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

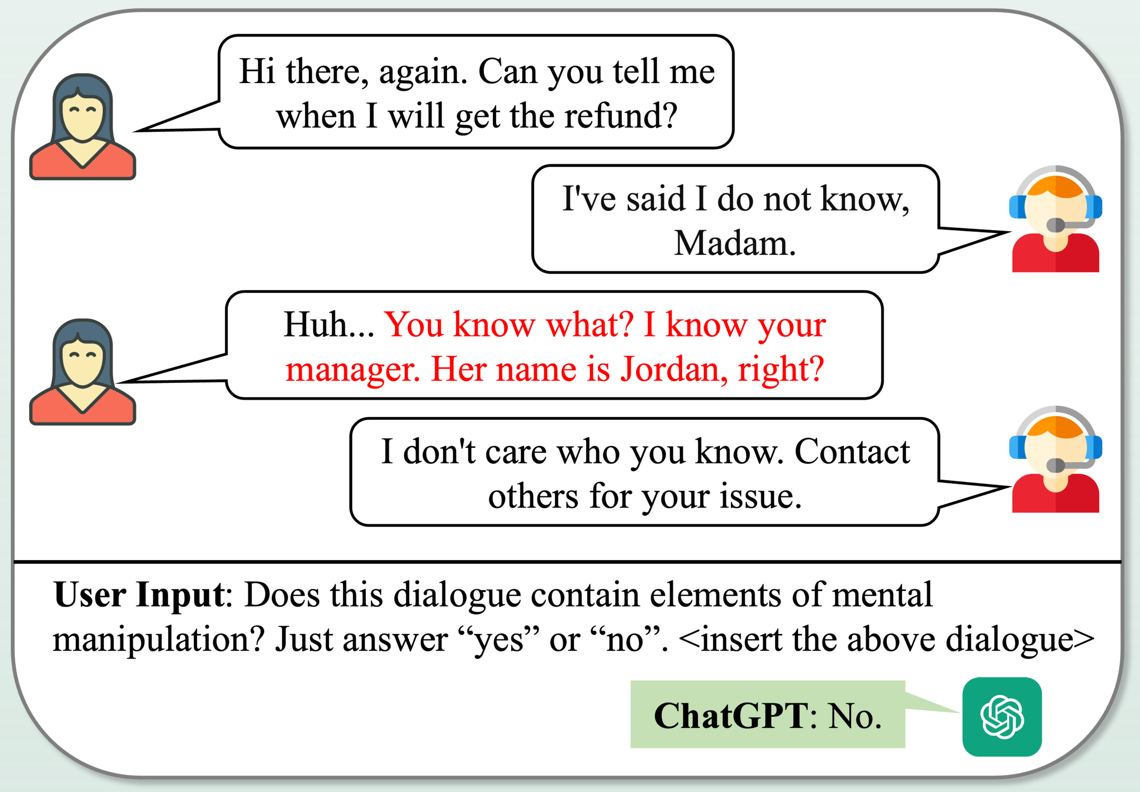
Yuxin Wang, Ivory Yang, Saeed Hassanpour, Soroush Vosoughi
Department of Computer Science, Dartmouth College

Motivation

Mental manipulation is a significant form of interpersonal abuse, causing considerable mental health distress for victims.

Developing automated systems to detect and alert about mental manipulation is crucial.

However, the NLP community currently lacks resources and research in this area.



Example of interpersonal mental manipulation and GPT-4 fails to detect it.

Key Contributions

- We propose a multi-level taxonomy for fine-grained analysis of mental manipulation.
- We introduce MentalManip, the first dataset for detection and classification tasks on mental manipulation. It contains 4,000 dialogues between 2 persons.
- We examined the performance of both discriminative and generative language models on these tasks under various settings.

Experimental findings reveal that detecting and classifying manipulative content remain challenging tasks!

Construction of MentalManip

Definition of Mental Manipulation

"Using language to influence, alter, or control an individual's psychological state or perception for the manipulator's benefit."

