MentalManip: A Dataset For Fine-grained Analysis of Mental Manipulation in Conversations

Yuxin Wang[1] **Ivory Yang**[2] **Saeed Hassanpour**[3] **Soroush Vosoughi**[4]
[1,2,4]Department of Computer Science, Dartmouth College
[3]Department of Biomedical Data Science, Dartmouth College
[1]`Yuxin.Wang.GR@dartmouth.edu`
[4]`Soroush.Vosoughi@dartmouth.edu`

Abstract

Mental manipulation, a significant form of abuse in interpersonal conversations, presents a challenge to identify due to its context-dependent and often subtle nature. The detection of manipulative language is essential for protecting potential victims, yet the field of Natural Language Processing (NLP) currently faces a scarcity of resources and research on this topic. Our study addresses this gap by introducing a new dataset, named MENTALMANIP, which consists of $4,000$ annotated fictional dialogues. This dataset enables a comprehensive analysis of mental manipulation, pinpointing both the techniques utilized for manipulation and the vulnerabilities targeted in victims. Our research further explores the effectiveness of leading-edge models in recognizing manipulative dialogue and its components through a series of experiments with various configurations. The results demonstrate that these models inadequately identify and categorize manipulative content. Attempts to improve their performance by fine-tuning with existing datasets on mental health and toxicity have not overcome these limitations. We anticipate that MENTALMANIP will stimulate further research, leading to progress in both understanding and mitigating the impact of mental manipulation in conversations.

1 Introduction

Mental manipulation is a deceptive strategy aimed at controlling or altering someone's thoughts and feelings to serve personal objectives (Barnhill, 2014). Facilitated by digital technologies, mental manipulation has gained unprecedented capability to target and influence individuals via interpersonal interactions and public dissemination of information (Ienca, 2023), causing significant mental health distress to victims (Hamel et al., 2023). Compared to overt verbal toxicity and abuse, such as hate speech, manipulation is more insidious, deceitful, and context-dependent, posing challenges for individuals and automatic moderation tools to discern. For years, the NLP community has witnessed advancements in verbal toxicity and abuse detection, but most of those works focus on context-free content and face challenges in identifying implicit toxicity (Wiegand et al., 2021; Mishra et al., 2020; Yin and Zubiaga, 2022; Deng et al., 2023). Existing works in dialogue systems have targeted context-aware toxicity, but are limited to direct toxicity, such as profanity, condescension and forms of hate speech (Baheti et al., 2021; Gao and Huang, 2017; Wang and Potts, 2019). We argue that existing toxicity detection resources are insufficient for developing automatic systems to detect and properly handle verbal mental manipulation. Additionally, current state-of-the-art Large Language Models (LLMs) are not well-positioned to address this issue, as demonstrated in Figure 1.

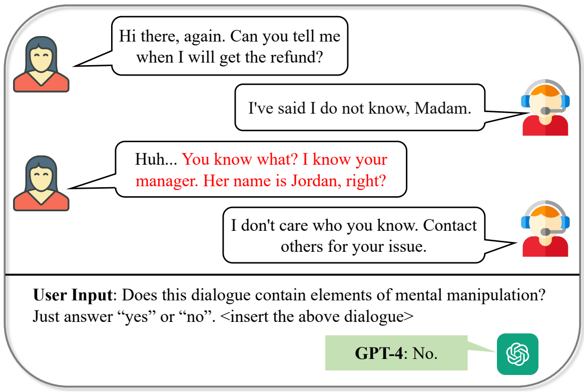

Figure 1: An example dialogue that contains elements of mental manipulation which GPT-4 fails to identify. The manipulative parts are highlighted in red.

This paper introduces MENTALMANIP, a dataset with multi-level annotations for mental manipulation detection and classification. We define mental manipulation as *using language to influence, alter, or control an individual's psychological state or perception for the manipulator's benefit.*

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3747–3764
August 11-16, 2024

Dataset	Research Scope	Dialogical	#Sample	#Avg. Utterance	Data Source	Label Scheme
Dreaddit (Turcan and McKeown, 2019)	Mental Stress	No	$3,553$	1	Reddit posts	Binary
SDCNL (Haque et al., 2021)	Depression & Suicide	No	$1,895$	1	Reddit posts	Binary
ToxiGen (Hartvigsen et al., 2022)	Hate Speech	No	$274,186$	1	GPT-3	Binary
DetexD (Yavnyi et al., 2023)	Delicate Text	No	$1,023$	1	Online forums	Multi-level
Fox News (Gao and Huang, 2017)	Hate Speech	Yes	814	2.0 (± 0.1)	News comments	Binary
TalkDown (Wang and Potts, 2019)	Condescension	Yes	$4,992$	2	Reddit comments	Binary
ToxiChat (Baheti et al., 2021)	Offensiveness	Yes	$2,000$	2.3 (± 0.5)	Reddit comments	Binary
MDRDC (Zhang et al., 2021)	Malevolence	Yes	$6,000$	4.8 (± 1.9)	Twitter threads	Multi-level
MENTALMANIP (ours)	Mental Manipulation	Yes	$4,000$	6.5 (± 5.3)	Movie scripts	Multi-level

Table 1: Comparison of properties and statistics: MENTALMANIP dataset versus existing datasets in verbal toxicity and mental health problem detection. The columns "#Sample" and "#Avg. Utterance" represent the number of instances/dialogues and the average number of utterances per dialogue, respectively. Numbers in round brackets are standard deviations.

This definition aligns with one of the explanations of manipulative influence in Barnhill's work on the philosophy of online manipulation (Barnhill, 2022). MENTALMANIP dataset contains $4,000$ multi-turn fictional dialogues between two characters extracted from online movie scripts. To enable fine-grained analysis, we proposed a labeling taxonomy covering three dimensions: presence of manipulation, manipulation technique, and targeted vulnerability, which are illustrated in Figure 2. Our taxonomy aids in the precise detection of mental manipulation and facilitates a nuanced classification of the techniques used by manipulators, as well as the vulnerabilities targeted in victims.

We conducted three classification tasks on MENTALMANIP to detect the existence of mental manipulation and its elements. Our experiments spanned state-of-the-art generative and discriminative LLMs across multiple settings. To investigate the effectiveness of existing datasets in toxic speech and mental health for our objectives, we fine-tuned LLMs with seven relevant datasets and conducted evaluations. Experimental results reveal that these LLMs are limited in understanding mental manipulation, as shown by a significant number of misclassified "manipulative" dialogues. Moreover, fine-tuning LLMs with these relevant datasets did not enhance their detection and classification capabilities on manipulative language. These findings highlight the importance of our dataset, and suggest an avenue for future studies in mental manipulation detection and analysis.

Our MENTALMANIP dataset, along with the code for statistical analysis and experiments in this paper, is available in our GitHub repository: `https://github.com/audreycs/MentalManip`.

2 Related Works

2.1 Mental Health Detection

Leveraging NLP technologies for the early detection and intervention of mental health issues stands as a valuable endeavor. In recent decades, there has been considerable research identifying specific mental health concerns, such as stress (Guntuku et al., 2019; Nijhawan et al., 2022), depression (Eichstaedt et al., 2018; Xu et al., 2019), and suicide (De Choudhury et al., 2016; Coppersmith et al., 2018). Scalable and accessible data resources for these issues have been proposed. (Turcan and McKeown, 2019; Haque et al., 2021; Naseem et al., 2022). More recently, research has shown that LLMs, like GPT-4, exhibit promising yet limited performance on tasks related to mental health (Yang et al., 2023; Xu et al., 2023). Researchers have pointed out a lack of explainability for the detection results of LLMs, and highlighted the importance of domain-specific fine-tuning on LLMs for better performance. These findings underscore the need for data resources featuring nuanced annotations and targeting unaddressed mental health issues.

2.2 Toxic Speech Detection

In NLP, "toxic speech" is an umbrella term referring to language that is rude, disrespectful, or offensive, potentially harming conversation quality and negatively impacting recipients (Dixon et al., 2018). Lots of benchmark datasets have been developed to detect explicit and implicit toxic speech in online posts and comments, including those focusing on racism and sexism (Waseem, 2016; Basile et al., 2019; Hartvigsen et al., 2022; Yavnyi et al., 2023), online harassment (Hosseinmardi et al., 2015; Rosa et al., 2018), and trolling (Miao et al., 2020). Recent works have also investigated performance of state-of-the-art LLMs on toxic speech (Wang et al., 2023). Although many mental manipulations, such as intimidation, fall under toxic speech, their subtle and complex nature create challenges beyond the capability of context-free toxicity detection methods. Existing works in dialogue systems address context-aware toxicity detection (Wang and Potts, 2019; Baheti et al., 2021; Zhang et al., 2021), but they focus on explicit toxicity and overlook implicit verbal manipulation.

