: "unhappy; behaved; negative; mischievous; badly"
- Method 2: first attribute of first event
  Example: "unhappy"
- Method3: first attribute of all events
  Example; "unhappy; reckless'
- Method4: first event + first attribute
  Example: "PersonX gets a warning: unhappy"
- Method5: first event + all attributes
  Example: "PersonX gets a warning: unhappy; behaved; negative; mischievous; badly

We determine the best method for each dataset as a hyperparameter when tuning BERT-AUG on the dev set.

4.3 Knowledge Extraction

We extract ATOMIC knowledge for both HurricaneEmo and Stories2Insights, illustrated in the *Data Extraction* section of Fig. 2. In this section, the tweets of HurricanEmo and the text of Stories2Insights are both referred to as the dataset D, as we follow the same extraction process for both datasets.

Following Lin et al. (2018), our extraction procedure is decomposed into two phases: one for speed and one for precision. The first phase serves as a coarse-grained filter to efficiently collect an initial pool of ATOMIC knowledge candidates P_i for each datapoint $d_i \in D$: if there is word overlap between d_i and a ATOMIC candidate event, the ATOMIC candidate is added to P_i. Next, we use TF-IDF

Model	AGR	LOV	SBM	AWE	DSP	CNT	RMR	AVG
BERT (Desai et al., 2020)	67.6	54.0	67.4	68.3	55.7	66.8	58.5	62.6
BERT	**75.6**	65.1	**72.8**	**73.7**	58.0	**77.9**	63.2	69.5
BERT-AUG	75.0	**68.8**	70.5	72.7	**64.4**	75.9	**65.2**	**70.4**

Table 2: HurricaneEmo Test Results (Accuracy), including aggressiveness (agr), love (lov), submission (sbm), awe (awe), disapproval (dsp), contempt (cnt), remorse (rmr), and average (avg) across all binary tasks. We recompute the baseline after extra data preprocessing. Best results are bold.

Model	Emotional	Epistemic	Functional	Indigenous	Intrinsic	Social	Avg
BERT	81.0	87.7	**84.6**	84.6	84.4	80.1	83.7
BERT-AUG	**86.9**	87.7	82.6	**89.4**	**87.2**	**82.8**	**86.1**

Table 3: Stories2Insights Test Results (Accuracy). Best results are bold.

to find the top n (we choose n=50) most similar ATOMIC candidates (considering ATOMIC event only) to d_i in P_i, and add these to a smaller pool S_i. Then, for precision we score each ATOMIC candidate in S_i with d_i for semantic match, using BERTScore, an evaluation metric for text similarity (Zhang et al., 2019). In this step, we use both the ATOMIC event and attributes to ensure that the causal social knowledge of each event is captured. We rank S_i by score and return the top k (where k=3) ATOMIC candidates for each datapoint.

To determine how precise this retrieval is, we manually examine the knowledge-datapoint matches for each dataset, demonstrated in Table 1. We analyzed 30 events from the dev set of each binary dataset (7 for HurricaneEmo, 6 for Stories2Insights; totaling 1260 datapoints), and examined matches between 1) the datapoint and ATOMIC event, and 2) the dataset label and ATOMIC attributes. This analysis investigated both whether 1) the ATOMIC events matched events in the datapoint and 2) the causal relationship between the event and its attribute is the same as the relationship between the datapoint and its label. We observe that ATOMIC matches both the events and causal relations in Stories2Insights better than those in HurricaneEmo. See Section A.4.1 in the appendix for more details.

5 Event Influence & Bias Discovery

We aim to (1) identify influential social and causal events in each dataset, and (2) analyze their respective inferential relationships to discover attitudes toward target social structures.

To identify influential events in the datasets, we first extract underlying ATOMIC events from each dataset via *event selection* and then measure how influential these events are via *leave-K-out* experiments (illustrated in *event selection* and *leave-K-out* in the *Knowledge Influence* section of Fig. 2). We then explore the following types of social and cultural structures in our datasets: *religion, economy, family, government, education* and *technology*. We consider a dataset to be biased if there exists a positive or negative attitude toward one of these structures in the influential underlying events.

