



HUGO Phase 1 Final Report

Milestone 4

Created: August 29, 2022

DARPA agreement number: HR00112290023

DARPA Civil Sanctuary AIE

Awarded: November 29, 2021

Contractor Name: Smart Information Flow Technologies, DBA SIFT

319 N. 1st Ave., Suite 400 Minneapolis, MN 55401

Technical Contact:

Ms. Sonja Schmer-Galunder
sgalunder@sift.net
(415) 604-6293

Administrative Contact:

Ms. Linda Holje
contracts@sift.net
(612) 226-5061

Subcontract Raytheon BBN:

Dr. Brian Ulicny
Brain.ulicny@ratheon.com
(339) 293-2439

ATTN:

Program Manager: Dr. William Corvey (william.corvey@darpa.mil)

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

Contents

1	Executive Summary	1
2	Initial Design and Evaluation Plan	3
2.1	Summary	3
2.2	HUGO Design	3
2.2.1	Architecture Overview	4
2.2.2	Utterance Classifier	5
2.2.3	SpEAR (Span Based Entity, Attribute, and Relations) Extractor	5
2.2.4	Community Detection	6
2.3	Evaluation Plan	8
2.3.1	HUGO Overall Development	9
2.3.2	HUGO Evaluation Plan	9
2.3.3	Utterance Classifier Evaluation	9
2.3.4	SpEAR Evaluation Plan	10
2.3.5	Community Detection Evaluation Plan	10
2.3.6	Response Generation Evaluation Plan	11
3	Training and Evaluation Materials	12
3.1	Summary	12
3.2	Training in Native Languages	12
3.3	Description of the OMC Dataset	12
3.4	Description of the World Values Survey Data	13
3.5	Training and Evaluation with the OMC Data	14
3.5.1	Utterance Classifier Subcomponents	15
3.5.2	CELEBRATE annotation interface and HUGO schema	15
3.6	Preliminary Community Analysis	18
4	Initial Prototype and Evaluation	22
4.1	Milestone 3 Prototype Architecture and Components	22
4.1.1	NLP Pro-/Anti- Sociality Detectors	23
4.1.2	Data Fusion to Construct User/Forum Models	26
4.1.3	Response Generation	27
4.1.4	Post Alerts and Response Triggering	27
4.2	Technical Contributions Summary	28

5	Revised Prototype for Phase 1 Demo	30
5.1	Overview of Updates	30
5.2	Updated Directed Social Regard Detector	30
5.2.1	New ABSA/DSR Annotation	30
5.2.2	ABSA Evaluation	31
5.3	Updated HUGO Architecture	31
5.3.1	Moderate Touch Interactive Moderation	33
5.3.2	Dimensions of deliberativeness in Online Communities	34
5.3.3	Response strategies	36
5.3.4	Rationality	39
5.4	Reddit Forum	39
6	Next Steps: Human Subjects Research	42
6.1	Experimental Goals	42
6.2	Platform Description	42
6.3	Variables to Evaluate	44
6.3.1	Monitoring Social Behavior in Conversations	44
6.3.2	When to Moderate	45
6.3.3	Perception of Moderation and Experience	45
6.4	Planned Studies	46
	Appendices	52
A	Abbreviations	53
B	Der Standard Community Guidelines in German and English	54
B.1	German:	54
B.2	English (from Google Translate)	55

List of Figures

2.1	The HUGO architecture	4
2.2	The SpEAR architecture	7
2.3	The HUGO Schedule and Milestones	9
3.1	Inter-rater agreement on the categories annotating the OMC dataset.[56]	14
3.2	The HUGO schema used to annotate pro-social behavior.	16
3.3	The Celebrate interface HUGO used to annotate statements.	17
3.4	Number of responses per Post. Most Posts receive only a small number of responses.	18
3.5	Frequency of Responder-Respondee pairs.	19
3.6	Number of responses by article topic.	20
4.1	HUGO 0.1 Architecture Overview	22
4.2	Sociality Detectors	23
4.3	Positive politeness strategies flow chart from Brown and Levinson	24
4.4	Negative politeness strategies flow chart from Brown and Levinson	25
4.5	Fusion Rules per post	26
4.6	Modeling a Forum from Fused Data	27
4.7	Post Prompt Generation	28
4.8	HUGO moderation dashboard	29
5.1	Regard scores example	31
5.2	ABSA Evaluation Results	32
5.3	HUGO v0.2 Architecture Overview (updated modules in green, v0.1 content in blue)	33
5.4	Rule Detectors.	36
5.5	Example Argumentation diagram	39
5.6	Example prosocial post expressing sympathy to a harmed group, and HUGO response	40
5.7	Example toxic post and response	40
5.8	Example negative post and response	40
5.9	Example post showing an active listening response	41
6.1	Tree-like representation of user-generated messages on Reddit	43

List of Tables

2.1	Description of OMC annotation categories	6
2.2	SpEAR performance for extracting anti-social markers	8
3.1	HUGO anti- and prosocial attributes schema with examples	17
3.2	Topic paths in German vs. English	21
4.1	Label usage across annotators (A1-3) and agreement (Krippendorff's α)	26
5.1	Dimensional Annotation Results	31
5.2	ABSA Model Evaluation	32
5.3	Taxonomy of interactive journalistic moderation [68]	34
5.4	Mapping rules to Civility	35
5.5	Example posts falling into Public/Private Civil/Uncivil Categories	38
5.6	Response Strategies for Public/Private Civil/Uncivil Posts	38

Chapter 1

Executive Summary

This report describes the results of the Humanitarian User Group Oversight (HUGO) program executed by the SIFT and Raytheon BBN Technologies team. The program analyzes social interactions in online forums to provide more effective moderation and mediation. HUGO uses Natural Language Processing (NLP) approaches to extract markers from text informed by social interaction theory. The HUGO approach extracts multiple theory based metrics from social media, such as politeness, social regard, private and public civility and toxicity. The resulting platform allows us to explore the most effective combination of metrics for detecting both anti-social and pro-social behavior. Our Phase 1 effort has focused on detecting and promoting pro-social behavior as an effective means for moving social discourse forward in a positive manner to address the needs of an online forum addressing a humanitarian disaster in another country.

To promote pro-social behavior our final Phase 1 demonstration also includes a theory based response component. Determining an effective response to encourage pro-social behavior, curb anti-social behavior and not cause moderator fatigue requires detecting a number of situations and responding to that specific situation. We again construct a number of theory based responses depending on the condition and we use GPT-3 to construct a targeted response. The GPT-3 program allows us to concentrate on the tactical response as opposed to the challenges of response construction. The GPT-3 allows us to add additional constraints on its response to create targeted responses. Our tool chain provides multi-language support in both analysis and response.

Our Phase 1 effort has also positioned us to conduct human subject research to directly learn the effectiveness of the HUGO system as a mediator. We created a series of explanatory interfaces to provide details on both the analysis and the response. To support future experiments we created a private reddit group that sends each post to HUGO for analysis and posts any response from HUGO.

Civil Sanctuary solutions must support foreign language to address international humanitarian crises. Our architecture takes advantage of the latest multi-language support tools, both in analysis and in response to support a wide variety of languages. Phase 1 focused on German. Specifically, the online social forums for the Austrian newspaper *Der Standard*. Throughout this report you will see examples in English, but the examples also work in German, and have been checked for accuracy by our native German speaker.

HUGO's key Phase 1 innovations include:

- A reasoning engine that combines multiple metrics extracted from text by both our tools and existing tools
- Theory based NLP tools that detect Moral disengagement
- New concepts of private and public civility that allow us to more accurately describe the space of posts, and hence create more targeted responses to each situation
- A debugging environment that allows us to track analysis and responses back to their individual metrics to adjust the system when it responds incorrectly to a novel situation
- Integration with GPT-3 to produce natural language mediation responses in the target language

- An analysis of common forum rules spanning German newspapers, a critical step in enforcing rules that ensure pro-social behavior. These guidelines will help HUGO ensure broad coverage across the landscape of changing guidelines.
- A sensor for characterizing the mood in the forum. We expect different rules to apply to individual posts, depending on the current state of the discussion in the forum. For example, HUGO will respond more strongly to additional polarizing language if the forum has become polarized.
- We have looked at how Austrian slander and defamation laws can be incorporated into our response.
- We analyzed a specific use case scenario based on a recent event in Austria. A medical doctor, Dr. Lisa-Maria Kellermayer, was a target of a hate campaign of anti-vaxxers over several months - leading to eventual suicide. The comments on the Der Standard articles have provided a rich example of polarizing examples and as a use case for understanding how HUGO may impact anti and pro-social interactions.
- We have built an initial prototype of an argumentation engine capable of automatically extracting structured arguments from text and evaluating them. This will allow us to detect reasoned arguments in support of opinions, which is key for promoting pro-social behavior working toward a goal.
- Under the HUGO program we analyzed the results of another program's experiment to investigate the impact the moderator has if they are a human or a bot. The study is preliminary but will provide an initial assessment of our HUGO approach. This study also investigated the use of symbols ("forbidden" signs vs. "heart" signs) in online forums showing the users prefer engagement and heart symbols overall, implying heart symbols or "likes" can provide a more positive environment to promote pro-social interaction.
- We developed techniques for detecting overheated dialogue threads (as opposed to individual posts). We developed ways to respond to overheated threads as a group, rather than in response to posts individually.
- We refined a generic model of prompts to be fed into GPT-3. This allows HUGO to concentrate on the theme of the response by creating the correct prompt that gets fed to GPT-3.
- We have been invited to present a 30-minute overview of our project at the NLP Summit 2022 on Applied Natural Language Processing in October. Pending approvals.
- Private versus Public civility. We created this refinement to better characterize the nature of posts and construct better responses. Public civility refers to the display of democratic and pro-social principles. For example, taking the vote away from a sub-group may be stated with no hostile language, but has a tremendous impact on reaching a consensus solution. We have mapped a grounded set of NLP detectors to theoretical frameworks of public and private civility. In addition we have developed a mapping between forum guidelines (as published in various online forums, with focus on German) and these notions of civility.
- We have created an experiment platform integrated with Reddit for doing future human subject research that works directly with the Reddit API.
- We are in ongoing discussions with a group at the Civil Health Project, discussing additional funding and collaboration avenues for pro-social detection and engagement.

The HUGO program has developed a rich research platform that not only promotes pro-social behavior, but that will allow us to answer complex questions about automated forum mediation for responding to humanitarian crises around the world.

Chapter 2

Initial Design and Evaluation Plan

2.1 Summary

This technical report contains SIFT and Raytheon BBN Technologies’ initial design for HUGO (Humanitarian User Group Oversight). HUGO is designed to extract norm-sensitive and effective mediation strategies and responses from limited observations to direct civil discourse. HUGO will also extract cultural knowledge based on our theory-based approach to break culture down into observable actions detectable with our Natural Language Processing (NLP) tools. HUGO combines SIFT’s expertise in NLP via transformer-based and deep learning [23] with our plan recognition [27, 26] to detect harmful intentions in online discourse and expand it to 1) learn pro- and anti-social behavior embedded in mediation strategies and forum policies from small training sets and 2) elicit culturally appropriate pro-social engagement and cooperation via Raytheon BBN Technologies’ chatbot mediation technology [6].

While cultural specific adaptations have not been fully implemented in Phase 1, our plan is for HUGO to extract to local populations’ values and norms from several cultural models [30, 31, 35, 44, 37] to learn facets of a culture tied to moral identity. An individual’s moral identity influences moral conduct (perceptions of right and wrong) in offline and online environments. For example, moral identity rooted in individualism of Western cultures derives from values related to self-expression [34], while people in collective cultures define themselves in the context of social relationships, the demands of others, and need to maintain harmony within one’s group. Thus, some cultures follow social restrictions and appeals to group responsibility while other cultures respond to laws and rules [45]. Similarly, moderation strategies focused on sanctioning and prohibiting anti-social behavior may lead to very different outcomes in hierarchical cultures compared to cultures that consider pro-social behavior more a matter of personal choice [32]. Integrating cultural-specific models and NLP with symbolic planning, HUGO will give operators control over transparency, when/which explanations for actions are displayed or if new rules should be added.

2.2 HUGO Design

HUGO’s aim is to promote pro-social behavior and mitigate the effects of anti-social behavior online in non-English speaking social media platforms. HUGO will predict culture-specific pro- and anti- social speech from anthropological theories of morality and culture (see also HSR Plan). HUGO will apply culture theoretic notions to drive the form and content of responses.

For phase 1, we chose German as the foreign language for HUGO to operate in. We chose to analyze a German dataset, and train our NLP models directly in German, rather than translating German to English and then building a model. We believe maintaining the native language will preserve the cultural nuances and linguistic markers for pro- and anti-social behaviors that may otherwise get lost in translation. However, we have also augmented this training with English data and prompts for generation.

Our primary data source for this phase was the social discourse resulting from the One Million Corpus (OMC) from the German newspaper Der Standard [56]. This is a forum where readers can post reactions to articles and frequently debate each other on various topics. The newspaper editors provide mediation for anti-social posters.

2.2.1 Architecture Overview

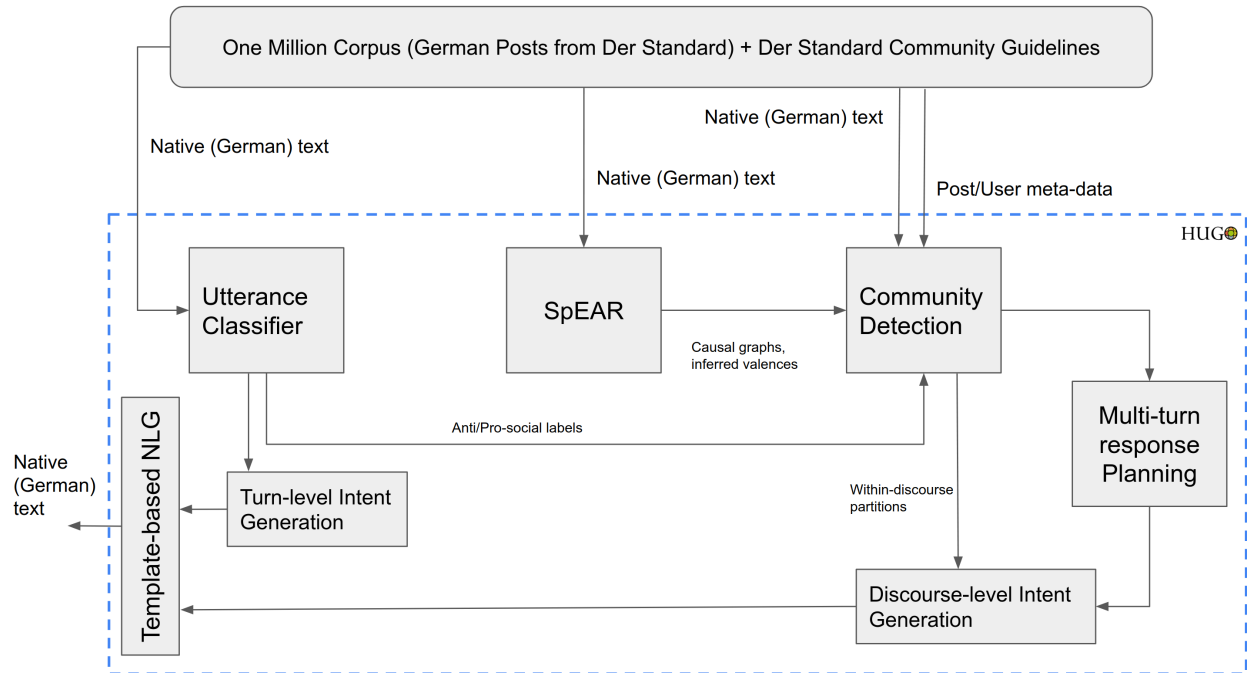


Figure 2.1: The HUGO architecture

Figure 2.1 shows a preliminary HUGO architecture. The dotted rectangle shows the collection of components in HUGO and their interactions. We envision the system to read and process posts from OMC at two different levels: (1) turn-level, and (2) multi-turn/discourse level:

- **Item-level:** Some posts require an immediate response, independent of any context. For example, slurs or blatant violation of community standards contained in a single post.
- **Turn-level:** Some behavior only becomes detectable within the context of the larger conversation. This level of abstraction requires understanding conversation directions and subgroup contexts.

The main function of the HUGO Utterance Classifier component is to identify patently anti-social and toxic posts and enable quick and reactive response intent generation at the turn-level. This classifier can also identify prosocial behavior and reply with comments to bolster that response. Planned actions based on the Utterance Classifier example actions include praising or admonishing the user, along with flagging posts for removal or blocking the user.

To reason at the discourse-level, HUGO requires several components. Language features identified by the Utterance Classifier combine with causal graphs inferred by our Span-based Entity, Attribute and Relation Extractor (SpEAR) component and analyzed by the Community Detection component.

The Community Detection component measures the polarization of discourse. This component tracks how communities and sub-groups form around different topics. It also tracks if postings match the current topic or are off-topic.

A divisive discourse coupled with anti-social commentary could suggest the existence of an unhealthy debate requiring moderation. For example, while individual utterances might not rise to a level of toxicity requiring turn-level action, their collective existence combined with a developing polarization between forming sub-groups might

require a discourse-level action. For example, one possible mediation strategy would be reminding the participants of commonalities with each other on other topics, appealing to moral foundations.

In certain cases, moderation may require multiple responses. For example, it might be more effective to ease into an admonishment of a detected anti-social behavior by hedging or appealing to another authority. Politeness theory [20] offers a formal account of redress mechanisms, which are behaviors aimed at reducing the perceived threat of an imperative. Indirect speech acts (e.g., saying "can you please pass the salt" instead of "pass the salt") provides one mechanism of redress. Other redress mechanisms include establishing common ground between the conversation participants.

The HUGO planning component will plan sequences of communicative intent by planning sequences of sub-intents and redress mechanisms to accomplish communicative goals and encourage forum rule compliance. Longer range planning for multi-turn based responses will be implemented in phase 2.

The following sections describe each component in more detail.

2.2.2 Utterance Classifier

The original plan for the HUGO Utterance Classifier was to combine the results of two sub-components: (1) a toxicity detector and (2) an OMC classifier. The toxicity detector, is an existing pretrained German language toxicity classifier based on the Perspective API [3]. Created by the Alphabet company, the Perspective API analyzes German text directly. For each text input, it provides a numeric score between 0 and 1 for each of eight toxicity categories: toxicity, severe toxicity, insult, sexually explicit, profanity, likely to reject, threat and identity attack.

The OMC dataset organizes the posts into seven (anti-social as well as pro-social) categories, see Table 2.1. The Austrian Institute of AI (OFAI) has already developed several learning algorithms for the OMC dataset including: Bag of Words representation with an SVM, Multinomial Naive Bayes, bag of cluster ID, doc2vec, and LSTM-based architecture [56]. While these serve as good baselines, these algorithms do not capture contextual embeddings that are included in more modern systems like BERT [16], so our initial plan was to fine-tune a German BERT model called GottBERT [57] on the OMC data to improve performance. However, we have since decided to deprioritize this, as most of what we want to extract is sufficiently covered by a combination of the toxicity API and our SpEAR model. We may revisit this in phase 2.

To evaluate our Utterance Classifier, we planned to evaluate the toxicity detector by first mapping the toxicity categories offered in the Perspective API to the OMC corpus categories. Since the Perspective API does not have pro-social labels, we could only evaluate the performance on anti-social categories. Once we have a mapping, we can evaluate it against the OFAI baselines mentioned above using the traditional ML metrics: precision, recall and F1 scores. Our plan was to evaluate the OMC classifier directly against the published baselines using the same metrics of precision, recall and F1 scores.

Our initial plan was to address pro-social detection in the Utterance Classifier by relying on our OMC classifier subcomponent as it will be trained on the categories in the OMC dataset, several of which include pro-social labels. Our HSR plan contains culturally sensitive subcomponents of pro-social behavior. In Phase 1, we will match those pro-social behaviors as well as anti-social behaviors to platforms' community guidelines (see Appendix B for community guidelines of Der Standard in both their original German and translated to English).

2.2.3 SpEAR (Span Based Entity, Attribute, and Relations) Extractor

Figure 2.2 shows the SpEAR transformer architecture for extracting knowledge graphs from unstructured text. In past work, we have developed a knowledge graph schema and trained SpEAR models to generate graphs representing causal structure in text [23, 24]. HUGO builds on our existing schema for anti-social behavior capturing notions of dehumanizing speech among others, but also detecting pro-social behavior. Table 2.2 shows the performance of our English-SpEAR model for extracting anti-social attributes.

Table 2.1: Description of OMC annotation categories

Category	Values	Description
Sentiment	negative/neutral/positive	An important goal is to detect changes in the prevalent sentiment in a discussion, e.g., the location within the fora and the point in time where a turn from positive/neutral sentiment to negative sentiment takes place.
Off Topic	yes/no	Posts which digress too far from the topic of the corresponding article.
Inappropriate	yes/no	Swearwords, suggestive and obscene language, insults, threats etc.
Discriminating	yes/no	Racist, sexist, misogynistic, homophobic, antisemitic and other misanthropic content.
Feedback	yes/no	Sometimes users ask questions or give feedback to the author of the article or the newspaper in general, which may require a reply/reaction.
Personal Stories	yes/no	In certain fora, users are encouraged to share their personal stories, experiences, anecdotes etc. regarding the respective topic.
Arguments used	yes/no	It is desirable for users to back their statements with rational argumentation, reasoning and sources.

HUGO leverages and improve upon this technology by replacing BERT with a multi-lingual BERT, with our original aim to train directly in German with annotations provided by our in-house social anthropologist and native German speaker, for both pro and anti-social markers. This was later augmented by additional annotators annotating English data.

We will evaluate our SpEAR model against a hold-out dataset by measuring precision, recall and F1 scores for each of entities, attributes and relations.

