Title

ADViRDS: Assessment of Domestic Violence Risk Dataset and Scale on Social Media

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Tong, Chengwei
Guo, Mengzhuo
Tian, Yao
et al.

Publication Date

2024

Peer reviewed

ADViRDS: Assessment of Domestic Violence Risk Dataset and Scale on Social Media

Chengwei Tong[1] (cwtong@mail.ustc.edu.cn), Mengzhuo Guo[1] (mzguo@mail.ustc.edu.cn), Yao Tian[1] (tyzkd@mail.ustc.edu.cn), Mengzhu Zhang[2] (zhangmengzhu2021@163.com), Yangyang Li[3] (liyangyang@live.com), Chunyan Zhu[2] (ayswallow@126.com), Jie Bao[1] (baojie1996@mail.ustc.edu.cn), Rongrong Sheng[2] (psyshengrr@126.com), Qianqian Li[2] (lqqian187@163.com), Yong Liao[1] (yliao@ustc.edu.cn)

[1]University of Science and Technology of China, Hefei, China
[2]Anhui Medical University, Hefei, China
[3]Academy of Cyber, Beijing, China

Abstract

This study presents ADViRDS, an innovative scale and dataset specifically developed for examining the psychological traits of domestic violence (DV) perpetrators. Recognizing the critical need to understand the psychological dynamics of perpetrators, our research shifts the focus from the experiences of DV victims to the characteristics of the perpetrators. Our approach involves a six-dimensional scale designed to detect the psychological traits of DV perpetrators, formulated with insights from established DV research and psychologists. To complement this scale, we constructed a detailed dataset containing 574 individual entries from the Chinese social media platform "Zhihu." Each entry was carefully annotated by experienced professionals, ensuring a high degree of accuracy and relevance. We conducted a comprehensive analysis using a range of models, including Zero-Shot classification, GPT series, and fine-tuned pre-trained models, to evaluate their effectiveness in identifying individuals with psychological predispositions to DV. The findings reveal significant insights into the models' capabilities, highlighting the nuances in detecting DV tendencies through psychological profiling. Our research offers a new paradigm in DV studies, focusing on the psychological traits of perpetrators for a comprehensive understanding of DV dynamics and prevention.

Keywords: Domestic Violence; Psychological Traits; Social Media

Introduction

Domestic violence (DV) has become an increasingly pressing concern in recent years, with its severity reaching alarming levels. While the COVID-19 pandemic and associated stress factors have exacerbated DV cases (Kourti et al., 2023; Piquero, Jennings, Jemison, Kaukinen, & Knaul, 2021), the core issue often lies within the complex psychological makeup of the perpetrators(McCauley et al., 1995). The repercussions of ignoring these psychological aspects are grave, leading to increased depression, heightened suicide risks, and a general decline in life satisfaction(M. Liu et al., 2021). Therefore, understanding the connection between psychological attributes and DV is vital for better control and management of abusive behaviors in the home(Chandan et al., 2020; Foa, Cascardi, Zoellner, & Feeny, 2000).

Currently, there is a range of studies addressing DV, including analyses focusing on the mental health of DV victims and investigations of the extent of DV reflected in social media posts(M. Liu et al., 2021; Salehi, Ghahari, Hosseinzadeh, & Ghalichi, 2023; Homan, Schrading, Ptucha, Cerulli, & Ovesdotter Alm, 2020). However, these studies predominantly centered on the experiences and perspectives of the victims. As a result, they are, strictly speaking, limited in their ability to effectively assess and predict the potential tendency for DV. Thus, it is essential to develop a more holistic approach that encompasses perpetrator perspectives to fully understand and assess DV risks and patterns.

Furthermore, even if we assume the existence of sufficient data on DV, few studies indicate the criteria for determining the psychological traits of DV perpetrators, which is markedly different from the detection of other psychological health problems with gold standards, such as depression and suicide(Zhang, Chen, Wu, & Zhu, 2022; Gaur et al., 2021). This lack of a clear diagnostic benchmark poses a substantial challenge in the creation of a reliable and representative dataset for studying DV.

To address the issue mentioned above, we have meticulously developed ADViRDS, a comprehensive scale and a corresponding dataset, each annotated with insights from psychological experts. Here are our key contributions:

- We proposed a six-dimension scale to detect psychological traits of DV perpetrators. Each dimension, formulated with psychologists, draws from established traits in prior DV research, as shown in Section 3. We believe this integrated approach can further enhance DV-related research.
- We constructed a specialized dataset to facilitate additional experiments. This dataset comprises 574 individual entries, meticulously gathered from the Chinese social media platform "Zhihu." Each entry has been carefully annotated by experts in psychology, ensuring a high level of accuracy for our research. The dataset is released to the public at `https://github.com/DwToretto/ADViRDS`.
- We conducted a series of experiments to assess the accuracy of existing models in detecting the psychological traits of DV perpetrators on our dataset, in order to evaluate its applicability.

