Toward Transformer-Based NLP for Extracting Psychosocial Indicators of Moral Disengagement

Scott Friedman, Ian Magnusson, Sonja Schmer-Galunder, Ruta Wheelock,
Jeremy Gottlieb, Pooja Patel, Christopher Miller
{sfriedman, imagnusson, sgalunder, rwheelock, jgottlieb, ppatel, cmiller}@sift.net
SIFT, 319 N 1st Ave., Minneapolis, MN 55401 USA

Abstract

Moral disengagement is a mechanism whereby people distance or disconnect their actions from their moral evaluation. This work presents a novel knowledge graph schema, dataset, and transformer-based NLP model to identify and represent indicators of moral disengagement in text. Our graph schema is informed by Albert Bandura's psychosocial mechanisms of moral disengagement, including dehumanization, victimization, moral condemnation and justification, and attribution (or displacement) of responsibility. Our preliminary dataset is comprised of online posts from five different communities. We present initial evidence that (1) our theory-based schema can represent moral disengagement indicators across these communities and (2) our transformer-based NLP model can identify indicators of moral disengagement in text. As it matures, this thread of computational social science research can help us understand the spread of morally-disengaged language and its effect on online communities.

Keywords: moral disengagement; dehumanization; computational social science; NLP; hate speech; social media

Introduction

People have the capacity for compassion and cruelty toward others—and both at the same time—depending on their moral values and on whom they include and exclude in their category of humanity (Bandura, 1999, 2016). These are matters of *moral disengagement*, the psychosocial mechanisms of selectively disengaging self-sanctions from inhumane or detrimental conduct. Its antecedents are widespread. A study in 2008 found that moral disengagement occurred more frequently in boys than girls in the context of bullying at 3 Midwestern US middle schools (Turner, 2008). Another study in 2010 of 50 adult employees of large companies in Malaysia finds only a weak statistical relation between gender and moral disengagement, but finds strong negative associations with measures of conscientiousness, extroversion, and organizational ethical climate (Saidon, Galbreath, & Whiteley, 2010).

Evidence of moral disengagement is present in modern hate speech: social media contains calls to violence against outsiders (Kennedy et al., 2018; Hoover et al., 2020); online forums dehumanize girls and women (Ging, 2019; Hoffman, Ware, & Shapiro, 2020); and the manifestos of violent actors justify their actions by dehumanizing and blaming others (Peters, Grynbaum, Collins, Harris, & Taylor, 2019). We have evidence that hate speech with these indicators increases prejudice through desensitization (Soral, Bilewicz, & Winiewski, 2018)—and that the frequency of this language is related to the frequency of violent acts in the world (Olteanu, Castillo, Boy, & Varshney, 2018)—so understanding moral disengagement has real-world importance.

Recent work in NLP provides tools and frameworks for helping us analyze moral disengagement. For instance, researchers have used syntactic and semantic patterns to analyze dehumanization (Mendelsohn, Tsvetkov, & Jurafsky, 2020). Meanwhile recent NLP advancements have in *transformer models* unlock even more complex extraction of relations, events, and attributes from text using broad linguistic context (Eberts & Ulges, 2020; Magnusson & Friedman, 2021; Devlin, Chang, Lee, & Toutanova, 2019).

In this work, we present a novel semantic graph schema for representing indicators of a prominent theory of moral disengagement (Bandura, 1999, 2016), we train a transformer-based NLP model to identify moral disengagement indicators from text, using a preliminary (378-example) dataset from five online communities. We demonstrate how our graph representation captures indicators of moral disengagement, and we present promising initial empirical cross-validation results on our preliminary dataset.

We continue with a a brief review of cognitive indicators of moral disengagement and related work in NLP. We then describe our graph schema, dataset, NLP model architecture, and empirical results. We close with a discussion of the early successes and challenges in capturing indicators of moral disengagement and our plans for future work.

Background

Psychological Mechanisms of Moral Detachment

Our approach to detecting moral disengagement in language is primarily informed by Bandura (1999, 2016), whose theory of moral disengagement involves psychosocial mechanisms for disengaging self-sanctions from inhumane conduct. According to this theory, our moral values constrain our behaviors to morally-acceptable boundaries, but if and when we *violate* these standards we are faced with either (a) acknowledging our own immorality or (b) disengaging to retain integrity even after compromising our moral standards. Bandura identified several mechanisms of moral disengagement. The first five of these are covered, in part, by our NLP *indicators*, described below. We address the last three in our conclusion as potential future work.