Table 1 summarizes some existing datasets addressing toxicity or mental health problems.

3 Constructing MENTALMANIP

3.1 Taxonomy

Establishing a structured labeling taxonomy when developing a dataset is crucial. Drawing inspiration from Simon's research on psychological manipulation (Simon and Foley, 2011), we crafted a multi-level taxonomy encompassing three dimensions:

- Presence of Manipulation: This level employs binary classification, indicating if a dialogue contains elements of mental manipulation.
- Manipulation Technique: This level identifies specific manipulation techniques used in conversation.
- Targeted Vulnerability: The last level indicates particular victim vulnerabilities exploited by the manipulator.

We present the detailed taxonomy in Figure 2, which contains 11 different techniques and 5 vulnerabilities. We provide the definition of each technique and vulnerability in Appendix A. To ensure clarity and comprehensiveness, we incorporated insights from a psychological expert and feedback from annotators.

3.2 Data Source and Preprocessing

We prioritize dialogues as our primary data format as they maintain original context, unlike standalone comments and posts. To guarantee a semantically rich and stylistically diverse dataset, we prioritize human-crafted content over LLM-generated material. We finally chose Cornell Movie Dialogs Corpus[1] (Danescu-Niculescu-Mizil and Lee, 2011) as the data source to construct MENTALMANIP. The

[1]`https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html`

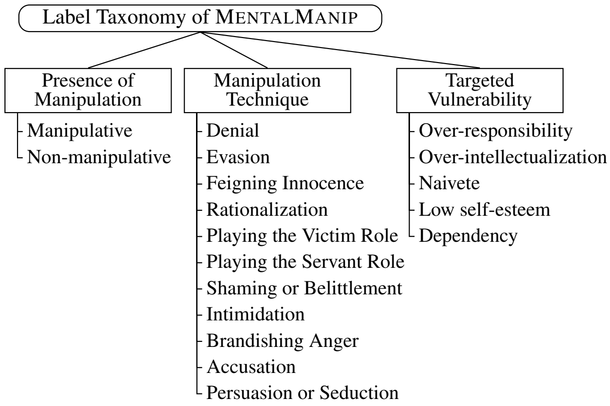

Figure 2: Multi-level taxonomy of MENTALMANIP.

Cornell Movie Dialogs Corpus contains $220,579$ conversational exchanges extracted from 617 raw movie scripts spanning a wide range of genres. The overwhelming majority of dialogues occur between two characters, which we standardized for. We replaced original speakers' names with "Person1" and "Person2" to eliminate potential biases.

Since manipulative language is relatively sparse in conversation, we need to filter the original data to get dialogues potentially containing elements of manipulation. We utilized two approaches to achieve this: 1) key phrase-based matching, and 2) BERT classification.

For key phrase-based matching, we sourced key phrases from online resources, selecting those that frequently occur in manipulative conversations, without restricting their n-gram size. After collection, we manually conducted tense conversion (converting all phrases to present tense), phrase simplification (e.g., "It's fine, nobody cares about me anyway" to "nobody cares about me"), and merging of similar phrases. Ultimately, we obtained a list of 175 cleaned key phrases.

Appendix B presents examples of the cleaned key phrases and details of the online resources we used. The full list of cleaned key phrases is available in our GitHub repository. To screen out candidate dialogues where key phrases are present, we implemented a length-adaptive matching criterion due to the lexical diversity of language. A dialogue is considered a match if any sentence contains at least $P\%$ tokens from a key phrase. The value of P is detailed in Table 2.

For BERT classification approach, we fine-tuned a pre-trained BERT model with a sequence classification head on top. Our goal was to get a classifier to differentiate manipulative dialogues from general toxic content. To prepare the training data,

Key Phrase Length l	$<= 4$	$<= 6$	$<= 10$	> 10
Matching Percentage P	100%	90%	80%	70%

Table 2: Length-adaptive matching criterion.

we inquired GPT-4 Turbo (Bubeck et al., 2023) by zero-shot prompting[2] on whole Cornell Movie Dialogs Corpus and obtained a set of "manipulative" dialogues flagged by GPT-4. We observed that GPT-4 generated a large portion of false positives for manipulative content. We examined $1,378$ "manipulative" dialogues flagged by GPT-4, and labeled only 464 dialogues as truly manipulative, with the remaining 914 being false positives. These $1,378$ labeled dialogues were then used to train the BERT classifier. Finally, we employed BERT classifier on all identified "manipulative" dialogues to obtain highly likely manipulative dialogues.

We initially identified $1,406$ dialogues through key phrase-based matching and $3,739$ dialogues using BERT classification, totaling $5,145$ dialogues. Following this, we eliminated duplicates and low-quality dialogues, including those that were extremely short or had broken contexts. Some dialogues were also rephrased to improve readability. After these adjustments, the total number of dialogues prepared for annotation was $4,876$.

3.3 Human Annotation

We established our annotation platform using Label Studio[3]. Each dialogue represents an annotation task. We recruited 17 college students, all native or fluent English speakers, to serve as annotators. The group of annotators reflects a diverse range of characteristics including gender (14 females, 3 males), ethnicity (11 Asians, 5 Whites, 1 Latino), educational backgrounds (majors such as English, Computer Science, and Physics), and cultural backgrounds (including both US-born and non-US-born individuals). During recruitment, applicants with an educational background in psychology or linguistics were preferred. We conducted tutorial sessions, required annotators to carefully read instructions, and monitored their annotation activities. Screenshots of the annotation platform and instructions are provided in Appendix I. To ensure annotation quality, we assigned three annotators to each task. During the task assignment process, we ensured that the same pairs of annotators were not assigned to evaluate the same dialogues. This approach helped to reduce the potential for bias when assessing inter-annotator agreement.

[2] API calling format is presented in Appendix C.

[3] https://labelstud.io/

Dataset	#Dialogue	Manip:Non-manip	Tech%	Vul%
MENTALMANIP$_{\text{con}}$	$2,915$	$2.24:1$	60.0%	20.8%
MENTALMANIP$_{\text{maj}}$	$4,000$	$2.38:1$	53.9%	18.3%

Table 3: Statistics of MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$, detailing dialogue counts (#Dialogue), the manipulative to non-manipulative dialogue ratio, and the percentages of dialogues labeled with techniques (Tech%) and vulnerability (Vul%). The exact numbers are provided in Table 15 in Appendix J.

In each task, annotators are presented a dialogue, then prompted to answer four questions:

- $Q1$ (binary choice): Does this dialogue contain elements of mental manipulation? (Options are "Yes" or "No".)
- $Q2$ (multiple choice): What techniques are used by the manipulator? (Options are techniques in Figure 2.)
- $Q3$ (binary choice): Are there any victims resulting from manipulation in this dialogue? (Options are "Yes" or "No".)
- $Q4$ (multiple choice): Which vulnerabilities are targeted in the victim? (Options are vulnerabilities in Figure 2.)

$Q2$ and $Q3$ are conditional upon $Q1$, and $Q4$ is conditional upon $Q3$. Annotators could choose at most three techniques and at most two vulnerabilities. To accommodate indecision, we included a "cannot decide" option in $Q2$ and $Q4$. Annotators were required to rate their confidence on a scale from 1 (not confident) to 5 (very confident). Furthermore, annotators could highlight sections they identified as manipulative to aid in our review. Appendix H provides an annotation example.

In total, we obtained more than 13K annotations. After quality review, the final size of well-labeled dialogues is $4,000$. Appendix D provides a detailed statistics of annotation quality, including the heat map of agreement between any two annotators, inter-gender agreement, scatter plot of agreement and confidence, and density distributions of agreement and confidence among annotators. We also calculated the inter-annotator agreement using Fleiss' Kappa (Fleiss and Cohen, 1973) based on their answers on $Q1$. The score was 0.596, indicating a moderate annotator agreement. This agreement level is as per our expectation, as the judgment of manipulation is very subjective. We name this dataset MENTALMANIP, and provide samples of it as a supplementary file.

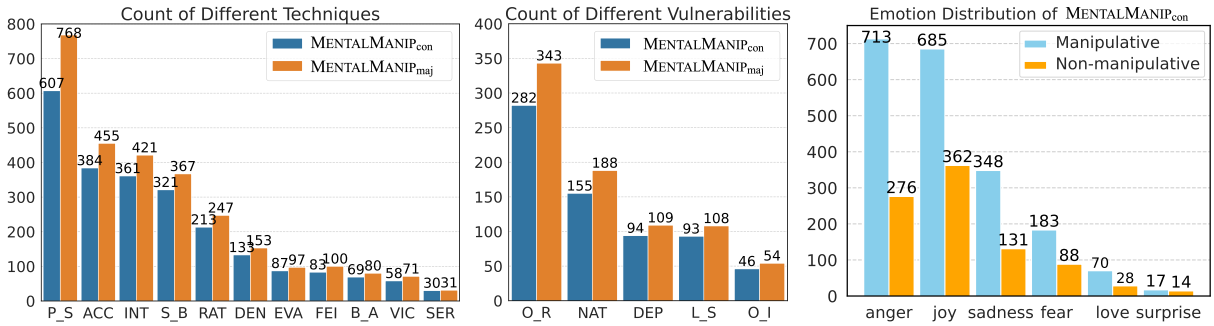

Figure 3: Statistics of MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$. The x-axis ticks in the left two panels are abbreviations for techniques and vulnerabilities (see Appendix A). The emotion distribution of MENTALMANIP$_{\text{maj}}$ dataset is in Appendix E.

