StereoKG: Data-Driven Knowledge Graph Construction for Cultural Knowledge and Stereotypes

Awantee Deshpande, Dana Ruiter, Marius Mosbach, Dietrich Klakow
Spoken Language Systems Group, Saarland University, Germany
{adeshpande,druiter,mmosbach,dietrich}@lsv.uni-saarland.de

Abstract

Analyzing ethnic or religious bias is important for improving fairness, accountability, and transparency of natural language processing models. However, many techniques rely on human-compiled lists of bias terms, which are expensive to create and are limited in coverage. In this study, we present a fully data-driven pipeline for generating a knowledge graph (KG) of cultural knowledge and stereotypes. Our resulting KG covers 5 religious groups and 5 nationalities and can easily be extended to include more entities. Our human evaluation shows that the majority (59.2%) of non-singleton entries are coherent and complete stereotypes. We further show that performing intermediate masked language model training on the verbalized KG leads to a higher level of cultural awareness in the model and has the potential to increase classification performance on knowledge-crucial samples on a related task, i.e., hate speech detection.

1 Introduction

Fairness, accountability, and transparency to fight model-inherent bias and discrimination have become a major branch of machine learning research in recent years. This includes studying cultural bias and stereotypes in datasets and language models. Stereotypes are cognitive schemas that aid in categorizing and perceiving other social groups (Hilton and von Hippel, 1996), and becoming conscious of this stereotyping can increase cultural knowledge and sensitivity (Buchtel, 2014). However, without mindfulness, stereotypes lead to inferring traits of individuals from their (e.g., socio-economic) status or social group (Hoffman and Hurst, 1990), which then leads to systemic discrimination. Stereotypes as inherent cognitive functions are equally present in human-generated content, e.g., text resources used to train machine learning algorithms, which then further propagate and lead to discrimination (Hovy and Spruit, 2016). Within the natural language processing community, bias reduction includes work in reducing gender (Bolukbasi et al., 2016), ethnic or religious bias (Manzini et al., 2019) in word embeddings or classification tasks (Dixon et al., 2018; Badjatiya et al., 2020; Mozafari et al., 2020). Nevertheless, these techniques often rely on predefined lexicons, which are mostly human-written and thus expensive in their creation. Instead, we present an entirely data-driven pipeline for the creation of a scalable knowledge graph (KG) of cultural knowledge and stereotypes. Our resulting knowledge graph, called StereoKG, consists of 4,722 entries about 10 different social groups, i.e., 5 religious groups and 5 nationalities. This knowledge graph has several use cases, ranging from analyzing existing stereotypical and cultural knowledge online, to removing ethnic and religious bias or increasing the cultural awareness of classifiers. In our experiments, we focus on the latter: integration of cultural knowledge to improve classification performance. Overall, our contributions are threefold:

- Development of a fully **data-driven** knowledge graph construction approach on Twitter and Reddit data.
- **Manual evaluation** and analysis of the resulting knowledge graph of cultural knowledge and stereotypes, highlighting the importance of multiple-mention entries in representing cultural stereotypes, which achieve higher quality than single-mention entries.
- Classification experiments showing that performing intermediate masked language model training on linearized stereotype knowledge can improve the **classification performance** on knowledge-crucial samples on a hate speech task.

The rest of the paper is structured as follows:

After describing the related work (Section 2), we present our knowledge graph creation technique (Section 3) which is then evaluated in a quantitative and qualitative fashion (Section 4). Section 5 describes the knowledge integration experiments, which constitute downstream task performance on hate speech detection and masked language modelling predictions of cultural content. We then discuss (Section 6) and conclude (Section 7) our findings. Ethical concerns are addressed in Appendix A.

2 Related Work

Cultural knowledge about different social groups and entities plays an important role in responding to contextual situations. In this work, we target cultural knowledge as a form of commonsense (LoBue and Yates, 2011). Incorporating cultural commonsense in reasoning tasks is an understudied practice in NLP. Anacleto et al. (2006) study the variation of cultural commonsense and how it affects computer applications. While there exist knowledge base resources for general commonsense (Lenat, 1995; Speer et al., 2017; Tandon et al., 2014), to the best of our knowledge, Acharya et al. (2020) have provided the only work targeting the construction of a cultural knowledge graph, which comprises various rituals and customs for two cultures. However, since it relies on crowdsourcing, it is limited in its coverage and is not easily extendable.

Cultural knowledge is largely correlated to **stereotypes**. Contrary to exhaustive research avenues analyzing gender and ethnic stereotypes, our work focuses on the lesser-studied nationality and religious stereotypes. Snefjella.B et al. (2018) have shown that national stereotypes could be grounded in the collective linguistic behavior of nations, while the Harvard Pluralism Project[1] stresses the importance of considering religion as a factor for prejudice. Because of the diversity of social groups and their behavioral traits, stereotypes and cultural attributes have unclear boundaries, making it difficult to distinguish between the two. Keeping this in mind, we treat cultural knowledge and stereotypes as interchangeable terms.

Stereotypes have been used to estimate **bias** in language models using curated datasets (Nadeem et al., 2021; Nangia et al., 2020). Stereotypical data has also been extracted from search engine autocomplete predictions using query prompts (Baker and Potts, 2013) and then used for analyzing how language models learn these concepts (Choenni et al., 2021). Bolukbasi et al. (2016) use minimal pairs of male-female terms to debias word embeddings.