1. Dehumanization. Our moral self-evaluation depends to some degree on how we regard the people who have been (or

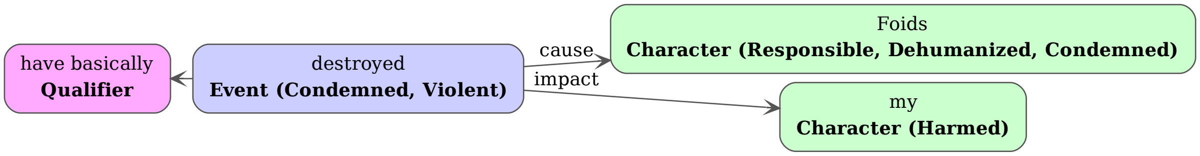
have basically
Qualifier
destroyed
Event (Condemned, Violent)
cause
impact
Foids
Character (Responsible, Dehumanized, Condemned)
my
Character (Harmed)

Figure 1: Knowledge graph corresponding to "*Foids have basically destroyed my life.*" This text was adapted from two separate Incel posts, to protect the privacy of individuals.

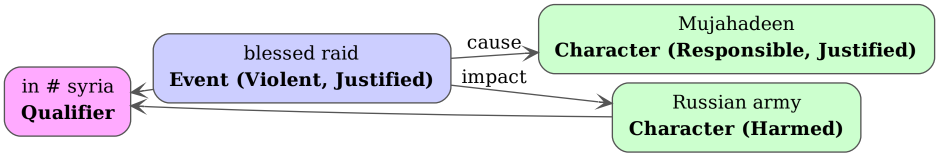

Figure 2: Knowledge graph corresponding to "*Breaking: blessed raid by Mujahadeen against Russian army in #syria.*" This text was adapted from three separate ISIS posts, to protect the privacy of individuals.

will be) harmed. Dehumanizing a person or group removes dignifying qualities and agency—and instead attributes subhuman, animalistic, mechanistic, or demonic qualities—in order to reduce self-punishment for harmful actions against them. Dehumanizing language is used to demarcate a morally superior *in-group* from an inferior *out-group*. Mendelsohn et al. (2020) offer a broad analysis of explicit and implicit manifestations of dehumanization in language.

2. Social and Moral Justification. Justifying the victimization of others, e.g., resolving that they "*deserved*" it or that one "*should*" enact harm, indicates that the out-grouping (e.g., via dehumanization, above) has been effective, due to the difficulty of inflicting harm on others perceived to be like us. This is evident in recent analyses of polarization and partisan bias (Van Bavel & Pereira, 2018). Those who morally justify violence may view themselves as righteous defenders of values and humanity (Bandura, 1999).

3. Victimization. By viewing *themselves* as victims in the world, perpetrators may see retaliatory actions as righteous (Bandura, 1999). This is evident in the frequently-voiced sentiment by anti-feminist *incels* ("*involuntary celibate*" males) that they have a right to sex and they are victims of the injustice of celibacy (Ging, 2019).

4. Attributing Blame / 5. Displacing Responsibility. Blame attribution and displacement of responsibility are tightly linked mechanisms. Attributing blame for an immoral act may help justify subsequent immoral actions against the offender and cast oneself as a victim (see above). Meanwhile, *displacing* blame of an immoral act jointly diminishes our own responsibility and justifies action against another party, e.g., accusing victims of bringing harm upon themselves. Escaping and displacing blame not only justifies harm upon others; it can also strengthen perceived moral rigorousness. Prominent examples of displacement include blaming victims of abuse for disobedient or seductive behaviors and blaming supernatural forces for possessing or overtaking one's agency.

6. Minimizing or Disregarding Injurious Effects. When harmful events are ignored, disputed, or minimized—or otherwise out of sight and mind—self-censure is not necessary. This type of mechanism might include casting aside the evidence that human activity causes climate change, thereby reducing the urgency for change and action.

7. Euphemistic Language. Language can sanitize an event and strip agency, e.g., to "*he was let go*" instead of *we fired him.* People behave more cruelly when detrimental practices are sanitized (Bandura, 1999).

8. Advantageous Comparison. Comparing one's own acts with more reprehensible acts of others (e.g., of the adversary) may reduce their blameworthiness and even make them appear righteous.

