Towards a Multi-Entity Aspect-Based Sentiment Analysis for Characterizing Directed Social Regard in Online Messaging

Joan Zheng, Scott Friedman, Sonja Schmer-Galunder, Ian Magnusson, Ruta Wheelock, Jeremy Gottlieb, Diana Gomez, Christopher Miller
SIFT
19 N 1st Ave., Suite 400, Minneapolis, MN 55401
{jzheng, friedman, sgalunder, imagnusson, rwheelock, jgottlieb, dgomez, cmiller}@sift.net

Abstract

Online messaging is dynamic, influential, and highly contextual, and a single post may contain contrasting sentiments towards multiple entities, such as dehumanizing one actor while empathizing with another in the same message. These complexities are important to capture for understanding the systematic abuse voiced within an online community, or for determining whether individuals are advocating for abuse, opposing abuse, or simply reporting abuse. In this work, we describe a formulation of directed social regard (DSR) as a problem of multi-entity aspect-based sentiment analysis (ME-ABSA), which models the degree of intensity of multiple sentiments that are associated with entities described by a text document. Our DSR schema is informed by Bandura's psychosocial theory of moral disengagement and by recent work in ABSA. We present a dataset of over 2,900 posts and sentences, comprising over 24,000 entities annotated for DSR over nine psychosocial dimensions by three annotators. We present a novel transformer-based ME-ABSA model for DSR, achieving favorable preliminary results on this dataset.

1 Introduction

The social media landscape is a complex, dynamic information environment where actors express advocacy, opposition, empathy, dehumanization, and various moralistic signals, with the intent—or sometimes the side-effect—of influencing others. A single message may also express multiple sentiments in one sentence, e.g., opposing one political candidate and endorsing another, or blaming one party for harming another, or dehumanizing one party and empathizing with another.

The complexity of multiple sentiments—which may comprise multiple strategies of influence—in a single message means that classifying an entire tweet's sentiment (Da Silva et al., 2014), or even quantifying it (Gao and Sebastiani, 2016), along a single dimension, is both at too high a granularity (i.e., we want to assess the author's perspective *on multiple topics*) and at too few dimensions (i.e., we want to assess the author's perspective *along multiple dimensions*).

Aspect-based sentiment analysis (ABSA) (Yang et al., 2018), allowing multiple dimensions of sentiment on a message, gets us part-way to a solution. Multi-entity ABSA (ME-ABSA) (Tao and Fang, 2020) gets us further in this direction by classifying along multiple dimensions across entities, but these models are frequently expressed as classification problems (e.g., **positive**, **neutral**, and **negative** predictions), and we desire a finer-grained numerical approach.

In the present work, we present a novel multi-entity transformer-based ABSA regression implementation of *directed social regard* (DSR), the prediction of social attitudes directed toward various actors and topics mentioned in the text. Social attitudes are modelled along nine continously-valued sentiment aspects: advocate, oppose, dehumanization, empathy, violent, condemn, justified, responsible, and harmed. Masked language modelling methods are utilized to support sets of aspects associated with each unique entity type. In the present work, DSR is computed for each *character* (i.e., human individual, human group, or ideology) in a message and each event that harms characters within a message. Also in the present work, the DSR dimensions are informed in part by Bandura's psychosocial theory of moral disengagement (Bandura, 1999, 2016), which we describe below.

To implement and validate our approach, three labelers rated nine dimensions of social regard for each character and event in a dataset of English-language social media posts sourced from curated Twitter datasets. To model DSR, we desitned a transformer-based regression architecture designed specifically for fine-grained sentiment analysis of

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 203 - 208

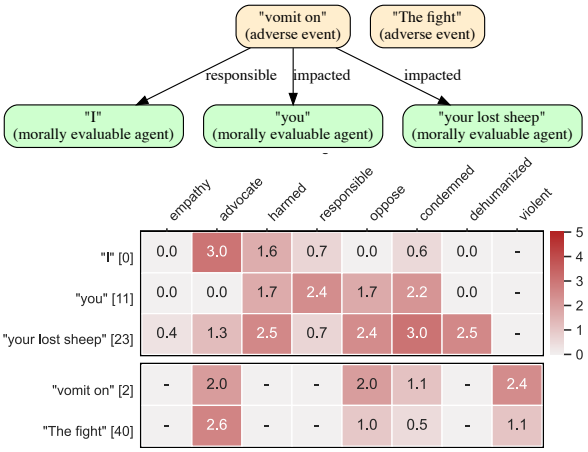

Figure 1: NLP output from "I vomit on you and your lost sheep. The fight is not over and never will be." adapted from a Kaggle social media dataset.

multiple entities.