In our work, we create a unified resource of cultural knowledge and stereotypes. Knowledge graphs serve as sources of representing knowledge in a structured format. Factual knowledge bases such as DBPedia (Auer et al., 2007), Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014) contain grounded knowledge about individual entities. **Knowledge graph construction** for commonsense reasoning has also been a common object of research (Lenat, 1995; Speer et al., 2017; Tandon et al., 2014). While some KGs comprise an if-then reasoning scheme (Sap et al., 2019; Forbes and Choi, 2017), some contain knowledge in the form of triples (Vrandečić and Krötzsch, 2014) or as simple natural language sentences (Bhakthavatsalam et al., 2020; Thorne et al., 2021). Crowdsourced KGs, e.g., Wikidata, result in good quality knowledge, but require large-scale manual annotation and resources. In contrast, KGs constructed in an automated manner have a lower cost in construction, are easily extendable, and have been shown to be useful in several downstream applications (Suchanek et al., 2007; Bhakthavatsalam et al., 2020). For example, Romero et al. (2019) use questions as prompts for learning commonsense cues from search engine query logs and question-answering forums and construct a commonsense knowledge base.

Explicit **knowledge integration** of knowledge resources into language models can be roughly categorized into fusion based approaches and language modeling based approaches. Fusion based approaches (Peters et al., 2019; Wang et al., 2021; Yan et al., 2021) typically perform knowledge integration by combining language model representations with representations extracted from knowledge bases. Compared to language modeling based approaches, as explored by us, they rely on aligned data and are typically applied during the pre-training stage. Language modeling based approaches commonly start from a pre-trained language model and perform knowledge integration via intermediate pre-training. For example, Bosselut et al. (2019) integrate commonsense knowledge by performing language modeling on triples ob-

[1] https://pluralism.org/stereotypes-and-prejudice

tained from ATOMIC and ConceptNet. Recently, Da et al. (2021) analyzed this approach in the few-shot training setting. In contrast to our study, both works consider autoregressive language models and use the resulting models for knowledge base construction, while we study the impact of knowledge integration on downstream task performance. Similar to our work, Lauscher et al. (2020) integrate commonsense knowledge via masked language modeling. They obtain sentences for intermediate pre-training by randomly traversing the ConceptNet knowledge graph. Unlike our work, they do not update the weights of the pre-trained model and train adapter layers instead. Moreover, while we focus on hate-speech classification as our downstream task, they evaluate on GLUE.

3 Knowledge Graph Construction

We focus our cultural KG on 5 religious (*Atheism*[2]*, Christianity, Hinduism, Islam, Judaism*) and 5 national (*American, Chinese, French, German, Indian*) entities. Previous work on automatic KG creation depended on external algorithms, i.e., autocompletion of search engine queries (Romero et al., 2019; Choenni et al., 2021; Baker and Potts, 2013). This dependency is limiting, as external providers may filter[3] outputs of their autocomplete algorithm, especially on sensitive topics such as *culture* and *identity*. Instead, we keep control over the whole KG creation process. The entire KG construction pipeline is illustrated in Figure 1.

Using statement and **question mining**, cultural knowledge and stereotypes regarding our entities of interest are collected from two social media platforms, Reddit and Twitter. For Reddit, we limit our search to subreddits relevant for the respective subjects (e.g. *r/germany* for Germans) together with common question-answering subreddits (e.g., *r/AskReddit*) using the PRAW[4] library. The complete list of queried subreddits is given in Appendix B. Similar to the commonsense mining approach by Romero et al. (2019) and Choenni et al. (2021), we use fixed question and statement templates (Table 1) to identify potential sentences containing cultural knowledge with the assumption that questions posted about various national and religious entities act as cues for underlying stereotypical notions about them. This results in 11,259 mined questions and statements. The questions are then **converted into statements** using `Quasimodo`[5] (Romero et al., 2019), as OpenIE does not process interrogative sentences.

Query Templates
Why is <*SUB*>
Why isn't <*SUB*>
Why are <*SUB*>
Why aren't <*SUB*>
Why can <*SUB*>
Why can't <*SUB*>
Why do <*SUB*>
Why don't <*SUB*>
Why doesn't <*SUB*>
How is <*SUB*>
How do <*SUB*>
What makes <*SUB*>
Why does <*SUB*> culture
<*SUB*> are so
<*SUB*> is such a

Table 1: Question-based (top) and statement-based (bottom) query templates.

To reduce redundancies in the KG triples, we **cluster** the mined sentences with similar content together using the fast clustering method for community detection implemented in the `SentenceTransformers`[6] (Reimers and Gurevych, 2019) library. This step results in 6,993 singletons and 610 clusters with more than one instance. We hypothesize that non-singleton clusters are better representatives of cultural knowledge and stereotypes, as these are based on questions that have been asked by several users, while singletons may be based on unique thoughts which do not represent a popular stereotype or cultural reality. The qualitative difference between singletons and clusters is evaluated in Section 4.2.

All assertions are then **converted into triples** using `OpenIE` (Mausam, 2016). As OpenIE outputs multiple triples which may be noisy or irrelevant, they are filtered using the following heuristics:

- Eliminate triples containing personal pronouns, e.g., *I*, *he*.
- Eliminate triples not containing the original subject entity.
- Remove colloquialisms (e.g, *lol*) and modalities (e.g., *really*) from triples.