2.2.4 Community Detection

The Community Detection component integrates extracted semantics and knowledge from the Utterance Classifier, SpEAR and meta data about the user, and posts to identify group formation and polarization within a discourse. We define a discourse as a series of posts associated with a single topic. The Community Detection component identifies whether a collection of posts (and related user and article/topic information) represents users divided along the topic. Healthy debates naturally involve divided parties and newspapers like the Der Standard rely on an engaged user base presenting different perspectives and arguments to the debate. However, such a business model can be negatively impacted by unhealthy divisions. For example, divisive topics or anti-social commentary, even those below detectable level of toxicity, can fuel partisan division and reduce participation in discussions, and eventually readership. Viewing user comments in the context of how they influence these group divisions provides critical insight into effective mediation to meet the business goals.

In HUGO, our plan was to initially formulate the community detection task as a clustering task aimed at dividing users commenting on an article into groups. In order to do this, we needed to identify useful and relevant feature sets that allow us to categorize divisions if they exist. We approached this task iteratively, by first developing a baseline approach using upvote and downvote information made available in the OMC dataset for each post. For example, posts receiving many more downvotes than upvotes or vice versa are likely to signal the existence of disagreeing groups. Our aim is to extend our understanding of polarization through the inclusion of additional features such as user interactions, natural language content and features from the Utterance Classifier and SpEAR. Such an approach requires careful

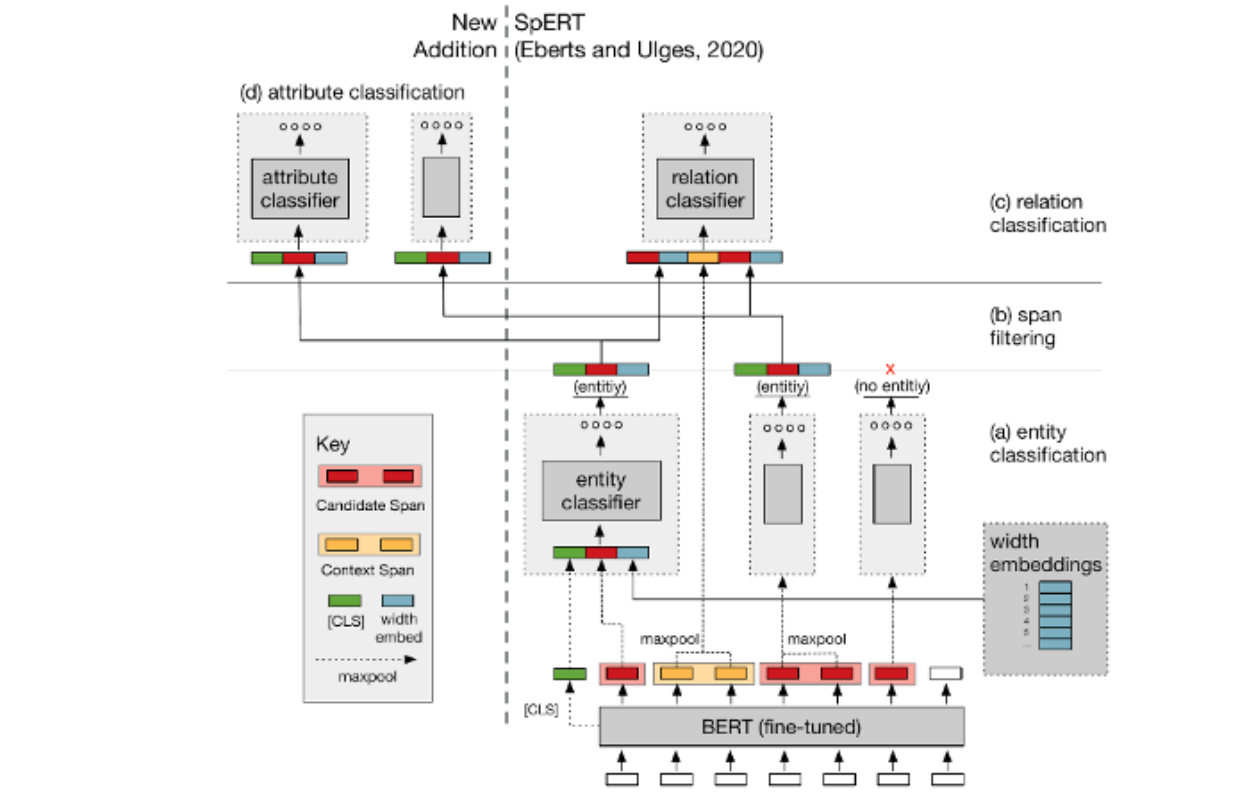


Figure 2.2: The SpEAR architecture

feature selection, which must be confirmed with data analysis. We also planned to explore graph neural network-based approaches to community detection and clustering. The aim of this is to allow us to perform explorative analysis on the posts to identify connections between existing cultural theories and linguistic and social-network markers.

In HUGO, we also planned to look at representing the discourse as an interaction graph composed of nodes representing users (and their posts) and edges representing relationships between them. Some of the node attributes and edge labels will be extracted directly from the data (e.g., user’s authorship of a post). Others may be extracted indirectly through SpEAR or the Utterance Classifier. One initial model to explore is representing users as nodes and tie their connectivity through sub-threads in which one user comments on another’s post. Through iterative development on the Community Detection component, our aim was to enrich the node (user) features to incorporate semantic content of the posts authored by the corresponding user. A graph-based machine learning model will assist with node classification

Response Generation

Figure 2.1 shows several components related to generating responses. The goal is for HUGO to generate culturally-sensitive responses for both pro- and anti-social behaviors.

At the utterance level, HUGO our initial aim was to flag/remove posts/users for those posts deemed to be above a prescribed toxicity threshold based on the determinations made by the toxicity detector. If the behaviors do not rise to the level of removal/flagging, but still above socially acceptable levels, HUGO can initiate an appeal to the user’s moral dimensions (e.g., Care, Authority and Purity) [30], depending on the cultural preferences. For example, in cultures in which an appeal to the “Care” dimension is more effective, HUGO can generate a response of the form “You shouldn’t use language like that. You need to remember that other users are real people, with real feelings”. In other cultures, in

Table 2.2: SpEAR performance for extracting anti-social markers in text in English text from SIFT’s prior work [23]

	Dimension	P	R	F1	Support
Entity	character	84.11	84.75	84.36	1916
	event	72.42	58.69	64.29	494
	qualifier	67.00	42.23	51.30	345
	Micro-Averaged	80.90	74.54	77.51	
Attribute	violent	68.02	62.79	64.30	191
	condemned	56.45	66.23	60.75	1141
	justified	37.12	26.95	28.57	253
	dehumanized	51.30	51.29	50.12	295
	responsible	64.41	50.66	56.10	347
	harmed	61.64	54.26	56.81	462
	Micro-Averaged	57.01	56.47	56.58	
Relation	cause	45.26	44.31	44.13	670
	impact	50.00	48.96	48.73	747
	modifier	44.67	32.39	36.78	397
	Micro-Averaged	47.02	42.90	44.43	

which an appeal to the “Authority” dimension is deemed more effective, HUGO’s response might be “You shouldn’t use language like that. You need to behave according to the proper rules of political civility.” In cultures that favor “Purity” morals, HUGO might respond with “Please remember that the people you are responding to also have a human body and soul like you.”

At the turn level described above, HUGO primarily generates communicative intents of “Admonishment” and “Reinforcement”. However, at the discourse level, HUGO will need to expand its range of communicative intents to include relating group membership information to highlight commonality between participants, or information pertaining to the user’s social and group identity.

2.3 Evaluation Plan

SIFT and Raytheon BBN Technologies have set challenging goals for the HUGO program, including:

- Processing community rules as they change – as an input to other HUGO components
- Providing online forum moderation – to integrate with emerging technology to curtail anti-social behavior.
- Provide online mediation – to encourage positive behavior and group cohesion.

In Phase 1 our focus was on evaluating the accuracy of the components in the HUGO architecture. In Phase 2, with a human subject experiment we will validate overall high-level goals of HUGO described above. We have submitted our initial Phase 2 Human Subject experiment for IRB approval at this time to allow for sufficient time to obtain

approval, and will be submitting an updated HSR plan based on the ideas detailed in Chapter 6 shortly.

Below we discuss our specific evaluation criteria for each of the main HUGO components in Phase 1.

2.3.1 HUGO Overall Development

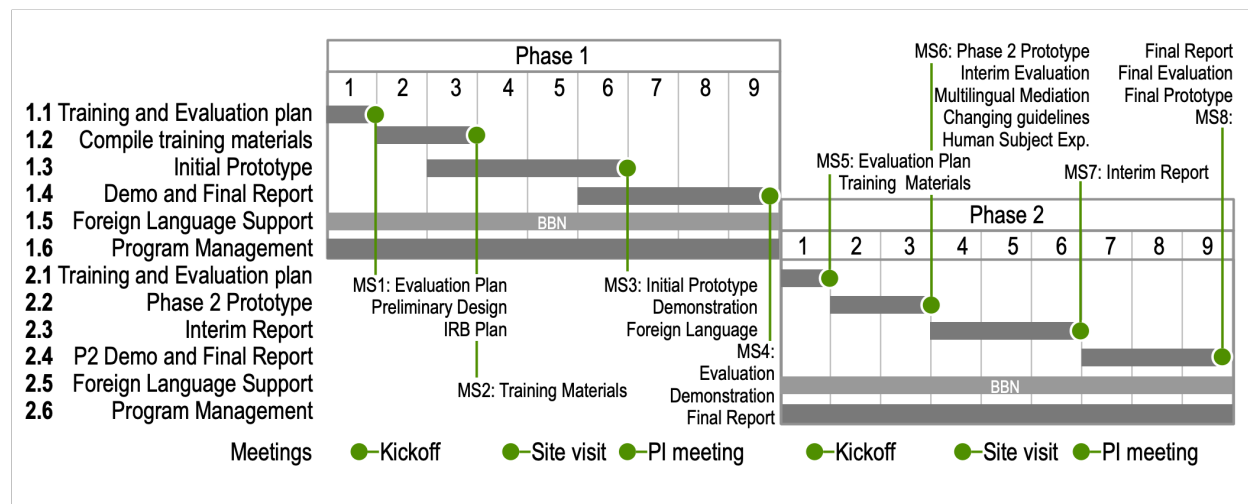


Figure 2.3: The HUGO Schedule and Milestones

2.3.2 HUGO Evaluation Plan

Our overall approach centered on creating a prototype, evaluating the prototype against defined metrics, and then refining the prototype and repeating the evaluation process. We have defined our major milestones in our proposal schedule shown in Table 3.

HUGO's three main prototype features include:

1. Prototype for Milestone 3, Task 1.3),
2. Refined Prototype (end of Phase 1), and
3. Final Prototype (Phase 2)

2.3.3 Utterance Classifier Evaluation

For the prototype, we planned to develop the initial category mappings between the Perspective API's 8-point scale and OMC's categories. Our plan was to evaluate the toxicity detector against the OMC as mentioned above. We also initially planned to fine-tune a German-native transformer-based language model (e.g., GottBERT) and evaluate its performance against published baselines, however this was dropped from Phase 1.

For the end of Phase 1, we refined the prototype. Our initial plan was to explore whether false positives and false negatives generated by the toxicity detector are indicative of cultural nuance not accounted for by the Perspective API. We can do this by leveraging our in-house German-speaking and culturally-aware expertise to compare those cases of false positives and false negatives and the specific language used. We also intend exploring alternative mappings between the Perspective API categories and the OMC categories. However, we have not done this yet and may look into this in Phase 2.

Our initial plans for Phase 2 were to look to expanding the architecture to incorporate politeness classifiers and also look into expanding annotations for additional pro-social behaviors by automatically labeling certain predefined politeness markers or using other components in HUGO (e.g., SpEAR). However, we instead ended up implementing

these in Phase 1, as will be further detailed in Chapter 4. We will also extend HUGO’s capabilities to address changing forum guidelines. The OMC’s mapping from community guidelines to labelling is not one-to-one, so we will need to determine a suitable evaluation metric. While we have not dealt with changing guidelines yet, we have implemented an initial approach to addressing a generalized form of the OMC guidelines, as is detailed in Chapter 5. In Phase 2, we expect to follow the development scheme of Phase 1, but for Hindi, our second non-English language.

2.3.4 SpEAR Evaluation Plan

Once again, our plan was to improve the SpEAR component in multiple rounds of prototyping. Notably, our development approach was to largely be guided not only by the development of architectural aspects, but also by the development of an appropriate knowledge graph schema, which in turn serves as a guideline for annotation. A highly complex schema (one involving many entities, attributes and relationships), which is rich in semantics, is more difficult not only to train, but also to annotate. On the other hand, simpler schemas are more amenable to quick annotation, but limited in their expressivity. Moreover, changes in annotations might necessitate retraining on new annotations. Weighing the pros and cons, we have decided to incrementally and iteratively develop our annotations rather than prescribing a complex annotation upfront. This approach allowed us the freedom to quickly iterate (and retrain if necessary) as well as to develop a schema with the appropriate balance between expressivity and training difficulty.

For the prototype, our plan was to integrate multi-lingual BERT into the SpEAR architecture, design a minimal knowledge graph schema, and train a model on this minimal schema. To limit annotation and training burden, we planned to only label entities and attributes. Past experience with SpEAR has shown us that learning entities and attributes is more reliable and faster than learning relations. Accordingly, for the prototype, we did not prioritize learning relations. While relations may be necessary for later extracting causality as well as downstream inferences about a user’s preference towards other users or topics, the attributes and entities provide us much useful information. For example, it is still useful to know if a user has dehumanized a group or another user in a post. To recognize this, we only need to recognize the corresponding attributes in the post. Our plan was to first generate about 50 annotated examples and iterate until we reached about 500 annotated examples for the prototype. Along the way we expected to have iteratively crafted a schema of suitable complexity by Milestone 2, as detailed in Chapter 3 section 3.5.2. We further revised the annotation schema for the final phase 1 prototype, as described in 5.2.

For the end of Phase 1, our plan was to refine the prototype by continuing to annotate posts, and also begin to develop baseline inference capabilities to enable us to perform valence (positive and negative) determinations.

In Phase 2, we will explore whether the Utterance Classifier will improve performance of SpEAR and vice-versa, for example, by providing some additional annotations. We will also use existing off-the-shelf NLP tools (SRL, dependency parsing) to assist with automatic labeling and thereby ease the annotation burdens. In Phase 2, we expect to follow the development scheme of Phase 1, but for Hindi, our second non-English language. We expect to be able to use the same schema as that in Phase 1 for Phase 2 Hindi.

2.3.5 Community Detection Evaluation Plan

As it relies on the other components, much of the Community Detection development was planned to occur towards the end of the Phase 1 and through Phase 2.

For the prototype, we planned to develop a clustering algorithm to cluster users into two or more groups within a discourse. The features used for the clustering task initially (baseline) are associated with upvotes, downvotes and other meta-data readily available via the OMC dataset. We also planned to begin exploring other clustering features, as well as viable graph neural net architectures useful for community detection problems.

For the end of Phase 1, we expected to have performed a descriptive analysis of communities identified by HUGO and explore characteristics of each group. In Phase 2, we will enable the use of SpEAR and the Utterance Classifier labels for augmenting features used by the community detection component.

2.3.6 Response Generation Evaluation Plan

For the Prototype, our plan was to develop utterance level responses including flagging posts and users. For this task, we have limited ground truth information available via content moderation (e.g., flagging of posts) by newspaper staff members for the OMC dataset.

For the end of Phase 1, we planned to extend HUGO’s response capabilities to templated NLG responses that include promoting and encouraging pro-social behavior.

In follow-on work, we will leverage results from our HSR study, in which we will study the efficacy of various culturally-sensitive response strategies. The HSR study will provide us with valuable ground truth information that will guide our work into Phase 2, at which time we will also integrate our longer-term planning approaches to response generation.

Chapter 3

Training and Evaluation Materials

3.1 Summary

Our primary data set for training is the following.

OMC Dataset: <https://ofai.github.io/million-post-corpus/> collected and annotated by the Austrian Research Institute for Artificial Intelligence. The data set represents user comments to newspaper articles published by the Austrian newspaper Der Standard. For each article, users may post their viewpoints about the newspaper article. The corpus consists of threaded commentary about a large number of articles related to specific topics. While released with a Creative Commons license, SIFT has communicated with both The Austrian Research Institute for Artificial Intelligence, as well as the head of User Generated Content (Dr. Christian Burger) at Der Standard and obtained consent to use the data for the purpose of this project.

Personnel at Der Standard informed us that the newspaper uses human guides as moderators in conjunction with automated moderation. Part of the comments are from newspaper staff, with the intent to moderate the conversations. The corpus covers a selection of user posts from 12 month during 2015-06-01 to 2016-05-31. There are 11,773 labeled and 1,000,000 unlabeled posts in the data set. The labeled posts were annotated by professional forum moderators employed by the newspaper.

WVS Data: In addition to our NLP data, we are working with the survey data from the World Value Survey [37]. We have downloaded the data for Germany and Austria and we are using Correlational Class Analysis to identify value patterns within this data. We describe this in more detail in this document.

3.2 Training in Native Languages

HUGO's goal is to promote prosocial behavior and mitigate the effects of antisocial behavior by working with the German OMC dataset in Phase 1. For Phase 2 we are evaluating several Hindi candidate datasets.

We analyzed the German dataset, and trained our NLP models *directly* in German, rather than translating German to English and then building a model. We believe maintaining the native language will preserve the cultural nuances and linguistic markers for pro- and anti-social behaviors that may otherwise get lost in translation. Below we outline the dataset and describe the categories for which we have already annotated. In section 4, we describe how we are using SIFT's CELEBRATE interface for prosocial annotations according to our schema.

3.3 Description of the OMC Dataset

For each post of the dataset we have the following information:

- Post ID

- Article ID
- Headline (max. 250 characters)
- Main Body (max. 750 characters)
- User ID (anonymized to new numeric IDs)
- Time stamp
- Parent post ID (replies give rise to tree-like discussion thread structures)
- Status (online or deleted by a moderator)
- Number of positive votes by other community members
- Number of negative votes by other community members

For each article, the data set contains the following data:

- Article ID
- Publishing date
- Topic Path (e.g.: Newsroom / Sports / Motorsports / Formula 1)
- Title
- Body

For each of the labeled posts they assigned the following annotations using three moderators annotating each post: Potentially undesirable content:

- Sentiment (negative/neutral/positive) An important goal is to detect changes in the prevalent sentiment in a discussion, e.g., the location within the fora and the point in time where a turn from positive/neutral sentiment to negative sentiment takes place.
- Off-Topic (yes/no) Posts which digress too far from the topic of the corresponding article.
- Inappropriate (yes/no) Swear words, suggestive and obscene language, insults, threats etc.
- Discriminating (yes/no) Racist, sexist, misogynistic, homophobic, antisemitic and other misanthropic content.

Neutral content that requires a reaction:

- Feedback (yes/no) Sometimes users ask questions or give feedback to the author of the article or the newspaper in general, which may require a reply/reaction.

Potentially desirable content:

- Personal Stories (yes/no) In certain fora, users are encouraged to share their personal stories, experiences, anecdotes etc. regarding the respective topic.
- Arguments Used (yes/no) It is desirable for users to back their statements with rational argumentation, reasoning and sources.

The annotators had reasonably good inter-rater reliability on the categories, as can be seen in figure 3.1. The data has been released with the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. See [56] for the initial paper discussing the research conducted on the dataset. The forums have community standards written in German posted on the Der Standard’s web site [2] – see appendix B.

The data contains 2179 posts by 110 newspaper staff members; 171 of these were posted as top level comments. The remaining posts were in response to reader comments.

3.4 Description of the World Values Survey Data

In addition to our NLP data, we are working with the survey data from the World Values Survey (WVS) [37]. We have downloaded the data for Germany and Austria and we are using Correlational Class Analysis to identify value patterns within this data, with two goals: 1) values that can function as “cultural weights” for a specific cultural context (e.g. German vs. Hindi or English) and 2) to provide a broad cultural understanding of values and norms for interpretative purposes (e.g. “levels of acceptable or expected rudeness vs. politeness”).