We next describe the psychosocial theory of moral disengagement. We then describe our approach and empirical results, closing with a discussion of limitations and future work.

1.1 Moral Disengagement

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.

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 et al., 2020); and the manifestos of violent actors justify their actions by dehumanizing and blaming others (Peters et al., 2019). We have evidence that hate speech with these indicators increases prejudice through desensitization (Soral et al., 2018)—and that the frequency of this language is related to the frequency of violent acts in the world (Olteanu et al., 2018)—so understanding moral disengagement has real-world importance.

2 Approach

We describe our knowlege graph and attibute schema, sources of textual data, annotation process, and our architecture for representing and scoring attributes of social regard.

2.1 DSR Schema

Our DSR schema for a single social media post includes (1) a simple knowledge graph representation adapted from previous work in social media NLP (withheld for review), and (2) nine numerical intensity ratings on said characters and events to capture the directed social regard of the author, which is the primary focus of this work. An example of the system's output for a public Kaggle dataset tweet is shown in Figure 1. This was not part of our training dataset, so this is a novel machine prediction. We use this example to describe our schema.

The knowledge graph contains two types of *entities*, each comprising a span (i.e., contiguous span of tokens) in the text: (1) **characters**, also known as **morally evaluable agents**, comprising the author, human individuals, ethnicities, organizations, religions, ideologies, and geopolitical entities, and (2) **adverse events** that may cause harm or be morally questionable as described by the author. In Figure 1, the characters are "I," "you," and "your lost sheep," since the latter was inferred to refer to people in this context. The events include "vomit on" and "the fight."

The DSR values capture sentiment according to dimensions of moral disengagement described above, in addition to sentiment analysis, as expressed by the author of the text. For each dimension we describe whether it was motivated by Bandura's (1999, 2016) moral disengagement theory ^{B} or by sentiment analysis ^{S} and whether it applies to characters ^{c} or events ^{e} or both.

1. **Advocate**: Endorsement or support of an entity by the author.^{S,c,e}
2. **Oppose**: Opposition or adversarial attitude to an entity by the author.^{S,c,e}
3. **Dehumanization**: Actor described with non-human or lesser-than-human attributes, diminishing their agency or humanity.^{B,c}
4. **Empathy**: Actor described with empathy, compassion, humanity.^{B,c}
5. **Violent**: Event described as having literal or metaphorical physical or sexual violence.^{B,e}
6. **Condemn**: Entity morally condemned.^{B,c,e}
7. **Justified**: Entity morally justified.^{B,c,e}
8. **Responsible (for harm)**: Actor described as causing harm to others or to themselves.^{B,c}
9. **Harmed**: Actor described as being harmed by themselves or others.^{B,c}

Each of the Bandura-motivated dimensions captures a factor of moral disengagement: diminishing or accentuating humanity indicates whether the author might include the target in their circle of humanity; descriptions of violence and responsibility for harm are indicators of blame or advocacy for violence; mention of harmed individuals (including oneself) is an indicator of victimization and potential justification of subsequent action; and moral condemnation and justification indicate a moral standpoint for adverse events.

The heat-map in Figure 1 shows the nine moral dimensions across all of the characters and events from this example, where "your lost sheep" are the ony ones dehumanized.

2.2 Dataset and Annotation Methodology

Documents were selected from text posts known to contain online abuse or hate speech, including the Moral Foundations Twitter Corpus (Hoover et al., 2020); the Gab Hate Corpus (Kennedy et al., 2018); *How ISIS Uses Twitter* dataset from Kaggle (Khuram, 2017); and Manosphere community text posts (Ribeiro et al., 2020).

To optimize for content eligible for fine-grained sentiment analysis, documents were considered only if they met three criteria: (1) written in 280 or fewer characters; (2) written in English words or emoticons; and (3) contained more content than user mentions, URLs, or links to images.