[2]Although atheism is not a religion, we still include it under the list of religious dispositions as a religious belief.

[3]In its battle against biased or hateful content, Google has imposed filters on its autocomplete predictions for targeted questions.

[4]https://github.com/praw-dev/praw

[5]https://github.com/Aunsiels/CSK

[6]https://www.sbert.net/examples/applications/clustering/README.html

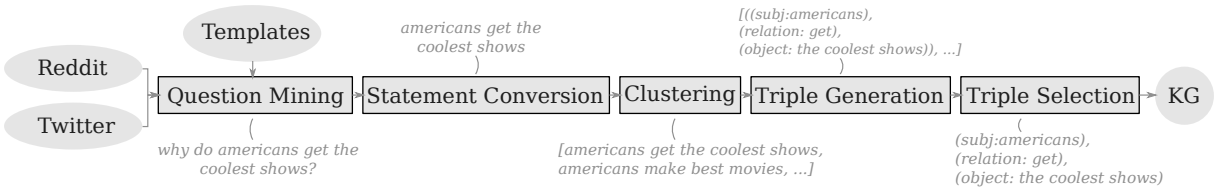

Figure 1: From noisy social media content to structured knowledge graph: the creation pipeline of StereoKG.

While most triples are singletons, many are part of a cluster. In order to **select the triple** to represent a cluster in the final KG, triples within a cluster are converted into sentences via concatenation of their subject-predicate-object terms. These are ranked on their grammaticality using a binary classification model[7] trained on the corpus of linguistic acceptability (CoLA) (Warstadt et al., 2019). Concretely, the rank of a sentence is the score assigned to the *grammatical* class by the classification model, and the triple with the highest rank is chosen as the representative for the entire cluster. Since CoLA and the resulting classifier are restricted to English, our triple selection currently only works for English data. However, our method provides an advantage over standard cluster representative selection methods such as centroid identification, since we ensure that the chosen representative triple is the most fluent choice in its cluster. This is important, since (grammatical) completeness is an important quality feature for a KG, which we also assess as part of our human evaluation.

4 Knowledge Graph Evaluation

The resulting KG consists of 4,722 entries, with Americans being the largest represented group (1,071 entries) and Jews (43) the smallest. The proposed pipeline can also be utilised to extend the KG with additional entities. In the following section, we describe the qualitative and quantitative evaluation of the generated KG.

4.1 KG Statistics

To gain insights into the sentiments and overall distribution of descriptive predicates, we evaluate the KG on two criteria.

Sentiment Analysis We perform a ternary (*positive*, *neutral*, *negative*) sentiment analysis over the KG triples by verbalizing them into sentences. We use a pre-trained sentiment classification model[8] (Barbieri et al., 2020) for this task. We observe that for certain subjects, e.g. *atheists*, the triples have a higher tendency to be negatively evaluated by the simple presence of the entity term. In order to mitigate this bias in the sentiment analysis classifier, we mask[9] the subject entities with their type, e.g. *"islam seems to be conservative"* → *"religion seems to be conservative"* and *"french culture is pure"* → *"nation culture is pure"*, and then perform classification.

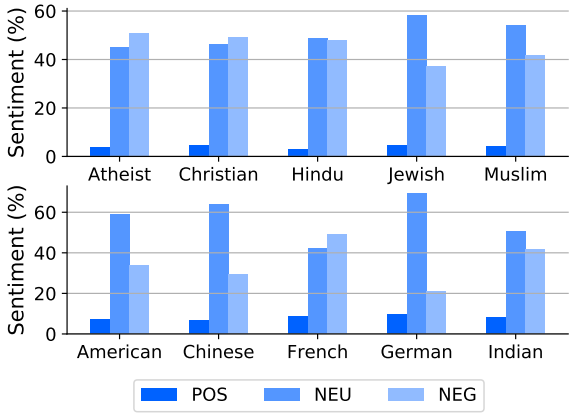

Figure 2: Percentage of POSitive, NEUtral and NEGatively evaluated triples per religious (top) and nationality (bottom) entity.

Pointwise Mutual Information (PMI) PMI $\pi(x, y)$ measures the association of two events. We calculate π between entities $E = e_1, ..., e_n$ and their co-occurring predicate and object tokens w as:

$$\pi(e, w) = \log \frac{p(e, w)}{p(e)p(w)} \quad (1)$$

[7] https://huggingface.co/textattack/distilbert-base-uncased-CoLA

[8] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[9] Note that the more generic term used to mask the specific religion or nationality terms may also have a biased representation in the pre-trained classifier. However, when applying masking via generic terms, we observe a large decrease in the negative classification of otherwise neutral/positive samples for certain subjects, indicating a decreased level of model bias.

Infrequent tokens co-occurring with a single entity will have higher PMI scores with the said entity. To focus our analysis on common tokens co-occurring with one entity while maintaining low co-occurrence with other entities, we use the following PMI-based **association metric** α:

$$\alpha(e, w) = (\pi(e, w) - \overline{\pi}(e, w)) \cdot f(e, w) \quad (2)$$

Where $f(e, w)$ is the frequency of w amongst all tokens co-occurring with e and

$$\overline{\pi} = \sum_{e_i \in E \setminus \{e\}} \pi(e_i, w) \quad (3)$$

Intuitively, Equation 2 mitigates the effect of infrequent tokens in the PMI calculation and gives a relative score across all the entities. We calculate α between entities and their co-occurring predicates and objects to identify trends in the contents of the triples.