Related Work

In the field of DV research, challenges such as the difficulty in data acquisition and concerns over privacy and ethics have led to limited research, particularly in China. Unlike the conventional machine learning tasks in other fields that benefit

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

from extensive and high-quality datasets with gold standard diagnoses, the study of DV is constrained. This is further complicated by cultural factors in China, where the protection of "face" regarding sensitive topics can impede the public disclosure of detailed information(Zhou, Wang, & Zimmer, 2022). These aspects not only restrict the comprehensive psychological analysis of DV perpetrators but also limit the feasibility of conducting research with clinically validated diagnostic information in this domain. Despite the significant challenges in collecting and analyzing DV data, valuable insights can still be gleaned from existing DV and other psychological studies, providing guidance for further exploration in this field.

Domestic Violence Study

An effective approach to gathering comprehensive data on DV is collaboration with governments or law enforcement agencies to access information on DV cases from criminal databases. This partnership enables researchers to obtain a more detailed and accurate picture of DV incidents, contributing significantly to the depth and quality of studies in this area. Hsieh et al. (2018) worked with Taipei City Government to improve the efficiency of DV prevention and risk management. Wijenayake et al. (2018) and Yu et al. (2023) independently utilized data from the Australian Re-offending Database and multiple national registries in Sweden, respectively. Their analyses focused on demographic and psychological factors of DV offenders to predict the risk of recidivism in such cases.

However, collaboration with official agencies to access their data often presents challenges due to privacy concerns and data access restrictions. Additionally, analyzing data from criminal databases mainly allows for the prediction of recidivism, rather than the potential risk of initial offenses. In contrast, social media data, being more readily accessible and encompassing daily interactions of individuals, holds immense value for predicting the broader risk of DV in society. Salehi et al. (2023) collected and classified posts from Twitter and Instagram using criteria developed by a DV expert, applying machine learning methods to predict DV-related content on social media. Subramani et al. (2019) constructed a novel 'gold standard' dataset with multi-class annotations from Facebook and employed deep learning models for recognition to support victims of DV. Xu et al. (2022) analyzed the emotional evolution trend of the Chinese public's response to DV incidents on Weibo from a temporal and geospatial perspective, demonstrating that social media data is extremely valuable for DV research.

Mental Health Issues on Social Media

Recent research in mental health has increasingly turned to social media as a vital source for addressing various mental health issues, such as depression(Zanwar, Wiechmann, Qiao, & Kerz, 2022; Kelley & Gillan, 2022; Shi et al., 2023) and suicide(O'Dea et al., 2015; Desmet & Hoste, 2018).

Wu et al. (2023) introduced the DepCov dataset to study the impact of COVID-19 on depression through Weibo's social media activities. Additionally, Zhang et al. (2022) introduced a psychiatric scale-guided method for identifying high-risk posts through an evolving algorithm for early depression detection (ERD). Gaur et al. (2021) conducted both time-variant and time-invariant models based on the Columbia-Suicide Severity Rating Scale (C-SSRS) (Posner et al., 2008) to assessed suicide risk. Moreover, Rabani et al. (2023) further advanced this field by categorizing suicide risk levels on Twitter and Reddit.

Inspired by these previous work, our study has decided to utilize social media data to identify psychological traits of potential DV perpetrators, contributing to the expanding research on mental health issues in digital spaces.

ADViRDS: Assessment of Domestic Violence Risk Dataset and Scale

In this section, we will provide a detailed overview of the data collection process, as well as the construction process of our ADViRDS scale and dataset. The relevant processes are illustrated in Figure 1.

Data Collection

To determine the most suitable source for in-depth data analysis, we initially did a preliminary investigation across four prominent online platforms in China: Weibo[1], Douban[2], Xiaohongshu[3] and Zhihu[4]. On Weibo, the DV super-topic was reviewed, which predominantly featured brief narratives from the perspective of victims. Similarly, Douban's DV support group was found to primarily consist of victim narratives and discussions. Xiaohongshu presented a different challenge; most of the content shared by its user groups uses pictures as the main carrier and has less text content, which is not conducive to extracting information. Consequently, these three platforms – Weibo, Douban, and Xiaohongshu – were deemed less suitable for comprehensive data scraping due to their structural limitations and the nature of the content.