Three English speakers were hired on the Prolific survey platform (Palan and Schitter, 2018) to score entities for DSR attributes. Out of our collected documents, 2,907 documents that met our criteria were annotated by at least two of our human annotators. These annotations contain a total of 24,425 unique entities. Annotators were asked to rate entities for each sentiment using a scale ranging from zero (not present) to five (most intense).

To measure inter-annotator agreement between our three human raters, we compute Krippendorff's α (Krippendorff, 2011) for each of the nine aspects, as shown in Table 1.

For drawing tentative conclusions, Krippendorff recommends using variables with reliabilities above $\alpha = 0.667$ (Krippendorff, 2018), which are achieved by our aspects **violent** and **oppose**. Both these aspects were labeled with intensity 4-5 more frequently compared to other aspects. For training and testing purposes, we identified annotations with high agreement as those where annotators

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

Table 1: Nonzero label usage comparision across our three annotators (A1-3) across 24,245 entities and nine aspect labels, along a five point intensity scale. Also includes Krippendorff's α.

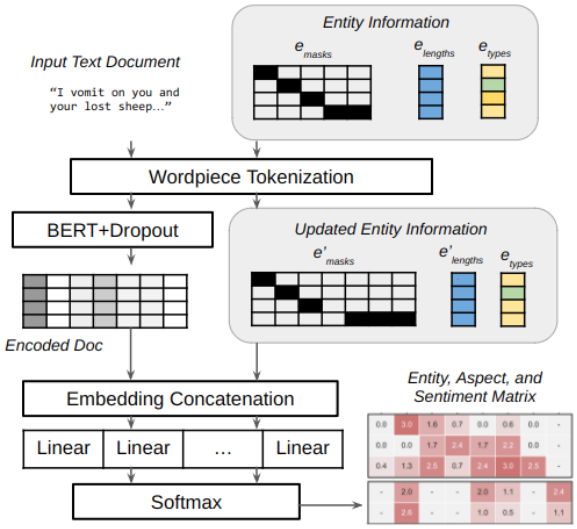

Figure 2: An overview of the ABSA architecture optimized for the DSR task.

falling within two standard units of each other, and with a maximum difference of two intensity units. These selection criteria limit disagreements while maintaining moderate-intensity aspects.

2.3 Architecture

We used two transformer-based NLP models: (1) an entity- and relation-extractor based on the SpERT architecture (Eberts and Ulges, 2020) to extract characters and entities comprising one or more continuous tokens in the text and (2) a novel ABSA-based model that scores each character or entity for the applicable DSR dimensions.

Importantly, for training and testing the DSR performance, we only use the human-annotated characters and events; we do not train or test the DSR model on machine-predicted entities, but this is how we envision applying the model on novel texts. We focus on the ABSA/DSR in this paper.

ABSA/DSR Architecture. The input for the DSR ABSA model is a text with entities annotated with (1) token start/end indices and (2) entity type (i.e., **character** or **event**). These may be either manually annotated (as we have done in our evaluation) or automatically predicted from a entity recognition system, e.g., (Eberts and Ulges, 2020; Friedman et al., 2021).

As shown in Figure 2, the text document is processed by a pre-trained BERT (Devlin et al., 2019) embedding layer using wordpiece tokenization. An interaction layer creates a fixed-dimentional pooled matrix, which contains a concatenation of BERT-encoded document and its entities represented as masked token sequences, the collection of masks for each entity type, and the lengths of each token span. These separate sequences are concatenated together as a matrix to support batch evaluation along multiple entities by the linear aspect classifiers.

This matrix representation feeds into a separate linear layers for each DSR aspect. Which entity gets graded by each linear layer is determined by the type of entity (e.g., as shown in Figure 1, an **event** entity does not have a **dehumanized** DSR aspect). This is implemented when multiplying the concatenated input matrix by the entity mask, which creates a matrix with nonzero inputs at the same indices as the linear layers it is eligible to be scored by. A softmax activation function calcuates the prediction associated with each aspect.

3 Experiment

We evaluated the DSR/ABSA architecture on the above dataset with the above DSR schema. We used human-labeled characters and events as inputs for this experiment in order to focus the evaluation on the DSR rather than the span extraction, but we report that on a 90/10 train/test split, the entity extractor scored F1 scores of 0.95 and 0.73 for extracting characters and events, allowing determiner mismatch, e.g., an event "the airstrikes" is allowed to match to "airstrikes."