Results Figure 2 shows the results of the sentiment classification. Overall, positively evaluated instances are rare across all entities, with most being neutral or negatively evaluated. The results of the association analysis are highlighted in Table 2. The most positively (4.7%) and least negatively (37.2%) evaluated religious group are *Jews*, where positive stereotypes include *strong* for Jewish women (α = 5.19). Most (58.1%) instances about Judaism are neutral reports of cultural practices, e.g., about *circumcision* (α = 6.78). Hindus have the smallest proportion of positive stereotypes (2.9%) and Atheists have the largest amount of negative evaluations (51.0%) which often include strong negative actions and emotions such as *attack* (α = 2.04), *angry* (α = 1.37) and *obnoxious* (α = 2.69). Nationalities tend to be more frequently positively evaluated than religious groups, with Germans being the most positively evaluated (9.5%) and the least negatively evaluated (21.0%) with most instances being neutral mentions of the countries role during *ww2* (α = 3.76). Chinese (6.7%) have the lowest proportion of positive stereotypes, however neutral sentiments are most common (63.9%) and are often about topics such as Chinese *food* (α = 2.77). The nationality with the largest proportion of negative stereotypes are the French (49.3%), which are mostly described with negative traits such as *elitist* (α = 5.09) or *vulgar* (α = 5.09), while neutral and positive mentions are often related to food, e.g., *croissants* (α = 5.09).

Since most stereotypical questions asked online have more negative connotations than positive, it confirms the premise that stereotypes can represent prejudicial opinions of different cultural groups.

4.2 Human Evaluation

We perform a **human evaluation** to gain insights into the quality of StereoKG. We focus on three quality metrics, namely *coherence* (COH), *completeness* (COM), and *domain* (DOM) evaluated on a nominal 3-point scale for negation (0), ambiguity (1), and affirmation (2) respectively. COH measures the semantic logicality of a triple, while COM measures if the grammatical valency of the predicate is fulfilled. DOM measures whether the triple belongs to our domain of interest, i.e., whether it can be considered a stereotype or cultural knowledge. We also measure two subjective *credibility* measures CR1 and CR2, where CR1 is a binary measure asking whether the annotator has heard of this stereotype/knowledge before, and CR2 asks whether they believe the information to be true on a scale of 0-4. To evaluate the overall quality of triples, we calculate the success rate (SUC), where a triple is considered successful if it achieves an above average (> 1) rating across all three quality metrics COH, COM, and DOM. The evaluation is performed on a total of 100 unique triples from the KG, where 50 triples each were randomly sampled from the subset of triples stemming from singleton and non-singleton clusters respectively. Each sample was annotated by 3 annotators, all of whom are students with different cultural backgrounds (*German (irreligious), Indian (Hindu), and Iranian (Muslim)*).

We assess **inter-annotator agreement** using the average observed agreement (OA) as calculated using the `NLTK agreement`[10] function, which does not penalize repeated entries of a single value[11] unlike other common metrics (e.g. Krippendorff-α). We observe high levels of agreement for both quality measures COH (0.82) and COM (0.74), while OA for DOM is lower (0.59) due to the subjective nature of what constitutes a *stereotype* (Table

[10] https://www.nltk.org/_modules/nltk/metrics/agreement.html

[11] Repeated entries of a single value are quite common in our annotations, since for most quality measures we use a 3-point or even 2-point scale.

Entity	#Instances	Top Tokens (α)
Atheist	731	*god, christians, annoying, believe, theists, obsessed, attack, vocal, angry, argue, troll, hate*
Christian	823	*obsessed, follow, bible, weird, hate, jesus, abortion, afraid, jewish, covid, non-christians*
Hindu	102	*men, india, hindustan, uc, muslim, caste, tolerant, babas, shameless, fool, jihads,marrying*
Jewish	43	*jew,wear, israel, circumcisions, conversion, discourage, evangelize, progressive, shiksas, leftist*
Muslim	842	*hate, countries, allowed, ex-muslims, obsessed, quran, eat, laws, allah, islamophobia, sharia*
American	1071	*culture, call, obsessed, pronounce, different, countries, afraid, healthcare, hate, british, soccer*
Chinese	277	*restaurants, companies, citizens, food, workers, students, tourists, menus, consumers*
French	138	*eat, speak, obsession, call, egg, pretty, croissants, depicted, proud, culture, exaggerate, elitist*
German	262	*obsessed, pronounce, words, ww2, water, war, nazi, prepare, berlin, love, disciplined, manual*
Indian	431	*culture, obsessed, hate, pakistanis, pictures, marriages, heads, defensive, afraid, stare, army*
Total	4722	

Table 2: Number of instances per entity and predicate/object tokens with highest association score α to entity.