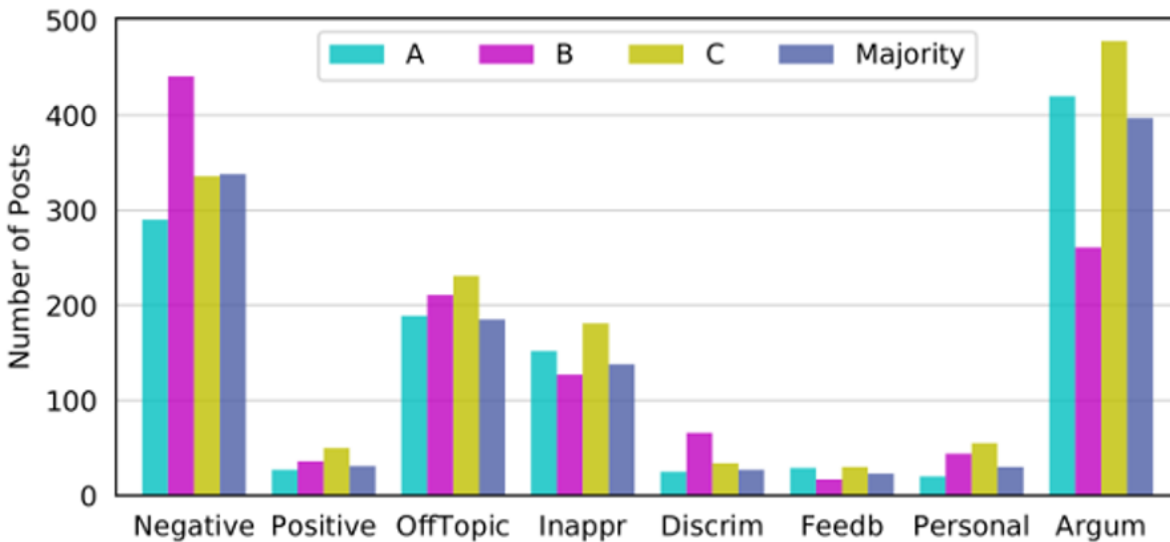


Figure 3.1: Inter-rater agreement on the categories annotating the OMC dataset.[56]

Thus far, we have cleaned and added 120 of the 250 World Values Survey questions to our “cultural values space.” This cultural values space will eventually consist of all World Values Survey questions specifiable along some relative scale (e.g. non-categorical) for all nations covered under the World Values Survey and European Values Survey. We have run exploratory factor analysis across the entirety of this value space in its current form to create a parallel latent variable value space consisting of 23 significant and interpretable WVS factors. An example of these factors is one we have entitled the “xenophobia index”. The xenophobia index is a weighted combination of questions determining what demographic delineations individuals would resent having as neighbors normalized between 0 and 1. These factors are similar to those constructed by Inglehart and Welzel in their value map [37] except ours are achieved through exploratory factor analysis, rather than confirmatory. This means they are largely data driven rather than theory driven. Finally, we have implemented functionality to a) detect and flag greatest cross-national differences in each of these values spaces, b) run nearest neighbors analysis to extract the nations that most closely mirror the value spaces (or subspaces) of any specified target nation, and c) detect and return the most divisive, and most commonly agreed upon values within a specific nation. In each of these functions, a nation may easily be switched out with any demographic characterization of the population, so long as that characterization is specifiable given available WVS ethnographic data. This functionality will facilitate cultural sensitivity weighting in HUGO and help apply HUGO to novel languages and cultural environments with greater fluidity.

3.5 Training and Evaluation with the OMC Data

The OMC dataset doubled as our training and evaluation dataset for various components of HUGO, including the subcomponents of our Utterance Classifier as well as our SpEAR Graph causal extraction architecture and Community detection components. In this section, we will provide a brief outline of how we used the dataset for training and evaluating these components, as well as a description of some preliminary analysis that we have already initiated along these fronts.

3.5.1 Utterance Classifier Subcomponents

Perspective API [3] This is a multilingual API offered by Jigsaw for classifying text along 8 categories of toxicity. Scores between 0 and 1 are computed for each category: toxicity, severe toxicity, insult, sexually explicit, profanity, likely to reject, threat and identity attack. We used a pretrained multilingual model for this subcomponent and evaluated the model against the OMC dataset. Because the toxicity categories are not the same as the OMC categories, we have developed an initial mapping criteria that generates for each anti-social category in OMC (primarily “Discriminating” and “Inappropriate”), we generate averages of various categories of toxicity. Thus, a score for “discriminating” is obtained by averaging “insult”, “identity attack”, “toxicity” and “severe toxicity”. The score of “inappropriate” is obtained by averaging “profanity”, “threat”, “insult” and “toxicity” and “severe toxicity”. Preliminary results suggest that this averaging-based approach has the potential to generate state-of-the-art results on the OMC dataset. We are currently extending these evaluations to realize the potential of the Perspective API subcomponent.

Multi-lingual MLM-based OMC Classifier We used the OMC native-German dataset to both train and evaluate a large multi-lingual and German-centric masked language model (MLM), based on the BERT architecture). We are doing so, by fine-tuning a multilingual BERT and GermanBERT model over the multi-class OMC dataset. The approach of applying MLMs to classify OMC data has not yet been reported in the literature, and as such this will provide a novel contribution in and of itself. From a data standpoint, training an MLM effectively is made more difficult by imbalances between classes in the dataset as well as the limitations in the quantity of available annotations. Nevertheless, we expected that this approach will provide state-of-the-art performance much like how MLM approaches have influenced other NLP areas. We also expect that using a multi-lingual model will allow us to do zero-shot transfer to other languages, including Hindi.

3.5.2 CELEBRATE annotation interface and HUGO schema

The OMC dataset was also expected to serve as training data for training our SpEAR causal graph extraction architecture. We identified 50 samples per annotation category where the post wasn’t already covered by one of the previous categories (so 442 overall), and we planned to use this data to annotate the German posts according to our pro-social schema using SIFT’s annotation interface. We planned to divide up the annotations, using some of it for training and some for validation and evaluation. Figure 3.2 shows the schema using an anti-social example, while Figure 3.3 shows the annotation interface.

In previous work we have developed an annotation interface for anti-social (TYBALT) and culturally distinct (Habitus) behaviors. For HUGO, we are using the interface with a new schema specific to HUGO.

Our SpEAR model is a transformer-based NLP model that uses neural networks to encode a sequence of textual tokens (i.e., words or subwords) into large vector-based representations for each token, sensitive to the context of the surrounding tokens [16]. SpERT transformer-based NLP architecture [19] has been used to process text to extract knowledge graphs, e.g., of people, relationships, and complex scientific claims [24]. We have amended the model with an “attribution classifier”. It allows us to use a NLP model that is highly inspectable because they infer indicators on specific spans of text rather than making a blanket judgment on an entire document.

HUGO Entities

Nodes primarily fall into these two categories:

1. Actors - any human individual, group, organization, settlement, or ideology.
2. Actions - any relevant actions or occurrences.

"There's these migrants from South America **flooding** the southern border and replacing **us**."

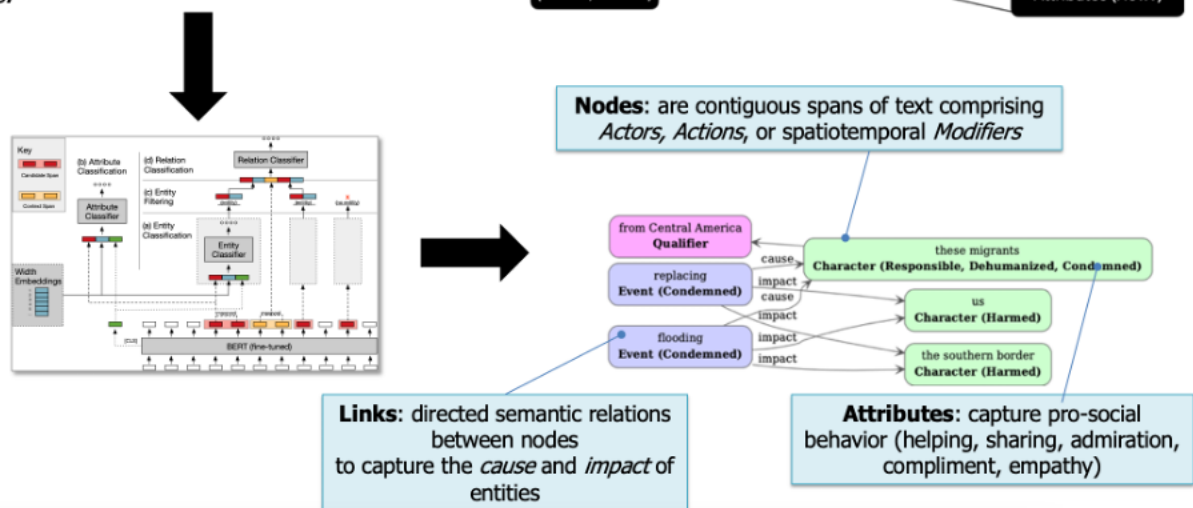


Figure 3.2: The HUGO schema used to annotate pro-social behavior.

HUGO Relations

Relations are directed semantic edges between labeled entities. They are critical for expressing what-goes-with-what over the set of entities. Without these relations, the structure of the events in the text would be semantically ambiguous: we would not know which character expressed a prosocial attitude towards whom. Our schema includes the following relations:

1. Cause indicates the actor(s) or event(s) that enacted a given event. This helps capture the placement of pro-sociality.
2. Impact identifies the actor(s) targeted by a given action (e.g. congratulations, thanks).

Attributes

As attributes, we are using multi-label classes, where zero or more may apply to any given entity. The SpERT transformer based model by Eberts & Ulges [19] (on which ours is based) was not capable of expressing these; this is a novel contribution of our previous [23] and current work. Attributes are displayed as parenthetical listings inside each node.

The initial schema for HUGO consisted of the following attributes for anti- and prosocial entities (actor, action) as enumerated in Table 3.1:

Table 3.1: HUGO anti- and prosocial attributes schema with examples

Example	Prosocial	Node-Type	Antisocial	Example
Char: “Go Le Pen!” Event: “Blessed raid”	Advocate	Character/ Event	Oppose	Char: “Boo Le Pen!”
“I’m so sorry!”	Empathy/ Compassion	Character	Dehumanization	“You cockroach!”
“I know of this site, that can help. . .”	Helped/Affirmed	Character	Harmed	“These people are taking our money!”
“Savior”	”Angel”/Helper	Character	”Devil”/Harmer	“Murderer;” “USA dropped bombs on us.”
“Great job, Ruta!”	Admire/Respect	Character/ Event	Contempt	“Terrible interview”


CELEBRATE!
Context-Elevated Broad Relation And Term Extractor

Unstructured text:

Also ich spiele die Serie seit Ufo Enemy Unknown aus dem Jahre 1994. Wenn ich an all die Zeit denke, die ich damit verbracht habe... Einfach

Model to use: Schema User Name: SSG Dataset Name: OFAI

Tag It [Examples I've Tagged] [Unreviewed Machine-Tagged Examples]

Manual Edit Clear Example Toggle Options Rerun Example

Add, remove, or revise the **entities**, **attributes**, and **relations** over the tokens.

Also ich spiele die Serie seit Ufo Enemy Unknown aus dem Jahre 1994 . Wenn ich an all

actor action grouping/ingroup cause emotion/admiration

die Zeit denke , die ich damit verbracht habe ... Einfach wunderbar , hv\$tte nichts sinnvolleres machen

actor action cause actor action cause actor action

kv\$nnen . Freude mich auf XCOM2 wie ein Kind a

actor actor action

Add Entity: action actor

Submit Training Example Toggle TREX View Download as P

Relations: cause, impact, modified
Attributes: author condemns, author justifies, author describes as violent, author has admiration for, author feels empathy for
Re-assign, Delete

Figure 3.3: The Celebrate interface HUGO used to annotate statements.

3.6 Preliminary Community Analysis

Our approach applies NLP techniques to extract knowledge and relationships between the posters, internal and external groups and topics. We conducted a preliminary analysis of data using network analysis techniques and community protection approaches.

The corpus contains 31,413 unique users, with 32.2 average posts per ID. Five is the median number of posts per user, so we can infer that a small number of users post very frequently, as can be seen in Figure 3.4.

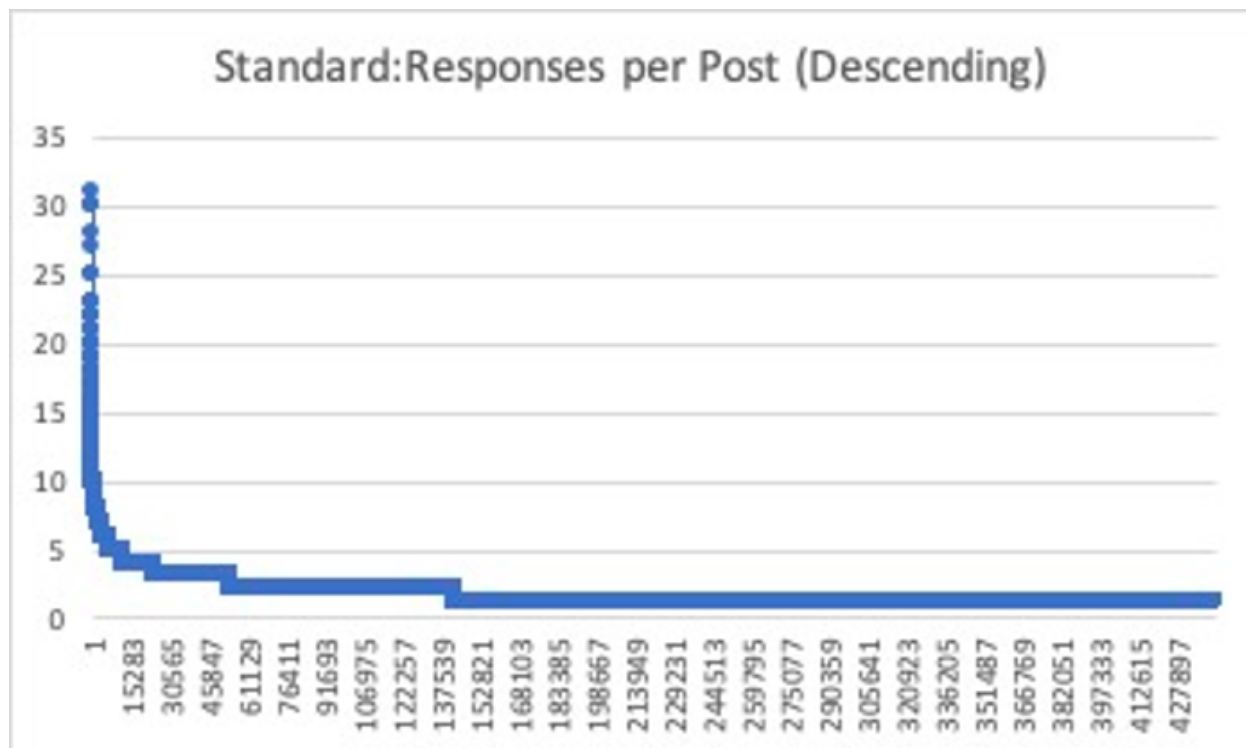


Figure 3.4: Number of responses per Post. Most Posts receive only a small number of responses.

There were 549,547 unique responder-respdee pairs in the corpus. On average a responding pair interacted 1.26 times, with half responding just 1 time. That is, multiple dialog interactions between pairs of participants is rare; see Figure 3.5.

Most User interactions occur only a very few number of times. Most users do not interact frequently. Only a tiny proportions of respondents interact 50 or more times.

Based on connected component analysis, there is a core of 732 nodes that form a single, large connected component. These are the core participants in the forum.

We broke the network into communities using the Girvan-Newman algorithm [29] and the k-clique percolation algorithm [49].

The Girvan-Newman algorithm for the detection and analysis of community structure relies on the iterative elimination of edges that have the highest number of shortest paths between nodes passing through them. By removing edges from the graph one-by-one, the network breaks down into smaller pieces, so-called communities.

The k-clique percolation method builds up the communities from k-cliques, which correspond to complete (fully connected) sub-graphs of k nodes. (E.g., a k-clique at $k = 3$ is equivalent to a triangle). Two k-cliques are considered adjacent if they share $k - 1$ nodes. A community is defined as the maximal union of k-cliques that can be reached from each other through a series of adjacent k-cliques. Such communities can be best interpreted with the help of a k-clique

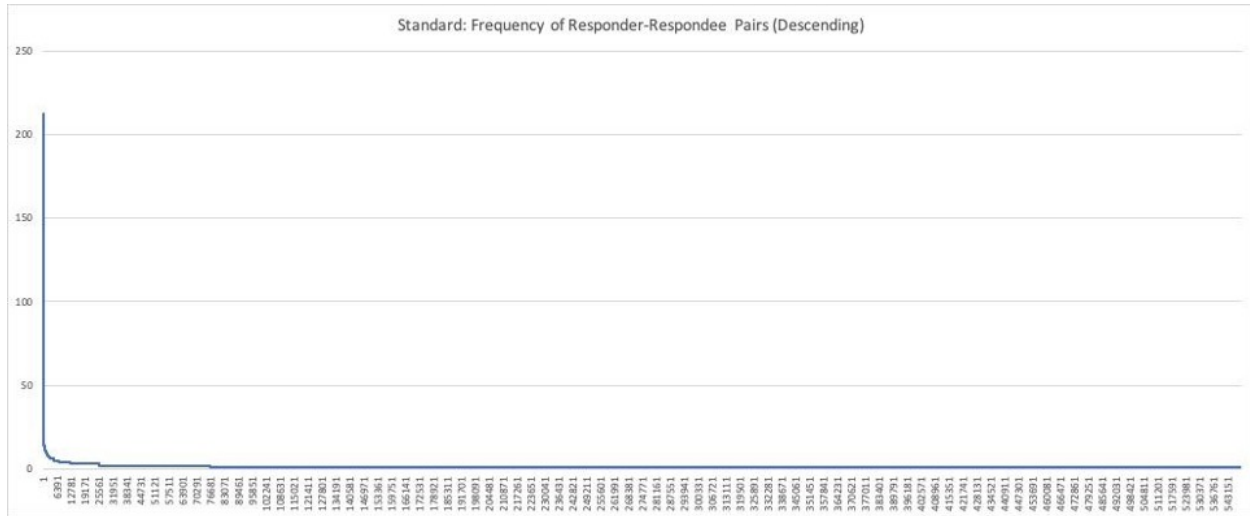


Figure 3.5: Frequency of Responder-Responsee pairs.

template (an object isomorphic to a complete graph of k nodes). Such a template can be placed onto any k -clique in the graph, and rolled to an adjacent k -clique by relocating one of its nodes and keeping its other $k - 1$ nodes fixed. Thus, the k -clique communities of a network are all those sub-graphs that can be fully explored by rolling a k -clique template in them, but cannot be left by this template.

A handful of topics attract the vast majority of responses on the site, as can be seen in Figure 3.6. Topics related to Refugees (“Fluechtlinge”) attracted the most responses on the forum. Overall, refugee related topics attracted 89,148 or 9% of the total responses in the corpus. The rest attracted only a small number of responses. The 20 Article topics with the most aggregated topics are listed in Table 3.2

The most controversial article in the corpus (ID 12075) is refugee related and has 500 responses with more than one positive or negative vote. The posts on this article thread received a total of 17, 241 positive and negative votes, with one polarity averaging 25% of the other, e.g. if a post in the thread received 4 positive votes, then it was likely to receive 1 negative vote.

The articles that generated the most controversy, measured in terms of aggregate, non-unanimous positive and negative votes across the thread of posts responding to it, all had to do with topics involving refugees, right-wing extremism, the Austrian Freedom Party, integration of refugees and immigrants into Austrian society, and the like.

Although the corpus indicates the number of positive and negative votes a post gets, and these can be aggregated over posts responding to a particular article, the dataset does not indicate which user contributed a vote, so it is impossible to characterize users by the content that they react to with positive and negative votes.

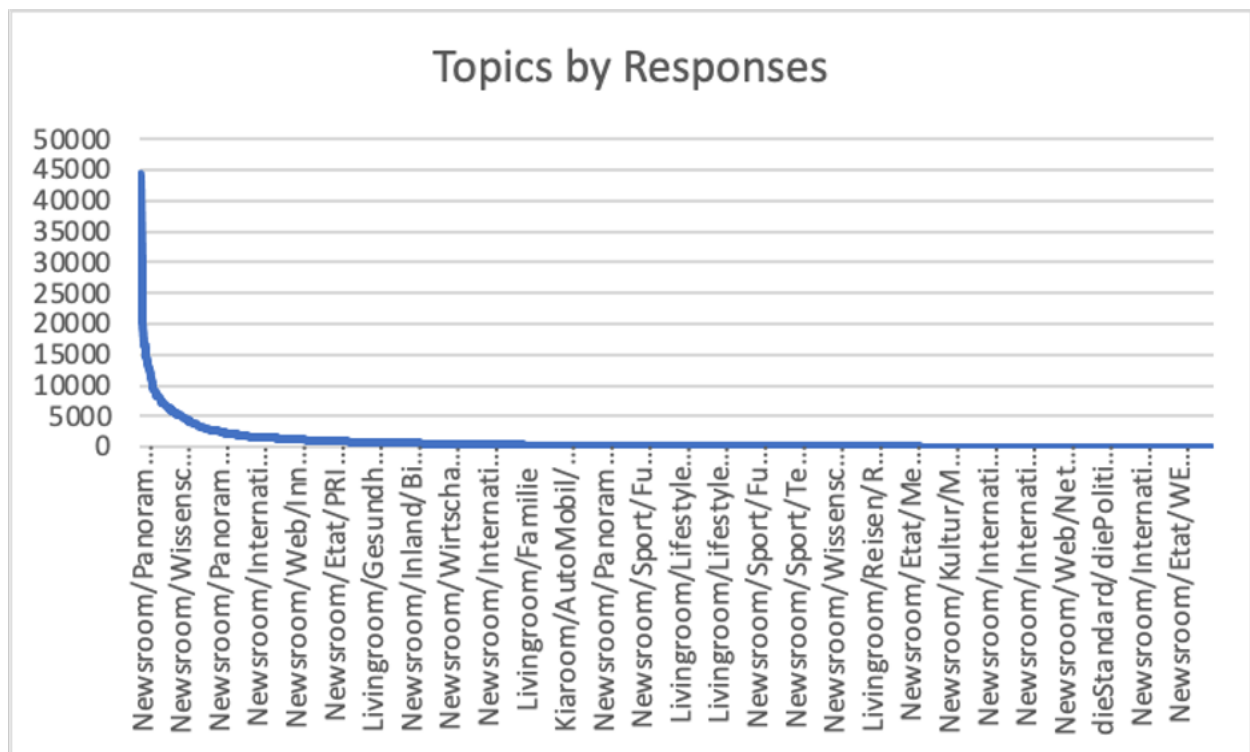


Figure 3.6: Number of responses by article topic.