5.1 Event Selection

We select underlying events in each dataset that potentially influence model behavior. To identify these events, we first fine-tune a separate BERT model for each binary dataset. Then, we represent the binary train data, dev data, and ATOMIC events with contextual embeddings using the respective BERT model for that dataset. Next, for each datapoint $d_i \in D$, we find its closest event e_j (where $e_j \in E$ and E denotes the ATOMIC events) using kNN. We then assign d_i to c_j, where c_j is the respective data cluster for e_j ($c_j \in C$ and C denotes all the clusters). If d_i is in train, we assign it to c_j^t (the event-specific train cluster), and if d_i is in dev, we assign it to c_j^d (the event-specific dev cluster). Thus, for each binary dataset, we have clusters of train and dev data for each ATOMIC event. This process is denoted as *kNN* in Fig. 2.

To identify the most representative underlying ATOMIC events learned by the model, we use the clusters of the top m (where m=50) most common ATOMIC events extracted for the train set (see Section 4.3 for details), and thus each binary dataset in HurricaneEmo and Sight2Insights has m clusters. To further identify which events have been learned best by the model, we then use cross-validation

to determine which clusters have the highest prediction accuracy (thus the dev set is not used to score clusters during event selection). We score clusters by their average prediction accuracy across each fold and select events whose clusters score highest. This process denoted as *Score Clusters* in Fig. 2. We use the top t (where t=5) events for our leave-K-out experiments.

5.2 Leave-K-Out

We evaluate whether the selected underlying events have a strong influence on model behavior (see *leave-K-out* in Fig. 2). Recall that each event e_j in a particular binary dataset has a cluster of event-specific datapoints c_j^t and c_j^d for train and dev. For each event e_j, we remove c_j^t from the train set and use c_j^d as an evaluation set. We compare results on c_j^d between a BERT baseline (discriminator trained on all train data in that binary dataset) versus BERT trained on the train set with c_j^t removed. We evaluate accuracy and look for prediction changes. We also include ablations that remove the same amount of randomly selected train data to establish a lower bound. See Section A.5.1 in the appendix for data sizes and a thorough discussion and interpretation of the respective results.

6 Results

6.1 Knowledge Integration

We examine knowledge integration in Table 2 and Table 3. Due to our additional HurricaneEmo data preprocessing (see Section A.1 in the appendix), we reran BERT baselines on this dataset. Overall, Stories2Insights has better performance improvements than HurricaneEmo.

For HurricaneEmo, we see visible improvement for LOVE, DISAPPROVAL, and REMORSE, which performed best using integration methods 2, 5, and 3, respectively, all of which are attribute focused. For Stories2Insights, we see visible improvements for datasets where the label is focused on social knowledge: EMOTIONAL, SOCIAL SIGNIFICANCE, INTRINSIC HUMAN, and INDIGENOUS, using integration methods 1, 3, 4, and 5, respectively.

6.2 Knowledge Influence

We find ATOMIC events that influence model performance in Table 4. Given the top t dataset events from event selection in Section 5, we demonstrate results for the events for which a change in performance occurred, i.e., the influential events (see Section A.5.2 in the appendix for the full list of events). We see that for all Stories2Insights events, and most events in HurricaneEmo, removing c_j^t decreases performance on c_j^d. Our ablations indicate a lower bound and thus highlight the cases where decreased performance is indeed significant. We underline all influential results that perform worse on c_j^d than both the baseline and ablation when the respective c_j^t is removed. These events are able to cluster highly relevant text and thus the c_j^t examples have strong prediction influence on c_j^d. See Section A.5.1 in the appendix for further discussion and interpretation of the results.