	COH (0-2)	COM (0-2)	DOM (0-2)	CR1 (0-1)	CR2 (0-4)	SUC (%)
SD	1.55	1.11	0.97	0.13	1.17	44.0
CD	1.70	1.42	1.18	0.29	1.56	59.2
All	1.63	1.26	1.07	0.21	1.36	51.5
OA	0.82	0.74	0.59	0.81	0.39	

Table 3: Human annotated COHerence, COMpleteness, DOMain and CRedibility metrics and SUCcess rate over the complete KG test sample (All) as well as its singleton-derived (SD) and cluster-derived (CD) subsamples. Average observed agreement (OA) given for each metric.

Corpus	Train	Dev	Test
OLID	3504/7088	894/1752	242/620
WSF	830/6662	105/965	261/1880

Table 4: Number of *hate/neutral* instances in the train, dev and test set of downstream tasks.

3). Similarly, OA for subjective measures CR{1,2} is mixed, as can be expected. To measure intra-annotator agreement, we duplicated 10 random samples. Intra-annotator agreement is high across all annotators (0.79, 0.95, 1.00).

The COH **quality metric** of the KG is high for both singleton (1.55) and non-singleton-derived entries (1.70), and COM is slightly lower (average COM=1.26). That indicates that the vast majority of entities are meaningful (COH), with some missing relevant information (COM). Overall, DOM is close to 1, suggesting that it was often not clear to annotators whether an entity can be considered a stereotype, which is also reflected in the overall lower inter-annotator agreement on this metric. Entities stemming from non-singleton clusters have a high success rate of 59.2, meaning that the majority of non-singleton-derived entities lean positively across all three quality metrics COH, COM, and DOM. Overall, non-singleton entities are of higher quality than singleton-derived entities (SUC $+15.2$), underlining the initial hypothesis that multiple occurrences of questions online are better indicators of a stereotype than unique questions. Moreover, stereotypical knowledge in non-singleton entities is more likely to be known (CR1 $+0.16$) and believed to be true (CR2 $+0.39$) by annotators.

5 Knowledge Integration

To explore how StereoKG can be used to integrate knowledge into an existing language model, we perform intermediate masked language modeling (MLM) on it in its structured (verbalized triple) and unstructured (sentence) form. The unstructured knowledge is more expressive and verbose, while the structured knowledge from triples is concise and less noisy as compared to the unstructured data. We then fine-tune and evaluate the language model performance on hate speech detection, a task for which we esteem stereotype knowledge to be of use.

5.1 Experimental Setup

Data We experiment with the effect of intermediate pre-training focusing on two kinds of downstream datasets for fine-tuning: one of the same domain as the pre-training corpus (Twitter), and another which is outside the domain data. We use the Twitter-based OLID (Zampieri et al., 2019) dataset as our in-domain dataset and the White Supremacy Forum (WSF) dataset (de Gibert et al., 2018) as our out-of-domain dataset. Both tasks are binary *hate/neutral* classification tasks. As OLID does not have an official validation set, we split off 20% of

samples from the training data for validation. Similarly, WSF is randomly split into 70-10-20% splits for training, validation, and testing respectively.

We manually identify 9 and 33 samples containing a stereotype or cultural knowledge about the subject entities of interest in the dev and test splits of OLID and WSF respectively. To analyze the effect of cultural knowledge integration on these samples exclusively, we use these to create dedicated stereotype test sets. To avoid breaking the exclusivity between validation and testing, we remove the samples found in the validation sets from the original validation splits. During our testing phase, we test the models on the complete test sets as well as the dedicated stereotype test sets. We give the final dataset statistics in Table 4.

Our unstructured knowledge (UK) comprises the original sentences from the clusters from which the triples are formed. Since pre-training requires a sentence format, we create our structured knowledge (SK) by verbalizing the triples from the KG with a T5-based (Raffel et al., 2020) triple-to-text conversion model (details in Appendix C).

Models For the **knowledge integration** experiments, we use the sequence classification pipeline in the `simpletransformers`[12] library. As baselines, we fine-tune two models: general-domain (BASE) RoBERTa[13](Liu et al., 2019) and domain-trained (DT) Twitter RoBERTa[14](Barbieri et al., 2020). Additionally, we continue MLM training of the baseline models before fine-tuning using i) unstructured (+UK) KG knowledge and ii) structured (+SK) verbalized triples to investigate the impact of stereotypical knowledge. All models are fine-tuned with early stopping (δ=0.01, patience=3) using the validation F1 score as the stopping criterion. We fine-tune 10 models for each configuration, each having a different random seed and report their averaged Macro-F1 with standard errors.

5.2 Knowledge vs. Domain

We fine-tune the BASE(+UK/SK) and DT(+UK/SK) RoBERTa models on the in-domain (OLID) and out-of-domain (WSF) training data and report Macro-F1 on the entire test set. To quantify the impact of injecting stereotypes, we also report results on the dedicated stereotype test set. Results on the complete test set and stereotype test set are shown in Table 5 (top) respectively.

Model	OLID (F1)		WSF (F1)	
	Complete	Stereotype	Complete	Stereotype
BASE	69.7±.7	65.1±2.3	60.5±.6	73.3±1.7
BASE+UK	70.6±.4	67.9±2.6	60.7±.5	72.7±1.3
BASE+SK	70.4±.6	66.9±2.0	59.5±1.2	67.5±3.2
DT	70.5±.4	72.5±1.7	60.8±.6	**77.7**±1.6
DT+UK	70.6±.4	73.4±3.4	**61.4**±.4	77.0±2.9
DT+SK	**71.2**±.2	**73.8**±1.8	60.6±.5	75.6±1.8
Our best	71.2	–	91.3*	–
Benchmark	80.0	–	78.0*	–

Table 5: Averaged Macro-F1 and standard errors of BASE and domain trained (DT) models with intermediate MLM training on unstructured (UK) and structured (SK) knowledge tested on OLID and WSF. Top results in **bold**. We compare our best model per test set against its corresponding OLID/WSF benchmark implementation. Values with * are accuracies.