Our focus then shifted to Zhihu, a platform that stood out for its in-depth user engagement and detailed responses. We concentrated on the question "What is the experience of being a victim of Domestic violence?[5]" posted on Zhihu. The comprehensive nature of the responses to this particular query provided an in-depth and wide-ranging insight into the behaviors of perpetrators and the experiences of victims. In contrast, answers to other similar questions were either limited in number or, despite appearing relevant, offered little substantive information for our analysis. Using a custom web crawler, we extracted 2012 answers from Zhihu and classified them into three categories: promotional content/narratives,

[1] https://weibo.com
[2] https://www.douban.com
[3] https://www.xiaohongshu.com
[4] https://www.zhihu.com
[5] https://www.zhihu.com/question/30644408

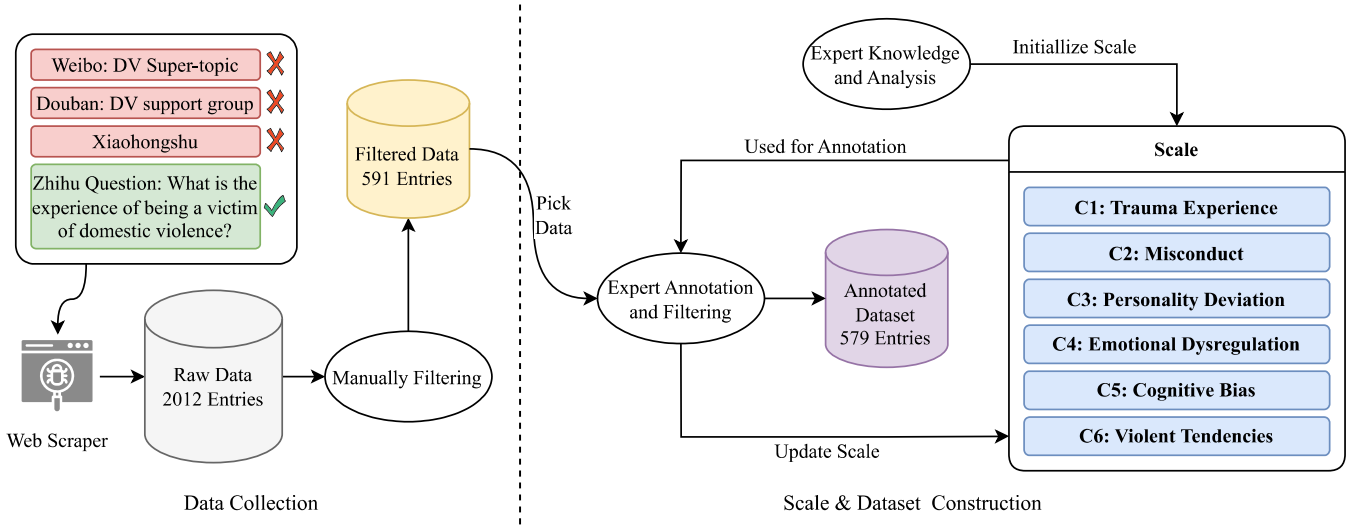

Figure 1: The steps for data collection and annotation, as well as scale construction.

basic descriptions of DV, and detailed responses outlining perpetrator traits.

Given our research focus on understanding the traits of perpetrators, we zeroed in on the third category. This led us to preserve 574 responses that specifically provided detailed descriptions of perpetrator traits. This approach allowed us to gather rich, nuanced data, facilitating a more in-depth analysis of DV patterns.

Scale Construction

Drawing upon C-SSRS, we collaborated with seasoned psychologists to construct an initial scale, informed by clinical experience and existing researches on DV perpetrators. This preliminary scale was then continuously iterated and refined through collected data, culminating in the final version: Assessment of Domestic Violence Risk Dataset and Scale (ADViRDS). The scale is designed with six dimensions to provide a comprehensive analysis of the psychological traits of DV perpetrators. These dimensions collectively encompass various aspects, offering a thorough and multidimensional understanding of their characteristics, as detailed below:

- **C1** ***Trauma Experience*** DV exhibits a significant generational transmission. Children who have experienced or witnessed DV during their childhood are more likely to rationalize and normalize violent behaviors, thereby increasing the likelihood of becoming perpetrators themselves in adulthood.(N. Liu, Zhang, & Cao, 2010; Zou, Zhang, & Zhang, 2001).

- **C2** ***Misconduct*** The majority of DV perpetrators tend to have lower educational attainment, and are more likely to be unemployed or laid-off compared to those in non-violent family groups. Additionally, incidences of alcohol abuse, severe smoking, gambling, mental illness, and disability are significantly higher among members of violent family groups. Moreover, perpetrators often have a higher proportion of the total household income(Zhao, Zhang, Li, Zhou, & Li, 2008).

- **C3** ***Personality Deviation*** DV perpetrators are characterized by neurotic personality traits, lower levels of mental health, less social support, and negative coping mechanisms. Those who commit severe physical violence display more antisocial and borderline personality traits, as well as a broader range of personality deviations or disorders(Zhao, Zhang, Li, Zhou, & Li, 2007).