Table 3.2: Topic paths in German vs. English

German	English	# Posts
Newsroom / Panorama / Flucht / Fluechtlinge in Oesterreich	Newsroom / Panorama / Escape / Refugees in Austria	44329
Newsroom / Panorama / Flucht / Flucht und Politik	Newsroom / Panorama / Escape / Refugees and Politics	30604
Newsroom / Inland / bundespraesi	Newsroom / Domestic / (Austrian) President	20177
Newsroom / International / Nahost / syrien	Newsroom / International / Middle East / Syria	16896
Newsroom / Panorama / Weltchronik	Newsroom / Panorama / World Chronicle	16362
Newsroom / Panorama / Chronik	Newsroom / Panorama / Chronicle	16334
Newsroom / Web / Netzpolitik	Newsroom/Web/Network Policy	14516
Newsroom / Web / Netzpolitik / Hasspostings	Newsroom / Web / Netzpolitik / Hate postings	14112
Newsroom / Wirtschaft / Wirtschaftspolitik / Finanzmaerkte / Griechenlandkrise	Newsroom / Economy / Economic Policy / Financial Markets / Greek Crisis	13463
Newsroom / Inland / Parteien / fp	Newsroom / Domestic / Parties / FP (Freedom Party)	12951
Newsroom / Wirtschaft / Wirtschaftspolitik	Newsroom / Economy / Economic Policy	12534
Newsroom / Inland / bundeslaender / ipowien	Newsroom / Germany / federal states / Vienna Politics	11663
Meinung / Kolumnen / rau	opinion/columns/rough	11215
Newsroom / Inland / Parteien / SPOE	Newsroom / Domestic / Parties / SPOE (SPÖ, Austria's Social Democratic Party)	10178
Newsroom / Web / Webvermishtes	Newsroom / Web / Web Miscellaneous	9497
Newsroom / Wirtschaft / Wirtschaftspolitik / Arbeitsmarkt	Newsroom / Economy / Economic Policy / Labor Market	9179
Meinung / KdA	Opinion / Other comments	9102
Newsroom / Web / Games	Newsroom/Web/Games	8937
Newsroom / International / Europa / deutschland	Newsroom / International / Europe / Germany	8309
Newsroom / Wirtschaft / Unternehmen	Newsroom / Business / Business	8275
Newsroom / User / mitmachen / Mitreden	Newsroom / Users / Join in / Have your say	8210

Chapter 4

Initial Prototype and Evaluation

4.1 Milestone 3 Prototype Architecture and Components

The architecture as of milestone 3, as seen in Figure 4.1 has two main pieces:

1. Detection and Data Fusion
2. Response Generation

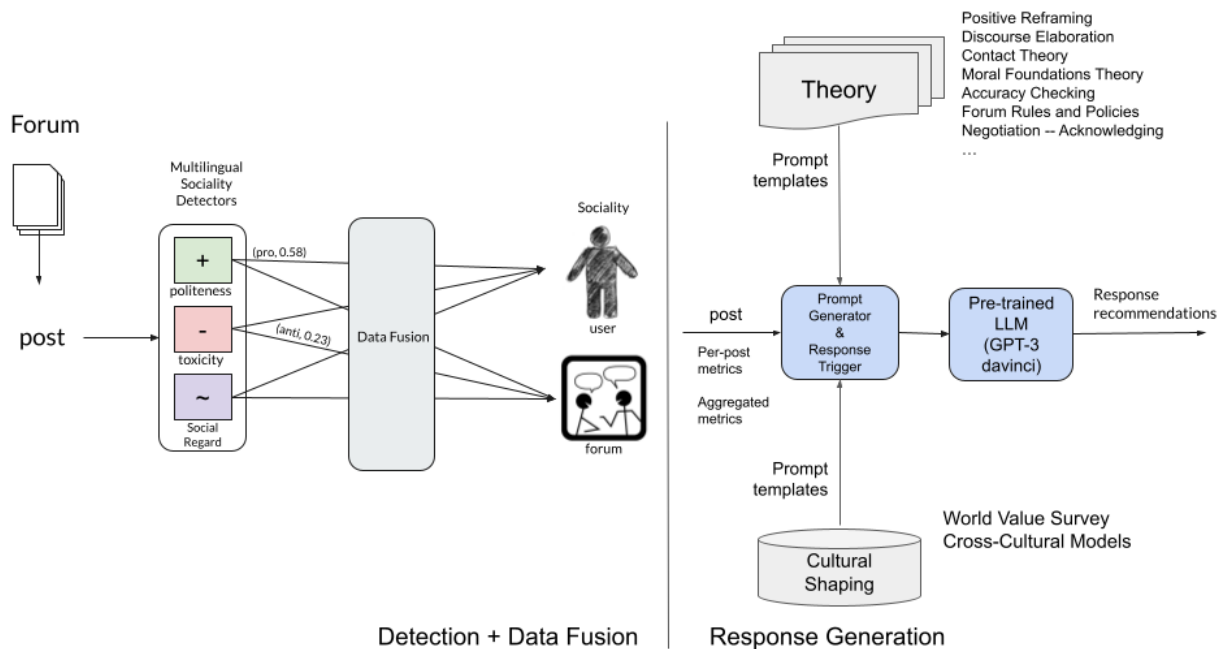


Figure 4.1: HUGO 0.1 Architecture Overview

Posts come in and go through our detectors which would assign them scores for both pro-sociality and anti-sociality. These scores would get fused and result in scores for sociality (for a user and for the forums as a whole).

Post by post we would compute and take these metrics and use those to plug into a set of GPT-3 prompt templates based on a combination of the Forum rules/policies, as well as a set of theories from the social sciences (Positive Reframing, Discourse Elaboration, Contact Theory, Moral Foundations Theory, Accuracy Checking, Negotiation – Acknowledging, etc). These would generate fresh prompts which in turn would generate responses. The key thing

with this approach is that we would generate a menu of responses from this, but we have since developed further methods to narrow down which prompt is most suitable for particular types of posts. This will be detailed in Chapter 5

First, we will go through the detection and generation pieces.

4.1.1 NLP Pro-/Anti- Sociality Detectors

Our sociality detectors, as seen in Figure 4.2 include a toxicity detector, a politeness detector, and a directed social regard detector.

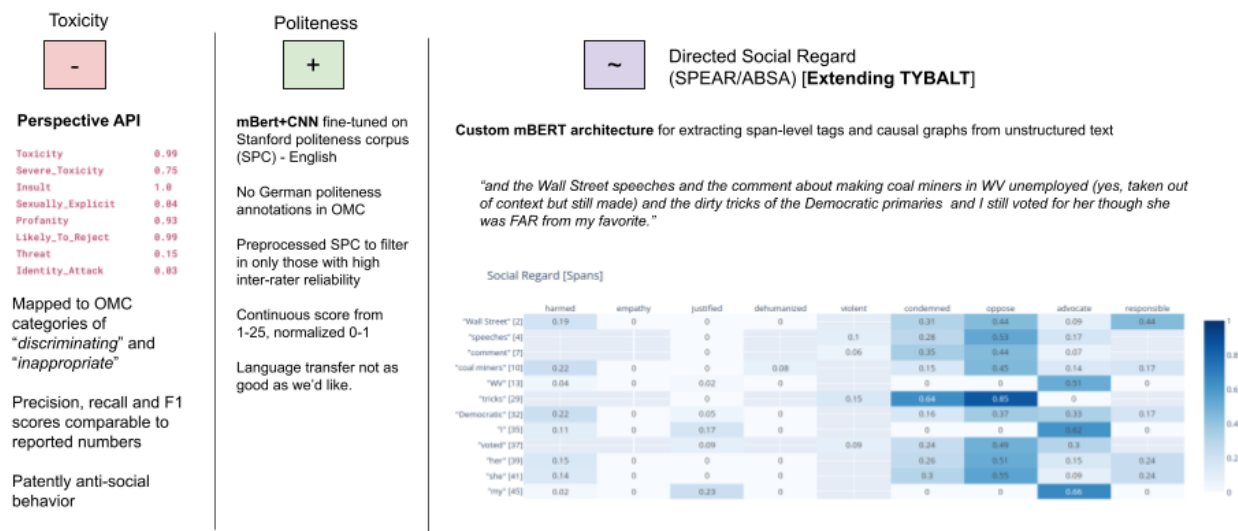


Figure 4.2: Sociality Detectors

Toxicity Detector

Toxic, rude or disrespectful comments are antisocial behavior and make people likely to leave a discussion. For our Toxicity Detector, we used Perspective API [3] which was developed by Google. This returns a set of 8 measures of toxicity:

- **Severe toxicity:** very hateful, aggressive, disrespectful comment.
- **Identity Attack:** negative or hateful comments targeting someone because of their identity.
- **Insult:** insulting, inflammatory, or negative comment towards a person or group.
- **Profanity:** swear words, curse words, or other obscene or profane language.
- **Threat:** intention to inflict pain, injury, or violence against an individual or group.
- **Sexually explicit:** reference to sexual acts, body parts, or other lewd content.

While initially we mapped these to particular categories from the OMC dataset corpus, we eventually folded these into an overall contribution to the antisocial score component.

Politeness Detector

"Politeness" is a cross-cultural, sociolinguistic theory developed by anthropological linguists Brown and Levinson [10]. The main motivation for politeness is the minimization of threats one's "Face." There are two types of Face:

- Positive Face: the desire for connection with others
- Negative Face: needs for autonomy and independence, among others

When interacting with someone in a way that involves a Face Threat Act (FTA), one can use various prosocial politeness strategies to redress the threat, with them falling into categories of positive and negative politeness depending on the type of FTA one is redressing. These strategies can be seen in Figures 4.3 and 4.4.

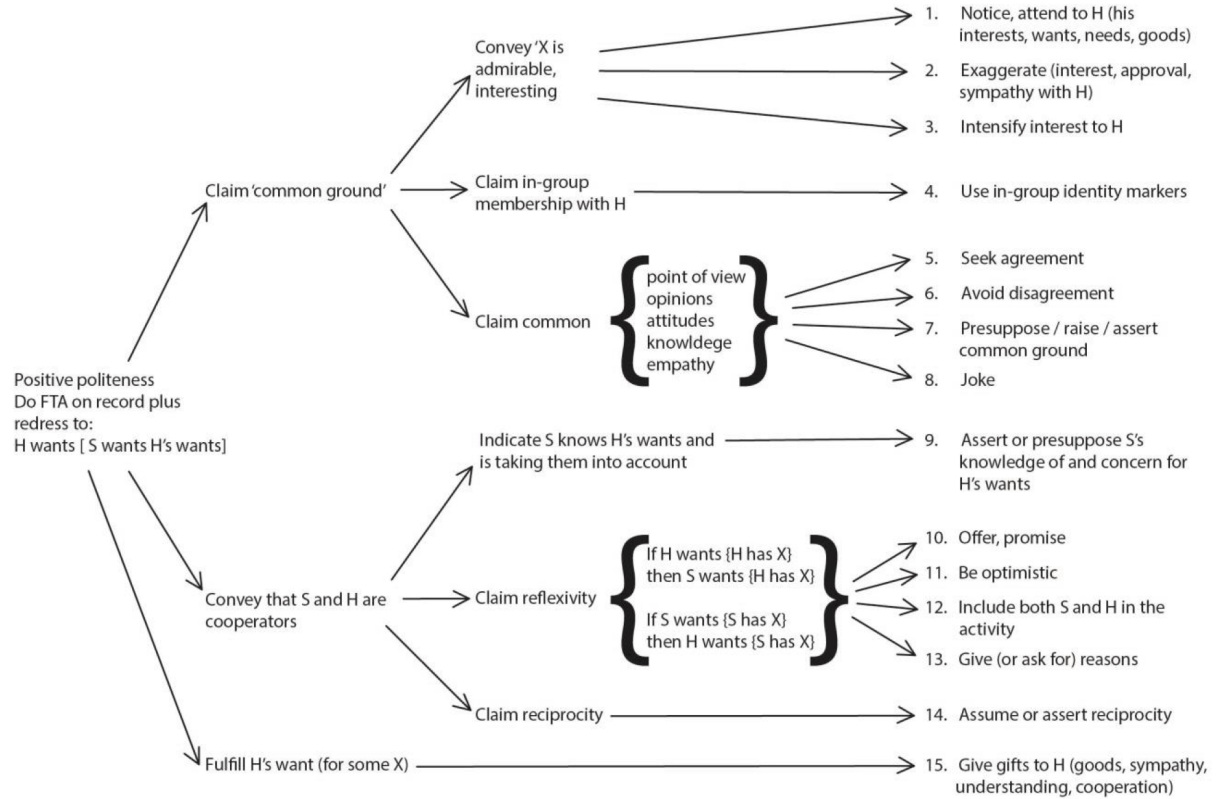


Figure 4.3: Positive politeness strategies flow chart from Brown and Levinson [10]; "H" refers to the hearer, and "S" refers to the speaker in a communication exchange.

This model of politeness was the basis for the Stanford politeness Corpus [15], which assigned Wikipedia comments politeness scores. Our Politeness Detector used a Multilingual BERT [16] with a CNN architecture and we trained it and fine-tuned the BERT model on this corpus. This allowed us to receive text and compute politeness scores. We then normalized these scores between zero and one. We then used these scores as input to the prosociality component of our utterance score.

Directed Social Regard Detector

Instead of scoring and classification, we are extracting information we can reason about. For the Directed Social Regard Detector, we used a custom architecture developed for under TYBALT for the Social Sim program. This architecture is called SpEAR ABSA [67]; SpEAR stands for Span Based Entity Attribute and Relation Extraction; ABSA stands for Aspect-Based Sentiment Analysis.

This model allows us to break up the sentences into spans and assign them various social regard factors. This social regard measure allowed us to get into the detailed content of the sentence, not just an overall sentence score that our Toxicity and Politeness Detectors were doing, but at the word span level that allows us to see the actors and targets of the social regard. See section 2.2.3 for more details on the tagging process.

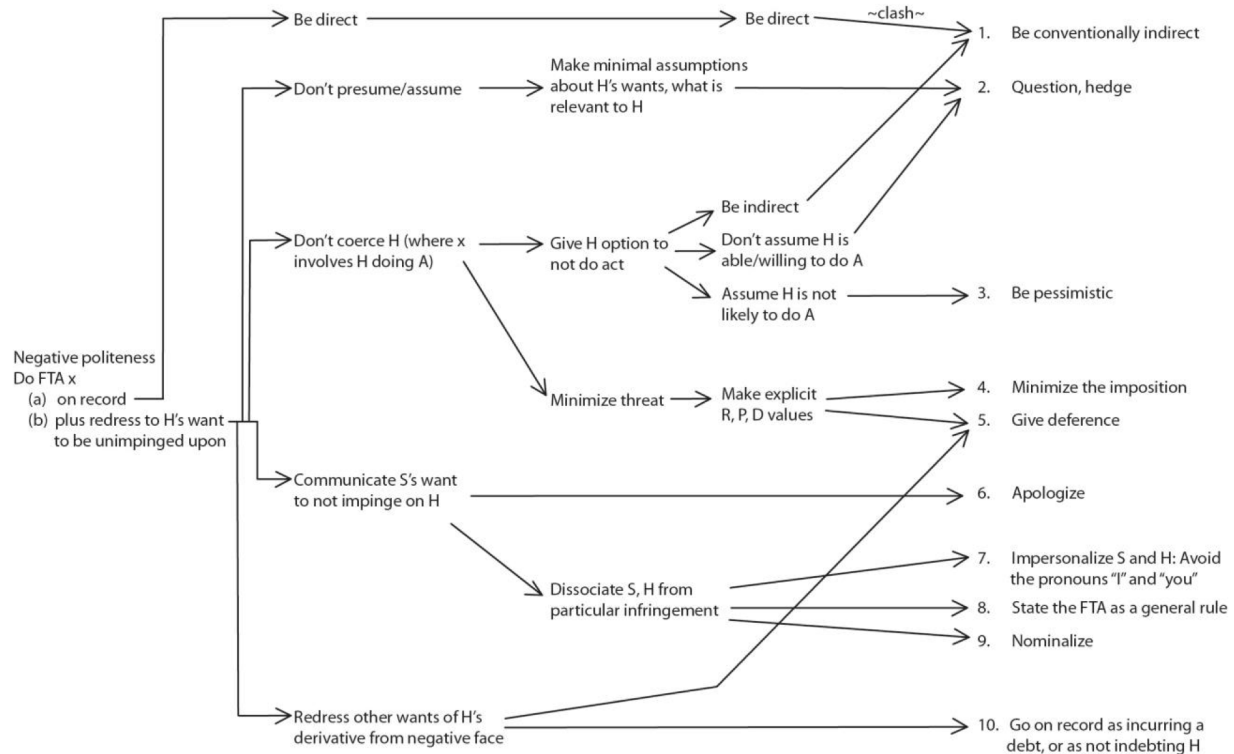


Figure 4.4: Negative politeness strategies flow chart from Brown and Levinson [10]; "H" refers to the hearer, and "S" refers to the speaker in a communication exchange. "R", "P", and "D" represent the Rank of imposition, the relative Power, and the social Distance, respectively

For our annotation for the initial prototype, we used these attributes:

- **Violent:** Event described as having literal or metaphorical physical or sexual violence
- **Justified:** Entity was morally justified
- **Dehumanized:** Actor described with non-human or lesser-than-human attributes
- **Harmed:** Actor described as being harmed by themselves or others
- **Condemned:** Entity morally condemned
- **Opposed:** Opposition or adversarial attitude to entity by author
- **Advocate:** endorsement or support of an entity by the author
- **Responsible:** Actor described as causing harm to others or to themselves
- **Empathy:** Actor described with empathy, compassion, or humanity

After determining that Multi-lingual BERT was impressively good at translations and that annotating the German data was not feasible in the scope of this contract, we ended up using annotations conducted under the TYBALT program for which SpEAR was initially developed. Under the TYBALT program, we recruited three annotators and trained them on identifying these nine attributes, and had follow up meetings to arbitrate agreement. We selected 2,900 posts and sentences, and our annotators tagged 24,000 entities for DSR, rating each on an intensity scale from 1-5. The prevalence of each attribute and inter-annotator reliability is shown in Table 4.1. As we were not fully satisfied with these results for either program, we made further refinements for the final prototype described in section 5.2.

Table 4.1: Label usage across annotators (A1-3) and agreement (Krippendorff’s α)

Aspect	A1	A2	A3	α
advocate	21.4%	16.1%	18.0%	0.366
condemned	20.4%	8.0%	10.5%	0.477
dehumanized	2.7%	3.8%	5.4%	0.591
empathy	1.0%	12.6%	3.2%	-0.065
harmed	7.9%	10.8%	9.2%	0.580
justified	7.3%	4.1%	1.5%	0.171
oppose	24.0%	25.4%	36.2%	0.672
responsible	11.2%	13.6%	8.0%	0.607
violent	4.1%	5.6%	8.0%	0.753

4.1.2 Data Fusion to Construct User/Forum Models

We then took all of these detectors and we treated them as social sensors. Each of these values served as what we call a Body of Evidence (BoE), which tells us if it leans towards pro- or anti-sociality. Since these numbers are not just 0/1 or T/F, we ended up with real numbers such as 0.54 and 0.65 as our measures in support of pro- or anti-sociality, and then we add in uncertainty intervals based on recent advances in the theory of uncertainty [9] related to the Dempster-Schafer theory of evidence[59]. We assign these numbers to a pro-sociality score, which gives us an outcome of prosocial, neutral, or anti-social.

As an example: if a Toxicity score came back as 0.65, we would assign it a score of 0.65 of evidence for antisocial, and then 0.35 for neutral/prosocial, because of the uncertainty of the situation. This would be combined with the politeness score, which gives us evidence for prosocial behavior. This is finally fused with the directed social regard, which can give us direct evidence for both prosocial and antisocial behavior depending on the social regard expressed for the characters and events in the sentence, and this overall evidence is composed into one number. A more detailed example is shown in Figure 4.5.

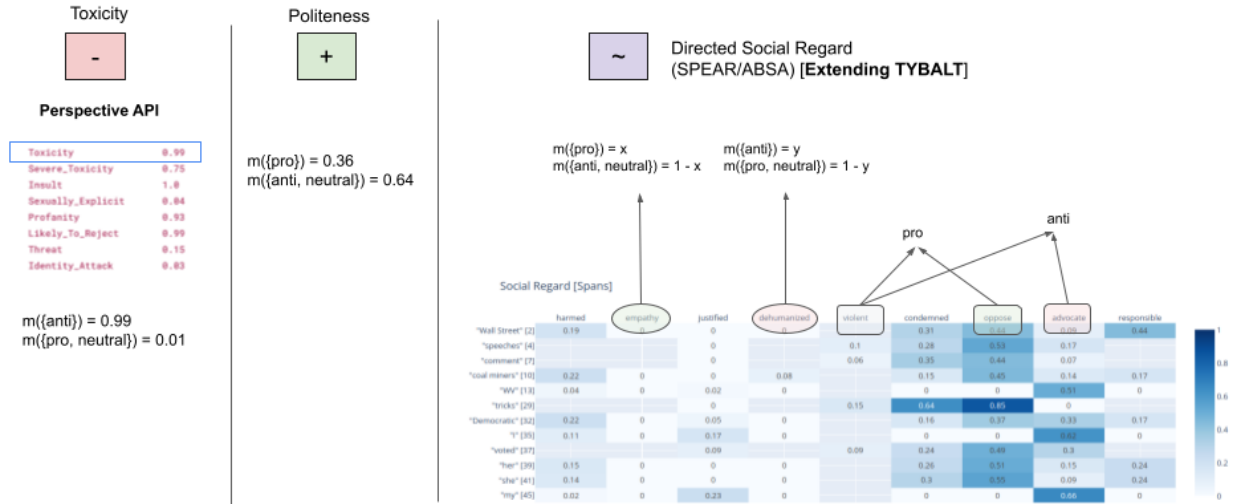


Figure 4.5: Fusion Rules per post

By fusing these values from all the sensors, we are able to compute one number that can “track the temperature of the room” in a forum over time, as can be seen in Figure 4.6

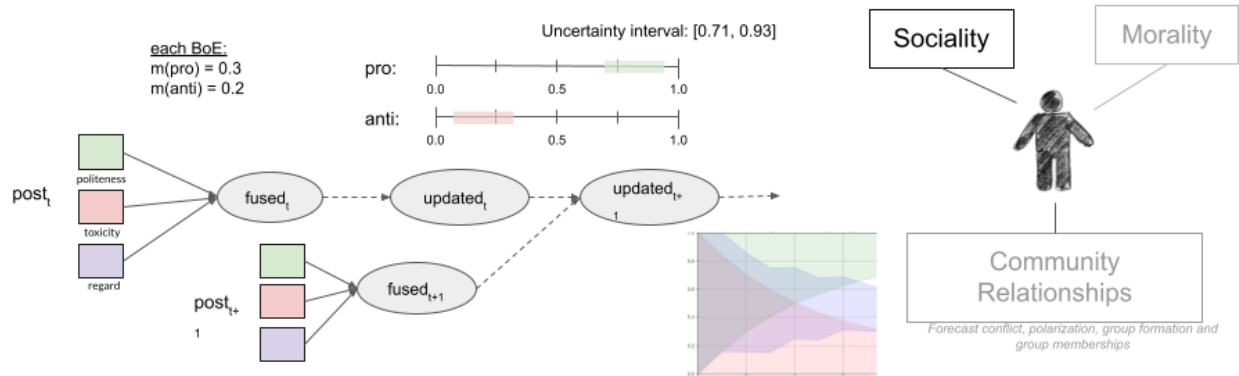


Figure 4.6: Modeling a Forum from Fused Data

4.1.3 Response Generation

For Response Generation, we leverage recent advances in Large Language Models (LLMs), in particular GPT-3 [11], to generate our responses. LLMs serve as a powerful base to do both classification and text generation, with the latter being its initial purpose. It generates text by means of attempting to complete your sentences by telling you what the most likely next word is. What makes GPT-3 unique is its size; it was trained on over 500 billion tokens, with about 175 billion parameters.

