Enhancing Detection of Relationship Abuse in Text with Multi-task Learning (Student Abstract)

Caroline Rinks
Vanderbilt University
2301 Vanderbilt Place, Nashville, TN 37235
caroline.p.rinks@vanderbilt.edu

Abstract

Intimate Partner Violence is a global, life-threatening public health issue that can be prevented by recognizing emotionally aggressive behaviors that signal the potential for future relationship abuse. To help identify these precursory unhealthy behaviors, this study proposes a Multi-task Learning framework for training robust models capable of detecting not only physically abusive behaviors but also emotionally abusive behaviors, such as belittling or manipulation, which historically precede physical abuse. Preliminary results indicate that Multi-task Learning can improve detection of emotional abuse and help tune detection models to particular kinds of relationship abuse.

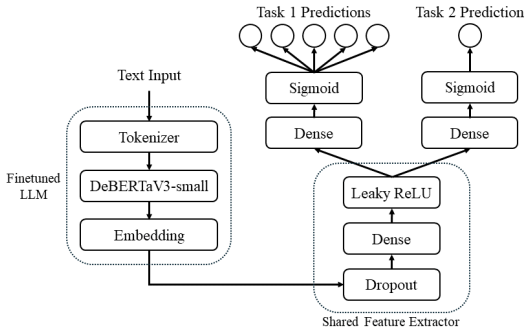

Figure 1: Multi-task learning model architecture.

Introduction

The CDC defines Intimate Partner Violence (IPV) as abuse or aggression that occurs in a romantic relationship, referring to both current and former spouses and partners, and can include physical violence, sexual violence, stalking, or psychological aggression. The CDC reports that slightly more than 2 in 5 women and 2 in 5 men in the U.S. have experienced severe physical violence from an intimate partner in their lifetime, and almost half of all women and 45.1% of men in the U.S. have experienced psychological aggression from an intimate partner in their lifetime. IPV is prevalent among all ages, with younger populations particularly at risk. Among those who reported sexual violence, physical violence, and/or stalking by an intimate partner, over 70% of women and 60% of men report their first incident occurred before the age of 25, and 27.1% of women and 21.4% of men report their first incident occurred before the age of 18 (Leemis et al. 2022).

Physical and sexual violence are often escalated acts that follow a history of psychological or emotional abuse (RAINN 2017). Therefore, many organizations like the One Love Foundation[1] dedicate their efforts to teaching others to recognize unhealthy emotional behaviors before they escalate to severe abuse. This study seeks to aid users in this task by developing a robust model for detecting both physical forms of abuse and precursory emotional abuse using Multi-task Learning (MTL). Multi-task Learning trains a model on multiple related tasks simultaneously, allowing for shared learning across tasks and potentially enhancing overall predictive performance (Crawshaw 2020). The preliminary results of this framework demonstrate the potential of leveraging MTL to tune detection models to specific kinds of relationship abuse.

[1] www.joinonelove.org/.

Related Work

Though many works have used Natural Language Processing (NLP) techniques to detect violence or offensive language (Jahan and Oussalah 2023; Dikwatta and Fernando 2019), there is a lack of work in applying NLP to detect violence or abuse in the context of intimate partner relationships specifically. One study, which is the most similar to this study, used machine learning and deep learning models to classify Twitter posts as self-reports of IPV (Al-Garadi et al. 2022). It differs from this study in that it considered self-reports exclusively, which are often retrospective in nature and, as a result, may already explicitly refer to intimate partners as abusive. This study seeks to train a robust model to detect abuse without such guarantees of language, corresponding to instances in which a user may not yet recognize behavior as abusive in nature. Furthermore, this study introduces a novel Multi-task Learning framework that leverages an additional emotional abuse classification task for providing robustness against more complex emotionally abusive behaviors.

Methodology

A MTL framework is introduced to improve detection of emotional abuse in a relationship. This study uses two training tasks related to relationship abuse. Task 1 targets emotional behavior specifically and is a multi-label classification task, in which text is labeled as hostile, antagonizing, dismissive, condescending, and/or sarcastic. The dataset for this task is the Unhealthy Comments Corpus (UCC) introduced by Price et al. (2020), which features over 44,000 online comments annotated by crowdworkers. This dataset was selected to target emotional abuse because of the resemblance of its labels to the 10 Signs of an Unhealthy Relationship outlined by the One Love Foundation and the CDC's definition of psychological aggression as behavior that "includes expressive aggression (insulting, humiliating, or making fun of a partner in front of others)" (Leemis et al. 2022). Task 2 in this study refers to a binary classification task, in which text is labeled as an instance of relationship abuse or not. The dataset for this task was introduced by Schrading et al. (2015). It contains 1,336 anonymous submissions of abuse posted to Reddit and details all types of relationship abuse.

Figure 1 presents an overview of the MTL model architecture designed for this study. A pre-trained DeBERTaV3 (He, Gao, and Chen 2021) large language model (LLM) is fine-tuned on both tasks during training. The text embeddings generated by the LLM are then passed to a shared feature extractor that consists of a single dropout layer followed by a dense layer with 64 units and Leaky ReLU activation. Finally, the extracted features branch into separate task-specific heads. All the parameters of the LLM, the shared feature extractor, and the task-specific heads are trained on both tasks simultaneously by optimizing the weighted sum of the loss of each task, L_1 and L_2:

$$L = \tau L_1 + (1 - \tau) L_2 \tag{1}$$

where the hyperparameter τ controls the weight of each loss.