For the **complete test set**, knowledge integration does not seem to have a significant effect, with most model variations being within the error bounds of each other. Only domain training positively affects the classification performance, with all DT models outperforming their BASE counterparts on the OLID dataset with gains of up to F1 +1.5. As expected, domain training does not have an effect on the performance for the out-of-domain WSF data.

While the effect of cultural knowledge integration is not significant on the full test sets, its effect becomes clearer when focusing only on the subset of instances that contain **stereotypes**. Firstly, domain training has a larger effect on these samples, with the DT model showing an increase of F1 +7.4 over BASE on OLID. When the DT model has additionally undergone intermediate MLM training on cultural knowledge, we observe further improvements in F1 for +UK and +SK respectively. While these improvements are within each other's error bounds, this suggests that the training on cultural knowledge can increase downstream task performance on knowledge-crucial samples, i.e., in our case, those that require cultural or stereotypical knowledge. A larger stereotype-containing test set is required to further verify this hypothesis by reducing error bounds. On the out-of-domain WSF data, we do not observe these trends, similar to the BASE model on OLID. This suggests that domain training is a prerequisite for effective knowledge integration.

To set our model results into perspective, we

[12] https://simpletransformers.ai/docs/classification-models/

[13] https://huggingface.co/roberta-base

[14] https://huggingface.co/cardiffnlp/twitter-roberta-base

compare our best models against the **benchmarks** provided by Zampieri et al. (2019) and de Gibert et al. (2018) for OLID and WSF, respectively (Table 5, bottom). On OLID, the benchmark model outperforms our best model by a large margin (F1 +8.8). However, their reported models are single runs without reported standard errors, thus it is unclear whether this specific run is representative for the underlying average model performance. For WSF, our best model outperforms the benchmark by a large margin (Acc +13.3), which is due to the simpler long short-term memory approach that constitutes this benchmark.

5.3 Cultural Knowledge Prediction

To further quantify the degree to which cultural and stereotype knowledge is encoded in the models, we compare their MLM predictions on **masked stereotypes**. We manually collected 100 sentences from the verbalized KG and masked tokens which require either cultural or stereotype knowledge to be completed. By taking into account the top 5 predictions and comparing them to the masked gold standard, we calculate the prediction accuracy at 5 (ACC@5)[15] and analyze common trends.

Our results in Table 6 show that both, the generic BASE and Twitter-based DT models have the same low level of **cultural awareness** (ACC@5=37%), with most predictions being vague e.g, *he, this, that*. However, adding 4,895 unstructured knowledge instances as intermediate MLM training data drastically improves results to 48% (BASE+UK) and 49% (DT+UK). Both +UK models show higher sensitivity to cultural correlations e.g., *Americans* and their struggle with *healthcare*, or *Muslims* and reading the *Quran*, which was not displayed by the baseline models. Further, adjective predictions about minorities tend to be more positive, e.g. *Jewish women are [strong]* →*beautiful*. The structured knowledge also improves cultural sensitivity to a large margin, i.e., +7% points (BASE+SK) and +4% points (DT+SK). However, their predictions are often more generic and less culture-specific than the +UK models, which may be due to the lack of variable context in which these stereotypes are seen due to the denoising factor of using SK.

[15] If the gold standard is present in the top 5 predictions, it is considered accurate.

6 Discussion

We create an automated pipeline to extract cultural and stereotypical knowledge from the internet in the form of queries. While this overcomes the limitations and expenses of crowdsourcing and is easily extendable to a large number of entities, several shortcomings still need to be addressed. Automated extraction results in irrelevant and noisy data, which is augmented by erroneous outputs during triple creation. This is also evidenced in the human evaluation that corroborates the existence of many incomplete triples in the resultant KG, which could also be due to the noisy OpenIE outputs. Other stages in the analysis, such as statement conversion, fast clustering, and triple verbalization give sufficiently good approximations.

Our knowledge integration experiments suggest that performing intermediate MLM training on (verbalized) cultural knowledge can improve the classification performance on knowledge-crucial samples. However, the sample of stereotypical examples in the test/dev sets of both hate speech corpora is low (9 for OLID and 33 for WSF), indicating that a more extensive dedicated hate speech test set focusing on stereotype entities is required to reduce error margins and verify results. Our experiments are limited to intermediate MLM training and we leave the exploration of other knowledge integration techniques for future work.

Our work serves as a preliminary research for studying stereotypes and cultural knowledge across different entities. Extending the KG for other entities than the one proposed in our work is easily done by plugging in new entities into our query templates (Table 1) and the pre-existing pipeline can be used to scrape data, create clusters and finally extract triples without the need of manual intervention. Nevertheless, the current version of StereoKG does not differentiate between (true) cultural knowledge and (untrue or stigmatizing) stereotypes. In reality, making this distinction is a challenge for human experts too, due to the fuzzy boundary between false "stereotypes" and perfectly true cultural "facts" because of the subjective nature of cultural knowledge.