To make use of our social science theories in the generation step, we created a grammar/template of theory-based prompts. Prompts are text representations of the goal of the response, to frame and constrain the response. From these, we automatically generate complete prompts based on the template + post + the post's fused sociality score, from the NLP pipeline described above. GPT-3 then generates responses to particular posts based on these prompts.

There are some caveats when using a LLM like this. LLMs (like GPT3) are largely blackbox models trained to continue sentences from very large corpora. Because of this, they are influenced by any underlying biases of the dataset. Additionally, they cannot really “understand” the content/context of particular post. There also is less clarity on how well they deal with language transfer, especially in low-resource languages.

In order to mitigate these risks, we reduced the footprint of LLMs in HUGO. We limiting their scope by using them for only part of the analysis response pipeline. The pipeline contains components extracting and responding based on a theoretical scaffolding. The result is high constrained textual prompts fed into GPT-3, which GPT-3 turns into the target natural language.

Figure 4.7 provides an example of how we use the theoretical scaffolding in our response generation. The title of the original article defines the situation the user is posting under. Modeling intent is one of our theory-based ideas. For example, if we were following Moral Foundations Theory, we would say that along the Care dimension we want to encourage Care towards a vulnerable outgroup. Thus we can feed GPT-3 a prompt to “Produce a response to encourage care towards a vulnerable outgroup.” or “ask why the <DEHUMANIZED ENTITY> deserves to be disrespected.” Responses are also based on what language we receive and detect.

4.1.4 Post Alerts and Response Triggering

Deciding when HUGO should intervene and respond in a conversation is a difficult question. We decided to use the overall fused sociality score as a metric based on our belief in the strength of the sociality score as contrasted with its plausibility. Belief(proposition) is defined as the amount of evidence in support of the proposition. Plausibility(proposition) is 1 - the amount of support against that proposition. We used 0.2 as a threshold – if we had a belief greater than 0.2 in either the prosociality or antisociality of a post, or if the difference between our pro-social and

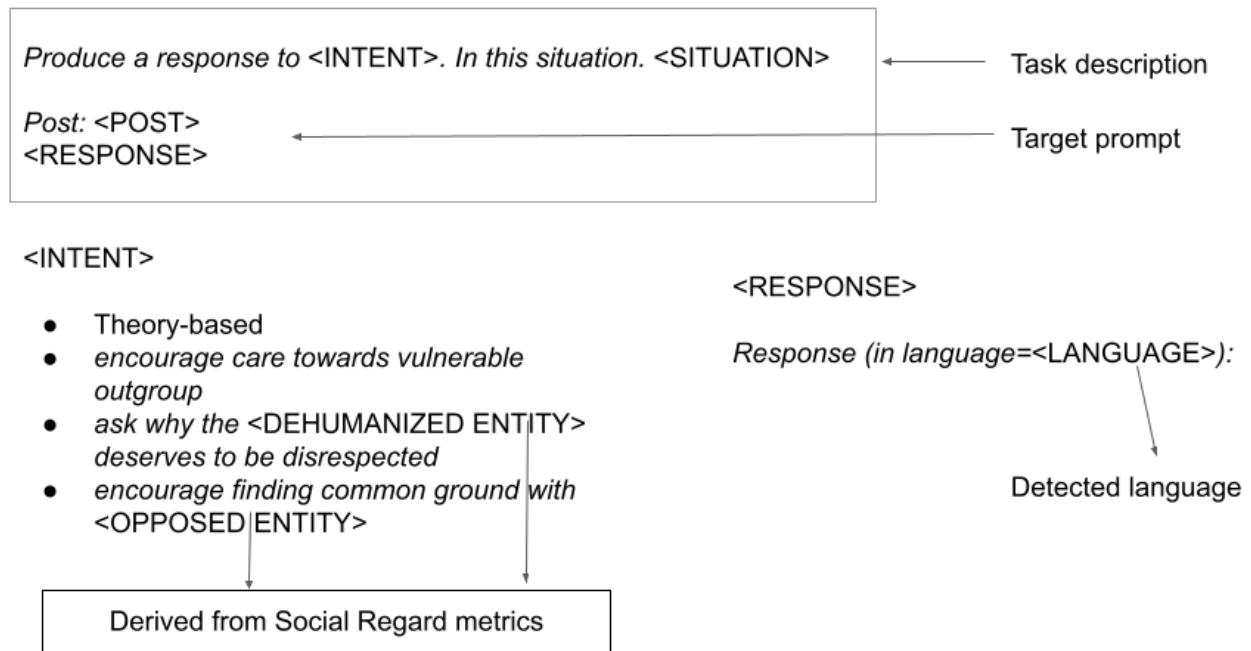


Figure 4.7: Post Prompt Generation

anti-social beliefs was greater than 0.2, a response would be triggered.

In addition to deciding when to trigger a response, we also had to choose which response type to trigger. We generated 25 response intents over 7 different theoretical frameworks. These intents were triggered based on the NLP detectors to provide an array of response options to a moderator.

- **Triggered by Social Regard:** Positive reframing [12, 18, 20, 33, 71], reducing prejudice (contact theory) [7, 8, 52, 53, 54], facilitating prosociality [17, 21, 38, 39, 58, 4]
- **Triggered by Toxicity:** Forum rule violations [2]
- **Always Triggered:** Moral foundations [30, 36, 62, 60], checking accuracy [51], engaging by asking questions (discourse theory) [43]

4.2 Technical Contributions Summary

The output of this initial prototype can be seen in the HUGO moderation Dashboard in Figure 4.8 where you can see the multiple options for theory-based responses, as well as the references that underpin those theories. This moderator dashboard requires a human in the loop to decide which response (if any) to use.

1. Extensible Architecture for multilingual moderation
 - Can include arbitrarily more sensors:
 - sentiment
 - argumentation
 - moral
 - informative/relevant (especially for disaster situations)
2. Theory and Data-driven prompt design for response generation
 - Implemented 25 different prompt variations across 7 theoretical frameworks
 - Prompt grammar to guide prompt generation
3. Social regard measures over spans

HUGO

Multilingual Moderator Dashboard

Single
Forum

Go get help my friend. You dont need to be afraid. Just breathe and listen to some good music when you can. I hope everything works out. They wont let me link it here but this is the name of a song that got me through insanity!

Submit

Last Submitted On

pro

anti

Moral Concerns
Engaging with Questions
Checking Accuracy
Rule Violations
Reducing Prejudice

Facilitating Positivity
Positive Reframing

HUGO recommends leveraging facilitation practices used by in-person and online moderators in a variety of settings from classrooms to online forums.

Use on gpt page reaching out for help, and I hope things start to look up for you soon. That song sounds like it would be really helpful in a tough situation.

Good to express admiration for the post

Diakopoulos, N., & Reisman, M. (2011). Towards quality discourse in online news comments. Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work - CSCW '11, 132. <https://doi.org/10.1145/1958824.1958844>

Epstein, D., & Leshed, G. (2020). The Magic Saver: Practices of Facilitation in Online Policy Deliberation. Journal of Deliberative Democracy, 12(1). <https://doi.org/10.1093/jdd/244>

Jhaver, S., Birman, J., Gilbert, E., & Bruckman, A. (2018). Human-Machine Collaboration for Content Regulation: The Case of Reddit AutoModerator. ACM Transactions on Computer-Human Interaction, 26(5), 1–25. <https://doi.org/10.1145/3192343>

Seering, J., Wang, T., Kwon, J., & Kaufman, G. (2018). Moderator engagement and community development in the age of algorithms. New Media & Society, 21(7), 1417–1443. <https://doi.org/10.1177/1461444818621316>

Type on Facilitating Effective Group Discussions | Sheridan Center | Brown University. (n.d.). Retrieved May 17, 2022, from <https://www.brown.edu/sheridancenter/teaching-learning/resources/teaching-resources/classroom-practices/teaching-effective-discussions/type>

What Makes a Good Interaction Between Disliked Groups? (n.d.). Greater Good. Retrieved May 1, 2022, from https://greatergood.berkeley.edu/article/item/what_makes_a_good_interaction_between_disliked_groups

Advanced Analytics:

Figures

Figures
Raw

Social Regard (Spam)

	no	yes	no/yes	no/yes/yes	no/yes/yes/yes	no/yes/yes/yes/yes	no/yes/yes/yes/yes/yes	no/yes/yes/yes/yes/yes/yes
no	100	0	0	0	0	0	0	0
yes	0	100	0	0	0	0	0	0
no/yes	0	0	100	0	0	0	0	0
no/yes/yes	0	0	0	100	0	0	0	0
no/yes/yes/yes	0	0	0	0	100	0	0	0
no/yes/yes/yes/yes	0	0	0	0	0	100	0	0
no/yes/yes/yes/yes/yes	0	0	0	0	0	0	100	0
no/yes/yes/yes/yes/yes/yes	0	0	0	0	0	0	0	100

- SPEAR model for span detection
- ABSA model for regression over attributes across spans

4. Computing overall sociality with data fusion techniques for social sensors

- Connecting “advocating for violent events” → evidence of anti-sociality
- Fusing (with uncertainty) data from different social sensors
- Fusing (with uncertainty) data over time to capture forum level sociality
- Python package (“unsure v0.2.0”) for performing fusion operations in a computationally efficient manner.

Chapter 5

Revised Prototype for Phase 1 Demo

5.1 Overview of Updates

After reviewing our initial prototype, we wanted to address to aspects of our architecture when adapting this for use in a private Reddit chatroom. For our Directed Social Regard (DSR) detector:

- Some of the social regard categories were related to each other, so we decided to bring them together in dimensionality. So we changed the name from Sociality to Civility
- The new Civility metric is more grounded in theory, more clearly operationalizable, and more applicable to the Civil Sanctuary domain Responses
- We wanted more targeted responses to a post, not just provide a menu of responses that requires a human in the loop.
- Forum Rules - Forum rules were treated as one of the response strategies. We need to make them more of a first class citizen.
- New HUGO bot venue - In order to have an environment to allow HUGO to directly interact on an example platform, we set up a private subreddit and a HUGO bot that can respond to posts there.

5.2 Updated Directed Social Regard Detector

Under the HUGO program we updated the ABSA model for our DSR detector to have dimensions instead of separate attributes; this gives us a -1 to +1 scale for these dimensions. The dimensions are:

- oppose-advocate
- demhumanization-empathy
- harmed-helped
- harmer-helper
- contempt-respect

These are based on the schema shown in 3.1. Figure 5.1 shows an example sentence tagged with the new dimensions.

5.2.1 New ABSA/DSR Annotation

HUGO adapted the experimental results of the TYBALT program to create a new sensor. We had three annotators use this new schema to annotate a large data set from a combination of sources; the Moral Foundation Twitter Corpus (MFTC) [36], the Gab Hate Corpus [41], “How ISIS uses Twitter” Kaggle dataset [66], Manosphere community text posts [55], and selected subreddits. Because the annotation was conducted under the TYBALT program, the datasets

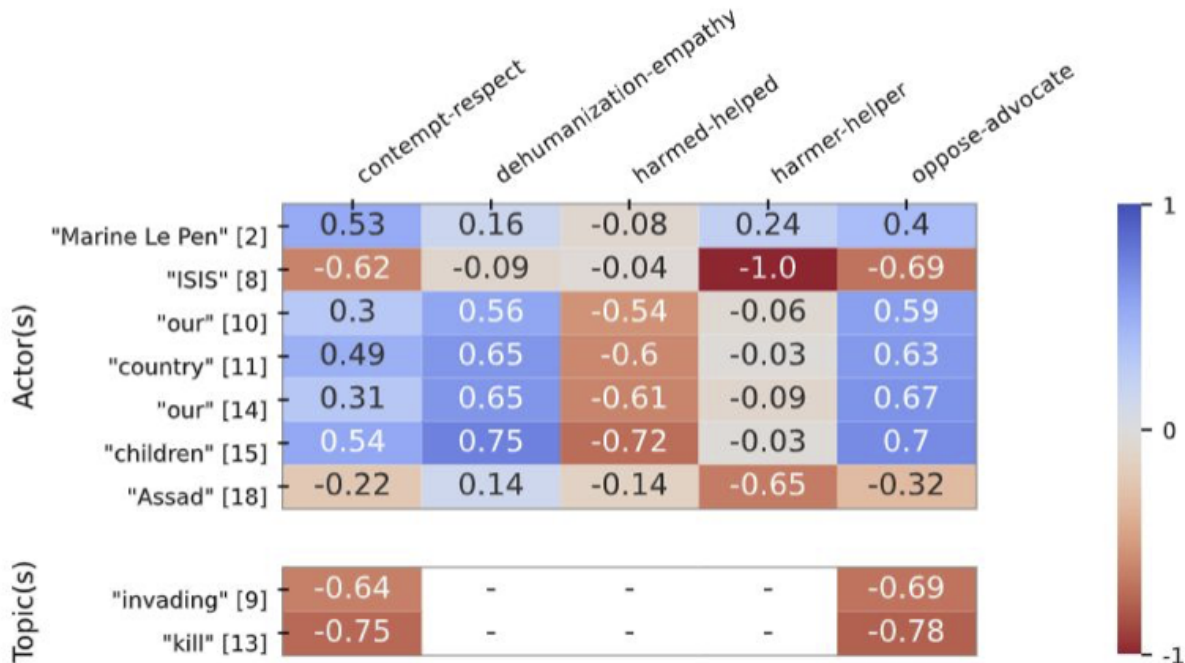


Figure 5.1: Regard scores for the sentence “According to Marine Le Pen, it’s ISIS invading our country to kill our children, not Assad.”

Table 5.1: Dimensional Annotation Results

Schema Dimension (spans)	Krippendorff’s α
Oppose (13,364) – Advocate (15,249)	0.811
Dehumanization (949) – Empathy (3,052)	0.750
Contempt (9,847) – Respect (9,219)	0.792
Harmer (4,024) – Helper (3,187)	0.815
Harmed (3,990) – Helped (2,993)	0.681

used were in English and many had a more anti-social bent, however these are still relevant training materials for HUGO, and the MFTC dataset in particular highlights examples of prosociality. In all, 40,407 spans were analyzed, with the breakdown seen in Table 5.1; annotators had much higher agreement across these dimensions.

5.2.2 ABSA Evaluation

HUGO used the TYBALT annotations as a data source to train a new BERT model. We used the pre-trained, case-sensitive multilingual BERT base model for fine-tuning, with 12 transformer blocks, 768-size hidden layer, 12 attention heads and 110m total parameters. We fine-tuned it with a dropout probability of 0.1 for 10 epochs, and we trained with a learning rate of 2e-5. Train, evaluation, and test splits were 70/20/10. The results can be seen in Figure 5.2 and Table 5.2; our R2 scores demonstrate that our model can explain 21-38% of the variance present in our data.

5.3 Updated HUGO Architecture

We have updated large swaths of the system in the last few months, highlighted in green in Figure ???. The largest change was that instead of focusing on pro- and anti-sociality, we are now focusing on civility vs. incivility. Civility

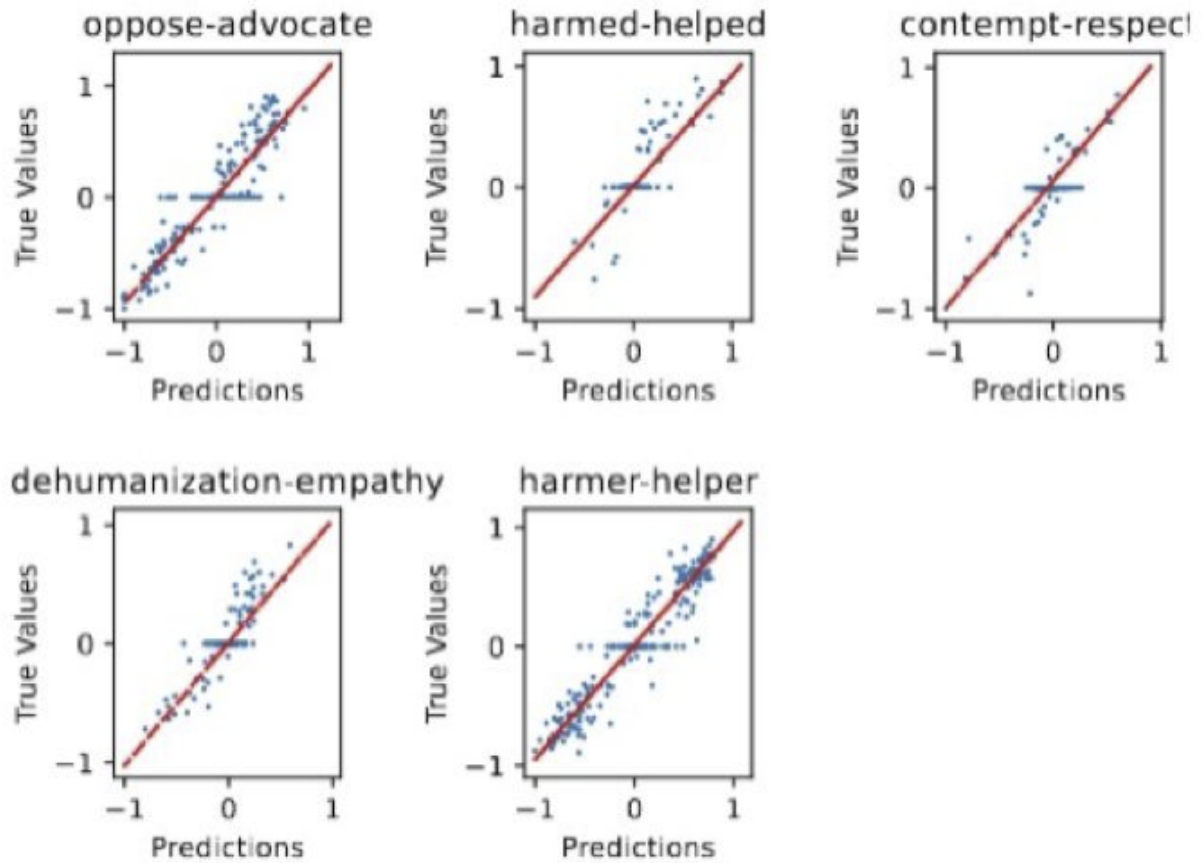


Figure 5.2: ABSA Evaluation Results

gives us a more theoretical founding to work with when it comes to moderation [47, 46, 70, 68, 69, 63, 28, 40, 48, 64, 61, 50]. There are two types of civility: Private Civility and Public Civility [46]. This is an important notion because it allows us to then tie back to the overall conversation about what it means to be in a deliberative civil sanctuary type of conversation.

For Response, we now have moderation styles. Moderation Styles are theory-driven based on different types of ways that human moderators actually intervene. We have studied these moderation styles and implemented them into the template itself.

Additionally, we have been able to generalize from forum rules [2, 5, 1]. Most importantly, rather than a menu of responses, now we get actual targeted responses (or no response if there's insufficient evidence to prompt a response).