Four MTL models and two single-task learning (STL) models were trained for 5 epochs with the default learning rate scheduler from the Hugging Face Transformers library (Wolf et al. 2020). The MTL Default model weights task 1 and task 2 equally, while MTL τ=0.4 and MTL τ=0.6 give greater weighting to task 2 or task 1 respectively. Since the UCC dataset was crowdsourced, an additional model with equal weighting, named MTL Confident, was trained on a higher quality version of the UCC dataset, which only included training samples where the confidence score given by the crowd workers was at least 0.6 (Price et al. 2020).

Results and Discussion

Table 1 reports the training loss, training accuracy, and testing accuracy for each of the models. Task 1 appears harder for all the models to learn and benefits from an MTL approach. The higher quality of the confident UCC training data also improves performance in general. All models achieve at least 95% testing accuracy on task 2.

The models' ability to detect different kinds of relationship abuse was analyzed by performing inference on task 2 with the inputs provided in Table 2. Examples 1 and 4 are from the evaluation dataset for task 2 and represent an instance of physical abuse and non-abuse respectively. Examples 2 and 3 are referenced from the One Love Foundation website[1], where they are used as examples of unhealthy relationship behavior, specifically manipulation and belittling. All models correctly classified example 1 as abuse and example 4 as not abuse with strong probabilities. Only the MTL τ=0.6 model and the STL model correctly classified example 2 as abuse. The MTL τ=0.6 model did so with a higher probability of 0.9307 compared to the STL model's probability of 0.8224. Furthermore, only the MTL models correctly identified example 3 as abuse. These results indicate that certain task weightings can improve the detection of specific types of abuse, and they demonstrate the potential for leveraging MTL to learn more complex representations of abuse. These methods could lay the groundwork for a system that delivers targeted interventions to victims based on the content and type of abuse. Future work would benefit from an expertly labeled emotional relationship abuse dataset, similar to the examples in Table 2, to enable further analysis of these methods.

		MTL Default	MTL τ=0.4	MTL τ=0.6	MTL Conf.	STL1	STL2
Training Loss	Both	0.337	0.241	0.374	**0.191**	n/a	n/a
	Task 1	0.331	0.556	0.573	**0.166**	0.380	n/a
	Task 2	**0.006**	0.030	0.076	0.025	n/a	0.022
Training Accuracy	Both	0.885	0.768	0.750	**0.961**	n/a	n/a
	Task 1	0.862	0.723	0.704	**0.955**	0.840	n/a
	Task 2	**0.998**	0.992	0.997	0.996	n/a	0.996
Testing Accuracy	Both	0.685	0.690	0.747	**0.754**	n/a	n/a
	Task 1	0.631	0.637	0.705	**0.713**	0.713	n/a
	Task 2	0.957	0.954	0.959	0.960	n/a	**0.963**

Table 1: Summary of final performance metrics.

	Abuse Type	Text
1	Physical	a friend told me a man pushed and choked her for not hooking up with him...
2	Emotional: Manipulation	give me your password. what do you have to hide? if you really loved me you'd give it to me
3	Emotional: Belittling	my boyfriend asked me why I posted this one pic. he said that I look really desperate for likes and that the pic is not flattering
4	Not Abuse	buying a wedding gift for best friend (guy) who got married, but not his wife. is that ok? one of my best friends...

Table 2: Custom examples used for inference.

Ethical Statement

The models trained in this work exhibit significant gender bias in regards to the portrayal of victims and abusers. When changing the gender in example 3, the probability of abuse changes significantly. Specifically, when a male is in the role

of abuser, the models output a higher probability of abuse than when a female is referenced as the abuser. This is likely due to higher female reports of abuse than male reports in general, and future research must address this issue.

Acknowledgments

The author gratefully acknowledges Ayan Mukhopadhyay for his guidance of this research as part of the AI for Social Impact course at Vanderbilt University.

References

Al-Garadi, M. A.; Kim, S.; Guo, Y.; Warren, E.; Yang, Y.-C.; Lakamana, S.; and Sarker, A. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, 15: 100217.

Crawshaw, M. 2020. Multi-Task Learning with Deep Neural Networks: A Survey. arXiv:2009.09796.

Dikwatta, U.; and Fernando, T. 2019. Violence detection in social media-review.

He, P.; Gao, J.; and Chen, W. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543.

Jahan, M. S.; and Oussalah, M. 2023. A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing*, 126232.

Leemis, R. W.; Friar, N.; Khatiwada, S.; Chen, M. S.; Kresnow, M.-j.; Smith, S. G.; Caslin, S.; and Basile, K. C. 2022. The national intimate partner and sexual violence survey: 2016/2017 report on intimate partner violence.

Price, I.; Gifford-Moore, J.; Fleming, J.; Musker, S.; Roichman, M.; Sylvain, G.; Thain, N.; Dixon, L.; and Sorensen, J. 2020. Six attributes of unhealthy conversation. *arXiv preprint arXiv:2010.07410*.

RAINN. 2017. Early Warning Signs of Dating Violence. https://www.rainn.org/news/early-warning-signs-dating-violence. Accessed: 2024-04-29.

Schrading, N.; Alm, C. O.; Ptucha, R.; and Homan, C. 2015. # whyistayed,# whyileft: Microblogging to make sense of domestic abuse. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1281–1286.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.