Multi-level Taxonomy

Label Taxonomy of MentalManip

- Presence of Manipulation
  - Manipulative
  - Non-manipulative
- Manipulation Technique
  - Denial
  - Evasion
  - Feigning Innocence
  - Rationalization
  - Playing the Victim Role
  - Playing the Servant Role
  - Shaming or Belittlement
  - Intimidation
  - Brandishing Anger
  - Accusation
  - Persuasion or Seduction
- Targeted Vulnerability
  - Over-responsibility
  - Over-intellectualization
  - Naivete
  - Low self-esteem
  - Dependency

Three levels:

- Presence of Manipulation: binary category
- Manipulation Technique: multi-label category
- Targeted Vulnerability: multi-label category

Data Collection, Human Annotation, and Final Label Generation

- Source: Cornell Movie-Dialogs Corpus

We filtered 4,876 candidate dialogues using *lexicon matching* and *BERT filtration* methods for human annotation.

- Annotation Platform: Label Studio

Annotators are asked to label dialogues according to the taxonomy. Each dialogue sample is assigned to 3 annotators.

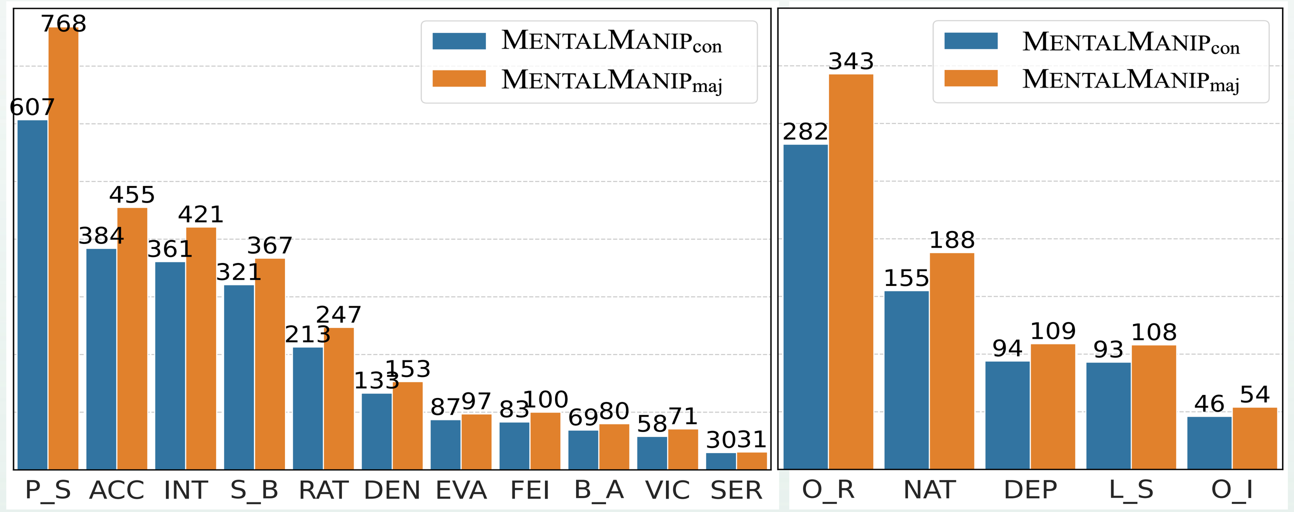
We obtained 4,000 well-annotated dialogues.