We use the pre-trained, case-sensitive BERT-base model for fine-tuning (12 transformer blocks, 768-size hidden layer, 12 attention heads, and 110M total parameters). We fine-tuned with dropout probability 0.1 for 3 epochs, and we trained with learning rate 2e-5. Train, evaluation, and test splits were generated from our social media dataset using by creating 60/20/20 splits.

Aspect	R^2	**RMSE**
advocate	0.257	1.285
condemned	0.259	1.293
dehumanized	0.130	0.649
empathy	0.150	0.752
harmed	0.194	0.968
justified	0.207	1.037
oppose	0.284	1.419
responsible	0.207	1.037
violent	0.114	0.572

Table 2: ABSA/DSR model performance: R^2 measures correlation between human and machine ratings and RMSE measures prediction error. Averaged RMSE is 1.00 out of five units of intensity.

Results. Results are shown in Table 2, with lowest error (i.e., RMSE) on **violent**, **dehumanized**, and **empathy** dimensions. As mentioned above, **violent** was one of the more intensely-rated aspects and had highest α score, so we believe this contributed to successful learning. The aspect **dehumanized**—and its dual, **empathy**—are central to Bandura's theory of moral disengagement.

The average RMSE across aspects was 1.00 of a 5-point intensity scale, and all R^2 results directly correlated, explaining between 11-29% of variance in annotators' intensity scores across aspects. We regard these results as preliminary but encouraging for continued work in this domain.

4 Discussion and Future Work

We have described an approach to encoding the directed social regard (DSR) of authors toward events and actors in their posts, informed by Bandura's (1999, 2016) psychosocial theory of moral disengagement. This helps characterize abuse and harm in online messaging, including the advocacy and opposition to said abuse and harm, by highlighting entities that are associated with aspects associated with moral disengagement.

Our transformer-based approach uses a multi-entity aspect-based sentiment analysis (ME-ABSA) treatment to represent and predict DSR across nine psychosocial dimensions. We provide empirical evidence that transformer-based architectures can detect relevant actors and events and then predict human DSR ratings within reasonable preliminary error bounds.

Limitations and Future Work. One factor likely reducing the performance of our DSR model is the imbalanced representation of sentiment labels in our dataset. There is a scarcity of examples

in our dataset of entities that are associated with some sentiments, particularly moderate to positive sentiments labels and sentiments with low to moderate degrees of intensity. As shown in Table 1, annotators used aspect labels **empathy** and **justified** less frequently than other sentiment aspects in our schema, and was not able to reach a reliably high degree of agreement when annotating these sentiments. To improve the capability of our directed social regard model for applications outside of the domain of online abuse and hate, it would be beneficial to learn from examples that contain a more diverse selection of sentiments expressed, such as examples associated with positive to neutral sentiments as well as examples that contain a balanced range of low, moderate, and high degrees of intensity.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001121C0186 and Contract No. FA86650-19-6017. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

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

Albert Bandura. 2016. *Moral disengagement: How people do harm and live with themselves.* Worth publishers.

Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. 2014. Tweet sentiment analysis with classifier ensembles. *Decision support systems*, 66:170–179.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. *24th European Conference on Artificial Intelligence*.

Scott Friedman, Ian Magnusson, Vasanth Sarathy, and Sonja Schmer-Galunder. 2021. From unstructured text to causal knowledge graphs: A transformer-based approach. In *Proceedings of the 2021 Conference on Advances in Cognitive Systems*.

Wei Gao and Fabrizio Sebastiani. 2016. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining*, 6(1):1–22.

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

Bruce Hoffman, Jacob Ware, and Ezra Shapiro. 2020. Assessing the threat of incel violence. *Studies in Conflict & Terrorism*, 43(7):565–587.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Zaman Khuram. 2017. How isis uses twitter.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Stefan Palan and Christian Schitter. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Jeremy Peters, Michael Grynbaum, Keith Collins, Rich Harris, and Rumsey Taylor. 2019. How the El Paso Killer Echoed the Incendiary Words of Conservative Media Stars.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2020. From pick-up artists to incels: a data-driven sketch of the manosphere. *arXiv preprint arXiv:2001.07600*.

Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, 44(2):136–146.

Jie Tao and Xing Fang. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7(1):1–26.

Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. 2018. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.