We observe some cases in HurricaneEmo where best performance is achieved by removing c_j^t, indicating mislabeling. We manually analyze the five HurricaneEmo event clusters in question. For each datapoint in c_j^d where the prediction was corrected after leave-K-out, we retrieve the removed c_j^t datapoints. We find that the train and dev manual label agreement between the two sets is 49.33%, averaged across all clusters. As the labels should agree, this indicates mislabeling and explains the performance improvements after removal (verified by the manual analysis in Section 6.4). Finally, there are cases in which the ablation performs worst. This seems related to the nature of the c_j^ds in question, such that the model is uncertain about these examples and thus predictions on this subset are highly sensitive to any changes in the train set.

We performed manual analysis to verify c_j (event-specific data) semantic similarity with e_j (the event) and also c_j^t and c_j^d (event-specific dev and train) similarity with each other. Two annotators analyzed approximately 190 datapoints across 13 different event clusters (with a Cohen kappa of 0.81 which is considered as 'almost perfect agreement' [1], see Section A.4.2 for more details). Table 6 shows that for HurricaneEmo, 87% of the event-specific data semantically match each other, and close to half of dev and train examples match the cluster's event. We also observe that for Stories2Insights, 100% of the event-specific data semantically match. Overall, the integration improvements, leave-K-out results, and final influential events suggest that ATOMIC represents a better cultural and social match for Stories2Insights, and allow us to clearly identify influential events in Stories2Insights. This is supported by our man-

[1] https://en.wikipedia.org/wiki/Cohen%27s_kappa

Dataset	Event	Full Train	Leave-K-Out	Ablation
HurricaneEmo				
AGR	PersonX drives from florida	**76.9**	74.4	76.9
LOV	PersonX checks the weather forecast	**60.4**	39.6	60.4
	PersonX uses — to avoid	40.0	**45.0**	35.0
SBM	PersonX sends — to the congress	82.4	**88.2**	76.5
AWE	PersonX prepares for the storm	60.6	**66.7**	60.6
	PersonX is still valid —	**69.2**	38.5	61.5
DSP	PersonX practices — in the state	57.1	**71.4**	57.1
	PersonX keeps PersonY in PersonY's prayers	**100.0**	50.0	100.0
CNT	PersonX sails close to the wind	**79.5**	74.4	79.5
RMR	PersonX crosses my heart and hope to die	82.3	**85.5**	80.7
	PersonX sends — to the congress	**69.1**	66.7	66.7
	PersonX keeps PersonY in PersonY's prayers	**80.0**	60.0	60.0
Stories2Insights				
Indigenous	PersonX offer — to the gods	**66.7**	33.3	66.7
	PersonX reads the bible	**87.5**	75.0	75.0
Intrinsic	PersonX provides — to children	**100.0**	80.0	100.0
Social	PersonX protects — from the effects	**71.4**	57.1	42.9
	PersonX sells — to the public	**88.2**	67.7	88.2

Table 4: Leave-K-Out Results. Best performance is bold and performance of most influential events is underlined.

Clustered Datapoints by Event
HurricaneEmo: PersonX prepares for the storm
even san antonio evac centers could get more than 1' of rain. tx gov. abbott suggests going farther inland to austi https://t.co/jsxrwwiu3y
puerto rico rations resources as hurricane maria approaches - https://t.co/ypziieca7a https://t.co/lzw1inirsu
cnn reports miami international airport & fort lauderdale-hollywood international airport are closed. latest updat https://t.co/repbpqdqsz
to all ga residents in hurricane irma 's path, stay safe & be careful! for shelter information, please visit: https://t.co/acytlq9orr
Stories2Insights: Person provides — for the children
We have many diseases which attack us at any including our children, so if medicines are around, we can always treat ourselves and do things which can bring for us money and our children will go to school and learn.
Motorcycle will help me to take children for treatment when they fall sick and also I can be taken for treatment using the motorcycle.
Chicken is good to have at home since it lays eggs which I use it in feeding my children.

Table 5: Examples of HurricaneEmo and Stories2Insights datapoints in ATOMIC event clusters.

