A model to identify Manipulative Language

Diletta Goglia, Davide Vega, Ece Calikus

UPPSALA UNIVERSITET
InfoLab

diletta.goglia@it.uu.se
dilettagoglia.bsky.social

IC2S2 NORRKÖPING 2025

TL;DR
- We develop a model to detect manipulative language (emotionally exploitative speech that bypasses rational consent).
- We annotated dialogues from Supernanny TV show and fine-tune a RoBERTa-base model via pseudolabeling.
- Our model outperforms general-purpose LLMs achieving human-level performance in identifying manipulation.

The use of emotion in persuasive language is problematic when employed to override a person's capacity for rational thought. In the context of **Social Cybersecurity**, this is particularly relevant: in digital environments individuals may experience undue pressure, coercion, or harmful influence.

Manipulation definition: Language used to exert social influence that does not rely on reasoned argument and voluntary acceptance of the receiver but rather on the exploitation of emotional appeals and linguistic strategies to achieve compliance.

Manipulative:	Non-manipulative:
You can't tell anyone by the way. About our way of working.	*Up for some fun? Come on, follow me!*
Why?	*What? There?*
Too many people won't see what we do here as normal.	*Yeah, why not? We can go out later and grab a drink it'll be like Edinburgh all over again.*

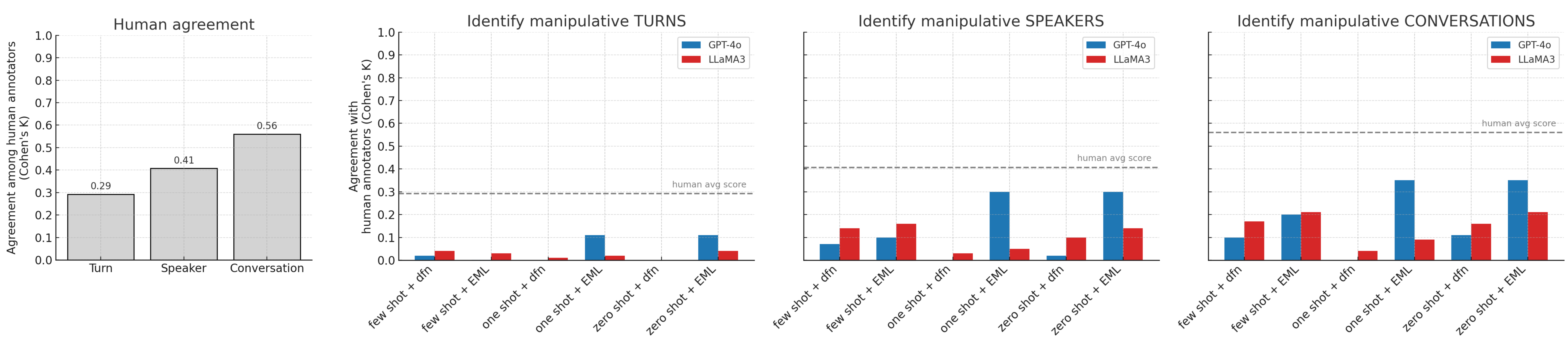
We collect **YouTube** videos from the `@officialsupernanny` channel. We **diarize** episodes by combining the PyTorch Voice Activity Detector model and the Whisper model for speech recognition. We built an *ad-hoc* **codebook** and provide it to annotators to manually label a portion of the dataset according to the following binary label: *does the turn contain manipulative language?*

Emotional Manipulative Language (EML) strategies:

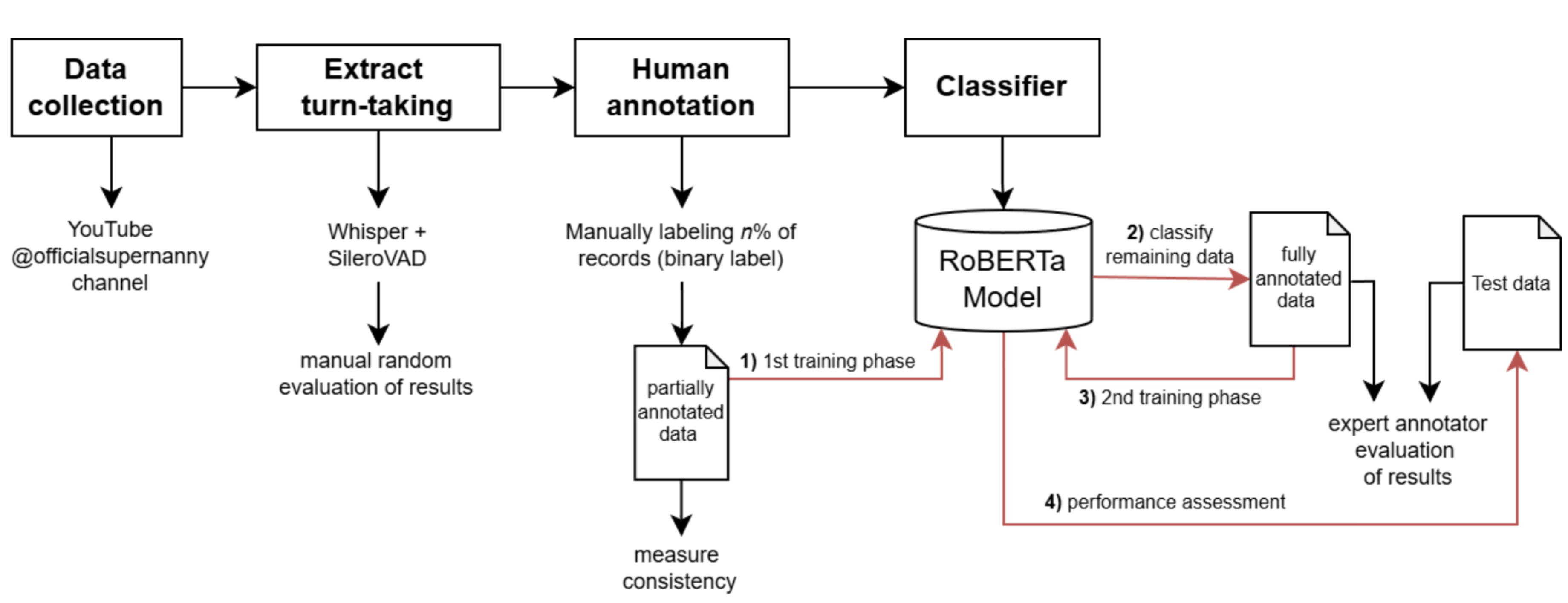
Minimization	Invalidating a person's feelings, opinions, or emotional experience (i.e., considering it weak, unaccepted, disrespected, or ineffective).
Power	Asserting dominance to control or intimidate others, exploiting elements like veiled threats, hierarchy, or authority.
Guilt	Blaming a person to make them feel responsible or bad about some wrongdoing, for example with accusations to state that they are at fault.
Shame	Language to make others feel inferior, unworthy, or embarrassed (e.g., including judgments, sarcasm, criticism, or put-downs).

Conversation ID	Turn	Speaker	Utterance ID		Manipulation
001	1	PERSON 1	1	you want to control everything	☐
001	2	PERSON 2	2	this time can you do it the right way	☐
001	3	PERSON 1	3	I don't remember these family conversations	☐
001			4	it's not hard to forget	
001	4	PERSON 2	5	you don't remember all the times that you cheated on me	☑

We compare the ability of two LLMs (OpenAI's **GPT-4o** and Meta's **LLaMA-3.3-70B**) to perform the same annotation under varying conditions. **LLMs do not achieve human-comparable performance in detecting manipulative language.** Even under **class-balanced** conditions, they exhibit a systematic bias toward predicting non-manipulative language and overlooking a substantial number of manipulative utterances.



We fine-tune a RoBERTa-base using a semi-supervised ML setting based on **pseudolabeling**. **Our model outperforms LLMs, reaching human-comparable performances** (Cohen's K: 0.984) on the identification of manipulative language. Results show relevant drop of false positives and false negatives, demonstrating the benefit of task-specific adaptation.



Identify TURNS		Label 0 (no manipulation)			Label 1 (manipulation)			Total	
Model		precision	recall	f1	precision	recall	f1	accuracy	macro_avg_f1
GPT-4o	fs-dfn	0.46	0.99	0.63	0.77	0.03	0.05	0.47	0.34
	fs-eml	0.46	0.99	0.62	0.55	0.01	0.02	0.46	0.32
	os-dfn	0.46	1	0.63	0	0	0	0.46	0.31
	os-eml	0.5	0.88	0.64	0.73	0.26	0.38	0.54	0.51
	zs-dfn	0.45	0.99	0.62	0	0	0	0.45	0.31
	zs-eml	0.5	0.88	0.64	0.73	0.26	0.38	0.54	0.51
Llama 3	fs-dfn	0.52	0.53	0.52	0.6	0.59	0.59	0.56	0.56
	fs-eml	0.52	0.45	0.49	0.59	0.65	0.62	0.56	0.55
	os-dfn	0.46	0.98	0.63	0.68	0.03	0.06	0.46	0.34
	os-eml	0.46	0.99	0.63	0.77	0.03	0.06	0.47	0.34
	zs-dfn	0.45	0.3	0.36	0.54	0.7	0.61	0.52	0.48
	zs-eml	0.51	0.5	0.5	0.59	0.59	0.59	0.55	0.55
Our model	RoBERTa-base	**1**	**0.98**	**0.99**	**0.99**	**1**	**0.99**	**0.99**	**0.99**

References
- Aboodi, R. (2021). What's wrong with manipulation in education? Philosophy of Education, 77(2), 66–80.
- Feidaros, L. (2016). Manipulation and persuasion. ANADISS, 11(21).
- Franke, M., & Van Rooij, R. (2015). Strategies of persuasion, manipulation and propaganda. In Models of Strategic Reasoning (Vol. 8972). Springer Berlin Heidelberg.
- Nettel, A. L., & Roque, G. (2012). Persuasive argumentation versus manipulation. Argumentation, 26(1), 55–69.
- Tillson, J. (2021). Wrongful Influence in Educational Contexts. Oxford Research Encyclopedia of Education.

What's next?
- Multi-label classification (identify the four EML strategies).
- Additional attention mechanism to infer context at conversational level
- More annotators

This work has been partly funded by eSSENCE, an e-Science collaboration funded as a strategic research area of Sweden. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.