- **C4** ***Emotional Dysregulation*** Perpetrators of severe physical violence exhibit personality traits such as a propensity for sadistic impulses, thrill-seeking, and emotional instability. Additionally, these individuals tend to have more pronounced traits of anger and impulsivity, coupled with poorer anger management skills(Ehrensaft, Moffitt, & Caspi, 2004).

- **C5** ***Cognitive Bias*** In cases of DV, families with authoritarian (where one person makes all decisions and others must comply) or laissez-faire (where each member acts independently without restrictions) structures are significantly more common than in control groups. Conversely, democratic (where decisions are made collectively) family structures are notably less common(MA, WU, & HONG, 2014). This suggests that irrational cognition and behavior among family members can easily lead to violent incidents.

- **C6** ***Violent Tendencies*** According to the experience of psychologists in clinical and police cases, DV perpetrators often struggle to control their violent behavior. In addition to targeting victims, they frequently vent their anger on objects and animals(Becker & French, 2004; Volant, Johnson, Gullone, & Coleman, 2008; Newberry, 2017).

Based on this scale, we annotated the collected data and established a comprehensive dataset.

Table 1: Examples of DV posts from ADViRDS along with the descriptions of each category.

Category	Description	Example
Trauma Experience	Poverty, physical abuse, sexual abuse, cold violence, emotional abuse, emotional neglect, etc.	*"I heard my grandfather beat my grandmother even worse. My dad also resents this; he used to tell us about how he was beaten as a child, how his mother was beaten..."*
Misconduct	Gambling, alcohol abuse, drug misuse, drug addiction, overspending, etc.	*"When I was young, my dad loved gambling and he would ask my mom for more money. If she didn't give him money, he would instantly rage, drag her into the bedroom, lock the door, and beat her brutally... "*
Personality Deviation	Obvious jealousy, suspicion, paranoia, coldheartedness, etc.	*"Now he still has a very strong desire to control, even to the extent of being paranoid about whom my sister dates... "*
Emotional Dysregulation	Impulsiveness, easy to anger, irritability, strong desire for control, poor self-control, etc.	*"He, like a madman, pressed me into a corner and hit my head with a hot water kettle. He is irritable and unable to control his emotions, often getting into conflicts with others... "*
Cognitive Bias	Significant deviation from mainstream understanding: black-and-white thinking, overgeneralization, absolutism, labeling, personalization, mind reading, shoulds and musts, emotional reasoning, etc.	*"After I started a relationship with my boyfriend, he always implied that he was superior and I was inferior. He belittled me to elevate himself, obstructed my normal social life, forbade me from seeing friends... "*
Violent Tendencies	History of violence, animal abuse, hitting people in childhood, threatening others, damaging property, etc.	*"Last night, because two cats at home fought, my boyfriend beat one of them severely. He even said he wanted to kill it. When I calmly told him to do it, he immediately went down to continue the beating... "*

Human Annotation

We invited four graduate students major in psychology to manually annotate the dataset. All annotators bring a wealth of experience in psychological research. To ensure the quality of the annotations, every sample in the dataset was reviewed by all annotators, with the most frequently assigned label being selected as the final one. After the preliminary annotation was completed,we sought the insights of two psychology experts to review the annotation. The average Fleiss' κ (Fleiss, 1971) of six-category is 0.708.

Data Analysis

After manual annotation, we finally introduced ADViRDS consists of 576 posts, each associated with one or more categories. Table 1 provides a board snapshot of the diverse categories of DV posts within the ADViRDS, along with detailed descriptions about each category. We used the *Google Translate API* to translate posts into English, and omitted some details, like names, places and irrelevant descriptions.

Furthermore, we conducted a statistical analysis on the category distribution of ADViRDS. As shown in Figure 2, C4 (*Emotional Dysregulation*) emerges as the most prevalent factor, constituting nearly 60% in total. It indicates close relationship between DV and emotional states of perpetrators.

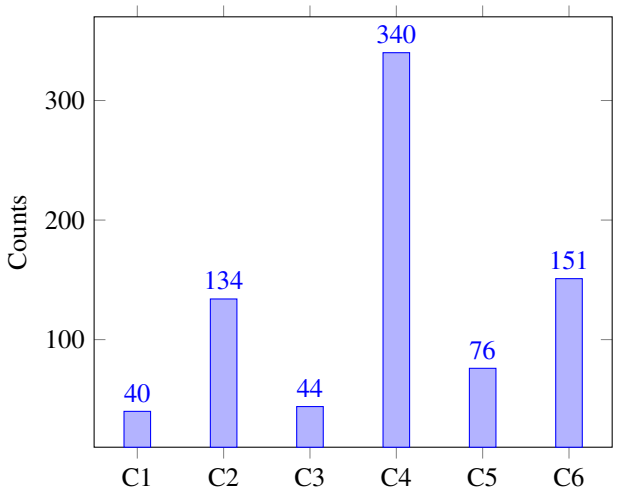

Figure 2: Category distribution statistics on ADViRDs.