Dataset	%dev	%train	%dev-train
HEmo	53	43	87
S2I	50	79	100

Table 6: HurricaneEmo and Stories2Insights Clustering Manual Analysis. %dev and %train show the match between dev/train clusters and the event. %dev-train shows the match between dev and train clusters. Higher % indicates more matches.

ual analysis which indicates that Stories2Insights events and causal relations are better captured by ATOMIC and that by using ATOMIC knowledge, we are able to get semantically matching clusters.

6.3 Bias Analysis

We examine bias with respect to the following cultural structures: *religion, economy, family, government, education* and *technology*. For HurricaneEmo, we find strong cultural biases for *religion* and *technology*. In Table 4, we see that religion is associated with several causal reactions, specifically disapproval and remorse. This may be due to the use of religion in text, marking particularly traumatic situations. We also observe that there is a general mixed reaction toward *technology*, with "driving" associated with perceived aggression and "checking weather forecast" associated with perceived love (often in the context of rapid information spread during a disaster). Finally, it seems that there is a negative bias towards *government*, in which references to "congress" tend to elicit remorse. Interestingly, the *economy*, *family* and *education* seem relatively non-influential.

For Stories2Insights, we see that the mention of *religion* is both influential and strongly associated with INDIGENOUS values (defined by social norms and religion). "Providing for children" is also a very influential event, demonstrating a strong

Model	LOV	DSP	SBM	AWE	RMR
Baseline	65.1	58.0	72.8	**73.7**	63.2
Reduced Train	**68.4**	**61.1**	**73.2**	72.1	**66.4**
Number Examples Removed	4	13	8	97	11

Table 7: HurricaneEmo test set performance after removing potentially mislabeled train datapoints.

bias for *family* in the dataset, which elicits the INTRINSIC HUMAN value, associated with health and quality of life. This sheds light on how attitudes toward religion, family, and childcare are valued in positive ways within this corpus, and indicates how these structures may play strong roles in values associated with social norms and quality of life. Finally, the *economy* also seems to be quite influential, as we see that "PersonX sells — to the public" is strongly associated with SOCIAL SIGNIFICANCE (i.e., identity, status) illustrating that the economy plays a strong role in this value.

Given the performance improvements when adding social causal knowledge and the discovered influential cultural biases described above, we see that these datasets contain implicit assumptions that, when acquired, improve a model's performance on each dataset. In particular, there are several influential cultural biases in the dataset that may be harmful when generalizing to another task. For example, the importance and meaning of *religion* may be different depending on the task. In HurricaneEmo, *religion* plays a major role as a reaction to traumatic events, whereas in Stories2Insights it plays a role in indigenous values. We observe that the context of data collection (e.g., natural disaster tweets, perceived value collection, etc) is particularly important in the type of attitudes towards social structure a dataset might encapsulate, and thus recommend this type of analysis to better understand implicit bias held in a dataset based on its application. We want to strongly emphasize that is very important to properly analyze the implicit cultural biases for any train set before applying a model trained on this dataset.

6.4 Detecting Mislabeled Events

Finally, we leverage our analysis to identify mislabeled datapoints in HurricaneEmo and improve our performance on the full test set. We use all datasets that contain events where the best performance on the target c_j^d is achieved by training on the removed train set, as illustrated in Table 4. This suggests that the c_j^t have been mislabeled and are negatively affecting c_j^d, which we have confirmed with manual analysis (see Section A.4.3 for more details). We refer to these events as mislabeled events. To mitigate this, we use the mislabeled events in a removal heuristic, where we remove all examples that extract the mislabeled events as their highest scoring event. We then evaluate our model on the full test set, see Table 7. Every considered dataset demonstrates improvement over the BERT baseline on the full test set, except AWE, most likely due to the large number of removed examples for AWE, which may interfere with the predictions of other datapoints.