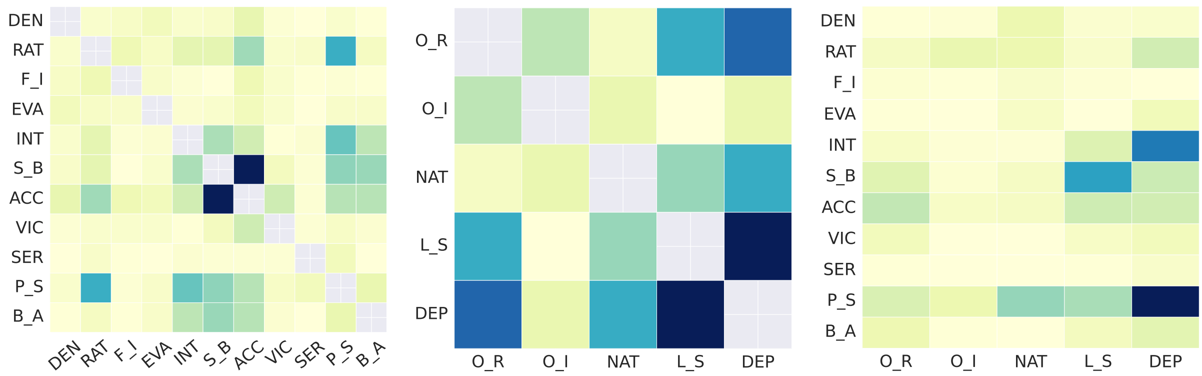

Figure 4: Co-occurrence heat maps among techniques (left), vulnerabilities (center), and techniques and vulnerabilities (right) in MENTALMANIP$_{\text{con}}$ dataset. Darker cell indicates a higher co-occurrence. The same figures showing results on MENTALMANIP$_{\text{maj}}$ dataset are in Appendix E.

3.4 Final Label Generation

To prepare the dataset for experiment, final labels need to be created. As each dialogue is annotated by three annotators, we adopted two strategies for generating the final labels:

- Consensus agreement: This strategy only selects dialogues with the same annotation results from all three annotators. The accordant result becomes the final label.
- Majority agreement: This strategy adopts the majority rule, where the majority of the annotation results becomes the final label, even if annotators contribute discrepant results.

Using these strategies on annotation results of question $Q1$, we obtained two versions of MENTALMANIP datasets. We denote the dataset generated using consensus agreement as MENTALMANIP$_{\text{con}}$ and the one using majority agreement as MENTALMANIP$_{\text{maj}}$.

We employed a specific strategy on both MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$ to generate the final labels for manipulative techniques and targeted vulnerabilities. If a technique or vulnerability is annotated by at least two annotators in one task, the technique or vulnerability will be added as the answer. This resulted in some dialogues lacking technique and vulnerability labels.

3.5 Dataset Statistics

In this section, we delve into the statistics of our datasets, MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$, as depicted in Table 3 and illustrated through Figures 3, 4, and 5. Our analysis utilizes a multi-class sentiment classification model, specifically the Distilbert-base-uncased-emotion model from Hugging Face, to determine the dominant emotion within each dialogue.

The analysis, presented in the left two panels of Figure 3, indicates a strong alignment in the distribution of manipulation techniques and vulnerabilities between MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$. Additionally, the same figure's right panel reveals that both manipulative and non-manipulative dialogues within MENTALMANIP$_{\text{con}}$ exhibit similar emotional distributions, with "joy" and "anger" being the two most common emotions. Figure 4 offers a heat

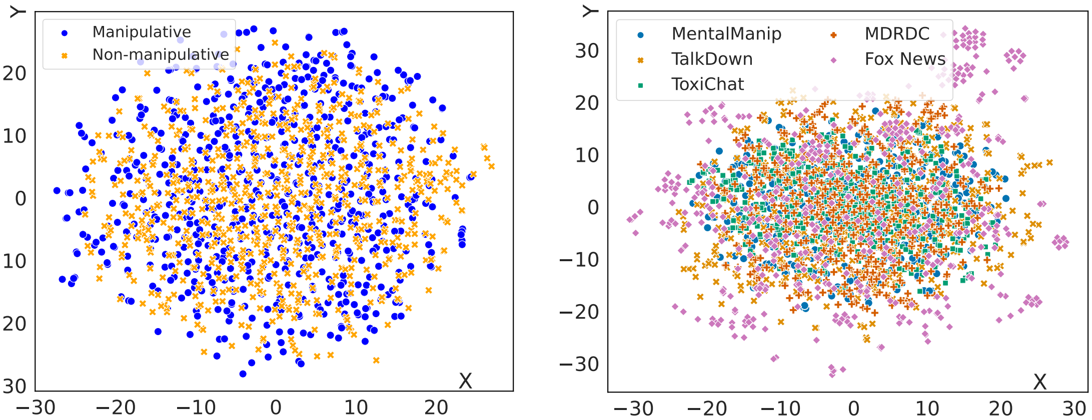

Figure 5: t-SNE visualization of Sentence Transformer embeddings of manipulative and non-manipulative dialogues in MENTALMANIP$_{\text{con}}$ (left) and the distribution of MENTALMANIP and other dialogical datasets (right).

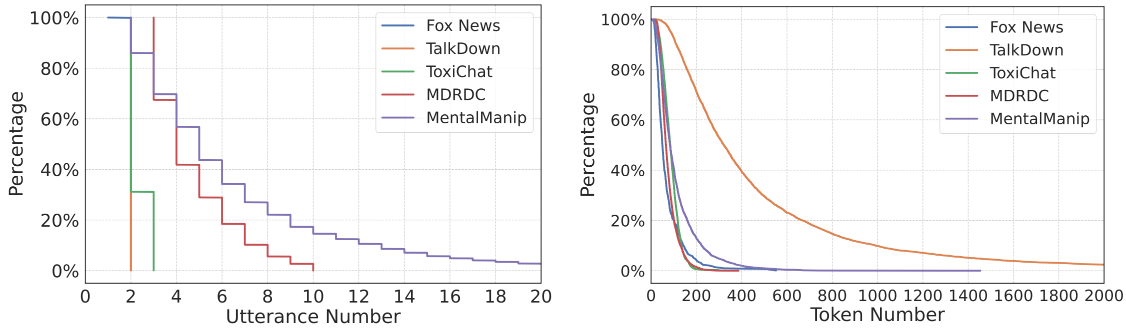

Figure 6: CCDF of utterance and token numbers per dialogue across MENTALMANIP and other dialogical datasets listed in Table 1.

map that elucidates the correlation between manipulation techniques and vulnerabilities, uncovering prevalent patterns like the association of accusations with shaming or belittling. Moreover, Figure 5's left panel showcases a t-SNE visualization of Sentence Transformer embeddings for both manipulative and non-manipulative dialogues within MENTALMANIP$_{\text{con}}$, using the all-MiniLM-L12-v2 model from Hugging Face. This visualization underscores the difficulty of distinguishing between manipulative and non-manipulative dialogues due to their intertwined embeddings.

Furthermore, we compare MENTALMANIP with other dialogical datasets listed in Table 1, noting that MENTALMANIP encompasses a greater volume of conversational exchanges, suggesting a richer dialogue context. The Complementary Cumulative Distribution Function (CCDF) for utterance and token counts of MENTALMANIP compared to other dialogical datasets is depicted in Figure 6. The right panel of Figure 5 visualizes the distribution of these datasets in the embedding space, illustrating significant overlap among them, except for the distinct clustering pattern of Fox News comments.

In summary, our analysis highlights the challenge of differentiating between manipulative and non-manipulative dialogues, indicating that reliance on emotion classification or conventional text embeddings alone is insufficient for this purpose. Moreover, our dataset's comparison with other datasets confirms its comprehensive distribution and diversity, aligning with the variety observed in related datasets.