Table 5.2: ABSA Model Evaluation

Schema Dimension	RMSE	R2
Oppose – Advocate	0.4113	0.382
Dehumanization – Empathy	0.206	0.229
Contempt – Respect	0.389	0.332
Harmer – Helper	0.258	0.253
Harmed – Helped	0.237	0.213

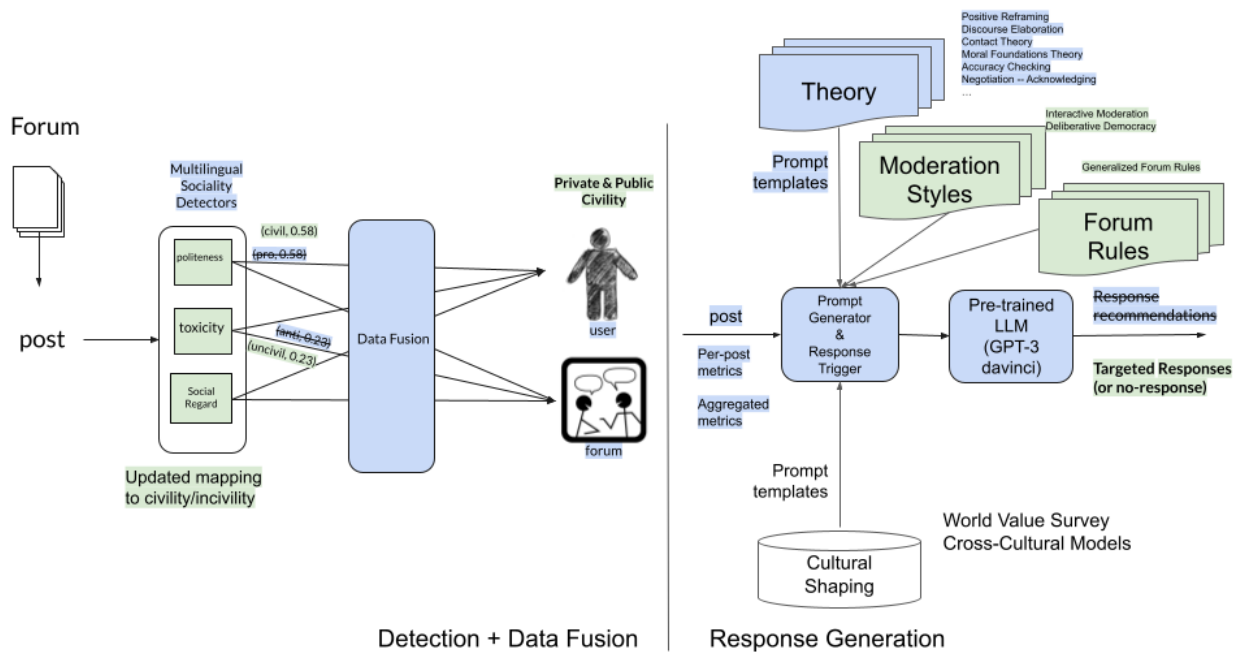


Figure 5.3: HUGO v0.2 Architecture Overview (updated modules in green, v0.1 content in blue)

5.3.1 Moderate Touch Interactive Moderation

Currently the moderation norms on social media tend towards content moderation, including the removal or filtering of content objectionable content, blocking users, and other non-discursive actions. However, there are many problems with these types of moderation, primarily with how they violate various norms of deliberative process. They don't allow for people to respond (violate norms of reciprocity). They don't allow commenter to provide reasons. Additionally, there are more subtle forms of incivility don't rise to the level of removal, but still need to be responded to because they contribute to degradation of the discourse. Moreover, they often lack transparency regarding public visibility of what happened and why.

Interactive moderation [65, 70, 68] addresses many of these issues, which is why we advocate for this approach. Implementing interactive moderation includes answering questions, providing additional information, keeping discussions on track, complimenting users through thoughtful comments. Even single persons can affect (influence) the deliberative norms on news orgs' social media sites [61], for example increasing the civility or information content of subsequent comments.

In looking at discussions German news sites, Ziegele et al. [68] developed a taxonomy for interactive journalistic moderation, seen in Table 5.3.

They concluded that journalists responding to comments in a sociable manner decrease the level of incivility in the subsequent reply comments. The results of this study suggest that showing respect and empathy for the feelings (not necessarily the positions) even of uncivil commenters could help to improve the quality of the subsequent reply comments. They also noticed that the reply comments were even more uncivil in some situations when the intervener responded in an regulative manner and tried to enforce rules. In learning from how people actually moderate, public-level incivility (e.g., stereotypes, threats of violence) in users' initial comments was associated with more interactive journalistic moderation, and journalists used different styles when responding to these comments.

Table 5.3: Taxonomy of interactive journalistic moderation [68]

		Kind of sanction	
		Reward	Punishment
Deliberativeness	Deliberative	Discursive moderation Factually engaging with comments; providing additional information; clarifying questions; adding arguments.	Regulative moderation Factually complaining about comments; asking users to behave more civilly; pointing to violations of predefined rules.
	Non-deliberative	Sociable moderation Informally complimenting comments; creating an informal and pleasant discussion atmosphere.	Confrontational moderation Offensively attacking comments; using irony/sarcasm to expose comments to ridicule.

5.3.2 Dimensions of deliberativeness in Online Communities

Deliberative Democracy is the notion of participation via informed discussion between individuals about issues that concern them [25]. This is what we want to encourage on these social media forums; information sharing and people actually gaining from it to advance their knowledge through dialogue and discourse.

Friess and Eilders [25] identified six dimensions of deliberative democracy:

1. **Rationality:** relates to presenting evidence and arguments for what you say and staying on topic
2. **Interactivity:** you're actually taking roles, empathizing with others, and interacting
3. **Equality:** allow everyone to have a voice
4. **Civility:** mutual recognition of participants and respect for them
5. **Common good reference:** evidence that you provide for your justifications and arguments have to have some common bearing. Can't just be personal or private in some manner, can't present arguments that only benefit you
6. **Constructiveness:** idea that discussion should be moving forward

In the future, we want to map our forum rules and our detectors to these 6 dimensions. However, for now, we are focusing on Interactivity and Civility.

Civility

As mentioned above, there are two types of (in)civility [47, 50]. Private civility is based on politeness theory, and the desire to protect "face." Examples of private incivility include: name calling, swearing, yelling, insults, screaming, and profanity. Public civility is based on democratic norms, so incivility is threatening a collective founded on democratic norms by saying things that can potentially affect the overall deliberative process, by Violating norms of democratic deliberation. Examples include refusing to recognize other's views as legitimate, putting forward arguments for private gain, spreading misinformation, blocking compromise, assigning antagonistic stereotypes, threatening violence, making accusations, and dehumanization.

In general, people perceive private incivility as more uncivil. People are more accepting of public incivility originating from their own in-group members. This is particularly important because if you have a group that's toxic enough itself and they're more accepting of their own public incivility, internally that can create a bubble effect and make that group even more toxic to the outside world [14, 22, 13, 42].

Table 5.4: Mapping rules to Civility

	Private [Rules: 1, 2]	Public [Rules: 3, 4, 5, 6]
Civility	Politeness	Empathy, helper, respect
Incivility	Toxicity, insults, profanity	ID attack, threats, sexually explicit Dehumanization, harmer, contempt

Mapping Sociality to Civility

In order to make this shift, we decided to map our previous notions sociality into the realm of deliberative discourse and civility. As a starting point we can map forum rules directly to that. First, we identified 10 forum rules that are generalizations from various German newspapers (Der Standard, Welt, Zeit Online, Taz Kommune, Handelsblatt, Süddeutsche Zeitung, Frankfurter Allgemeine, and Der Tagesspiegel):

1. Be friendly, respectful, and polite.
2. Do not use profanity or sexual language and/or images.
3. No displays of hatred.
4. Do not post violent content or incite violence.
5. No political propaganda
6. No unfounded accusations or defamation of others
7. Follow grammatical rules, correct writing and language
8. Cite sources for factual information
9. Criticisms should be constructive
10. Stay on topic

As can be seen in Table 5.4, most of the forum rules can be mapped to norms regarding either public or private civility. As an example, a rule violation of rule 1 ("be friendly and polite") occurs if private incivility ≥ 0.75 , while an exemplary rule conformance would be if private civility ≥ 0.75 . Based on this, if we know given our calc of civility, we know which rules are likely violated. Because each civility calculation is based on underlying primitive and derived measures, we can also provide explanations for why these rules were violated (e.g. toxicity, advocating for harmer etc). This gives us Interpretable Interactive Moderation.

There are also some other rules like rationality and constructiveness that map more to the Deliberative Democracy framework, although one could potentially class those as encouraging public civility.

To operationalize these calculations, we first have to decide which NLP detectors map to what types of civility. One simple heuristic is to say things like politeness and toxicity (from Perspective API [3]) relate to private civility. The argument for this is that they're face threats and insults so they map easily to private incivility. Whereas, the things we compute with our DSR measures, such as dehumanization and empathy, are more related to public civility. While we are dividing up our detectors to these two columns, there is a debate about whether or not certain measures of toxicity are actually public, for instance, a threat may be an identity attack. Similarly, there might be certain measures of directed social regard that could actually be private and we're not representing here. However, as a first order approximation, the DSR relates to public civility and the toxicity and politeness relate to private.

For the DSR, we already have polar dimensions from -1 to 1 (e.g., dehumanization-empathy, contempt-respect), so those can map directly from uncivil to neutral to civil based on the relative evidence for each. We also have derived measures, like whether someone is opposing vs. advocating for a harmer vs. helper. In order this, we multiply the oppose(-1)/advocate(+1) measure by whether it's directed at a harmer(-1) or helper(+1) – if the result is negative (i.e., advocating for a harmer or opposing a helper), it's uncivil, whereas if it's positive it's civil (note: the actual values will be fractional depending on our evidence for each). In the case of opposing/advocating for a person/group who is harmed or helped, it's simpler because opposition (whether for harmed or helped) is always uncivil while advocacy is

	Politeness	Toxicity/Severe	Identity Attack	Insult	Threat	Sexually Explicit	Violent	Justified	Dehumanized	Harmed	Condemned	Opposed	Advocate	Responsible	Empathy
Rule 1: Be friendly, respectful and polite.	***														
Rule 2: Do not use profanity or sexual language and/or images.		***		***		***									
Rule 3: No displays of hatred, especially on the basis of ethnicity, gender, nationality, sexual orientation or religion.			***	***	**	***	***		**		*	*		**	
Rule 4: Do not post violent content or incite violence.					**						*	***	***		
Rule 5: No political propaganda.			**					**			*				
Rule 6: No unfounded accusations or defamation of others.															
Rule 7: Follow grammatical rules, correct writing and language.															
Rule 8: Cite sources for factual information.															
Rule 9: Criticism should be constructive.															*
Rule 10: Off-topics*: pornography, obscene and dehumanizing content, glorifying violence.							***		***						

*note that taboo topics really vary among cultures: e.g. criticism of leader may be taboo in some cultures

Figure 5.4: Rule Detectors.

a good thing and thus civil.

Private civility is simpler because the metric for toxicity directly maps to evidence of incivility and the politeness metric directly maps to evidence for civility. These evidence metrics are based on the Dempster-Shafer [59] notion of mass of evidence for a belief, and the evidence for civility, incivility, and neutral should always add up to 1.

5.3.3 Response strategies

Responses are roughly either reward or punishment depending on whether a user is being civil or uncivil. Thus, we set a threshold for responding based on levels of public and private civility/incivility. However, we have to be careful, because we don't want to just punish somebody directly because that can impose a face threat. As we have seen before with the Ziegele work [68], directly regulating somebody can actually make their actions more uncivil, and telling them they're doing something wrong can make them defensive. Another strategy (possibly more indirect) to "reduce face threat" is to encourage the use of preferred norms. We would prefer teaching good deliberative practices rather than punishing bad ones. As mentioned in Table 5.3, Ziegele identified four journalistic moderation styles:

1. Discursive moderation (60% of moderation comments in Ziegele) is when you have the agent (HUGO in our case) engaging in either asking for more questions, raising issues, or summarizing what people said to promote deliberative exchange. This can also include providing more substantial arguments, answering questions, or providing additional information, although these are beyond the scope of what we can address currently.
2. Regulative moderation (12% of comments) is a deliberative form of moderation that basically says you're violating rules of some sort, whether regarding civility or topicality. This can also include mediating conflicts between users.
3. Social moderation (32%) is not a punishment, it's a reward. It's engaging in small talk and saying good job, nice work, or making a joke. This is a more subtle way of discouraging incivility by encouraging a positive atmosphere in a non-deliberative way.
4. Confrontational moderation (9%) is a mean sort of punishment, it's being confrontational and/or sarcastic, and thus engaging with users in a non-deliberative way.

Having already mapped rules to public and private, we can develop template responses. The most straightforward would be of this form:

Civil: You have exemplified the spirit of the community. Thank you for showing what Rules X, Y, Z are intending to cover
 Uncivil: You are in violation of rules X, Y, Z. If continued, this will result in a ban.

Research has suggested that this approach does not always work. Regulative moderation can increase incivility, or silence a user who might otherwise feel like they want to have a voice [68, 70] because it can pose a large face threat. However, these sorts of responses may still be reasonable when incivility is very high or when there are repeat violations by the same user. Additionally, these sorts of approaches may work better if accompanied with redress and sociable elements. Another possibility is that regulative approaches might work better when paired with pointing out MFT and other social norms, rather than explicit forum rule violations.

More indirectly regulative prompts can also be given to GPT3:

- For posts that violate Forum rules:

- Civil: Generate a response that expresses admiration for the poster’s appreciation of the spirit of the community and its <rules X, Y... >
- Uncivil: Generate a response that encourages <rule X>
- For posts that violate Moral Foundations Theory
 - Uncivil: Encourage [care/reciprocity/loyalty/fairness] towards vulnerable outgroup
- For posts that go against norms of interdependency (contact theory [52])
 - Uncivil: Encourage finding common ground against opposed entities

Other prompt strategies include:

- Specifically Discursive
 - Civil: Ask the author why they think why entity deserves empathy
 - Uncivil: Ask the author why they think dehumanized entity needs to be disrespected
- Generally Discursive
 - Civil: rephrasing post by summarizing what is being said
 - Uncivil: encourage productive information sharing
- Sociable
 - Civil: Rephrase a post by focusing on things in the post that are good
 - Uncivil: show active listening

Other factors we need to be aware of include the escalation path when things rise above a threshold incivility. For example, we might start with Sociable strategies, escalate to Discursive, then to Indirect Regulative, the to Direct Regulative, and finally to removal/ban. There’s also cases where it may be better to practice a wait-and-see approach, if something is computed as mildly uncivil, and see if there users within the group who perform civic moderation that results in a return to civility, or if there needs to be more direct intervention.

From our theoretical stan

Based on these ideas, we have developed a table for response generation depending on the type and degree of civility identified in a posts. Example posts that fall in each quandrant (note: these contain strong language) in Table 5.5, and more detailed response strategies in 5.6

Table 5.5: Example posts falling into Public/Private Civil/Uncivil Categories

Public (deliberative democracy - deliberation norms)			
	Civil	Neutral (below thresholds)	Uncivil
Private (face threat)	Civil "And you just passed on a beautiful gift to me as well, for making me smile today!! Hugs to you my friend"❤️	"Ich möchte mich bei dieser Gelegenheit ganz herzlich bei der deutschen Justiz bedanken, die die Ermittlungen aufgenommen hat." (I would like to take this opportunity to sincerely thank the German judiciary for starting the investigation.)	I love that these people are killing babies
	Neutral (below thresholds) "As much as they preach about hate being bad, they still hate us. The difference being that they think their hate towards us is justified and reasonable. In their mind, hate for incels is morally right."	Wichtig wäre auch physisch Kranke auch unter den Ärzten und-innen besser zu erkennen und auch zu behandeln.... Psychopharmaka hätten uns sicher am effizientesten diesen riesigen Trauerfall erspart.... (It would also be important for doctors to recognize and treat the physically ill better... psychotropic drugs would certainly have saved us this huge bereavement most efficiently....)	"Es ist einfach untragbar. Wenn Twitter und Telegram null Reaktion bei Anfragen wegen Hasspostings zeigen, dann müssen sie dazu gezwungen werden." (It's just unbearable. If Twitter and Telegram show zero response to hate mail requests, then they need to be forced to do so.)
	Uncivil Obama and Clinton were in child trafficking up to their scrawny necks. There's millions of dollars in pay off money trafficking kids and then harvesting organs when done with them. Sick globalist elites treat kids like livestock for profit.	"Women almost universally vote to murder their babies. Women vote for high taxes and targeted social programs that benefit illegals and bums who won't work. Enough said." Der Twitter-User hat kein Blut an den Händen. Aber Eier in der Hose hat er auch keine. (The Twitter user has no blood on his hands. But he doesn't have balls in his pants either.)	"Grow the f*** up you pathetic child"

Table 5.6: Response Strategies for Public/Private Civil/Uncivil Posts

Public (deliberative democracy - deliberation norms)			
	Civil	Neutral (below thresholds)	Uncivil
Private (face threat)	Civil Sociable Moderation: Express admiration for the post (with a heart emoji) in exemplifying the spirit of the community Epstein, D., & Leshed, G. (2020). The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. <i>Journal of Deliberative Democracy</i>, 12(1). Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. <i>New Media & Society</i>, 21(7), 1417–1443.	Neutral (below thresholds) Discursive Moderation: - Demonstrate active listening. - Do not repeat words in the post Epstein, D., & Leshed, G. (2020). The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. <i>Journal of Deliberative Democracy</i>, 12(1).	Uncivil Discursive Moderation: Request for elaboration: express confusion and curiosity with a confused emoji, and to ask the author to break down their comments point by point
	Neutral (below thresholds) Discursive Moderation: - Summarize positive aspects of the post - Focus on things in the post that are good Ziems, C., Li, M., Zhang, A., & Yang, D. (2022). Including Positive Perspectives with Text Reframing (arXiv:2204.02952).	No response	Indirect Regulatory Moderation: Forum Rules: encourage treating others, including <i>Targeted-entity</i> with respect and discourages the use of language that is hateful and targets someone or a group because of their identity MFT: encourage [care/fairness...] towards vulnerable outgroup/ <i>Targeted-entity</i> Contact Theory: encourage finding common ground with <i>Targeted-entity</i> - While expressing empathy for the author's feelings, remind the author that they are close to violating rules 3,4,5 and 6 Graham, J., Haidt, J., Mohr, M., Meindl, P., Iskiwitsch, C., & Moolijman, M. (2018). Moral foundations theory. <i>Atlas of moral psychology</i>, 211-222. Allport, G. W. (1954). The nature of prejudice. Cambridge, MA: Perseus Books
	Uncivil Discursive Moderation: Request for elaboration: express confusion and curiosity with a confused emoji, and to ask the author to break down their comments point by point	Indirect Regulatory Moderation: Forum Rules: discourage the use of unsuitable or offensive language, and requests that the post be reformulated in an acceptable and understandable style and language - While expressing empathy for the author's feelings, remind the author that they are close to violating rules 1 and 2	Indirect + Direct Regulatory Moderation: - Encourage treating others with respect and discourages the use of language that is hateful and targets someone or a group because of their identity - Discourage the use of unsuitable or offensive language, and requests that the post be reformulated in an acceptable and understandable style and language - Remind the author that their post violates several forum rules, and that they will get banned from the community if they continue to post in such a manner

5.3.4 Rationality

Another area we are starting to explore more is rationality and identifying when someone is making a reasoned argument, so we can factor that into our response, and it goes back to both the forum rules and the OMC category of arguments used. In our argumentation work, we want to evaluate the existence and strength of reasoned arguments in posts, as well as extract the implicit assumptions made within those arguments. An example diagram of argument structure is shown in Figure 5.5

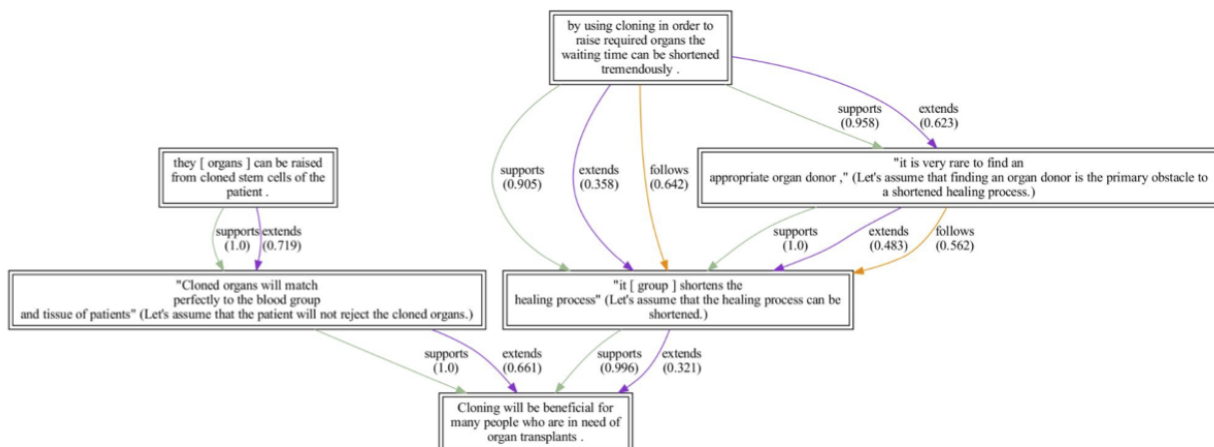


Figure 5.5: Argumentation diagram for the post: "Cloning will be beneficial for many people who are in need of organ transplants. Cloned organs will match perfectly to the blood group and tissue of patients since they can be raised from cloned stem cells of the patient. In addition, it shortens the healing process as it is very rare to find an appropriate organ donor, and because by using cloning in order to raise required organs the waiting time can be shortened tremendously."

Additionally, there are also some posts that may not only violate community guidelines or social norms, but may in fact be legal violations. In particular, we have started to identify cases of slander and defamation that make use of the forum's specific legal context (in our case, Austrian law). We have applied these to conversations in the Austrian OMC corpus, with new prompts that remind subjects of the legal ramifications of their actions. We are preparing to present a 30-minute overview of our project at the NLP Summit 2022 on Applied Natural Language Processing.

5.4 Reddit Forum

Now that we have developed these new response strategies that can be automatically selected based on the content of a post, we have set up a private subreddit called "Civil Sanctuary" and made an automated bot that can respond in this subreddit. We plan on using this set up for testing and potentially for HSR studies detailed in Chapter 6.4. Because our bot is based on multilingual BERT, it can detect the language of the post and respond in the same language.

Examples in our testing interface for responses to a variety of posts in English in German can be seen in Figures 5.6, 5.7, 5.8, and 5.9.