The content used for the construction of StereoKG stems from English-speaking Twitter and Reddit. This comprises a specific demographic which is only a subset of our global society. The stereotypes and cultural knowledge included in StereoKG therefore also underlie this sampling

Model	ACC@5 (%)	Example	Pred (top 3)
BASE	37	*Muslims are turning away [science].*	*too, now, again*
BASE+UK	48	*Americans don't have free [healthcare].*	*healthcare, lunch, tuition*
BASE+SK	45	*Americans are voting for [Trump].*	*freedom, democracy, them*
DT	37	*Atheists unilaterally support [abortion].*	*fascism, abortion, terrorism*
DT+UK	49	*Muslims compare apostasy to [treason]*	*treason, sin, genocide*
DT+SK	41	*Chinese toilets are [dirty].*	*disgusting, awful, shit*

Table 6: Cultural MLM prediction accuracy at 5 (ACC@5) of different models together with example instances with masked [gold standard] token and the top 3 predictions of the model.

bias. Extending the KG to other languages as well as data sources could yield a more global view on stereotypes regarding a specific entity.

7 Conclusion

This study presents StereoKG, a scalable data-driven knowledge graph of 4,722 cultural knowledge and stereotype entries spanning 5 religions and 5 nationalities. We describe our automated KG creation pipeline and evaluate the resulting KG quality through human annotation, showing that the majority of cluster-derived entries in the KG are of high quality (success rate 59.2%) and more common and credible than their singleton counterparts. The KG can easily be extended to include other nationalities as well as genders, sexual orientations, professions, etc., as the underlying subjects. Further, performing intermediate MLM training on verbalized instances of StereoKG greatly improves the models' capabilities to predict culture-related content. This improvement of cultural awareness has a positive effect on knowledge-crucial samples, where we observe a slight improvement in classification performance on a related downstream task, i.e., hate speech detection. Future work should focus on differentiating between cultural facts that should be represented in language models and stigmatizing stereotypes that should not be present in language models.

We make StereoKG and the code of our KG creation pipeline available under `https://github.com/uds-lsv/StereoKG`.

Acknowledgements

We thank our annotators for their keen work as well as the reviewers for their valuable feedback. This study has been partially funded by the DFG (WI 4204/3-1), EU Horizon 2020 project ROXANNE (833635) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Anurag Acharya, Kartik Talamadupula, and Mark A. Finlayson. 2020. An atlas of cultural commonsense for machine reasoning. *CoRR*, abs/2009.05664.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Junia Anacleto, Henry Lieberman, Marie Tsutsumi, Vânia Neris, Aparecido Carvalho, Jose Espinosa, Muriel Godoi, and Silvia Zem-Mascarenhas. 2006. Can common sense uncover cultural differences in computer applications? In *Artificial Intelligence in Theory and Practice*, pages 1–10, Boston, MA. Springer US.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2020. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. *CoRR*, abs/2001.05495.

Paul Baker and Amanda Potts. 2013. Why do white people have thin lips? Google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies*, 10(2):187–204.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Emma E. Buchtel. 2014. Cultural sensitivity or cultural stereotyping? positive and negative effects of a cultural psychology class. *International Journal of Intercultural Relations*, 39:40–52.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emilie Colin, Claire Gardent, Yassine M'rabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, Edinburgh, UK. Association for Computational Linguistics.

Jeff Da, Ronan Le Bras, Ximing Lu, Yejin Choi, and Antoine Bosselut. 2021. Analyzing commonsense emergence in few-shot knowledge models. In *3rd Conference on Automated Knowledge Base Construction*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

John F. Dovidio, Miles Hewstone, Peter Glick, and Victoria M. Esses. 2010. Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In *The SAGE handbook of prejudice, stereotyping and discrimination*, pages 3–28. SAGE Publications Ltd.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

James L. Hilton and William von Hippel. 1996. Stereotypes. *Annual Review of Psychology*, 47(1):237–271. PMID: 15012482.

Curt Hoffman and Nancy Hurst. 1990. Gender stereotypes: Perception or rationalization? *Journal of Personality and Social Psychology*, 58(2):197–208.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4074–4077. AAAI Press.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8):e0237861.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense properties from query logs and question answering forums. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1411–1420, New York, NY, USA. Association for Computing Machinery.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Snefjella.B, Schmidtke.D, and Kuperman.V. 2018. National character stereotypes mirror language use: A study of canadian and american tweets. *PLoS One*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 697–706, New York, NY, USA. Association for Computing Machinery.

Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 523–532, New York, NY, USA. Association for Computing Machinery.

James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2021. From natural language processing to neural databases. *Proc. VLDB Endow.*, 14(6):1033–1039.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ruiqing Yan, Lanchang Sun, Fang Wang, and Xiaoming Zhang. 2021. K-xlnet: A general method for combining explicit knowledge with language model pretraining. *CoRR*, abs/2104.10649.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Ethics Statement

Human Evaluation We perform a human evaluation using human raters. After making an internal call for participation that included a task description and the amount of compensation, we selected participants based on their timely response to our call. The chosen raters were compensated fairly.