4 Experiments

4.1 Experiment Setting

We conducted experiments of three classification tasks on MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$ to assess performance of state-of-art models in detecting mental manipulation: Manipulation Detection (Section 4.2), Technique Classification (Section 4.3), and Vulnerability Classification (Section 4.3). We analyzed the performance of GPT-4 Turbo (Bubeck et al., 2023), Llama-2-7B, Llama-2-13B[4] (Touvron et al., 2023), and RoBERTa-base (Liu et al., 2019) across three experimental settings: zero-shot

[4]Both Llama-2-7B and Llama-2-13B are Chat versions.

prompting, few-shot prompting, and fine-tuning. For zero-shot prompting, we presented a dialogue to LLMs to assess if it contained elements of mental manipulation. In few-shot prompting, aside from instructions, we randomly provided one non-manipulative and two manipulative dialogues with true answers as examples. In fine-tuning, Llama-2-13B and RoBERTa-base were fine-tuned on specific datasets, with Llama-2-13B undergoing instruction tuning and RoBERTa-base receiving traditional supervised fine-tuning. Formats for zero-shot and few-shot prompting are detailed in Appendix C. For Llama's training on different datasets, instructions were adapted to fit respective tasks. GPT-4 Turbo's implementation followed OpenAI's official cookbook[5]. Talkdown dataset was ignored due to its lengthy dialogues which far surpass the input token limit of RoBERTa-base.

For experiment data, we randomly split MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$ into training, validation, and test sets with a ratio of 6:2:2. We ensured proportional representation of manipulative and non-manipulative dialogues, and consistent inclusion of each technique and vulnerability across all sets. All experiments were performed on three Quadro RTX 6000 GPUs. We set the temperatures of GPT-4 Turbo and LLaMA-2 to 0.1 and 0.6, respectively. At these levels, the models exhibit more consistent and less random responses.

We seek to elucidate the following aspects:

- The effectiveness of LLMs in identifying and categorizing mental manipulation based on their inherent knowledge.
- The performance of LLMs when prompted with examples.
- The performance of LLMs post fine-tuning on relevant datasets.

4.2 Manipulation Detection

This task is framed as a binary classification task. In our interactions with ChatGPT and GPT-4, we found it tends to mistakenly classify non-manipulative dialogues as manipulative if they feature general toxicity, like profanity, without actual manipulative intent. Thus, we were keen to investigate the over-reactivity of LLMs when identifying mental manipulation.

Hypersensitivity of LLMs: We examined GPT-4 Turbo, Llama-2-7B, and Llama-2-13B on

[5] https://github.com/openai/openai-cookbook/

Predictions	GPT-4 Turbo	Llama-2-7B	Llama-2-13B
Manipulative	312	895	879
Non-manipulative	587	4	20
Accuracy	0.653	0.004	0.022

Table 4: Out of 899 non-manipulative dialogues in MENTALMANIP$_{\text{con}}$, the number of dialogues predicted as manipulative and non-manipulative.

the manipulation detection task using 899 non-manipulative dialogues in MENTALMANIP$_{\text{con}}$. Prediction results are detailed in Table 4. GPT-4 Turbo incorrectly identified 312 dialogues as manipulative. Both Llama-2-7B and Llama-2-13B exhibited poor accuracy, mis-classifying almost all non-manipulative dialogues, with Llama-2-13B showing slightly better performance. These results indicate Llama-2's limited capability in accurately discerning mental manipulation.

Then, we conducted manipulation detection on the entirety of MENTALMANIP$_{\text{con}}$ and MENTALMANIP$_{\text{maj}}$. Note that the distribution of manipulative and non-manipulative dialogues in both datasets is imbalanced, with manipulative dialogues being more prevalent, as detailed in Table 3. We evaluated the models based on binary Precision, binary Recall, Accuracy, micro F_1, and macro F_1. Because of binary classification, the accuracy has the same score as micro F_1.

Experiment results are presented in Table 5 and Table 6. It is observed that MENTALMANIP$_{\text{maj}}$ poses a greater challenge for prediction, as we expected. In zero-shot and few-shot prompting, Llama-2-13B classifies nearly all dialogues as manipulative, causing high recall rates. Few-shot prompting improves Accuracy and F_1 scores for both GPT-4 Turbo and Llama-2-13B. For GPT-4 Turbo, few-shot prompting increases its Recall, making it more likely to identify dialogues as manipulative. For Llama-2-13B, few-shot prompting makes it less sensitive and produces fewer manipulative predictions. Appendix F provides the confusion matrices for prediction results of GPT-4 Turbo and Llama-2-13B under zero-shot and few-shot prompting on MENTALMANIP$_{\text{con}}$. For fine-tuning, Llama-2-13B on Dreaddit gives the best performance among all finetuning results on existing datasets. Note that Dreaddit is about detecting Mental Stress. However, fine-tuning Llama-2-13B on all existing relevant datasets does not notably enhance performance beyond zero-shot or few-shot prompting outcomes. RoBERTa-base overall exhibits inferior Accuracy compared to Llama-2-13B.

Experiment Setting	Training Dataset	GPT-4 Turbo					Llama-2-13B					RoBERTa-base				
		P	R	Acc	F_1^{mi}	F_1^{ma}	P	R	Acc	F_1^{mi}	F_1^{ma}	P	R	Acc	F_1^{mi}	F_1^{ma}
Zero-shot prompting		.788	.682	.657	.657	.629	.693	.997	.696	.696	.450	–	–	–	–	–
Few-shot prompting	MentalManip$_{con}$	.802	.792	.724	.724	.683	.735	.912	.715	.715	.602	–	–	–	–	–
Fine-tuning	Dreaddit	–	–	–	–	–	.721	.982	.727	.727	.559	.864	.208	.435	.435	.422
	SDCNL	–	–	–	–	–	.698	.995	.702	.702	.471	.684	.822	.619	.619	.488
	ToxiGen	–	–	–	–	–	.693	.999	.696	.696	.446	.717	.864	.674	.674	.559
	DetexD	–	–	–	–	–	.696	.992	.698	.698	.465	.803	.215	.427	.427	.416
	Fox News	–	–	–	–	–	.690	.997	.691	.691	.434	.000	.000	.312	.312	.238
	ToxiChat	–	–	–	–	–	.689	.999	.691	.691	.429	.791	.333	.483	.483	.483
	MDRDC	–	–	–	–	–	.695	.999	.700	.700	.457	.743	.749	.651	.651	.595
	MentalManip$_{con}$	–	–	–	–	–	.828	.835	.768	.768	.731	.786	.904	.766	.766	.700

Table 5: Results of manipulation detection task on **MentalManip$_{con}$**. P, R, Acc, F_1^{mi}, and F_1^{ma} stands for binary precision, binary recall, accuracy, micro F_1, and macro F_1 respectively.

Experiment Setting	Training Dataset	GPT-4 Turbo					Llama-2-13B					RoBERTa-base				
		P	R	Acc	F_1^{mi}	F_1^{ma}	P	R	Acc	F_1^{mi}	F_1^{ma}	P	R	Acc	F_1^{mi}	F_1^{ma}
Zero-shot prompting		.816	.632	.632	.632	.602	.722	.997	.721	.721	.432	–	–	–	–	–
Few-shot prompting	MentalManip$_{maj}$	.812	.710	.672	.672	.627	.732	.979	.726	.726	.486	–	–	–	–	–
Fine-tuning	Dreaddit	–	–	–	–	–	.742	.960	.731	.731	.533	.814	.191	.386	.386	.378
	SDCNL	–	–	–	–	–	.726	.983	.720	.720	.458	.696	.565	.510	.510	.459
	ToxiGen	–	–	–	–	–	.723	.997	.723	.723	.436	.731	.734	.615	.615	.521
	DetexD	–	–	–	–	–	.727	.988	.724	.724	.460	.792	.225	.400	.400	.396
	Fox News	–	–	–	–	–	.722	.997	.721	.721	.432	.000	.000	.280	.280	.218
	ToxiChat	–	–	–	–	–	.721	.998	.721	.721	.428	.797	.348	.467	.467	.466
	MDRDC	–	–	–	–	–	.724	.998	.725	.725	.441	.779	.682	.632	.632	.581
	MentalManip$_{maj}$	–	–	–	–	–	.809	.851	.748	.748	.673	.791	.875	.743	.743	.651

Table 6: Results of manipulation detection task on **MentalManip$_{maj}$**. P, R, Acc, F_1^{mi}, and F_1^{ma} stands for binary precision, binary recall, accuracy, micro F_1, and macro F_1 respectively.