7 Conclusion

We used causal social commonsense knowledge to discover influential events and relationships that explain model behavior and pinpointed instances of cultural bias. First, we found that using large-scale language models augmented with causal social knowledge improved our social classification tasks, illustrating that the knowledge made the underlying social assumptions in the dataset explicit. Then, we identified underlying events in each dataset by clustering data around ATOMIC knowledge, to pinpoint cultural biases that the dataset may exhibit. We found that some of this knowledge strongly influenced model behavior through leave-K-out experiments, providing a dataset-level understanding of influential events and causal social commonsense relationships and allowing an analysis of the datasets' implicit influential cultural biases. Finally, we used these underlying and influential events to identify mislabeled train examples and thus improve training and performance.

Acknowledgements

We thank the reviewers for their useful feedback. This work was supported by DARPA MCS Grant N66001-19-2-403, NSF-CAREER Award 1846185, and an NSF Graduate Research Fellowship. The views are those of the authors and not of the funding agency.

8 Ethical Considerations & Limitations

We first address ethical considerations and limitations with respect to potential biases in our methods and resources, and then ethical considerations with respect to data use.

Bias: Emotions and values, and their expression and perception, are not universal. We use ATOMIC as, at the time of this work, it is the largest English source of social commonsense knowledge, however it is important that we further the creation and use of resources that are not limited to Western norms, developed countries, and the English language, especially when applying them to data outside of these domains. We anticipate that a knowledge graph better suited for representing different cultural attitudes will yield more coverage. For example, the events in ATOMIC may not cover important or representative cultural events in Uganda (the source of our second dataset), as it largely contains Western-centric social and cultural norms found in mostly developed countries.

It is also crucial to note that it is not the intention of this work to draw conclusions about various cultures that these datasets may derive from. The cultural attitudes we discover in a single dataset are not an accurate reflection of the entire culture that this dataset may derive from and we would find this conclusion to be particularly harmful. We instead aim to explore biases with respect to social structures in a particular dataset and are only able to discover and examine cultural attitudes that are particular and limited to the target dataset and are additionally limited by our knowledge recall. A precision-recall tradeoff exists based on the KG coverage, and thus the retrieved datapoints may not all be biased (precision) nor will all biases be retrieved (recall) using the available KG. Thus, we also stress the importance of using manual analysis to identify and confirm biases in the extracted datapoints. We hope this will further encourage the development of knowledge graphs that explicitly represent more annotated social/cultural commonsense knowledge by illustrating its usefulness in gaining a corpus-level understanding of datasets.

Similarly, we are additionally limited by the use of transformer models that were trained on mostly Western text and may be prone to capture Western cultural information even if fine-tuned on a dataset with different cultural attitudes. Providing a solution to this problem is beyond the scope of this paper, but future work could explore more culturally diverse data for pre-training models.

Data Use: Twitter data and interview transcripts are sensitive data and thus require strong considerations about the use and release of the data. The publicly released data for HurricaneEmo is fully anonymized to protect the identity of users. Stories2Insights data has also been fully anonymized, but is not publicly released and has been kept private to ensure the safety of the communities and to prevent harmful use of the data. Following suit, we do not release any data for this dataset to limit potential harm or misuse. For more details on the ethics concerning the collection and intended use of either of these datasets, please refer to each dataset's original paper.

Future Work: Our contribution focuses on the novel combination of knowledge graph relations, interpretability methods, and clustering to identify both influential and biased commonsense relationships. We aim to use this as an opportunity to encourage work in the curation of resources that explicitly annotate cultural attitudes or "commonsense" biases by illustrating how such resources can facilitate a corpus-level understanding of implicit influential cultural biases. Our work is limited by the methods we explored, and thus we encourage further investigation of certain elements in our experimental pipeline, in particular other datasets, clustering method variations, other MLMs, more diverse KGs, and other knowledge models.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Samyadeep Basu, Xuchen You, and Soheil Feizi. 2020. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pages 715–724. PMLR.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out*

(DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 74–79.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 646–651.