Modeling Stereotypes Stereotypes are fundamentally cognitive schemas that help the perceiver process the dynamics of different groups. They are made up of a collection of traits that are ascribed to a given social group (Dovidio et al., 2010). If made conscious, they can aid in improving cultural sensitivity (Buchtel, 2014). However, in most cases, these are unconscious beliefs and can then lead to bias and discrimination (Hoffman and Hurst, 1990). Human-written content reflects these cognitive biases, and when natural language processing (NLP) models are trained on this biased data, they can further propagate stereotypes and discrimination (Hovy and Spruit, 2016). Mitigating bias in NLP has thus become a major research direction. These works often require structured knowledge or lists about biased terms, e.g., Bolukbasi et al. (2016) rely on a list of male-female minimal pairs. Our work's contribution is to automatize this process by exploiting social media users' beliefs about social groups, i.e., we collect assertions and questions about social groups which appear often in both Reddit and Twitter data. In this sense, our approach can be described as a similar process that occurs in humans as they become aware of their own mental processes, including stereotypes (Buchtel, 2014). If we are aware of stereotypes, we can use them to improve cultural sensitivity and mitigate the effects of bias and discrimination.

StereoKG could be used to generate stereotypical content (e.g., through verbalization). While verbalized stereotypes can improve the downstream task performance on knowledge crucial samples (Section 5), they could, however, also be **misused** in a hurtful manner, e.g., by using stereotypical knowledge in question-answering systems. However, this is a general issue pertaining to language models which we are trying to mitigate through our work: if trained on bias(ed) data, they could be misused to generate harmful content.

Environmental Impact Our models are trained on Titan X GPUs with 12GB RAM. In order to economize the energy use, we did not perform any extensive hyperparameter exploration.

Entity	Subject-specific	Generic
Atheist	*r/TrueAtheism*, *r/religion*, *r/DebateReligion*, *r/atheism*	*r/explainlikeimfive*, *r/AskReddit*, *r/TooAfraidToAsk*, *r/NoStupidQuestions*
Christian	*r/religion*, *r/DebateReligion*, *r/TrueChristian*, *r/DebateAChristian*, *r/AskAChristian*, *r/atheism*, *r/Christianity*, *r/Christian*, *r/Christianmarriage*, *r/Bible*	*r/AskReddit*, *r/NoStupidQuestions*, *r/explainlikeimfive*
Hindu	*r/India*, *r/hindusim*, *r/librandu*, *r/IndiaSpeaks*, *r/awakened*, *r/IAmA*, *r/atheismindia*, *r/india*, *r/AskHistorians*	*r/explainlikeimfive*, *r/AskReddit*, *r/TooAfraidToAsk*, *r/NoStupidQuestions*
Jewish	*r/Judaism*, *r/AskHistorians*, *r/religion*, *r/DebateReligion*, *r/AskSocialScience*	*r/explainlikeimfive*, *r/AskReddit*, *r/TooAfraidToAsk*, *r/NoStupidQuestions*, *r/Discussion*
Muslim	*r/religion*, *r/DebateReligion*, *r/TraditionalMuslims*, *r/progressive_islam*, *r/atheism*, *r/islam*, *r/exmuslim*, *r/Hijabis*, *r/indianmuslims*, *r/AskSocialScience*	*r/AskReddit*, *r/NoStupidQuestions*, *r/explainlikeimfive*, *r/ask*
American	*r/AskAnAmerican*	*r/explainlikeimfive*, *r/OutOfTheLoop*, *r/TooAfraidToAsk*, *r/offmychest*, *r/NoStupidQuestions*, *r/linguistics*, *r/AskReddit*
Chinese	*r/shanghai*, *r/China*, *r/asianamerican*, *r/HongKong*, *r/Sino*	*r/explainlikeimfive*, *r/AskReddit*, *r/TooAfraidToAsk*, *r/NoStupidQuestions*
French	*r/French*, *r/france*, *r/AskAFrench*, *r/AskEurope*	*r/explainlikeimfive*, *r/AskReddit*, *r/NoStupidQuestions*
German	*r/germany*, *r/German*, *r/europe*, *r/AskGermany*, *r/AskAGerman*	*r/explainlikeimfive*, *r/AskReddit*, *r/offmychest*, *r/TooAfraidToAsk*, *r/NoStupidQuestions*
Indian	*r/India*, *r/india*, *r/indiadiscussion*, *r/IndianFood*, *r/indianpeoplefacebook*, *r/ABCDesis*	*r/explainlikeimfive*, *r/retailhell*,*r/AskReddit*, *r/TooAfraidToAsk*, *r/NoStupidQuestions*

Table 7: Section 3 - Subreddits for Reddit extraction

B List of Subreddits

We gather data from several subject-specific and generic subreddits as listed in Table 7.

C Triple Verbalization

The triple verbalization technique takes inspiration from KELM (Agarwal et al., 2021). We use the WebNLG 2020 (Colin et al., 2016) corpus to fine-tune a T5-base[16] model for 5 epochs and then apply it to triples in StereoKG. It results in a corpus of verbalized triples in sentence form:

> *<jewish men, get, circumcisions> → "Jewish men get circumcisions."*
> *<american culture, obsessed with, novelty> → "The American culture is obsessed with novelty."*

These sentences constitute the structured knowledge (SK) and are used for intermediate MLM pre-training of the baseline models.

[16] https://huggingface.co/t5-base