Following this are C2 (*Misconduct*) and C6 (*Violence Tendencies*), which illustrate how specific objective environmental factors, such as alcoholism or gambling can escalate the risk of DV from different perspectives. The least common categories are C1 (*Trauma Experience*) and C3 (*Personality Deviation*), accounting for 6.9% and 7.6% respectively. This

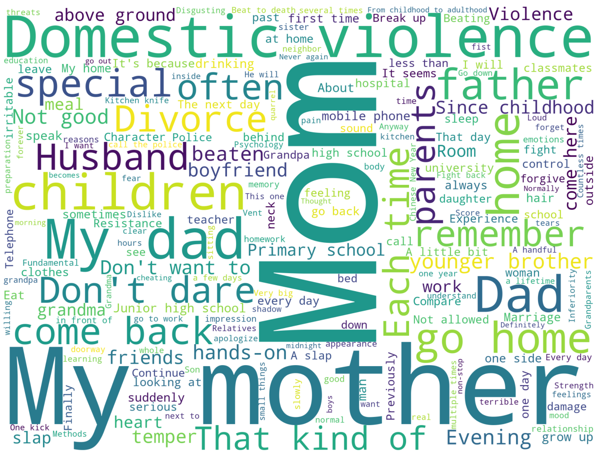

Figure 3: Wordcloud generated based on ADViRDS.

might be because childhood psychological trauma and personality issues behaviors may be difficult to identify from the narratives of the victims. These findings underscore the diversity of ADViRDS, making it a valuable resource for studying various aspects of DV behavior.

The word cloud derived from ADViRDS offers a compelling visual depiction of the most prominent terms (refer to Figure 3). Terms such as *"father"*, *"mother"*, *"boyfriend"*, and *"girlfriend"* are prevalent, showing that DV not only occurs in parent-child relationships, but also intimate partner violence emerging as a prevalent manifestation of DV. Additionally, the frequent appearance of words like *"beaten"*, *"slap"*, and *"don't dare"* highlights the diversity of DV behaviors, which include not only physical violence, but also verbal and psychological violence.

Experiments

The experiments are designed to address two questions: How feasible is it to detect the DV perpetrator traits with the current resources and technology? How well can large language models perform on our ADViRDS dataset how much do they outperform small models?

Baselines and Metrics

At the onset of our experiments, we outlined our choice of three model types tailored for different scenarios: plug-and-play models, fine-tuned pretrained language models(PLMs), and large language models(LLMs).

- **mDeBERTa-v3-base-XNLI-multilingual-NLI-2mil7** (Laurer, Van Atteveldt, Casas, & Welbers, 2024) is a multilingual model fine-tuned from mDeBERTa-v3-base on the XNLI and multilingual-NLI-26lang-2mil7 datasets. With proficiency in natural language inference across 100 languages, it is intended for tasks like multilingual zero-shot classification.
- **XLM-RoBERTa-large-XNLI** is based on XLM-RoBERTa-large (Conneau et al., 2019), fine-tuned using NLI data from 15 languages. Tailored for zero-shot text classification, it is designed for tasks that involve classification across various languages without explicit language-specific training data.
- **RoBERTa-wwm**(Cui et al., 2019) is a large-scale Chinese language model based on RoBERTa, incorporating whole word masking for natural language understanding.
- **MacBERT**(Cui et al., 2020) is an improved Chinese BERT with novel MLM as correction pre-training task, which mitigates the discrepancy of pre-training and fine-tuning.
- **BART**(Shao et al., 2021) is a versatile, base-sized transformer model adapted for the Chinese language leveraging the BART architecture's denoising autoencoder capabilities for effective language understanding and generation.
- **GPT-3.5-Turbo** is the most capable and cost effective model in the GPT-3.5 family, which is a set of language models improved on GPT-3(Brown et al., 2020). It is optimized specifically for chat, but also performs excellently in traditional tasks.
- **GPT-4**(OpenAI, 2023) is a large multimodal model which outperforms GPT-3.5 in solving complex problems with higher precision, attributed to its broader general knowledge and enhanced reasoning capabilities.