Costanza Conforti, Stephanie Hirmer, David Morgan, Marco Basaldella, and Yau Ben Or. 2020. Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. *arXiv preprint arXiv:2004.12935*.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. *arXiv preprint arXiv:2004.14299*.

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*.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

R Ezhilarasi and RI Minu. 2012. Automatic emotion recognition and classification. *Procedia Engineering*, 38:21–26.

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

Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2470–2481.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *IJCAI*, pages 5003–5009.

Stephanie Hirmer and Peter Guthrie. 2016. Identifying the needs of communities in rural uganda: A method for determining the 'user-perceived value'of rural electrification initiatives. *Renewable and Sustainable Energy Reviews*, 66:476–486.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. Swat-mp: the semeval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313.

Humayun Kayesh, Md Islam, Junhu Wang, et al. 2019. On event causality detection in tweets. *arXiv preprint arXiv:1901.03526*.

Humayun Kayesh, Md Saiful Islam, Junhu Wang, Shikha Anirban, ASM Kayes, and Paul Watters. 2020a. Answering binary causal questions: A transfer learning based approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Humayun Kayesh, Md Saiful Islam, Junhu Wang, ASM Kayes, and Paul A Watters. 2020b. A deep learning model for mining and detecting causally related events in tweets. *Concurrency and Computation: Practice and Experience*, page e5938.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Lucy H Lin, Scott B Miles, and Noah A Smith. 2018. Natural language processing for analyzing disaster recovery trends expressed in large text corpora. In *2018 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–8. IEEE.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 421–430.

Saif Mohammad. 2012. # emotional tweets. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 2(31):301–326.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. Neural feature extraction for contextual emotion detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 785–794.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. 2020. Explaining and improving model behavior with k nearest neighbor representations. *arXiv preprint arXiv:2010.09030*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Abraham C Sanders, Rachael C White, Lauren S Severson, Rufeng Ma, Richard McQueen, Haniel C Alcântara Paulo, Yucheng Zhang, John S Erickson, and Kristin P Bennett. 2021. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of covid-19 twitter discourse. *medRxiv*, pages 2020–08.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. In *Conference on Empirical Methods in Natural Language Processing*.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2019. How is your mood when writing sexist tweets? detecting the emotion type and intensity of emotion using natural language processing techniques. *arXiv preprint arXiv:1902.03089*.

Avirup Sil, Fei Huang, and Alexander Yates. 2010. Extracting action and event semantics from web text. In *AAAI Fall Symposium: Commonsense Knowledge*. Citeseer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 HurricaneEmo Pre-processing

We performed the following pre-processing steps for HurricaneEmo (released by Desai et al. (2020)). First, we de-duplicated each train, dev, and test file (per the authors suggestion). Then, we de-duplicated across splits, removing data from the train set if it was present in the dev or test sets. We used these files as our input for the pipeline described in our paper. We chose not to include the OPTIMISM category in our experiments since we were not able to reproduce a baseline result sufficiently close to the original paper on the fully deduplicated dataset. After preprocessing, the train/dev/test splits for each of the files are: AGGRESSIVENESS: 1695/493/495, AWE: 2942/868/868, CONTEMPT: 1507/452/443, DISAPPROVAL: 2363/706/707, LOVE: 1028/307/304, REMORSE: 3104/910/908, SUBMISSION 2432/724/721.

A.2 Stories2Insights Pre-processing

We performed the following pre-processing steps for Stories2Insights. We received the dataset from the authors, with the T3 labels described in the original paper (Conforti et al., 2020). We instead use the T1 labels defined in their paper as binary datasets. To create these datasets, we followed the T3 label groupings shown in Appendix A of their paper. We then created datasets with a 1:1 positive-to-negative datapoint ratio for the train, dev, and test sets. These negative datapoints were collected evenly from all other T1 datasets. We repeated this process for all T1 labels and every set (train, dev, and test). After preprocessing, the train/dev/test splits are: EMOTIONAL: 805/75/84, EPIST: 553/82/65, FUNCTION: 2551/302/345, INDIGENOUS: 695/64/104, INTRINSIC HUMAN: 2317/261/320, SOCIAL SIGNIFICANCE: 1241/134/151. If interested in using this data, please contact the original authors for access.