Experiment Setting	Model	Technique					Vulnerability				
		P^{mi}	R^{mi}	Acc	F_1^{mi}	F_1^{ma}	P^{mi}	R^{mi}	Acc	F_1^{mi}	F_1^{ma}
Zero-shot prompting	GPT-4 Turbo	.311	.618	.111	.414	.376	.373	.786	.092	.506	.423
	Llama-2-13B	.174	.448	.025	.250	.233	.164	.366	.000	.227	.222
Few-shot prompting	GPT-4 Turbo	.387	.533	.224	.449	.394	.429	.626	.269	.509	.370
	Llama-2-13B	.324	.283	.205	.302	.193	.157	.183	.042	.169	.162
Fine-tuning	Llama-2-13B	.349	.821	.029	.490	.384	.265	.756	.008	.393	.280
	RoBERTa-base	.479	.470	.264	.475	.334	.532	.496	.445	.513	.250

Table 7: Results of technique and vulnerability multi-label classification on **MentalManip$_{con}$**. P^{mi}, R^{mi}, Acc, F_1^{mi} and F_1^{ma} stands for micro precision, micro recall, accuracy, micro F_1 and macro F_1, respectively.

Specifically, fine-tuning it on Fox News dataset results in badly degraded performance. This decline may stem from the broader semantic distribution of the Fox News dataset, as illustrated in Figure 5.

4.3 Technique and Vulnerability Classification

Here we examined these models on multi-label classification tasks to identify manipulative techniques and victim vulnerabilities. We present the experiment results on MentalManip$_{con}$ in Table 7, where we report micro Precision and micro Recall. For few-shot prompting on both classification tasks, we provided 2 randomly chosen examples. We can observe that GPT-4 Turbo performs better than Llama-2-13B under zero-shot and few-shot prompting, and few-shot prompting increases their accuracy. Fine-tuning Llama-2-13B on MentalManip$_{con}$ still gives better performance than fine-tuning RoBERTa-base. Precision and recall scores for each technique and vulnerability category are detailed in Table 8, specifically for zero-shot prompting using GPT-4 Turbo and

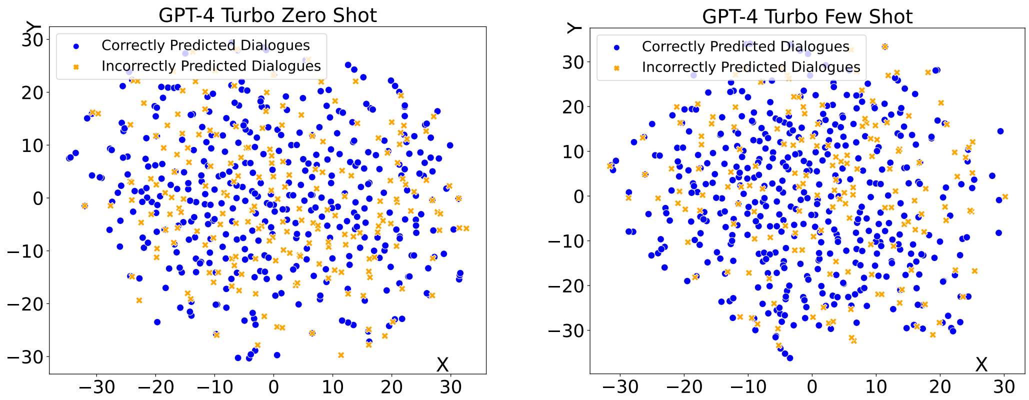

Figure 7: The t-SNE visualization of Sentence Transformer embeddings for dialogues in MENTALMANIP$_{con}$'s test set, which were correctly or incorrectly predicted by GPT-4 Turbo under both zero-shot and few-shot settings.

Llama-2-13B.

4.4 Analysis on Incorrect Predictions

In this section, we explore whether there are significant semantic differences between dialogue instances where models correctly predicted the label and those where they failed. We extracted dialogues from MENTALMANIP$_{con}$'s test set that were correctly and incorrectly predicted by GPT-4 Turbo. Under the zero-shot setting, GPT-4 Turbo accurately predicted 383 dialogues and inaccurately predicted 200. In the few-shot setting, the numbers were 422 correct predictions and 161 incorrect. We used t-SNE visualization of Sentence Transformer embeddings to analyze the semantic distributions of these dialogues, which are presented in Figure 7. The visualizations show that the dialogues, whether correctly or incorrectly predicted, are semantically indistinguishable, underscoring the difficulty of differentiating manipulative language based solely on lexical or semantic features.

5 Conclusion and Future Studies

This study introduces MENTALMANIP, a pioneering dataset aimed at identifying and classifying mental manipulation in a fine-grained level. We assessed GPT-4 Turbo, Llama-2-13B, and RoBERTa-base across three classification tasks under various settings. Experiment results reveal that models' understanding of mental manipulation do not align well with human perspectives. LLMs tend to incorrectly identify general toxicity as manipulation, a challenge particularly pronounced in smaller LLMs such as Llama-2-7B and Llama-2-13B. In future work, it would be worthwhile to expand the dataset sources beyond the Cornell Movie Dialog Corpus and incorporate real-case interpersonal interaction data. We also recognize that the performance of LLMs on more complex prompting paradigms, such as chain-of-thought, can be investigated.

Detecting and finely classifying mental manipulation in conversations is a challenging task due to the implicit and context-dependent nature of the language often used. Furthermore, the subjectivity of human judgment complicates the alignment of AI models' predictions with human choices. We have made the MENTALMANIP dataset publicly available for future studies and hope it will inspire and foster further research in various NLP tasks and applications.

Technique/Vulnerability	GPT-4 Turbo		Llama-2-13B	
	P	R	P	R
Denial	0.180	0.857	0.085	0.900
Evasion	0.208	0.714	0.060	1.000
Feigning Innocence	0.184	0.823	0.063	0.563
Rationalization	0.204	0.789	0.178	0.568
Playing the Victim Role	0.056	0.875	0.071	0.625
Playing the Servant Role	0.138	1.000	0.000	0.000
Shaming or Belittlement	0.473	0.709	0.304	0.688
Intimidation	0.476	0.861	0.500	0.467
Brandishing Anger	0.538	0.259	0.208	0.200
Accusation	0.450	0.529	0.353	0.358
Persuasion or Seduction	0.778	0.395	0.610	0.217
Over-responsibility	0.180	0.692	0.109	1.000
Over-intellectualization	0.200	0.222	0.136	0.667
Naivete	0.234	0.833	0.187	0.944
Low self-esteem	0.384	0.909	0.200	0.182
Dependency	0.635	0.810	0.750	0.103

Table 8: Precision and Recall of each technique and vulnerability category under zero-shot setting.

6 Limitations

We recognize that MENTALMANIP dataset has several limitations:

Language and Format The MENTALMANIP is limited to English-language content and focuses exclusively on dialogues between two individuals. Real-world interactions, however, are frequently more multifaceted. Consequently, this restriction may limit the dataset's applicability to more complex scenarios.

Data Source The MENTALMANIP is derived from online movie scripts, which means the speech style presented may not accurately reflect natural, real-life communication.

Data Annotation The process of annotation is inherently subjective, which can introduce uncertainties in the precision of labeling. Additionally, the selection of annotators could lead to significant biases. For instance, despite the lack of notable differences in inter-annotator agreement across genders, an imbalanced gender demographic among our annotators could still influence the results. We recognize that despite our best efforts, assembling an annotator pool that perfectly mirrors the general population remains a challenging endeavor.

7 Ethics and Broad Impact

Before annotating, we noted that many dialogues, especially from R-rated movies, contained profanity that might upset annotators. To protect their well-being, we rephrased these instances into milder language while keeping the original context. When recruiting annotators, we emphasized ensuring a diverse team in terms of race and gender. Throughout the annotation phase, we actively encouraged annotator feedback, as summarized in Appendix G.

The MENTALMANIP dataset contains a range of uncensored sensitive materials, including hate speech, violence, threats, mental health issues, sexual content, profanity, and more. While our dataset is primarily designed for detection and classification tasks, we recognize the potential for misuse, particularly in the training of malicious generative AI systems. For example, there is a risk that the data could be used to create automated chatbot systems that employ manipulative language for unethical purposes like scams. It is crucial to address these risks to ensure responsible use.

Acknowledgments

We thank the Label Studio Platform for providing complimentary access to the Enterprise Cloud Platform through their Academic Program. We also thank the annotators for their time and effort. This research was supported in part by grants from the US National Library of Medicine (R01LM012837 & R01LM013833) and the US National Cancer Institute (R01CA249758).

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862. Association for Computational Linguistics.

Anne Barnhill. 2014. What is manipulation? In *Manipulation: Theory and Practice*. Oxford University Press.

Anne Barnhill. 2022. How philosophy might contribute to the practical ethics of online manipulation. In *The philosophy of online manipulation*, pages 49–71. Routledge.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1178222618792860.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems. CHI Conference*, volume 2016, pages 2098–2110.

Jiawen Deng, Jiale Cheng, Hao Sun, Zhexin Zhang, and Minlie Huang. 2023. Towards safer generative language models: A survey on safety risks, evaluations, and improvements.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA. ACM.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. In *Proceedings of the National Academy of Sciences*, volume 115, pages 11203–11208. Proceedings of the National Academy of Sciences.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 260–266. Incoma Ltd. Shoumen, Bulgaria.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C. Eichstaedt, and Lyle H. Ungar. 2019. Understanding and measuring psychological stress using social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 214–225.