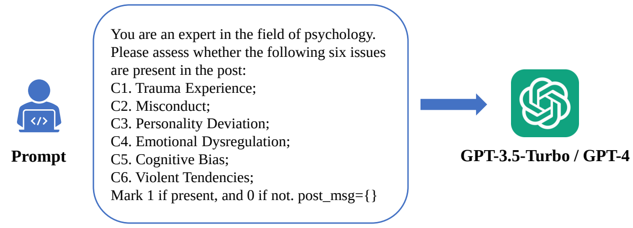

Figure 4: Use GPT-3.5-Turbo or GPT-4 to generate assessments for the dimensions in our scale.

Experiments Setup

All experiments are conducted on Ubuntu 20.04 with Intel Xeon Gold CPU@ 3.40GHz, and 8 NVIDIA GeForce RTX 3090 GPUs with 24GB RAM, using PyTorch 2.1.0 as the primary deep learning framework.

For fine-tuned models, we train the models using Adam optimizer with a 10 warm-up steps for a maximum of 30 epochs with patience of 10 epochs. We search the learning rate in {1e-5, 2e-5, 5e-5} and the batch size in {16, 32} with a grid search.

To evaluate the classification performance of LLMs, we designed a sophisticated prompt to obtain assessment results for the six dimensions on our scale. Figure 4 shows the process with specific examples of each dimension omitted.

We abstract the whole pipeline as a multi-label classification task, employing a 0.7:0.15:0.15 ratio for random partitioning of data into training, validation, and test sets. All

model evaluations are carried out on the test set. Following previous experimental configurations(Roy et al., 2023), we choose macro F1 scores and average AUC-ROC as our evaluation metrics.

Results Analysis

As shown in Table 2, most models in experiments struggle to achieve very high performance on ADViRDs, which indicates the dataset's inherent complexity and challenges it poses. Among the plug-and-play models, mDeBERTa performs better with a F1 score of 43.6% and an AUC-ROC of 57.9%, rivaling some of fine-tuned models. MacBERT demonstrates best capability, attaining the highest AUC-ROC of 60% in small models, which means it performs better in classifying negative cases. This can likely be attributed to its unique approach of applying MLM as a correction task for Chinese, which enables it to discern subtle features within the text. Meanwhile, as we expected, large language models especially GPT-4, showcase remarkable performance, achieving the best F1 score of 65.9% and an AUC-ROC of 67.6% in all. LLMs' talent in synthesizing and understanding contextually rich information significantly reflects their advanced levels in complex circumstances, such as in psychology, internal behaviors, and specific speaking styles on social networks.

Table 2: Overall F1 score and average AUC-ROC on ADViRDS, where the best results are highlighted in bold.

Model	Macro F1(%)	AUC-ROC
XLM-RoBERTa	39.3	42.1
mDeBERTa	43.6	57.9
BART	34.9	55.8
RoBERTa-wwm	34.1	57.2
MacBERT	42.3	**60.0**
GPT-3.5-Turbo	59.3	54.0
GPT-4	**65.9**	58.9

Table 3 depicts the AUC-ROC scores for each category on ADViRDS. Nearly all models struggled to accurately classify C1 (*Trauma Experience*). In contrast, categories C2 (*Misconduct*) and C4 (*Emotional Dysregulation*) saw more promising results, especially from mDeBERTa, which achieved an impressive AUC-ROC of 77.9% in C4. This could be attributed to the majority of samples explicitly mentioning traits (e.g. alcohol abuse or being easily angered). C3 (*Personality Deviation*) and C5 (*Cognitive Bias*), which demand a nuanced understanding of psychology to deduce mental states from observable behaviors, proved more challenging, posed a significant challenge. LLMs like GPT-4 excelled in C6 (*Violent Tendencies*), which requires the ability to discern subtle behaviors and draw inferences (e.g. history of violence). Overall, the models displayed varying sensitivities to different categories, demonstrating that ADViRDS tends to be a substantial challenge for simple models, relatively. It also shows that large language models offers new perspectives for detecting DV perpetrator traits and other psychological researches.

Table 3: The AUC-ROC of each category on ADViRDS, where the best results are highlighted in bold.

Model	C1	C2	C3	C4	C5	C6
XLM-RoBERTa	33.1	41.6	35.6	38.9	50.7	53.0
mDeBERTa	26.6	49.6	**67.8**	**77.9**	**66.8**	58.8
BART	49.4	71.9	50.0	51.9	51.1	60.6
RoBERTa-wwm	48.7	**76.5**	48.7	61.6	50.0	58.0
MacBERT	50.0	75.7	50.0	67.5	57.5	59.0
GPT-3.5-turbo	34.9	60.7	56.3	52.4	58.1	61.7
GPT-4	**56.3**	69.4	50.4	57.0	57.2	**63.0**