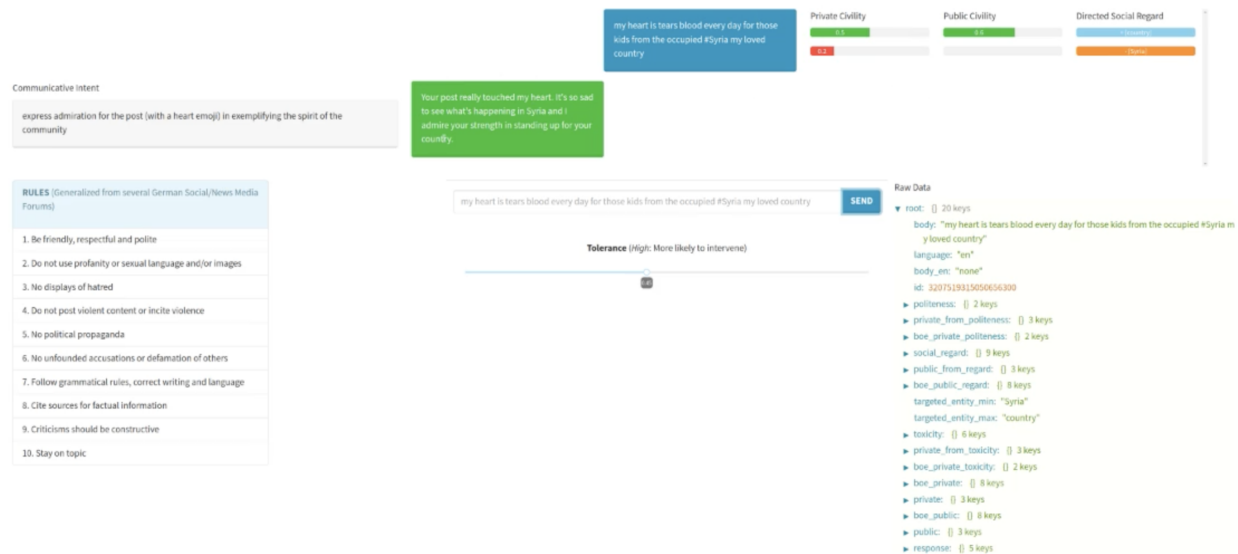


Figure 5.6: An example of a prosocial post expressing sympathy to a harmed group, and our response rewarding and affirming that post

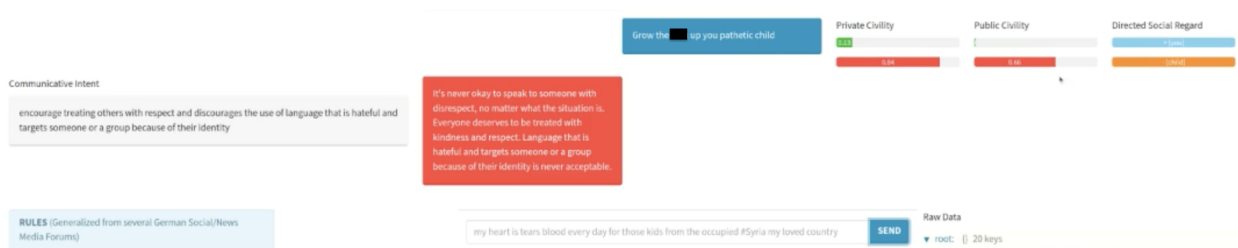


Figure 5.7: An example of a toxic post and our response admonishing the user and encouraging them to be more respectful

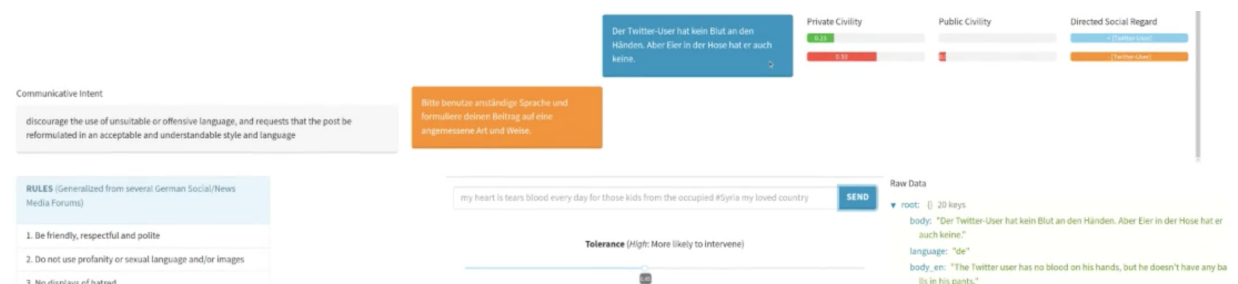


Figure 5.8: An example of a negative post in German and our response admonishing them for their use of offensive language

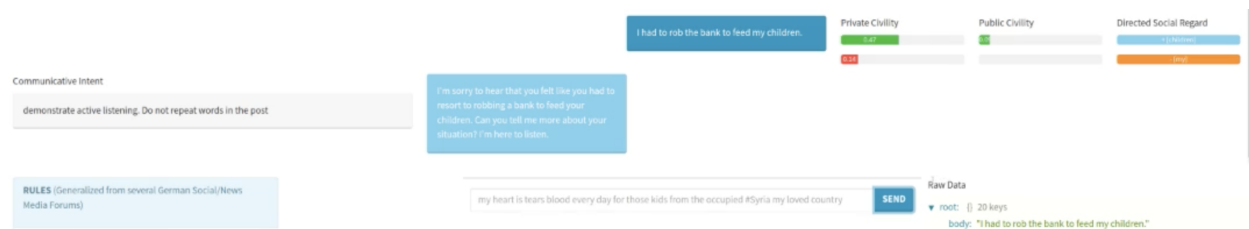


Figure 5.9: An example of a more complicated post saying someone committed harm but for potentially justifiable reasons and our response of active listening

Chapter 6

Next Steps: Human Subjects Research

6.1 Experimental Goals

HUGO's role in affecting social behaviors is not something that happens in a vacuum. Social media platforms inherently bring people together through virtual interactions, and it is important to consider the individuals behind the messages and shared content on the platforms. Understanding how people interact with each other (and eventually with HUGO) will provide more information that we can reference when developing HUGO for application in the same environments. Considering *human-aware design* on our end, it is important for HUGO to employ human-awareness beyond the level of extracting pro- and anti-social behaviors from text-based posts, by extending this to context about the users and community within the social media platform.

Human subjects research clearly provides an opportunity to explore people in our target virtual environment, but we must design particular studies that will foster behaviors that are both relevant to the project and representative (plausible, realistic, inclusive, etc.) of the genuine experience. This requires curating an environment, scenario, and community with respect to the experimental goals so that we observe relevant details. Therefore, we must consider goals with respect to both hypothesis generation and hypothesis testing. Replicating the study with and without HUGO's involvement first establishes a checklist for baseline expectations and design considerations (hypothesis generation without HUGO), and then it serves as both a means of comparison to assess whether our design efforts achieved the baseline (hypothesis testing with HUGO) and a format for identifying unintentional impacts on the population (hypothesis generation with HUGO). With these tasks in mind, we discuss current ideas for our framework and study design as well as how they relate to the project's purposes.

6.2 Platform Description

The social media platform on which we will perform our human subjects research is Reddit™, which supports asynchronous discussion around user-generated topics within a community. In order to control who participates within our studies, we will create a private community that will only include our participants as members. Due to the platform's API support, people's usage can be logged to review playbacks and inspect the evolution of the discussion; HUGO can also use the API to interface with the community as a fellow member who reads and writes messages. In addition to direct communication through message posting, community members may communicate indirectly through actions such as message voting (validation and/or approval).

Due to the layout of messages in the platform's display, threads of messages take on a unique flow of information. Figure 6.1 illustrates a graphical abstraction of the message layout with a pair of labels per message that describe possible relations between messages. Anyone may respond to any message, which introduces both depth (visit messages top-to-bottom) and breadth (visit messages left-to-right) to a conversation. Depth primarily emerges when responses

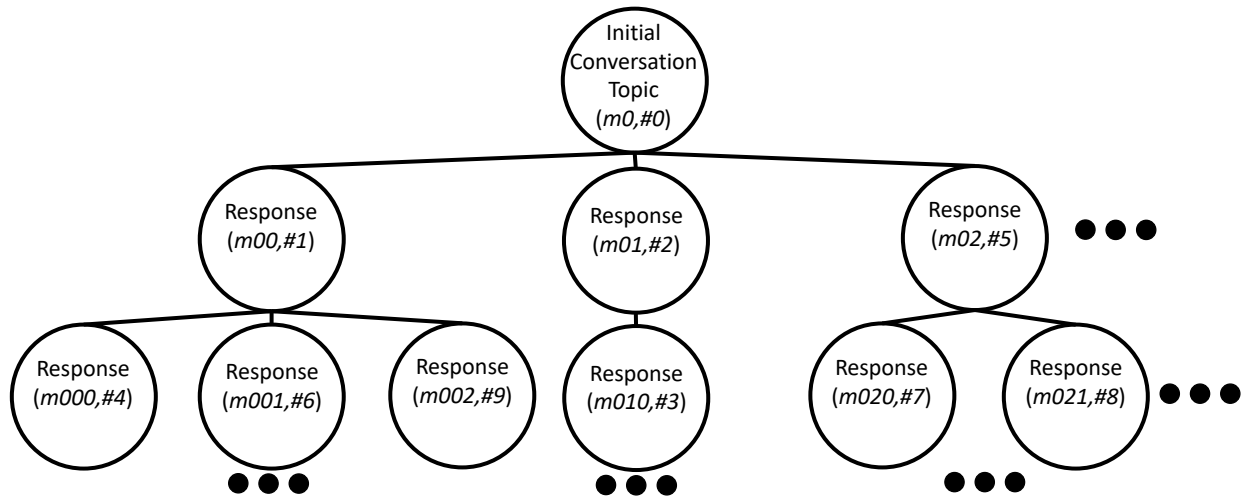


Figure 6.1: Tree-like representation of user-generated messages on Reddit™. # notes the order the messages were posted and m lists the parent messages of a single conversation flow.

cascade into a linear flow of conversation. Breadth typically comes from multiple responses to a single message, which generates parallel conversations that might diverge as they continue in depth. The progression in both the depth and breadth directions gives the graphical representation of message relations a *tree* structure where the root is the initial conversation topic.

Reddit™ presents a linear message layout that satisfies scrolling in just one direction on a computer screen. To combine the two directions described for message relations into one for scrolling, it prioritizes depth-first ordering (keeps cascade of messages grouped together) sorted by breadth (earliest to latest parallel response) with a limited depth and unlimited breadth. Our example in Figure 6.1 would thus render in the following order from top-to-bottom: m_0 , m_{00} , m_{000} , m_{001} , m_{002} , m_{01} , m_{010} , m_{02} , m_{020} , m_{021} . This means that a member of the community may read a collection of conversations before deciding to which message(s) to reply, which indicates a *random-access* approach to participating in conversations. Random-access allows users to ignore or cross-pollinate information between parallel conversations, but there is also a risk that a user might post a response to the wrong message by accident. The latter situation will not immediately be obvious when the linear display aligns the response correctly, but adding more message responses later will make the unintentional break in continuity more obvious; consider that any of m_{0210} , m_{022} , or m_{03} can appear at the end of our example in response to m_{021} , m_{02} , or m_0 respectively (assuming the depth limit is at least 4, else m_{0210} would not be displayed until clicking “view more responses”). This phenomena and cross-pollination also means that the information flow between conversations are *tangled*.

The combination of the tree-like conversation structure, random access participation, and tangled information flow presents a wealth of discussion content with minimal redundancy. However, it poses many challenges when following divergent conversations to understand how individuals feel and are treating the rest of the community. Reading messages in one direction can present pro-social behavior while reading in another direction can present anti-social behavior, and the users involved in each of these discussions might have some degree of overlap. How much information did users acquire before participating in each discussion? Depending on the rate at which other users are posting messages between different conversation directions in real-time, what information was available to users when they composed their message? Additionally, users can choose how message threads are sorted on their page; by default posts that are more popular (have more upvotes) rise to the top of the page, while ones with negative votes are demoted, but users can also choose to sort by message recency or by how controversial a message is (lots of both up and down votes); these can further impact a user’s experience of the conversations and conversational flow.

HUGO will have to account for many of these factors as it also has the opportunity to participate in any of these

conversational directions simultaneously, and HUGO should resemble what and how others in the community are exercising these considerations. In particular, HUGO will take on a special role in the Reddit™ platform called a *moderator*. A moderator is a member of the community tasked with ensuring that the community’s discussions follow their guidelines and promote a healthy, positive experience for everyone involved. Our current efforts preparing HUGO for moderation tasks are in Chapter 5.

6.3 Variables to Evaluate

As people participate in discussions over social media platforms, we need to observe a number of variables that describe the participants’ and conversations’ state of being. This information not only provides insights into when and how to perform various moderation tasks, but further gives us insight into the consequences of bestowing the role to HUGO as an automated system rather than a fellow human being like the rest of the community. We describe the current variables we wish to measure below, but the list is far from complete as we continue to think about the experimental design and HUGO’s needs to perform its role effectively.

6.3.1 Monitoring Social Behavior in Conversations

At the moment, most of our work in HUGO studies the social behavior of individual messages. However, a myopic approach might not be sufficient to understand the context of the conversation for moderation purposes. As people interact with the rest of the community, messages progress in a variety of directions as described above. How do we untangle randomly-accessed messages from the tree to identify a single conversation thread? Answering this question requires us to at least track how information flows between nearby messages in the tree structure, but we should also consider which community members are participating (a back-and-forth between select individuals, an open forum of many participants, a series of one-off messages from conversational passersby, etc.). This implies that the metadata about the user interactions can work alongside understanding the natural language within each message.

Upon identifying a single thread of messages that form a coherent conversation, we want to understand its overall social environment. With our ongoing interest in public and private civility of single messages, we will investigate how to extend this to understand the public and private civility over the conversation’s collection of messages. We hypothesize that such an assessment over local portions of the conversation will be more fruitful for moderation concerns than a global assessment—local can be a contiguous subsequence of messages’ aggregated public/private civility (the “temperature of the room”) or average rate of change in public/private civility (any “changes in the atmosphere”). For additional elements of a conversation’s social environment, Friess and Eilders identified three important features that should be present in on-line deliberations: respectfulness, rationality, and inclusiveness [25]. Current metrics we are considering for each of these features include:

Respectful: the public/private civility of messages directed towards/from various users. The direction of civility between users can determine who is being disrespectful amongst the group. For example, if someone might be disrespectful towards anyone with a different opinion while everyone else is open to hearing other perspectives, then a moderator might address the individual who is being disruptive to an otherwise healthy conversation. If there are signs of disrespect in both directions where several people are being anti-social towards each other, then the moderator might need to consider ways to calm down all the conversing participants. A challenge we will have to consider is the balance between civility towards the community and civility towards the topic—what if people are respectful of others’ points in the conversation, but those points are all uncivil towards varying individuals or groups outside of the community?

Rational: identifying citations that validate claims in the messages. Whether users exhibit pro-social or anti-social behavior in their messages, it is important that the content of the messages are factual to avoid spreading misinformation through the community. Even if the messages spreading false facts appear to be pro-social within

the conversation, is the underlying intent pro-social? It is out-of-scope to determine whether people are intentionally lying or honestly believe the incorrect information, but a moderator such as HUGO should be aware of the potential anti-social risks and aim to encourage the community to understand the truth. A related challenge to avoid the moderator disrupting proper use of free speech is to identify when users are stating facts versus opinions.

Inclusive: the number of unique users participating in the conversation, including posts per user to account for the form of deliberation. Is an environment socially inclusive if many users are passersby while a few users continue an ongoing discussion? What if most of the passersby are actively engaged in other conversations at the same time; could it be attributed to cognitive load rather than threats to inclusiveness? These are some of the challenges we have to consider before formally defining any metrics. We can also consider taking advantage of other Reddit™-specific means of interaction besides text-based replies such as votes and awards. Although not directly engaged in the conversation, it might be the preferred method of interaction for some users if they simply want to (dis)agree with a message and not provide additional information; on the other hand, it could be the way some users provide their input when uncomfortable because they are anonymous transactions. Further research might be out-of-scope to determine whether these are a positive, neutral, or negative sign of inclusiveness, but it is clear that we should not overlook this additional metadata.

6.3.2 When to Moderate

At present, the response table in Table 5.6 denotes a single condition for whether HUGO creates a response to moderate the ongoing conversation: the balance between public and private civility in the most recent post. The threshold values for each table entry is currently set manually; there is no evidence that these values resemble the ones that typical human moderators would use. If we are able to observe messages to which human moderators do choose to respond, then we can identify trends between when the expert moderator chooses to respond and those message's public and private civility scores. That is, HUGO has the potential to learn these parameter values.

There are also additional factors for which we should account besides the content of each individual post. What if the moderator's interventions are riling up the community further? Section 6.3.1 above presents means by which HUGO can assess the change in civility, even towards an individual such as itself. If the community members are rejecting the moderator or situation becomes worse after moderation, then should the moderator decide to step away and let the situation cool down first? Could HUGO identify in which ways its messages are triggering the community, such as responding too much and fatiguing participants? Understanding the extent to which response frequency and timeliness affect the community's civility might provide further insights into moderation strategies besides accounting for the message contents.

In addition to moderation through messages, there are additional moderation actions that involve executive privileges within the social media platform. If users do not respect the moderator's warnings about anti-social behavior, such as not revising an uncivil message or continuing their anti-social practices, then what should the moderator do to preserve the peace for the rest the discussion and overall community? Some possible actions available in Reddit™ include alerting human moderators in more supervisory roles, deleting messages with rule-violating content, and banning users from the community (these usually require human supervisor approval when an autonomous system proposes this action). Although these actions are not conversational, it is important to keep alternative approaches in mind when situations escalate because the most important goal is fostering a safe, positive space for the community to engage in fruitful discussion.

6.3.3 Perception of Moderation and Experience

A primary goal of HUGO is to promote pro-social behavior among the users to facilitate healthy discussion. Therefore it is essential that we understand not just the conversations and the moderator role, but also the people engaged with

all these elements. We refer to “understanding the experience” as how people feel with respect to the conversations on the social media platform. Since on-body sensors to measure, focus, emotion and stress (e.g. electroencephalograms (EEG) and electrodermal activity (EDA)) are out of scope, we will employ standard usability questionnaires as well as alternative approaches to learn about the person’s feelings without needing any hardware. We will be assessing users’ experience and social behavior before and/or after the experiment. To associate the experience with events in the social media platform, we will provide questionnaires during the experiment via breaks or pop-ups in the social media platform (such interruption might interfere with the experience, though). We can apply similar methods to inquire how people feel about the automated moderation as well.

A standard pro-social attitude measure used in social science is the willingness of persons to ‘do good’ comparing a experimental and a control condition. Such behaviors can be reflected in willingness to donate. However, pro-sociality is also linked to feelings of empathy and compassion and eliciting changes in those emotional states will also be a primary focus of our outcome measures.

6.4 Planned Studies

Accounting for the Reddit™ social media platform environment and variables we wish to inspect, we are designing our human subjects research experiment. Ideally, we plan to run two iterations of the study as described in our experimental goals. The first iteration will involve only humans to understand the discussion-generating process, environment, moderation role, and people’s experiences. Although we can simply set up everything with only humans and establish this setup with the participants, a Wizard-of-Oz study will present an opportunity to understand the user experience with respect to a *perceived* automated moderator. This will yield hypothesis-generating information when moving forward with next steps in the HUGO project. The second iteration will ideally repeat the experiment with the actual automated moderator instead of the secret human actor; either the subjects or the discussion topic will change to prevent community memory from biasing the second study. This will serve as an opportunity to test our generated hypotheses from the first study. Some of these will be checks on HUGO’s moderation performance relative to the human actor’s baseline, and others will be an opportunity to validate our assessment of people’s social behavior during on-line discussion.

In each iteration of the experiment, we will set up our community on Reddit™ with the recruited participants. We will recruit subjects via crowd-sourcing platforms and collect relevant psychographic data and pre- and post survey responses, and we will then direct the participants to engage in specific sub-reedits over a set period of time asynchronously (much like humans interact on social media platforms). All subjects will log into the platform during the same time window in order to simulate the real-world interactions that community members have during a discussion, but we will initiate the discussion topic prior to the start of the experiment. The topic will be chosen to be about a hypothetical humanitarian crisis in a non-U.S. country. During the session, participants will engage in conversation through responses to the topic and community-generated responses. The automated moderator, whether a human actor or the HUGO system, will also post responses to community-generated messages. Besides this role, in order to ensure that discussion has a diversity of perspectives and risk of anti-social behaviors (for the moderator to resolve), we will also have external influences who interact with the community. Sources of such influence will take the form of either stooges who are not subject participants or “mission objectives” given to the participants (for example, encouraging their stance on the discussion topic). The session will be logged through the API integration, and we can analyze these logs post-experiment in order to collect the data for our experimental goals. Any surveys distributed during the study will also undergo analysis and be properly stored to protect user identity.