NLP and Graph Extraction with Transformers

Transformer-based methods for NLP utilize 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 (Devlin et al., 2019). This is widely regarded as a state-of-the-art methodology for NLP. Our approach is built on the SpERT transformer-based NLP architecture (Eberts & Ulges, 2020), which has been used to process text to extract knowledge graphs, e.g., of people, relationships, and complex scientific claims (Magnusson & Friedman, 2021). Many existing transformer models—similar to the model presented in this paper—require hundreds (sometimes thousands) of labeled training examples to reach high proficiency.

Our approach is also informed by recent work in NLP for computational social science. Recent work has detected and characterized linguistic patterns of dehumanization in large news corpora (Mendelsohn et al., 2020), and other work has extracted linguistic indicators of interpersonal respect and social distance using language patterns (Voigt et al., 2017). Like the present approach, these computational social science methods are highly inspectable because they infer indicators on specific *spans* of text rather than making a blanket judgment on an entire document. Unlike the present approach, these methods do not infer broad indicators of moral disengagement, and they primarily use lexicon-based and pattern-based analyses instead of transformer-based methods.

Approach

We next describe our knowledge graph schema, preliminary dataset, and NLP architecture for representing and extracting indicators of moral disengagement. Where possible, we refer to the labeled examples shown in Figures 1-4, which were run through our NLP model. The text of these examples is manually synthesized by our team by combining content of real examples to protect the privacy of the actual authors. To be sure, the examples in our dataset were not authored by provably violent, morally-disengaged individuals; however, the same rhetoric has appeared in manifestos of violent individuals (Peters et al., 2019), and hate speech may increase violent prejudices through desensitization (Soral et al., 2018).

Knowledge Graph Schema

Our knowledge graph schema represents *entities* (i.e., textual spans describing an element of interest), *relations* (i.e., semantic connections between entities), and *attributes* (i.e., multi-label tags on a span). We describe the entities, attributes, and relations of the schema, referencing the graphed examples rendered by our system in Figures 1-4.

Entities. Entities are labeled spans within the textual examples. The same exact span cannot correspond to more than one entity type, but two entity spans can overlap. Entities comprise the nodes of Figures 1-4upon which attributes and relations are asserted.

Our schema includes the following three entity types:

1. **Characters** are any human individual, group, organization, settlement, or ideology. In Figure 1, "*my*" is a character designating the author, as is "*foids*" (short for the incel pejorative "*femoid*" to portray women and girls as machines). In Figure 3, "*the southern border*" (part of a human settlement) is designated a character as well.
2. **Events** are any harmful actions or occurrences. In Figure 1, "*destroyed*" is designated a harmful event, as is Figure 2's "*blessed raid*" and Figure 3's "*flooding*" and "*replacing.*" Events frequently have responsible parties (who performed the act) and impacted parties (who were victimized).
3. **Qualifiers** are spatial, temporal, or epistemic constraints on events and characters. These often indicate when or where an event might (or did) happen, or whether the author believes an event *should* happen. In Figure 1, the qualifier "*have basically*" indicates that this is a past event, and the qualifiers in Figures 2 and 3 are spatial qualifiers on events and characters.

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 caused the harmful event, which was the victim, and which character or event is spatially or temporally qualified. In Figures 1-4, relations are directed arcs, and the unlabeled arrows are all *modifier* relations, left blank to avoid clutter. Our schema includes the following relations:

1. **Cause** indicates the character(s) or event(s) that enacted a given event. This helps capture the placement—or displacement—of responsibility (Bandura, 1999). We label causes for mentions of direct action (e.g., the raid "*by*" the Mujahadeen in Figure 2) as well as financial intervention (e.g., "*funded by*") or command (e.g., "*X calls us to...*" or "*X ordered Y to...*"), since these are linguistic moves by the author to place responsibility. Some events do not have an explicit cause.
2. **Impact** identifies the character(s) targeted by a given event. This helps capture the victimization of a group, individual, or the author themselves (e.g., in Figure 1), which may justify subsequent counter-action (see Figures 1 and 3). It also captures the target of *justified* victimization, e.g., the harm to the "*Russian army*" is encouraged by the author in Figure 2. Some events do not have an impacted character.
3. **Modifies** indicates a link between a character or event and a qualifier. These relations constrain characters' and events' existence, time, and location.