A.3 Reproducibility & Hyperparameters

We utilized BERT-base for all of our experiments (Devlin et al., 2018). We followed hyperparameter settings for BERT as described in Desai et al. (2020). To train BERT for HurricaneEmo, we use batch size 8 and learning rate 2e-5. To train BERT for Stories2Insights, we use batch size 8 and learning rate 1e-5. Training for both was completed in 3 epochs. We train each model using 1 GeForce GTX 1080 Ti GPU.

A.4 Manual Analysis Details and Further Analysis

A.4.1 Data Extraction

Manual analysis for examining the knowledge-datapoint matches for each dataset was completed by an expert author, since this is time-consuming, fine grained verification analysis (as opposed to model evaluation). Details on the setup of the analysis are in the main paper Section 4.3.

A.4.2 Cluster Match

This section describes the manual analysis used to analyze the cluster matches via our methods in Section 6.2 in the main paper (Table 6). Manual analysis was completed by 2 expert authors since this is time-consuming, fine grained verification analysis (as opposed to model evaluation). We obtain high agreement, with a Cohen kappa of 0.81 (which is considered as 'almost perfect agreement'; see https://en.wikipedia.org/wiki/Cohen%27s_kappa). We selected 10 dev and 10 train datapoints for each event cluster, and because not all events had a full 10 datapoints in dev or train, this leads to total of 190 datapoints.

Using this data, we performed an analysis to identify whether the event-specific data sets that were extracted for a certain ATOMIC event (1) semantically matched the events and (2) semantically matched each other. For (1), we wanted to see whether the semantic meaning in the ATOMIC event was reflected in the clustered data. For (2), we wanted to see whether the semantic meaning in event-specific dev (c_j^d) and event-specific train (c_j^t) sets was matched. The results were computed as follows. For (1), if the majority of selected datapoints matched the event, we considered this to be a positive data−event match. We then compute: # events with semantically matching data−event clusters / # total events. For (2), if the majority of selected datapoints matched across dev and train, we considered this to be a positive train−dev match. We then compute: # events with semantically matching train−dev clusters / # total events. Majority is calculated such that a majority of both train and dev had to be semantically similar.

We see that the results for (2) are very high in Table 6, which supports our findings that there is mislabeling in the dataset (i.e, the labels are different between dev and train sets but the semantic

Dataset	Event	%Event-Specific Train	%Event-Specific Dev
HurricaneEmo			
AGR	PersonX drives from florida	4.7	7.9
LOV	PersonX checks the weather forecast	24.0	36.2
	PersonX uses — to avoid	5.5	6.5
SBM	PersonX sends — to the congress	1.2	2.3
AWE	PersonX prepares for the storm	4.6	7.6
	PersonX is still valid —	0.8	1.5
DSP	PersonX practices — in the state	0.6	1.0
	PersonX keeps PersonY in PersonY's prayers	0.1	0.3
CNT	PersonX sails close to the wind	7.2	8.6
RMR	PersonX crosses my heart and hope to die	5.9	6.8
	PersonX sends — to the congress	3.2	4.6
	PersonX keeps PersonY in PersonY's prayers	0.2	0.5
Stories2Insights			
Indigenous	PersonX offer — to the gods	4.7	4.7
	PersonX reads the bible	4.7	12.5
Intrinsic	PersonX provides — to children	1.1	1.9
Social	PersonX protects — from the effects	2.0	5.2
	PersonX sells — to the public	21.9	25.4

Table 8: Leave-K-Out Cluster Size. %Event-Specific Train is the % of train examples removed during Leave-K-Out training and %Event-Specific Dev is the % of dev examples the trained model is evaluated on.

meaning are similar, thus there is mislabeling). We also see that (1) can be low for some datasets, indicating that while the train and dev data may be semantically matched, this semantic meaning may differ from the original ATOMIC commonsense relation for some datasets more than others. This may be due to limited coverage of ATOMIC for some events and is intended to show transparency in the limitations of our approach, which we hope will encourage the development of a KG that explicitly represents this type of cultural knowledge (which was not available at the time of this work).