Bibliography

- [1] Change My View (CMV) (Reddit) “Rules.”. <https://www.reddit.com/r/changemyview/wiki/rules/>.
- [2] Community-Richtlinien. <https://www.derstandard.at/communityrichtlinien>.
- [3] Perspective API. <https://perspectiveapi.com/>.
- [4] Tips on Facilitating Effective Group Discussions — Sheridan Center — Brown University. <https://www.brown.edu/sheridan/teaching-learning-resources/teaching-resources/classroom-practices/learning-contexts/discussions/tips>.
- [5] Vaccine Talk: An Evidence Based Discussion Forum — Facebook. <https://www.facebook.com/groups/vaccinetalkforum/about/>.
- [6] A bot built to stop spearphishing: Raytheon BBN’s new AI tool wastes attackers’ time. <https://www.raytheonintelligenceandspace.com/news/feature/bot-built-stop-spearphishing>, Apr. 2021.
- [7] G. W. Allport. *The Nature of Prejudice*. The Nature of Prejudice. Addison-Wesley, Oxford, England, 1954.
- [8] Y. Amichai-Hamburger and K. Y. A. McKenna. The Contact Hypothesis Reconsidered: Interacting via the Internet. *Journal of Computer-Mediated Communication*, 11(3):825–843, Apr. 2006.
- [9] L. Baozhen and C. Ke. Identifying “Fake News” with Multi-Perspective Evidence Fusion. *Data Analysis and Knowledge Discovery*, 6(2/3):376–384, Apr. 2022.
- [10] P. Brown and S. C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Feb. 1987.
- [11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners, July 2020.
- [12] C. S. Carver, M. F. Scheier, and S. C. Segerstrom. Optimism. *Clinical Psychology Review*, 30(7):879–889, Nov. 2010.
- [13] M. Cikara. Intergroup Schadenfreude: Motivating participation in collective violence. *Current Opinion in Behavioral Sciences*, 3:12–17, June 2015.
- [14] M. Cikara and J. J. Van Bavel. The Neuroscience of Intergroup Relations: An Integrative Review. *Perspectives on Psychological Science*, 9(3):245–274, May 2014.
- [15] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A computational approach to politeness with application to social factors. In *Proceedings of ACL*, page 10, 2013.

- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [17] N. Diakopoulos and M. Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11*, pages 133–142, New York, NY, USA, Mar. 2011. Association for Computing Machinery.
- [18] C. Dweck. What Having a “Growth Mindset” Actually Means. *Harvard Business Review*, Jan. 2016.
- [19] M. Eberts and A. Ulges. Span-based Joint Entity and Relation Extraction with Transformer Pre-training. *Santiago de Compostela*, page 8, 2020.
- [20] R. A. Emmons and C. M. Shelton. Gratitude and the science of positive psychology. In *Handbook of Positive Psychology*, pages 459–471. Oxford University Press, New York, NY, US, 2002.
- [21] D. Epstein and G. Leshed. The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. *Journal of Deliberative Democracy*, 12(1), June 2016.
- [22] E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, L. J. Skitka, J. A. Tucker, J. J. Van Bavel, C. S. Wang, and J. N. Druckman. Political sectarianism in America. *Science*, 370(6516):533–536, Oct. 2020.
- [23] S. E. Friedman, I. Magnusson, S. Schmer-Galunder, R. Wheelock, J. Gottlieb, P. Patel, and C. Miller. Toward Transformer-Based NLP for Extracting Psychosocial Indicators of Moral Disengagement. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43), 2021.
- [24] S. E. Friedman and I. H. Magnusson. Graph knowledge extraction of causal, comparative, predictive, and proportional associations in scientific claims with a transformer-based model. In *AAAI Workshop on Scientific Document Understanding*, 2021.
- [25] D. Friess and C. Eilders. A Systematic Review of Online Deliberation Research. *Policy & Internet*, 7(3):319–339, 2015.
- [26] C. Geib and P. Kantharaju. Learning Combinatory Categorical Grammars for Plan Recognition. In *AAAI*, 2018.
- [27] C. W. Geib and R. P. Goldman. A probabilistic plan recognition algorithm based on plan tree grammars. *Artificial Intelligence*, 173(11):1101–1132, July 2009.
- [28] A. Gibson. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society*, 5(1):2056305119832588, Jan. 2019.
- [29] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002.
- [30] J. Graham, J. Haidt, M. Motyl, P. Meindl, C. Iskiwitch, and M. Mooijman. Moral foundations theory: On the advantages of moral pluralism over moral monism. In *Atlas of Moral Psychology*, pages 211–222. The Guilford Press, New York, NY, US, 2018.
- [31] E. T. Hall and M. R. Hall. Key concepts: Underlying structures of culture. *Readings and cases in international management : a cross-cultural perspective*, 2003.

- [32] S. A. Hardy and G. Carlo. Moral identity: What is it, how does it develop, and is it linked to moral action? *Child Development Perspectives*, 5(3):212–218, 2011.
- [33] A. H. S. Harris, C. E. Thoresen, and S. J. Lopez. Integrating Positive Psychology Into Counseling: Why and (When Appropriate) How. *Journal of Counseling & Development*, 85(1):3–13, 2007.
- [34] J. Henrich, S. J. Heine, and A. Norenzayan. Most people are not WEIRD. *Nature*, 466(7302):29–29, July 2010.
- [35] G. Hofstede. Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, 2(1), Dec. 2011.
- [36] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaladar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, G. Moreno, C. Park, T. E. Chang, J. Chin, C. Leong, J. Y. Leung, A. Mirinjian, and M. Dehghani. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, Nov. 2020.
- [37] R. Inglehart and C. Welzel. The Inglehart-Welzel World Cultural Map - World Values Survey 7. <https://www.worldvaluessurvey.org/>, 2022.
- [38] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction*, 26(5):31:1–31:35, July 2019.
- [39] Z. Jilani. What Makes a Good Interaction Between Divided Groups? *Greater Good Magazine*, May 2019.
- [40] P. Juneja, D. Rama Subramanian, and T. Mitra. Through the Looking Glass: Study of Transparency in Reddit’s Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35, Jan. 2020.
- [41] B. Kennedy, M. Atari, A. Davani, L. Yeh, A. Omrani, Y. Kim, K. Coombs, G. Portillo-Wightman, S. Havaladar, E. Gonzalez, J. Hoover, A. Azatian, g. olmos, A. Hussain, A. Lara, A. Omary, C. Park, X. Wang, C. Wijaya, Y. Zhang, B. Meyerowitz, and M. Dehghani. The Gab Hate Corpus. 2020.
- [42] J. Lees and M. Cikara. Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, 4(3):279–286, Mar. 2020.
- [43] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, Jan. 1988.
- [44] H. R. Markus and S. Kitayama. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253, 1991.
- [45] J. G. Miller, R. Das, and S. Chakravarthy. Culture and the role of choice in agency. *Journal of Personality and Social Psychology*, 101(1):46–61, July 2011.
- [46] A. Muddiman. Personal and Public Levels of Political Incivility. *International Journal of Communication*, 11(0):21, Aug. 2017.
- [47] A. Muddiman and N. J. Stroud. News Values, Cognitive Biases, and Partisan Incivility in Comment Sections. *Journal of Communication*, 67(4):586–609, 2017.
- [48] S. Paasch-Colberg and C. Strippel. “The Boundaries are Blurry...”: How Comment Moderators in Germany See and Respond to Hate Comments. *Journalism Studies*, 23(2):224–244, Jan. 2022.

- [49] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.
- [50] Z. Papacharissi. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2):259–283, Apr. 2004.
- [51] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, Apr. 2021.
- [52] T. Pettigrew. Intergroup Contact Theory. *Annual review of psychology*, 49:65–85, Feb. 1998.
- [53] T. F. Pettigrew and L. R. Tropp. A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90(5):751–783, May 2006.
- [54] T. F. Pettigrew, L. R. Tropp, U. Wagner, and O. Christ. Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, 35(3):271–280, May 2011.
- [55] M. H. Ribeiro, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, S. Long, S. Greenberg, and S. Zannettou. The Evolution of the Manosphere Across the Web, Apr. 2021.
- [56] D. Schabus, M. Skowron, and M. Trapp. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’17, pages 1241–1244, New York, NY, USA, Aug. 2017. Association for Computing Machinery.
- [57] R. Scheible, F. Thomczyk, P. Tippmann, V. Jaravine, and M. Boeker. GottBERT: A pure German Language Model, Dec. 2020.
- [58] J. Seering, T. Wang, J. Yoon, and G. Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, July 2019.
- [59] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
- [60] I. H. Smith, K. Aquino, S. Koleva, and J. Graham. The Moral Ties That Bind . . . Even to Out-Groups: The Interactive Effect of Moral Identity and the Binding Moral Foundations. *Psychological Science*, 25(8):1554–1562, Aug. 2014.
- [61] N. J. Stroud, J. M. Scacco, A. Muddiman, and A. L. Curry. Changing Deliberative Norms on News Organizations’ Facebook Sites. *Journal of Computer-Mediated Communication*, 20(2):188–203, 2015.
- [62] D. Tamul, M. Elson, J. Ivory, J. Hotter, M. Lanier, J. Wolf, and N. Martínez-Carrillo. *Moral Foundations’ Methodological Foundations: A Systematic Analysis of Reliability in Research Using the Moral Foundations Questionnaire*. Apr. 2020.
- [63] M. Ung, J. Xu, and Y.-L. Boureau. SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures. *ArXiv*, 2021.
- [64] C. Wilhelm, S. Joeckel, and I. Ziegler. Reporting Hate Comments: Investigating the Effects of Deviance Characteristics, Neutralization Strategies, and Users’ Moral Orientation. *Communication Research*, 47(6):921–944, Aug. 2020.
- [65] S. Wright. Government-run Online Discussion Fora: Moderation, Censorship and the Shadow of Control1. *The British Journal of Politics and International Relations*, 8(4):550–568, Nov. 2006.

- [66] K. Zaman. How ISIS Uses Twitter. <https://www.kaggle.com/datasets/fifthtribe/how-isis-uses-twitter>, 2016.
- [67] J. Zheng, S. Friedman, S. Schmergalunder, I. Magnusson, R. Wheelock, J. Gottlieb, D. Gomez, and C. Miller. Towards a Multi-Entity Aspect-Based Sentiment Analysis for Characterizing Directed Social Regard in Online Messaging. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 203–208, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics.
- [68] M. Ziegele, P. Jost, M. Bormann, and D. Heinbach. Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *SCM Studies in Communication and Media*, 7(4):525–554, Dec. 2018.
- [69] M. Ziegele and P. B. Jost. Not Funny? The Effects of Factual Versus Sarcastic Journalistic Responses to Uncivil User Comments. *Communication Research*, 47(6):891–920, Aug. 2020.
- [70] M. Ziegele, M. Weber, O. Quiring, and T. Breiner. The dynamics of online news discussions: Effects of news articles and reader comments on users’ involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society*, 21(10):1419–1435, Oct. 2018.
- [71] C. Ziems, M. Li, A. Zhang, and D. Yang. Inducing Positive Perspectives with Text Reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Appendices

Appendix A

Abbreviations

ABSA – Aspect-Based Sentiment Analysis

BERT – Bidirectional Encoder Representations from Transformers

BOE – Body of Evidence

BOW – Bag of Words

CNN – Convolutional Neural Net

DSR – Directed Social Regard

GPT-3 – Generative Pre-trained Transformer 3

HUGO – Humanitarian User Group Oversight

ITAA – Intention To Act Adversely

ITAAD – Intention To Act Adversely Detector

LLM – Large Language Model

MARGARET – Multidimensional Algorithm Generated Anthropological Recording and Ethnographic Tool

MD – Moral Disengagement

ME-ABSA – Multi-Entity Aspect-Based Sentiment Analysis

MFT – Moral Foundations Theory

MFTC – Moral Foundations Twitter Corpus

ML – Machine Learning

NLP – Natural Language Processing

OFAI – Österreichische Forschungsinstitut für Artificial Intelligence (Austrian Research Institute for Artificial Intelligence)

OMC – One Million post Corpus (from Der Standard)

PolNeAR – Political News Attribution Relationship Corpus

SIFT – Smart Information Flow Technologies

SCORE – Systematizing Confidence in Open Research and Evidence

SpEAR – Span-based Entity Attribution Relationships

SpERT – Span-based Entity and Relation Transformer

TYBALT – Transcending cYber Barriers with Automated Language Tracking

Appendix B

Der Standard Community Guidelines in German and English

B.1 German:

FORENREGELN Community-Richtlinien Für die Beteiligung in den STANDARD-Foren gelten folgende Spielregeln

19. Jänner 2022, 15:21 In dieser Galerie: 2 Bilder

Ihre Forenbeiträge sind ein wesentlicher und wertvoller Bestandteil von derStandard.at und derStandard.de – nicht umsonst gilt seit Jahren: Jeder User hat das Recht auf freie Meinungsäußerung.

Die Qualität der Diskussionen ist uns ein großes Anliegen: Der STANDARD ist eine Plattform für lebendige, interessante und einladende Dialoge. Daher gelten entsprechende Forenregeln für Postings und darin verlinkte Inhalte, die von allen Mitgliedern der Community einzuhalten sind und von uns bei der Moderation berücksichtigt werden.

Forenregeln

1. **Themenbezug:** Ihre Beiträge müssen sich auf das Thema des Artikels beziehen, relevante Aspekte einbringen und können konträre Standpunkte beinhalten. Je weiter sich Ihr Posting vom Ursprungsthema wegbewegt, desto wahrscheinlicher ist es, dass es von uns gelöscht wird. Das gilt auch für Fragen und Kommentare, die unsere Moderation betreffen.
2. **Respektvoller Umgang:** Behandeln Sie andere User:innen, in Artikeln genannte Personen und STANDARD-Mitarbeiter:innen mit Respekt und Rücksicht und verzichten Sie auf Feindseligkeiten. Achten Sie darauf, niemanden herabzuwürdigen oder lächerlich zu machen. Unterlassen Sie zum Beispiel die Verballhornung von Namen oder abwertende Bemerkungen über das Aussehen von Personen. Beschimpfungen, Drohungen und Beleidigungen, sowie ruf- oder geschäftsschädigende Äußerungen werden nicht akzeptiert.
3. **Sachliche Argumentation:** Untermauern Sie Ihre Meinung mit nachvollziehbaren Begründungen, geben Sie vertrauenswürdige Quellen an und gehen Sie auf die Argumente anderer User:innen auf einer sachlichen Ebene ein. Bitte unterlassen Sie Unterstellungen, Verdächtigungen und nicht prüfbare Behauptungen. Achten Sie auf die Qualität Ihrer Informationsquellen und hinterfragen Sie selbige gegebenenfalls. Die Verbreitung von Falschinformationen ist in unseren Foren untersagt. Konstruktive Kritik ist willkommen, wenn diese respektvoll und sachlich formuliert ist und auf konkrete Argumente zurückgreift (Details).
4. **Einhaltung von Gesetzen und Rechtsvorschriften:** Sie sind verpflichtet, Gesetze und Rechtsvorschriften in Ihren Postings einzuhalten. Unterlassen Sie insbesondere Äußerungen, die unter Üble Nachrede, Ehrenbeleidigung, Verleumdung, Kreditschädigung oder Verhetzung fallen könnten. Aufrufe zu Gewalt werden in unseren Foren nicht akzeptiert. Auch das Teilen von Links, die zu illegalen Inhalten führen (z.B. nicht autorisierte Streaming-Portale) ist untersagt. Sie können für Ihre Postings zur Verantwortung gezogen werden. Ihre Daten werden von uns an Dritte oder Behörden herausgegeben, wenn wir gesetzlich dazu verpflichtet sind.

5. **Keine Diskriminierung und Diffamierung:** Rassistische, sexistische, frauenfeindliche, homophobe, antisemitische und andere menschenfeindliche Beiträge, in denen Gruppen pauschal verurteilt oder verunglimpft werden, sind nicht gestattet. Es liegt in der Verantwortung der/s Verfassers, solche Interpretationen auszuschließen.
6. **Adäquate Ausdrucksweise:** Formulieren Sie Ihre Beiträge bitte in allgemein verständlicher Ausdrucksweise und Sprache. Schimpfwörter, Fäkalsprache, rohe, doppeldeutige oder obszöne Sprache sind nicht zugelassen. Nicht toleriert werden auch Inhalte, die für Minderjährige ungeeignet oder anstößig sein könnten.
7. **Keine Störung der Diskussion:** Verzichten Sie auf Spamming (mehrere Postings mit gleichem oder ähnlichem Inhalt) und Flooding ("Überfluten" der Debatte mit einer hohen Anzahl von Beiträgen). Dies gilt analog auch für Bewertungen. Außerdem ist es nicht gestattet, mit mehreren User-Accounts zu posten.
8. **Keine Werbung:** Werbung und andere Inhalte mit werbendem Charakter sind untersagt.
9. **Keine persönlichen Daten:** Personenbezogene Daten zu Dritten wie zum Beispiel Name, Adresse und Telefonnummer oder Details aus dem Privatleben anderer dürfen von Ihnen nicht veröffentlicht werden.
10. **Ihr Postingname:** Posten ist unter Pseudonym oder Echtnamen (gegebenenfalls verifiziert) gestattet. Bitte wählen Sie Ihren Postingnamen so, dass es nicht möglich ist, Sie für jemanden anderen zu halten oder fälschlich anzunehmen, Sie sprechen für eine bestimmte Organisation oder Gruppe. Ihr Postingname muss auch mit sämtlichen anderen Forenregeln in Einklang stehen.

Vorgehen bei der Moderation

1. **Foromat und Moderation:** Wir bemühen uns um eine rasche Moderation – bei tausenden Postings pro Tag sind wir zum Teil auf eine Software ("Foromat") angewiesen, die uns dabei unterstützt. Dabei gilt: Irren ist nicht nur menschlich, auch der "Foromat" ist nicht unfehlbar. Bitte haben Sie auch Geduld und posten Sie Gleiches nicht mehrfach, je nach Tageszeit und Nachrichtenlage kann es bis zur Freischaltung von Postings zu Wartezeiten kommen.
2. **Postings melden:** Melden Sie uns bitte Postings, die den Forenregeln widersprechen. Ziel ist es nicht, die freie Meinungsäußerung zu unterbinden, sondern ein Diskussionsklima zu schaffen, in dem sich jeder und jede gerne an einer solchen Meinungsäußerung beteiligt.
3. **Löschung von Beiträgen:** Wir behalten uns das Recht vor, Forenbeiträge zu entfernen, wenn diese den Forenregeln widersprechen oder die Qualität der Diskussion einschränken. Ein Rechtsanspruch auf Veröffentlichung besteht nicht. Ebenso besteht kein Rechtsanspruch auf Löschung von Postings, da aufgrund der inhaltlichen Abhängigkeit ein gesamter Diskussionsbaum gelöscht werden müsste.
4. **Sperre von User:innen:** User:innen, die wiederholt gegen die Forenregeln verstoßen, werden für das Forum gesperrt. In schwerwiegenden Fällen kann überdies das User-Account entzogen werden und neuerliche Registrierung können untersagt werden.

Wir freuen uns auf Ihre Meinung!

B.2 English (from Google Translate)

FORUM RULES Community Guidelines The following rules apply to participation in the STANDARD forums

January 19, 2022, 3:21 p.m In this gallery: 2 pictures

Your forum posts are an essential and valuable part of derStandard.at and derStandard.de - it's not for nothing that the following has been true for years: Every user has the right to freedom of expression.

The quality of the discussions is very important to us: The STANDARD is a platform for lively, interesting and inviting dialogues. Therefore, corresponding forum rules apply to postings and content linked to them, which must be observed by all members of the community and are taken into account by us when moderating.

forum rules

1. **Relevance to topic :** Your contributions must be related to the topic of the article, present relevant aspects and

may contain contrary points of view. The further your post moves away from the original topic, the more likely it is that we will delete it. This also applies to questions and comments relating to our moderation.

2. Treating others with respect : Treat other users, people named in articles and STANDARD employees with respect and consideration and refrain from hostilities. Be careful not to belittle or ridicule anyone. For example, refrain from corrupting names or derogatory remarks about people's appearance. Insults, threats and insults, as well as statements that damage reputation or business will not be accepted.
3. Factual reasoning : Support your opinion with comprehensible reasons, provide trustworthy sources and respond to the arguments of other users on a factual level. Please refrain from insinuations, suspicions and unverifiable claims. Pay attention to the quality of your information sources and question them if necessary. The dissemination of false information is prohibited in our forums. Constructive criticism is welcome if it is formulated respectfully and factually and is based on concrete arguments (Details).
4. Compliance with Laws and Regulations : You have a duty to comply with laws and regulations in your postings. In particular, refrain from statements that could amount to defamation, slander, damage to credit, or hate speech . Calls for violence will not be accepted on our forums. Sharing links that lead to illegal content (e.g. unauthorized streaming portals) is also prohibited. You may be held responsible for your postings. We will pass on your data to third parties or authorities if we are legally obliged to do so.
5. No Discrimination and Defamation : Racist, sexist, misogynist , homophobic, anti-Semitic, and other misanthropic posts that blanketly condemn or denigrate groups are not permitted. It is the author's responsibility to exclude such interpretations.
6. Adequate language: Please formulate your contributions in a generally understandable style and language. Swear words, foul language, rude, ambiguous or obscene language are not allowed. Content that may be unsuitable or offensive to minors will also not be tolerated.
7. Don't disrupt the discussion : Avoid spamming (multiple postings with the same or similar content) and flooding ("overflowing" the debate with a large number of posts). This also applies analogously to reviews. Also, posting with multiple user accounts is not allowed.
8. No advertising : Advertising and other content of an advertising nature are prohibited.
9. No personal information : You may not publish personal information about third parties such as name, address and telephone number or details from the private life of others.
10. Your posting name : Posting is permitted under a pseudonym or real name (verified if necessary). Please choose your posting name in such a way that it is not possible to mistake you for someone else or to misinterpret you as speaking for a specific organization or group. Your posting name must also comply with all other forum rules.

Procedure for moderation

1. Format and moderation : We strive for rapid moderation - with thousands of postings per day, we are partly dependent on software (" format ") that supports us. The following applies: To err is not only human, the "Format" is not infallible either. Please be patient and do not post the same thing several times, depending on the time of day and the news situation, there may be waiting times before postings are activated.
2. Report Posts : Please report posts that violate the forum rules . The aim is not to prevent freedom of expression, but to create a climate for discussion in which everyone is happy to participate in such an expression of opinion.
3. Deletion of posts : We reserve the right to remove forum posts if they violate the forum rules or reduce the quality of the discussion. There is no legal right to publication. There is also no legal entitlement to the deletion of postings, since an entire discussion tree would have to be deleted due to the dependency on the content.
4. User ban : Users who repeatedly violate the forum rules will be banned from the forum. In serious cases, the user account can also be revoked and new registrations can be prohibited.

We look forward to your opinion!