Attributes. Attributes are multi-label classes, where zero or more may apply to any given entity. The SpERT transformer-based model (Eberts & Ulges, 2020) on which ours is based was not capable of expressing these; this is a novel contribution of our work, as described below. Attributes are displayed as parenthetical listings inside each node in Figures 1-4. Our schema includes the following attributes:

1. **Dehumanization** may manifest as describing a character as an animal, disease, toxin, disaster, demon, or machine, effectively removing a foundation for empathy (e.g., "*foids*" in Figure 1). This is an extreme form of out-grouping an individual or group.
2. **Violent** is attributed to events that entail physical or sexual violence (e.g., "*blessed raid*" in Figure 2), either literal or as a metaphor for emphasis.
3. **Condemned** is attributed to events or characters to which the author expresses a negative moral valence. The author expresses an implicitly negative moral sentiment about the "*foids*" and their destruction in Figure 1.
4. **Justified** applies to events or characters to which the author ascribes positive moral valence or necessity of action, such as the "*blessed raid*" and its performers in Figure 2.
5. **Responsible** is attributed to characters in which the author has placed responsibility for an explicit or implicit action.

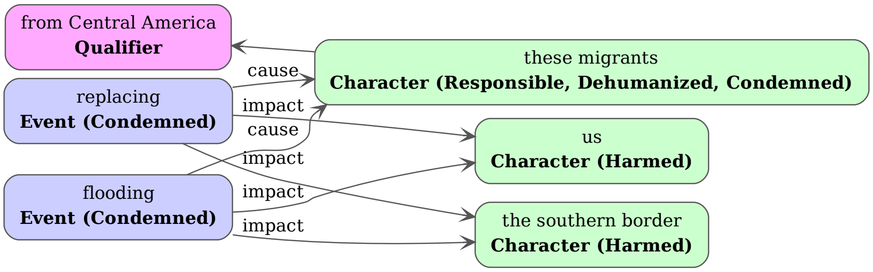

Figure 3: Knowledge graph corresponding to "*There's these migrants from Central America flooding the southern border and replacing us.*" This text was adapted from two conservative news quotes, to protect the privacy of individuals.

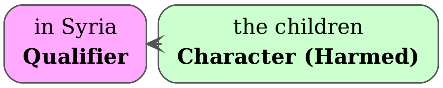

Figure 4: Knowledge graph corresponding to "*Weep for the children in Syria.*" This text was adapted from two separate ISIS posts, to protect the privacy of individuals.

In Figures 1-3, all characters that perform the harmful actions are marked as *Responsible*, but this also applies to text without explicit events.

6. **Harmed** applies to characters who are described as victims by the author, irrespective of whether a harmful event is mentioned, as demonstrated in Figure 4 where "*the children*" are inferred as harmed due to the surrounding context "*weep for the children.*"

As shown above, attributes express a diverse variety of categories on characters and events. One critical feature is the *cumulative* semantics of the attributes. The example in Figure 1 illustrates how three attributes applied to the same character "*foids*" captures different psychosocial indicators of moral detachment: this character is responsible for harm and is morally condemned (blame attribution) and is dehumanized to diminish their capacity for suffering. In the same example the author is the subject of harm, indicating self-victimization.

A separate combination of attributes tells a very different story in Figure 2 on the "*blessed raid*" event: the event is violent, but this violence is *morally justified* by the author. Similarly, the "*Mujahadeen*" in Figure 2 is tagged as *responsible* for harm, but *justified* in causing harm. This responsibility of harm is therefore not a case of blame (as with the "*foids,*" above); rather, as noted by Bandura (1999), those who morally justify violence may see themselves as protecting cherished values, fighting ruthless oppressors, preserving peace, saving humanity from subjugation, or honoring righteous commitments.

In Figure 3, "*these migrants*" is ascribed the "dehumanized" attribute by the NLP model primarily due to the "*flooding*" action ascribed to the character, which is a destructive action of a natural disaster and not of a human. This exemplifies how our transformer-based model assigns attributes by leveraging its BERT (Devlin et al., 2019) sentence context.