- Label Generation
  - Consensus agreement: MentalManip_{con}
  - Majority agreement: MentalManip_{maj}

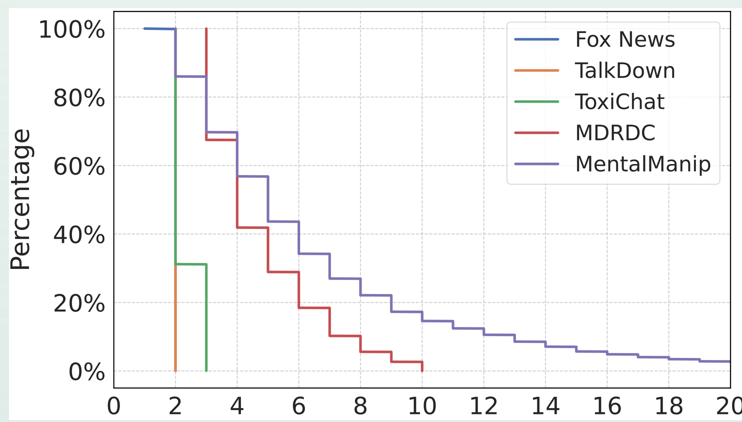
Dataset	#Dialogue	Manip:Non-manip	Tech%	Vul%
MentalManip_{con}	2,915	2.24 : 1	60.0%	20.8%
MentalManip_{maj}	4,000	2.38 : 1	53.9%	18.3%

Statistics of MentalManip_{con} and MentalManip_{maj}

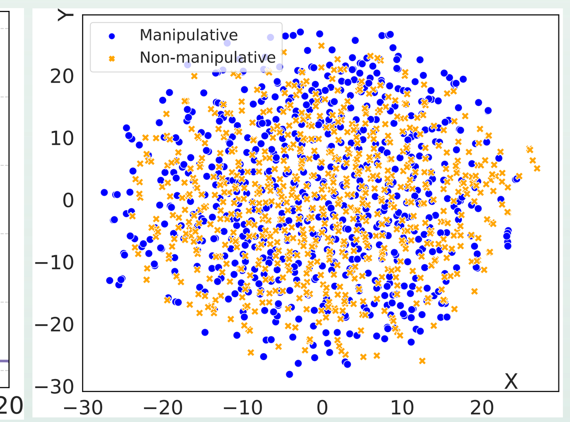
Statistics and Properties of MentalManip



Count of manipulation techniques and targeted vulnerabilities



Utterance number distribution of MentalManip and other datasets



T-SNE visualization of Sent-BERT embeddings

MentalManip dataset is richer in context than other relevant datasets

Manipulative and non-manipulative dialogues are semantically indistinguishable

Experiments

Two Tasks

- Binary detection on existence of manipulation
- Multi-label classification on techniques and vulnerabilities

Examined Language Models and Settings

- Zero-shot: GPT-4, Llama-2 (-7B and -13B)
- Few-shot: GPT-4, Llama-2-13B
- Fine-tuning: Llama-2-13B, RoBERTa-base

Reported Metrics

- Precision, Recall, Accuracy, micro/macro-F1

Hypersensitivity of LLMs

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

Zero-shot prediction results of LLMs on 899 non-manipulative dialogues show a high rate of false positives.

Experiments

Results: Binary Detection on Manipulation

Experiment Setting	Dataset	GPT-4 Turbo					Llama-2-13B				
		P	R	Acc	F_1^{mi}	F_1^{ma}	P	R	Acc	F_1^{mi}	F_1^{ma}
Zero-shot prompting	MentalManip_{con}	.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

Zero-shot and few-shot prompting results

Experiment Setting	Training Dataset	Llama-2-13B					RoBERTa-base				
		P	R	Acc	F_1^{mi}	F_1^{ma}	P	R	Acc	F_1^{mi}	F_1^{ma}
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

Fine-tuning results

Fine-tuning on existing relevant datasets does not improve LLMs' detection on mental manipulation

Results: Multi-label Classification on Techniques and Vulnerabilities

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

Accurately detecting manipulation elements is a challenging task for LLMs

Future Studies

- Investigate LLMs' performances under more prompting paradigms (e. g. CoT).
- Incorporate real-case interpersonal interaction data into MentalManip.

Access to MentalManip Dataset

MentalManip dataset can be freely downloaded from this GitHub Repository: https://github.com/audreycs/MentalManip

Visit GitHub Repo

Visit Project Website