Conclusion and Future Work

Leveraging expert knowledge, we present ADViRDS, a refined scale and dataset specifically designed for detecting psychological traits of DV perpetrators. The data was extracted from Zhihu and annotated using our scale, which was also updated progressively during the annotation process. Utilizing the annotated dataset, we embarked on a comparative analysis employing various models to assess their efficacy in identifying the nuanced psychological traits of DV perpetrators. The results, however, revealed a notable shortfall in the models' performance, with even the most advanced GPT-4 model achieving a Macro F1 score of merely 65.9%. We must emphasize that there are currently not many studies applying large models in the field of psychology, especially in the field of domestic violence. Our experimental results further highlight the need for groundbreaking psychological traits modeling approaches, potentially integrating knowledge infusion for improved accuracy and interpretability. Additionally, recent research has explored using multiple LLM agents for stance detection in social media posts(Lan, Gao, Jin, & Li, 2023), a method that could be adapted for DV propensity detection in future studies. A limitation of our study is the reliance on a single data source, underscoring the need for future endeavors in data augmentation and sourcing from a variety of social media platforms to broaden and diversify our dataset, thereby enhancing the robustness and generalizability of our findings.

Acknowledgments

This work is supported by the National Key Research and Development Program of China(2021YFC3300500).

References

Becker, F., & French, L. (2004). Making the links: Child abuse, animal cruelty and domestic violence. *Child abuse review: Journal of the British Association for the Study and Prevention of Child Abuse and Neglect*, *13*(6), 399–414.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chandan, J. S., Taylor, J., Bradbury-Jones, C., Nirantharakumar, K., Kane, E., & Bandyopadhyay, S. (2020). Covid-19: a public health approach to manage domestic violence is needed. *The Lancet Public Health*, *5*(6), e309.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., & Hu, G. (2019). Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Desmet, B., & Hoste, V. (2018, May). Online suicide prevention through optimised text classification. *Information Sciences*, *439–440*, 61–78. doi: 10.1016/j.ins.2018.02.014

Ehrensaft, M. K., Moffitt, T. E., & Caspi, A. (2004). Clinically abusive relationships in an unselected birth cohort: men's and women's participation and developmental antecedents. *Journal of abnormal psychology*, *113*(2), 258.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Foa, E. B., Cascardi, M., Zoellner, L. A., & Feeny, N. C. (2000). Psychological and environmental factors associated with partner violence. *Trauma, Violence, & Abuse*, *1*(1), 67–91.

Gaur, M., Aribandi, V., Alambo, A., Kursuncu, U., Thirunarayan, K., Beich, J., ... Sheth, A. (2021, May). Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS. *PLOS ONE*, *16*(5), e0250448. doi: 10.1371/journal.pone.0250448

Homan, C. M., Schrading, J. N., Ptucha, R. W., Cerulli, C., & Ovesdotter Alm, C. (2020). Quantitative methods for analyzing intimate partner violence in microblogs: Observational study. *Journal of medical internet research*, *22*(11), e15347.

Hsieh, T., Wang, Y.-H., Hsieh, Y.-S., Ke, J.-T., Liu, C.-K., & Chen, S.-C. (2018). Measuring the unmeasurable—a study of domestic violence risk prediction and management. *Journal of Technology in Human Services*, *36*(1), 56–68.

Kelley, S. W., & Gillan, C. M. (2022). Using language in social media posts to study the network dynamics of depression longitudinally. *Nature communications*, *13*(1), 870.

Kourti, A., Stavridou, A., Panagouli, E., Psaltopoulou, T., Spiliopoulou, C., Tsolia, M., ... Tsitsika, A. (2023). Domestic violence during the covid-19 pandemic: a systematic review. *Trauma, violence, & abuse*, *24*(2), 719–745.

Lan, X., Gao, C., Jin, D., & Li, Y. (2023). Stance detection with collaborative role-infused llm-based agents. *arXiv preprint arXiv:2310.10467*.

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, *32*(1), 84–100.

Liu, M., Xue, J., Zhao, N., Wang, X., Jiao, D., & Zhu, T. (2021). Using social media to explore the consequences of domestic violence on mental health. *Journal of interpersonal violence*, *36*(3-4), NP1965–1985NP.

Liu, N., Zhang, Y., & Cao, Y. (2010). Association between personality characteristics of adult male severe physical abuser and childhood abuse. *Chinese Journal of Public Health*, *26*(6), 733–734.

MA, C., WU, L., & HONG, W. (2014). Influencing factors for male physical violence in intimate partners. *Chinese Journal of Public Health*, *30*(1), 38–42.

McCauley, J., Kern, D. E., Kolodner, K., Dill, L., Schroeder, A. F., DeChant, H. K., ... Derogatis, L. R. (1995). The "battering syndrome": prevalence and clinical characteristics of domestic violence in primary care internal medicine practices. *Annals of internal medicine*, *123*(10), 737–746.