Problem Definition

We define the multi-attribute knowledge graph extraction task as follows: for a text passage $\mathcal{S}$ of n tokens $s_1, \ldots, s_n$, and a graph schema of entity types $\mathcal{T}_e$, attribute types $\mathcal{T}_a$, and relation types $\mathcal{T}_r$, predict:

1. The set of entities $\langle s_j, s_k, t \in \mathcal{T}_e \rangle \in \mathcal{E}$ ranging from tokens s_j to s_k, where $0 \leq j \leq k \leq n$,
2. The set of relations over entities $\langle e_{head} \in \mathcal{E}, e_{tail} \in \mathcal{E}, t \in \mathcal{T}_r \rangle \in \mathcal{R}$ where $e_{head} \neq e_{tail}$,
3. The set of attributes over entities $\langle e \in \mathcal{E}, t \in \mathcal{T}_a \rangle \in \mathcal{A}$.

This defines a directed multi-graph without self-cycles, where each node has zero to $|\mathcal{T}_a|$ attributes.

Dataset

Our preliminary dataset is comprised of 378 examples, each containing at least one sentence of plain text. Our dataset is comprised of selected examples from the following sources:

- The Kaggle *How ISIS Uses Twitter* dataset.[1]
- The Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2020).
- The Gab Hate Corpus (Kennedy et al., 2018).
- New York Times collection of news quotes related to the El Paso shooter's manifesto (Peters et al., 2019).
- Posts from an online forum for self-identified incels.

Three cognitive scientists (two anthropologists and one psychologist) selected examples from the above sources, three NLP-trained researchers separately labeled entities and relations in examples, and then two cognitive scientists labeled attributes on the entities, all adhering to the knowledge graph schema described above. Figure 5 plots the distribution of tokens, entities, relations, and attributes over the dataset. Some examples were *true negatives* with no indicators of moral disengagement, others contained attribute-level indicators but no relations or discrete events, and some others contained no attributes due to using objective language without explicit moral valence or dehumanization.

[1] https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

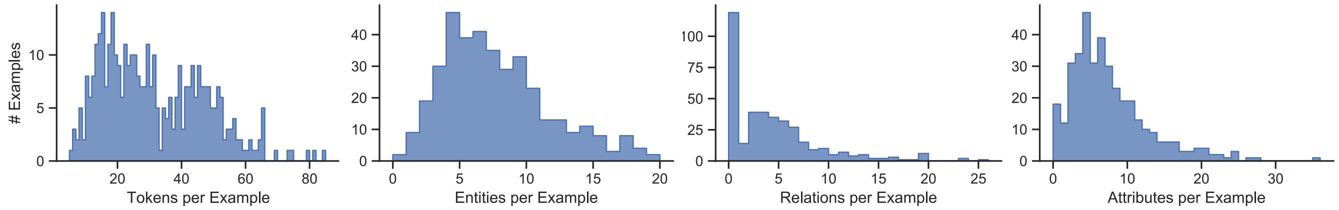

Figure 5: Dataset statistics for the number of tokens, entities, relations, and attributes per example.

This early, growing dataset is presently an order of magnitude smaller than many NLP datasets for event extraction. However, the ability of the schema to express examples from the above diverse sources (see also Figures 1-4) is evidence that our schema is general enough to express these psychosocial indicators across domains.

Model Architecture

Our model architecture extends SpERT with an attribute classifier. The original architecture provides components (Figure 6 a–c) for joint entity and relation extraction on potentially-overlapping text spans. The parameters of the entity, attribute, and relation classifiers, as well as the parameters of the BERT language model (initialized with its pre-trained values) are all trained end-to-end on our dataset.

The tokens $s_1, \ldots, s_n$ of the text passage $\mathcal{S}$ are each embedded by BERT (Devlin et al., 2019) as a sequence $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of high-dimensional vectors representing the token and its context. BERT also provides an additional "*[CLS]*" vector output, $\mathbf{e}_0$, designed to represent information from the complete text input. For all possible spans, $span_{j,k} = s_j, \ldots, s_k$, up to a given length, the word vectors associated with a span, $\mathbf{e}_j, \ldots, \mathbf{e}_k$, are combined by maxpooling to produce a single vector, $\mathbf{e}(span_{j,k})$, where each element contains the maximum value across the token vectors for that dimension. The final span representation, $\mathbf{x}(span_{j,k})$ is made by concatenating together $\mathbf{e}(span_{j,k})$ and $\mathbf{e}_0$ along with a width embedding, $\mathbf{w}_l$, that encodes the number of words, l, in $span_{j,k}$. Each valid span length l looks up a different vector of learned parameters, $\mathbf{w}_l$.