John Hamel, Clare E. B. Cannon, and Nicola Graham-Kevan. 2023. The consequences of psychological abuse and control in intimate partner relationships. *Traumatology*.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine Learning*, pages 436–447. Springer International Publishing.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Marcello Ienca. 2023. On artificial intelligence and manipulation. *Topoi*, 42(3):833–842.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lin Miao, Mark Last, and Marina Litvak. 2020. Detecting troll tweets in a bilingual corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6247–6254, Marseille, France. European Language Resources Association.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Usman Naseem, Adam G. Dunn, Jinman Kim, and Matloob Khushi. 2022. Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 2563–2572, New York, NY, USA. Association for Computing Machinery.

Tanya Nijhawan, Girija Attigeri, and T. Ananthakrishna. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9(1):33.

Hugo Rosa, Joao P. Carvalho, Pável Calado, Bruno Martins, Ricardo Ribeiro, and Luisa Coheur. 2018. Using fuzzy fingerprints for cyberbullying detection in social networks. In *2018 IEEE International Conference on Fuzzy Systems*, pages 1–7, Rio de Janeiro, Brazil. IEEE Press.

George K Simon and Kevin Foley. 2011. *In sheep's clothing: Understanding and dealing with manipulative people*. Tantor Media, Incorporated.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing, pages 3709–3717, Hong Kong, China. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, volume 3, pages 116:1–116:33.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-llm: Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Serhii Yavnyi, Oleksii Sliusarenko, Jade Razzaghi, Olena Nahorna, Yichen Mo, Knar Hovakimyan, and Artem Chernodub. 2023. DeTexD: A Benchmark Dataset for Delicate Text Detection. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 14–28, Toronto, Canada. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2022. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210.

Yangjun Zhang, Pengjie Ren, and Maarten De Rijke. 2021. A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses. *Journal of the Association for Information Science and Technology*, 72(12):1477–1497.

A Description of Taxonomy

Definitions of 11 manipulation techniques, with their abbreviations inside parentheses:

1. Denial (DEN): The manipulator either denies wrongdoing or pretends to be confused about others' concerns.
2. Evasion (EVA): The manipulator refuses to pay attention to something or gives irrelevant or vague responses.
3. Feigning innocence (FEI): The manipulator implies that any harm caused was accidental.
4. Rationalization (RAT): The manipulator rationalizes their inappropriate behavior with excuses.
5. Playing the Victim Role (VIC): The manipulator portrays themselves as a victim to gain sympathy, attention, or divert focus from their misconduct.
6. Playing the Servant Role (SER): The manipulator disguises their self-serving motives as a contribution to a more noble cause.
7. Shaming or Belittlement (S_B): The manipulator uses sarcasm, criticism, and put-downs to make others feel inferior, unworthy, or embarrassed.
8. Intimidation (INT): The manipulator places others on the defensive by using veiled threats.
9. Brandishing Anger (B_A): The manipulator uses anger to brandish emotional intensity to shock the victim into submission.
10. Accusation (ACC): The manipulator suggests that the victim is at fault, selfish, uncaring, or leading an excessively easy life.
11. Persuasion or Seduction (P_S): The manipulator employs charm, emotional appeal, or logical reasoning to make others lower their defenses.

Definitions of 5 vulnerabilities targeted, with their abbreviations inside parentheses:

1. Naivety (NAT): The victim is easily trusting and struggles to accept that others might be malevolent.
2. Dependency (DEP): The victim has interest-based or emotional dependencies on the manipulator.
3. Over-responsibility (O_R): The victim is overly self-critical and sets high standards for themselves, often assuming undue blame and responsibility for the manipulator's actions.

4. Over-intellectualization (O_I): The victim rationalizes the manipulator's hurtful behavior by believing there is always a justified reason behind it.
5. Low self-esteem (L_S): The victim is self-doubting and unconfident in pursuing their own wants and needs.

B Key Phrase-based Matching

B.1 Key Phrases Examples

- "you make me do this"
- "how could you do this to me"
- "know your place"
- "you should not feel that way"
- "what more do you want"
- "i do not remember"
- "i do not like drama"
- "watch your step"
- "you always do this"
- "you are too sensitive"
- "it was not intentional"
- "you do not love me"
- "you would do it if you love me"
- "it is all in the past"

The complete set of key phrases can be found in our GitHub repository.

B.2 Online Resources

We collected as many different manipulative phrases as we could from a number of websites suggested by Google by searching *"phrases manipulative people use"* and likewise. Some of the websites browsed are below:

- https://geediting.com/10-phrases-manipulative-people-use-to-control-their-partner-in-a-relationship
- https://thevessel.io/phrases-manipulative-people-use-to-play-the-victim
- https://hackspirit.com/common-phrases-people-use-to-manipulate-you
- https://geediting.com/phrases-manipulative-people-use-to-blame-others-for-their-actions

We found that there was significant overlap across phrases as we collected phrases.

C Prompting Formats for GPT-4 and Llama-2

When designing the prompts, we aimed to craft phrases that were both straightforward and broad, reflecting inquiries that real users might pose.

C.1 Manipulation Detection

Zero-shot prompting format:

```
'''
I will provide you with a dialogue.
Please determine if it contains elements
of mental manipulation. Just answer with
'Yes' or 'No', and don't add anything else.

<insert dialogue>
'''
```

Few-shot prompting format:

```
'''
I will provide you with a dialogue.
Please determine if it contains elements
of mental manipulation. Just answer with
'Yes' or 'No', and don't add anything else.
Here are 3 examples:

Example 1:
<insert example_dialogue1>
<insert example_answer1>

Example 2:
<insert example_dialogue2>
<insert example_answer2>

Example 3:
<insert example_dialogue3>
<insert example_answer3>

<insert dialogue>
'''
```

C.2 Technique and Vulnerability Classification

Zero-shot prompting format:

```
'''
Here are the definitions of 11 different
mental manipulation techniques:
<insert definitions of 11 techniques>

Now, I will provide you with a dialogue
that contains elements of mental manipulation.
Please determine which manipulative
techniques are used by the manipulator.
Respond only with the names of the techniques,
and do not add anything else.

<insert dialogue>
'''
```

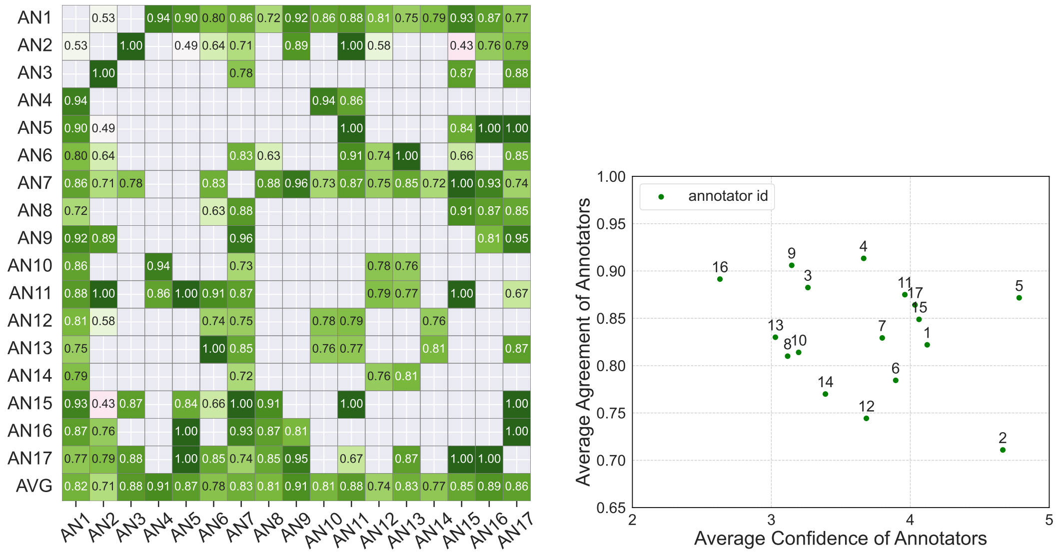

Figure 8: (left) Inter-annotator agreement of any two annotators based on their answers of whether a dialogue is manipulative. The last row is the average agreement score of each annotator. (right) Scatter plot of annotators' average confidence and inter-annotator agreement scores.