Newberry, M. (2017). Pets in danger: Exploring the link between domestic violence and animal abuse. *Aggression and Violent Behavior*, *34*, 273–281.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015, May). Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183–188. doi: 10.1016/j.invent.2015.03.005

OpenAI. (2023). Gpt-4 technical report. *ArXiv*, *abs/2303.08774*.

Piquero, A. R., Jennings, W. G., Jemison, E., Kaukinen, C., & Knaul, F. M. (2021). Domestic violence during the covid-19 pandemic-evidence from a systematic review and meta-analysis. *Journal of criminal justice*, *74*, 101806.

Posner, K., Brent, D., Lucas, C., Gould, M., Stanley, B., Brown, G., ... others (2008). Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, *10*, 2008.

Rabani, S. T., Ud Din Khanday, A. M., Khan, Q. R., Hajam, U. A., Imran, A. S., & Kastrati, Z. (2023, July). Detecting suicidality on social media: Machine learning at rescue. *Egyptian Informatics Journal*, *24*(2), 291–302. doi: 10.1016/j.eij.2023.04.003

Roy, K., Zi, Y., Gaur, M., Malekar, J., Zhang, Q., Narayanan, V., & Sheth, A. (2023). Process knowledge-infused learning for clinician-friendly explanations. *arXiv preprint arXiv:2306.09824*.

Salehi, M., Ghahari, S., Hosseinzadeh, M., & Ghalichi, L. (2023). Domestic violence risk prediction in iran using a machine learning approach by analyzing persian textual content in social media. *Heliyon*, *9*(5).

Shao, Y., Geng, Z., Liu, Y., Dai, J., Yang, F., Zhe, L., ... Qiu, X. (2021). Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Shi, Y., Tian, Y., Tong, C., Zhu, C., Li, Q., Zhang, M., ... Zhou, P. (2023). Detect depression from social networks with sentiment knowledge sharing. In *Chinese national*

conference on social media processing (pp. 133–146).

Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y., & Shakeel, H. (2019). Deep learning for multi-class identification from domestic violence online posts. *IEEE access*, *7*, 46210–46224.

Volant, A. M., Johnson, J. A., Gullone, E., & Coleman, G. J. (2008). The relationship between domestic violence and animal abuse: An australian study. *Journal of Interpersonal Violence*, *23*(9), 1277–1295.

Wijenayake, S., Graham, T., & Christen, P. (2018). A decision tree approach to predicting recidivism in domestic violence. In *Trends and applications in knowledge discovery and data mining: Pakdd 2018 workshops, bdasc, bdm, ml4cyber, paisi, damemo, melbourne, vic, australia, june 3, 2018, revised selected papers 22* (pp. 3–15).

Wu, J., Wu, X., Hua, Y., Lin, S., Zheng, Y., & Yang, J. (2023). Exploring social media for early detection of depression in covid-19 patients. *arXiv preprint arXiv:2302.12044*.

Xu, H., Zeng, J., Tai, Z., & Hao, H. (2022). Public attention and sentiment toward intimate partner violence based on weibo in china: A text mining approach. In *Healthcare* (Vol. 10, p. 198).

Yu, R., Molero, Y., Lichtenstein, P., Larsson, H., Prescott-Mayling, L., Howard, L. M., & Fazel, S. (2023). Development and validation of a prediction tool for reoffending risk in domestic violence. *JAMA network open*, *6*(7), e2325494–e2325494.

Zanwar, S., Wiechmann, D., Qiao, Y., & Kerz, E. (2022). Exploring hybrid and ensemble models for multiclass prediction of mental health status on social media. *arXiv preprint arXiv:2212.09839*.

Zhang, Z., Chen, S., Wu, M., & Zhu, K. Q. (2022). Psychiatric scale guided risky post screening for early detection of depression. *arXiv preprint arXiv:2205.09497*.

Zhao, X.-f., Zhang, Y.-l., Li, L.-f., Zhou, Y.-f., & Li, H.-z. (2007). A study of personalities of physical domestic violence on male perpetrators. *Chinese journal of clinical psychology*, *15*(5), 543–544.

Zhao, X.-f., Zhang, Y.-L., Li, L.-f., Zhou, Y.-f., & Li, H.-z. (2008). Logistic regression analysis of the psychosociology of physical domestic violence on male perpetrators. *Chinese journal of clinical psychology*, *16*(2), 210–212.

Zhou, Z., Wang, Z., & Zimmer, F. (2022). Anonymous expression in an online community for women in china. *arXiv preprint arXiv:2206.07923*.

Zou, S., Zhang, Y., & Zhang, Y. (2001). A study on sociodemographic and cultural factors of spousal violence. *Chinese Journal of Clinical Psychology*.