The span representation, $\mathbf{x}(span_{j,k})$, is classified into mutually-exclusive entity types by a multi-class linear classifier (Figure 6a). Only spans identified as entities move on to further analysis (Figure 6b). All pairings of the remaining entities are classified for relations by a multi-label linear classifier (Figure 6c), where pairs are represented by the concatenated vectors of the two spans with the "*[CLS]*" context vector replaced by the maxpool of the token vectors between the entities.

We implemented an additional subcomponent (Figure 6d) to infer multi-label attributes on the identified entities using $\mathbf{x}(span_{j,k})$ as input to another multi-label linear classifier. We take only identified entity spans as input to the attribute classifier, as this approach provided best performance and aligns with the finding by Eberts and Ulges (2020) that training on downstream tasks is best done on strong negative samples consisting of ground truth entities (i.e., teacher forcing).

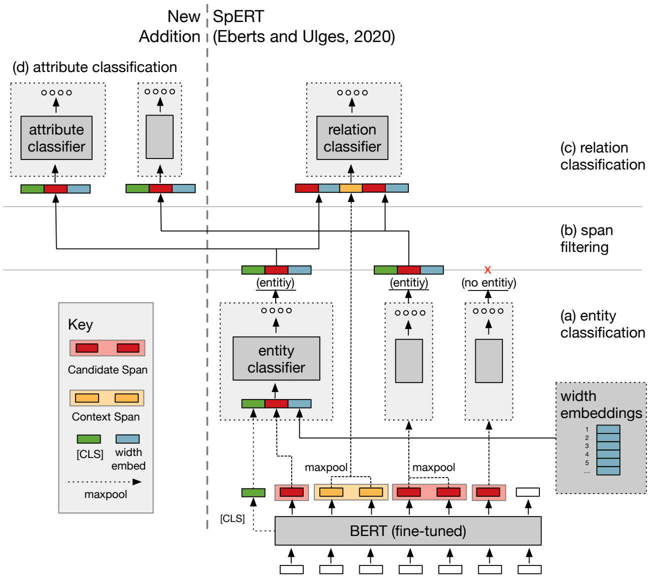

Figure 6: Our transformer-based model extends the SpERT components (a, b, and c) with attribute classification (d) that performs multi-label inference on identified entity spans.

Preliminary Results

After adapting the SpERT architecture (Eberts & Ulges, 2020) with the novel attribute-learning module, we assessed performance on our dataset with multiple transformer models and identified case-sensitive BERT (Devlin et al., 2019) as the best-performing pre-trained transformer variant.

We then conducted evaluations using 10-fold cross validation on the full dataset. Since we forego hyperparameter optimization until the dataset is completed, we do not report metrics on a held out test set. The per-class evaluations for our model are reported in Table 1. Despite the small size of our preliminary dataset, the model achieves promising results. While relation performance is notably lower, this can in part be explained the cascaded decision-making of the model wherein both entities paired by a relation need to be correctly extracted before the relation classifier can even attempt a correct prediction. Thus future efforts to improve entity extraction will raise this upper bound on relation performance.

	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	

Table 1: Precision, recall, F1, and support (i.e., occurrences in dataset) for each label using our extended SpERT model with the BERT case-sensitive language model.

Conclusion

This paper presents (1) a novel knowledge graph representation to capture linguistic indicators of moral disengagement informed by Bandura's (1999) theory, (2) a preliminary dataset from five online sources that label these indicators in text, and (3) a transformer-based NLP model that achieves promising initial results extracting indicators from text. The capacity of our knowledge graph schema to express moral disengagement indicators across domains (see Figures 1-4) is evidence that it can represent domain-general indicators.

Model Incompleteness

The work presented here does not capture all of Bandura's (1999) mechanisms, primarily due to the difficulty of reliably expressing them as spans of text or finding ample training data. Bandura's *euphemistic language* is a very subtle mechanism, where "sanitizing" language may be used to neutralize the negativity of an event. Recognizing sanitizing euphemistic language may require, in part, knowing that alternative, harsher language would have also described an incident, which involves another level of reasoning. The *disregarding of injurious effects* mechanism is likewise difficult, because this involves reasoning about what was *not* mentioned in the text. The *advantageous comparison* mechanism includes juxtaposition of an incident to more extreme, flagrant events to make it appear more benevolent. This is the most plausible mechanism of those we omitted, but we did not find sufficient examples of this in our initial search.