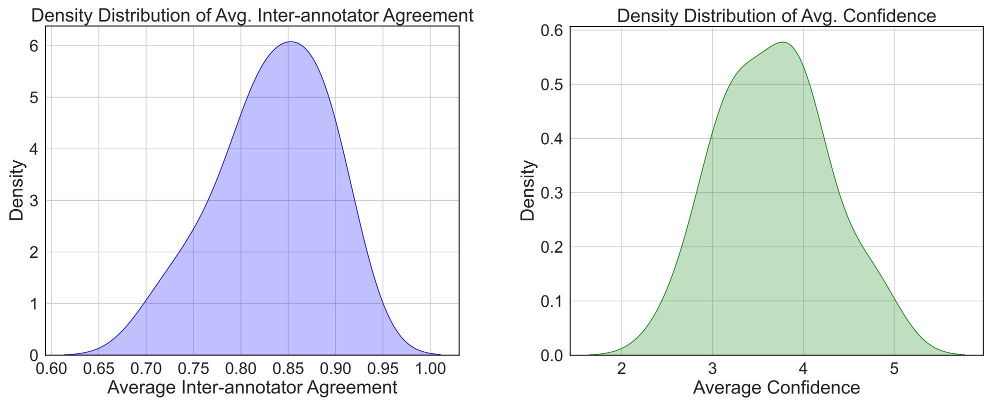

Figure 9: Density distribution of inter-annotator agreement and confidence of annotators.

Few-shot prompting format:

```
'''
Here are the definitions of 11 different
mental manipulation techniques:
<insert definitions of 11 techniques>

Now, I will provide you with a dialogue
that contains elements of mental manipulation.
Please determine which manipulative
techniques are used by the manipulator.
Respond only with the names of the
techniques, and do not add anything
else. Here are 2 examples:

Example 1:
<insert example_dialogue1>
<insert example_answer1>

Example 2:
<insert example_dialogue2>
<insert example_answer2>

<insert dialogue>
'''
```

For vulnerability classification, the prompting formats are similar.

D Analysis on Annotation Quality

Figure 8 is a heat map showing the inter-annotator agreement scores between any two annotators based on their answers for question $Q1$. For annotator ANi and annotator ANj, their agreement score is calculated as:

$$score_{ij} = \frac{\|\text{Annotations with same results}\|}{\|\text{Common tasks of } ANi \text{ and } ANj\|}$$

Gender (Count)	Female (14)	Male (3)	Avg. Conf.
Female (14)	0.82	0.83	3.61
Male (3)	0.83	0.91	3.96

Table 9: The average inter-annotator agreement scores across female and male annotators, and the average confidence scores of female and male annotators.

The last row is the average agreement score of each annotator. We can see that all annotators have a moderate to strong average agreement (≥ 0.7) with other annotators assigned with the same tasks.

Figure 8 is the scatter plot of annotators' average confidence and inter-annotator agreement scores. 16 out of 17 annotators reported an average confidence score above 3. We calculated the Spearman's rank correlation between annotators' inter-annotator agreement and confidence levels. The statistic value is -0.21 and P-value is 0.41, which reveals a very slight negative correlation between inter-annotator agreement and confidence levels, which is surprising for us. This observation may be attributed to several factors. Firstly, many annotators tend to assign lower or medium confidence scores (2 or 3) when they are not entirely certain of their decisions, regardless of their comprehension of the dialogue and the available options. Conversely, some annotators habitually assign high confidence scores (4 or 5) to most of their decisions, reflecting individual differences in confidence assessment. Secondly, the assessment of mental manipulation is inherently subjective and lacks a uniform standard. Variations in what constitutes manipulation among annotators—with some setting higher thresholds than others—further diminish the reliability of inter-annotator agreement as a measure of annotation quality.

We also analyzed the inter-annotator agreement and average confidence by gender, as detailed in Table 9. On average, male annotators exhibited higher confidence scores compared to their female counterparts. Furthermore, the inter-annotator agreement was higher among male annotators than among females. These differences could be significantly affected by the number of annotators of each gender and the volume and difficulty of tasks to which they were jointly assigned.

Figure 9 illustrates the density distributions of annotators' average agreement and confidence scores, both of which exhibit a normalized distribution.

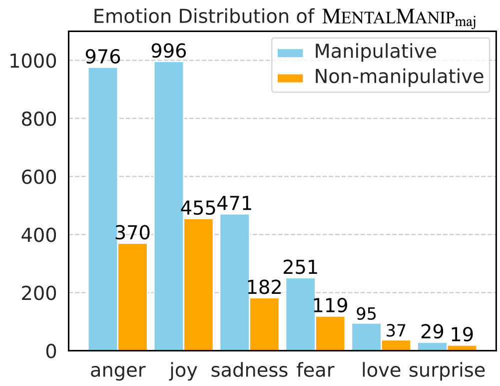

Figure 10: Emotion distribution of dialogues in dataset MentalManip$_{\text{maj}}$.

E More Statistics of MentalManip$_{\text{maj}}$

The emotion distribution of dialogues in MentalManip$_{\text{maj}}$ dataset is in Figure 10. The co-occurrence details are in Figure 11.

F Confusion Matrices

Please see Table 10, 11, 12, and 13.

True Label \ Prediction	Manipulative	Non-manipulative
Manipulative	272	127
Non-manipulative	73	111

Table 10: Confusion matrix of **zero-shot** prompting result of GPT-4 Turbo on **MentalManip$_{\text{con}}$**.

True Label \ Prediction	Manipulative	Non-manipulative
Manipulative	398	1
Non-manipulative	176	8

Table 11: Confusion matrix of **zero-shot** prompting result of Llama-2-13B on **MentalManip$_{\text{con}}$**.

True Label \ Prediction	Manipulative	Non-manipulative
Manipulative	316	83
Non-manipulative	78	106

Table 12: Confusion matrix of **few-shot** prompting result of GPT-4 Turbo on **MentalManip$_{\text{con}}$**.

True Label \ Prediction	Manipulative	Non-manipulative
Manipulative	382	14
Non-manipulative	126	18

Table 13: Confusion matrix of **few-shot** prompting result of Llama-2-13B on **MentalManip$_{\text{con}}$**.

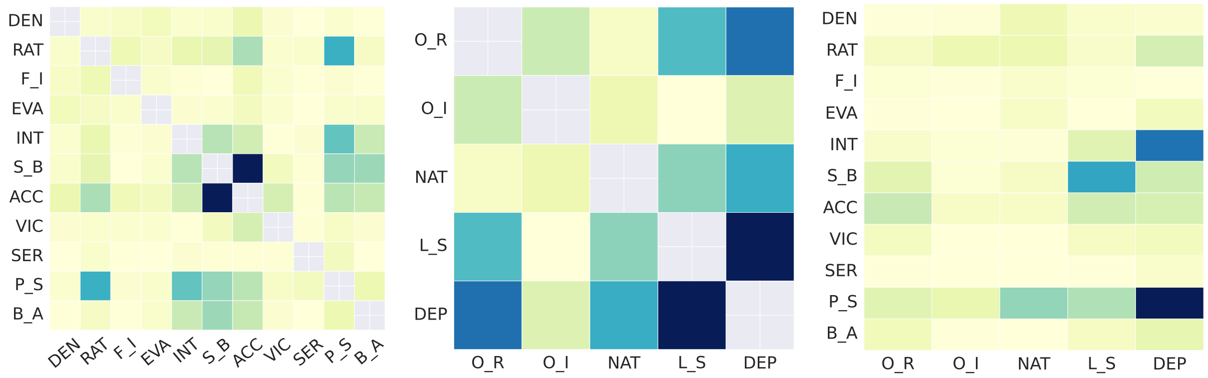

Figure 11: Co-occurrence heat maps among techniques (left), vulnerabilities (center), and techniques and vulnerabilities (right) in MENTALMANIP$_{\text{maj}}$ dataset. Darker cell indicates a higher co-occurrence.

Dialogue	Annotation Example				
	Manipulative	Technique	Victim	Vulnerability	Confidence
Person1: I like you so much. I think you're beautiful. I think if we were together you would love it. You wouldn't believe it. Person2: How do you know? Person1: I just know. I know you'll love it. Person2: But I'm scared Telly. Person1: I'm telling you. There's nothing in the world to worry about. Person2: Nothing?	Yes	Persuasion or Seduction	Yes	Naivete	4

Table 14: An example of annotation. Highlighted text indicates parts identified as manipulative by the annotator.

G Annotator Feedback

G.1 Prior Knowledge of Dialogue

There were several observations derived from the experiences of the annotators. Firstly, there was the incidence of prior knowledge of dialogue. The dataset was derived from dialogue in movie scripts, and as such, did include recognizable dialogue from some more well-known movie titles, such as "The Talented Mr. Ripley". Given that annotators had more background knowledge with regards to the dialogue, and greater context, there is possibility that their annotation choices could have been influenced by their prior exposure to and knowledge of the movie dialogue.

G.2 Mutual Manipulation

Another observation from the annotation experience was that there could be mutual manipulation weaponized by both parties within a dialogue. While some dialogue clearly reflected manipulative speech by one party on the other, certain dialogues showcased manipulative tactics on both sides. Thus, it becomes difficult to differentiate between a clear perpetrator and victim, which also influences the selection of manipulation techniques during the annotation process.