A.4.3 Mislabeling

This section describes the manual analysis used to confirm mislabeling in HurricaneEmo train examples. Manual analysis was completed by an expert author since this is time-consuming, fine grained verification analysis (as opposed to model evaluation). We completed manual analysis to identify mislabeling on a set of 30 train examples across different event clusters. The train examples were selected from cases where the performance improved on the event specific dev set when the event specific train set was removed.

A.5 Knowledge Influence Details

A.5.1 Leave-K-Out Cluster Size

We illustrate the sizes of the event-specific train (c_j^t) and dev (c_j^d) clusters with respect to the original train and dev dataset sizes for each of the leave-K-out results in Table 8. Some of the resulting datasets are very small and are more difficult to draw conclusions from. For this reason, we show the dataset sizes in Table 8 for an improved and transparent interpretation of the results. On the other hand, we can also see that several of the datasets are quite large and leave-K-out has a clear and significant impact on their performance with respect to the full train set (e.g., "PersonX checks the weather forecast" and "PersonX sells — to the public").

A.5.2 Selected Events

We show full top 5 events from the event selection step for each binary dataset in Tables 9 and 10.

Dataset	Event
Emotional	PersonX sells PersonY's — for money
	PersonX sleeps well —
	PersonX gets — at night
	PersonX devotes — to the study
	PersonX wakes up in the middle of the night
Epist	PersonX hears — on the radio
	PersonX checks the news
	PersonX supplies the — with food
	PersonX loves listening to music
	PersonX protects the — from injury
Functional	PersonX protects the — from injury
	PersonX uses — to prevent
	PersonX seeks god 's —
	PersonX teaches children —
	PersonX educates PersonX's children
Indigenous	PersonX offer — to the gods
	PersonX reads the bible
	PersonX pays a lot of money
	PersonX seeks god 's —
	PersonX treats the — with respect
Intrinsic	PersonX educates PersonX's children
	PersonX provides — to children
	PersonX works well in business to get
	PersonX pays a lot of money
	PersonX helps the — to understand
Social	PersonX uses — to prevent
	PersonX protects — from the effects
	PersonX sells —to the public
	PersonX uses — to support
	PersonX tries to use it

Table 9: Top 5 Events for Stories2Insights Binary Datasets

Dataset	Event
AGR	PersonX stays away from PersonY PersonX drives from florida PersonX doesn't have enough money PersonX doesn't want to go to school PersonX moves to texas
DSP	PersonX comes back to my house PersonX practices — in the state PersonX keeps PersonY in PersonY's prayers PersonX is trying to watch a movie PersonX sees PersonY's friends
CNT	PersonX says the wrong thing PersonX can't afford to fix it PersonX doesn't want to go to school PersonX supplies the — with food PersonX sails close to the wind
LOV	PersonX practices — in the state PersonX goes to the local animal shelter PersonX spends — with PersonX's families PersonX checks the weather forecast PersonX uses — to avoid
SBM	PersonX uses PersonX's — to help PersonX goes to the local animal shelter PersonX sends — to the congress PersonX organizes — in a way PersonX does n't want to go to school
AWE	PersonX thanks god PersonX provides — for the people PersonX uses — to protect PersonX is still valid — PersonX prepares for the storm
RMR	PersonX crosses my heart and hope to die PersonX strikes — into the hearts PersonX sends — to the congress PersonX keeps PersonY in PersonY's prayers PersonX doesn't want to go to school

Table 10: Top 5 Events for HurricaneEmo Binary Datasets