Interpretation and Implications of the Model

Mechanisms of moral disengagement reduce psychological feelings of discomfort when engaging in unethical behaviour (Bandura, 2002), so these mechanisms may be subtle and normalized. This means that detecting overt hate speech or toxic language is not a complete or accurate approach to detecting moral disengagement. Given this distinction, NLP detectors of moral disengagement may help identify and characterize harmful themes against groups and individuals.

Our knowledge graph schema and preliminary transformer-based model are designed to express and identify linguistic indicators of moral disengagement. None of these indicators—either alone or in conjunction—are sufficient (or designed) to categorize an individual or a group as morally disengaged. Rather, this approach has more potential for understanding how and why language changes over time, potentially to understand its possible detrimental impact on others (Peters et al., 2019; Soral et al., 2018).

Furthermore, our examples (e.g., Figures 1, 2, 3, and 4) illustrate that the graph-based representation of our model explicitly describes *who* has been dehumanized, blamed, and victimized, and the semantic linkage (i.e., relations) to violent events help describe the linkage across characters and events. This articulate approach facilitates human interpretation of the results, augmenting human-machine explanation of NLP outcomes. This graph-based approach stands in contrast to many document-level scoring models that rate the overall toxicity or sentiment of a document with a single feature value (e.g., a likelihood or intensity score).

Future Work

Our dataset is an order of magnitude smaller than most transformer-based approaches, so we are presently extending it in two ways. First, we are extending the dataset with new examples, since this will provide additional support for training the transformer model and increasing its F1 scores across the board. Second, we are gathering additional human ratings on existing examples to help us assess inter-rater agreement. Incorporating multiple ratings per example from multiple cultural perspectives will help us capture cultural idiosyncrasies in language and also identify areas of variable consensus.

We have also identified opportunities of improvement within our transformer-based model. We plan to augment the context representation of the relation classifier with broader token coverage from the BERT output, since the relation F1 is the weakest aspect of the model at present (see Table 1).

Another question is the deliberate use of mechanisms of moral disengagement over time, within and across communities. Shifts in these mechanisms may indicate shifts in what is considered acceptable (vs. radical) discourse (Astor, 2019). This question may be addressed using comparative studies over time, e.g., in parallel with historical events and social movements (Garg, Schiebinger, Jurafsky, & Zou, 2018). Our dataset and examples illustrate that mechanisms of moral disengagement are present in communications of fringe online communities, but they are also present in public discourse.

Finally, we plan to conduct transfer learning trials, e.g., to omit an entire source (e.g., ISIS or incels) from training and then use it for validation. This will help us characterize how linguistic indicators of moral disengagement generalize or vary across different communities.

Acknowledgments

This research was supported by funding from the Defense Advanced Research Projects Agency (FA8650-19-C-6025 P0000). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Astor, M. (2019). How the politically unthinkable can become mainstream. *New York Times*, *26*.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and social psychology review*, *3*(3), 193–209.

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of moral education*, *31*(2), 101–119.

Bandura, A. (2016). *Moral disengagement: How people do harm and live with themselves.* Worth publishers.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019.* Minneapolis, Minnesota: Association for Computational Linguistics.

Eberts, M., & Ulges, A. (2020). Span-based joint entity and relation extraction with transformer pre-training. *24th European Conference on Artificial Intelligence*.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities*, *22*(4), 638–657.

Hoffman, B., Ware, J., & Shapiro, E. (2020). Assessing the threat of incel violence. *Studies in Conflict & Terrorism*, *43*(7), 565–587.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., … others (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., … others (2018). The gab hate corpus: A collection of 27k posts annotated for hate speech.

Magnusson, I. H., & Friedman, S. E. (2021). 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.*

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, *3*, 55.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. (2018). The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12).

Peters, J., Grynbaum, M., Collins, K., Harris, R., & Taylor, R. (2019). *How the El Paso Killer Echoed the Incendiary Words of Conservative Media Stars.* The New York Times.

Saidon, I., Galbreath, J., & Whiteley, A. (2010). Antecedents of moral disengagement: Preliminary empirical study in malaysia. In *Proceedings of the 24th annual australian and new zealand academy of management conference.*

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, *44*(2), 136–146.

Turner, R. M. (2008). *Moral disengagement as a predictor of bullying and aggression: Are there gender differences?* The University of Nebraska-Lincoln.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, *22*(3), 213–224.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., … Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, *114*(25), 6521–6526.