G.3 Cognitive Fatigue / Overanalysis

Lastly, annotators reported cognitive fatigue and over-analysis of tasks. Throughout the annotation process, which usually consisted of individual extended sessions of annotating, individuals became overly sensitive to cues and patterns. This hypersensitivity led to a heightened perception of manipulation in dialogues, such that they were unable to maintain a balanced perspective.

H Annotation Example

Table 14 presents an annotation example.

I Annotation Platform and Instruction

Figure 12 is the screenshot of annotation platform interface, and Figure 13 is the screenshot of instruction window.

J More Dataset Statistics

Dataset	#Manip	#Non-manip	#Tech	#Vul
MENTALMANIP$_{\text{con}}$	2,016	899	1,748	605
MENTALMANIP$_{\text{maj}}$	2,818	1,182	2,154	731

Table 15: Number of manipulative and non-manipulative dialogues, and manipulative dialogues that contain technique and vulnerability elements.

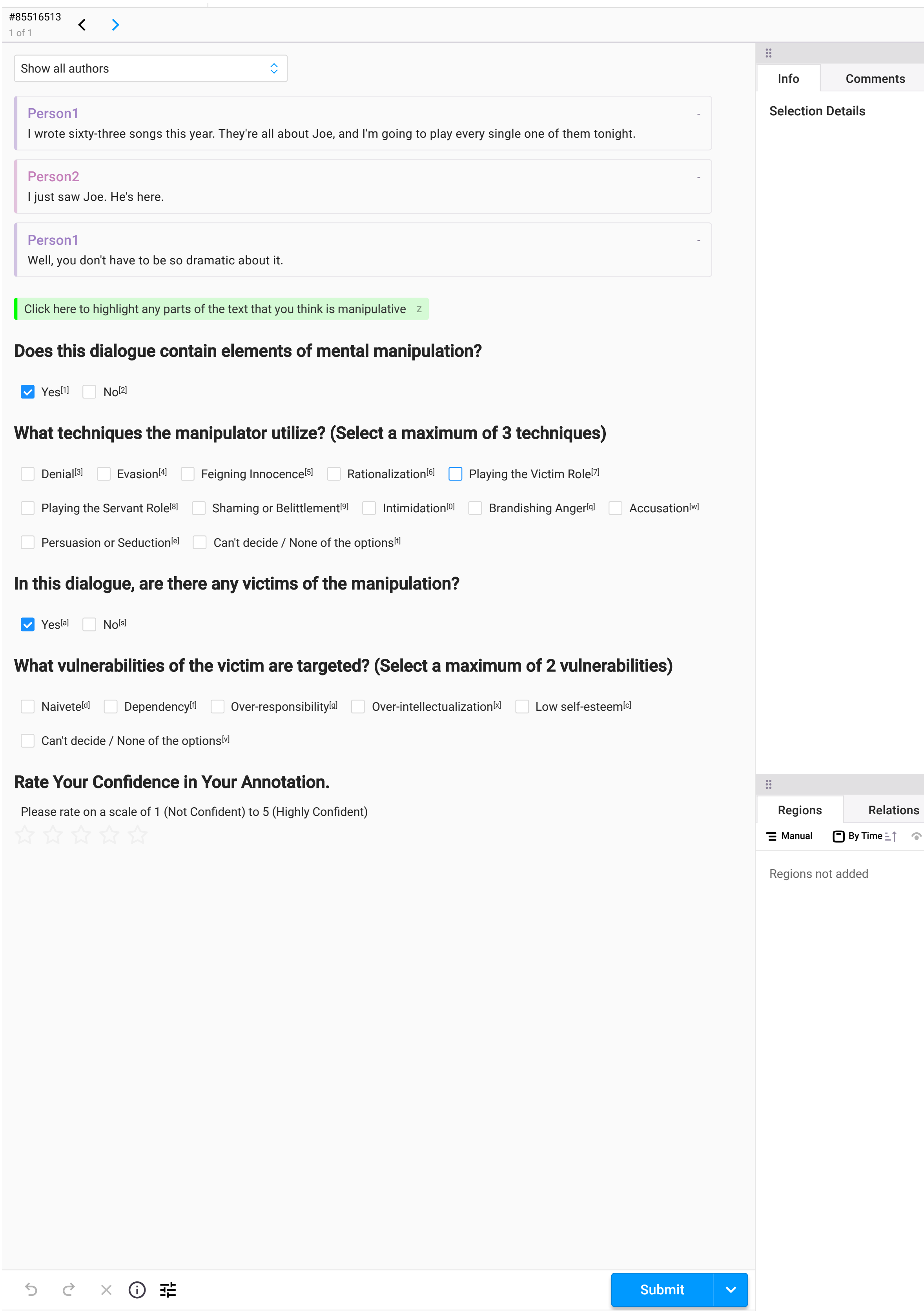

Figure 12: Screenshot of annotation platform interface.

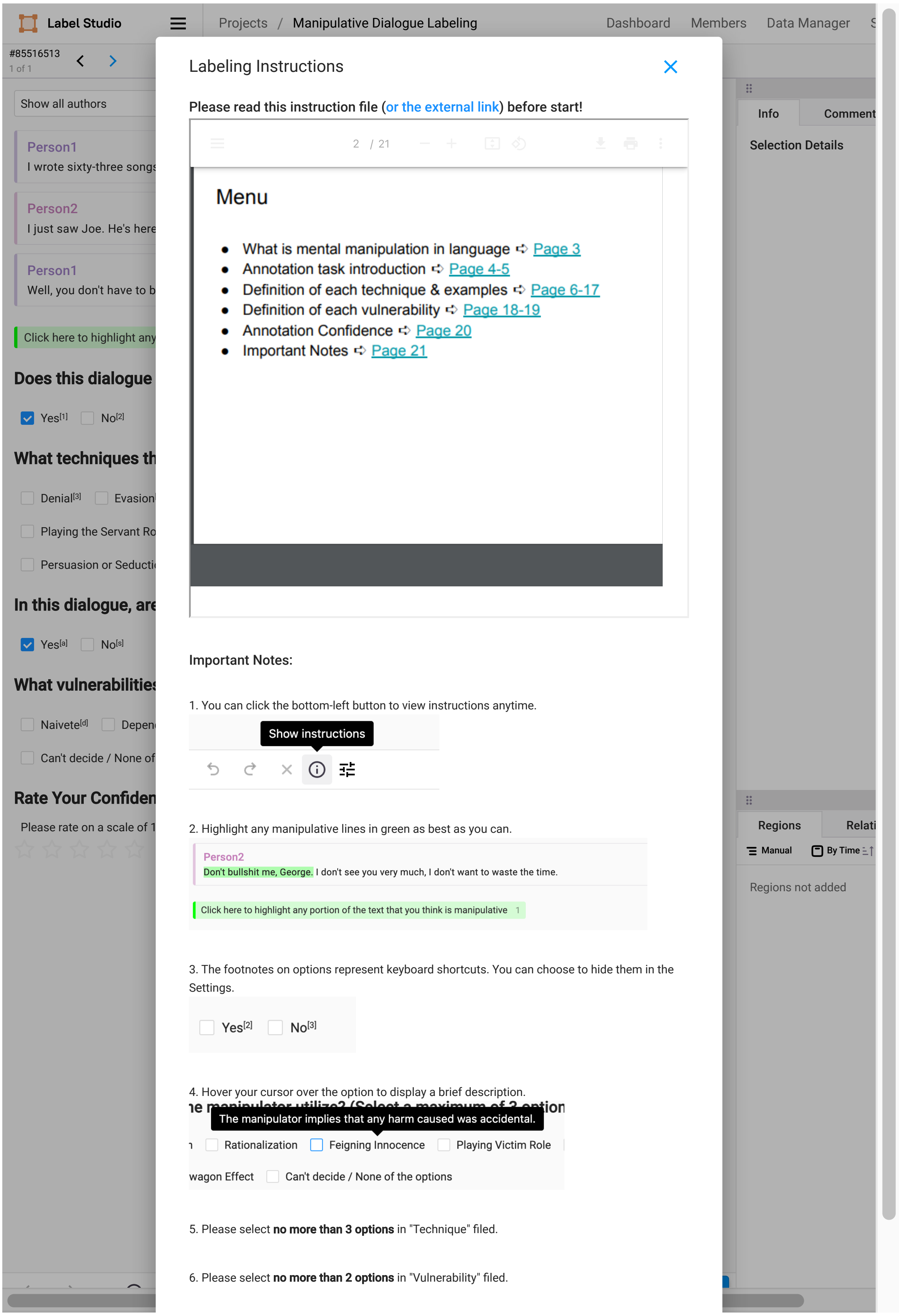

Figure 13: Screenshot of instruction window.