How would Stance Detection Techniques Evolve after the Launch of ChatGPT?

Bowen Zhang[1], Daijun Ding[1], Liwen Jing[2] *

[1]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China

[2]Faculty of Information and Intelligence, Shenzhen X-Institute, Shenzhen, China

介绍：

- 传统的基于深度学习的立场检测通常被定义为分类任务；
- 目前立场检测还存在着解释性差和标注数据少的问题；
- 在这项工作中，作者通过直接询问ChatGPT，来进行立场检测。这种方法可以看作是一种zero-shot的提示策略；
- ChatGPT可以在SemEval-2016和P-Stance等常用数据集上实现SOTA或类似的性能；

- 方法和结果：

Model	HC	FM	LA
Bicond† (Augenstein et al., 2016)	32.7	40.6	34.4
CrossNet† (Xu et al., 2018)	38.3	41.7	38.5
SEKT (Zhang et al., 2020b)	50.1	44.2	44.6
TPDG† (Liang et al., 2021)	50.9	53.6	46.5
Bert_Spc† (Devlin et al., 2019)	49.6	41.9	44.8
Bert-GCN† (Lin et al., 2021)	50.0	44.3	44.2
PT-HCL† (Liang et al., 2022)	54.5	54.6	50.9
ChatGPT	**78.0**	**69.0**	**59.3**

Table 1: Performance comparison (F1-avg) on SemEval-2016 dataset with zero shot setup.

80%的训练数据

Methods	FM		LA		HC	
	F1-m	F1-avg	F1-m	F1-avg	F1-m	F1-avg
BiLSTM (Augenstein et al., 2016)	48.0	52.2	51.6	54.0	47.5	57.4
BiCond (Augenstein et al., 2016)	57.4	61.4	52.3	54.5	51.9	59.8
TextCNN (Kim, 2014)	55.7	61.4	58.8	63.2	52.4	58.5
MemNet (Tang et al., 2016)	51.1	57.8	58.9	61.0	52.3	60.3
AOA (Huang et al., 2018)	55.4	60.0	58.3	62.4	51.6	58.2
TAN (Du et al., 2017)	55.8	58.3	63.7	65.7	65.4	67.7
ASGCN (Zhang et al., 2019)	56.2	58.5	59.5	62.9	62.2	64.3
Bert_Spc (Devlin et al., 2019)	57.3	60.6	64.0	66.3	65.8	69.1
TPDG (Liang et al., 2021)	67.3	/	74.7	/	73.4	/
ChatGPT	**68.4**	**69.0**	58.2	59.3	**79.5**	**78.0**

Table 2: Performance comparison on SemEval-2016 dataset with in-domain setup.

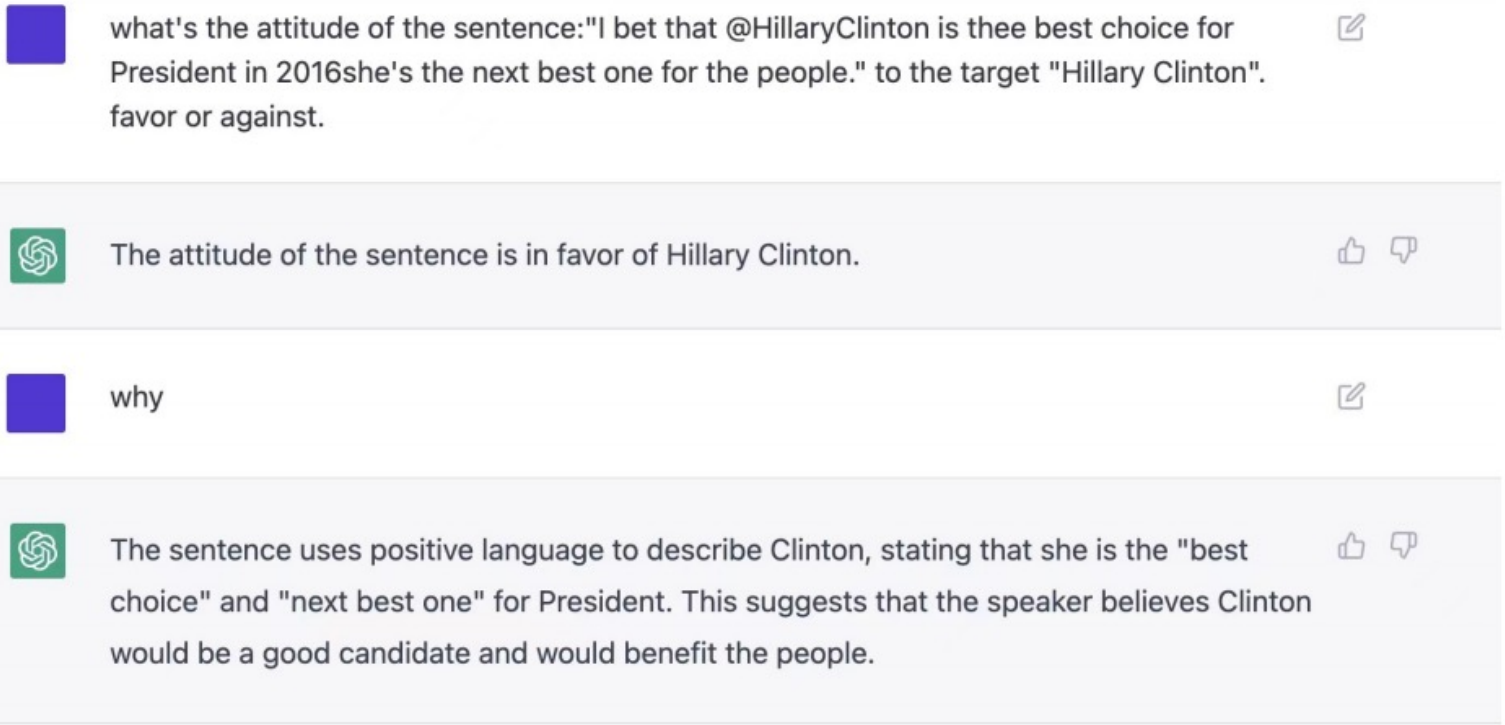
80%的训练数据

Methods	Trump		Biden		Bernie	
	F1-m	F1-avg	F1-m	F1-avg	F1-m	F1-avg
BiLSTM (Augenstein et al., 2016)	69.7	72.0	68.7	69.5	63.8	63.9
BiCond (Augenstein et al., 2016)	70.6	73.0	68.4	69.4	64.1	64.6
TextCNN (Kim, 2014)	76.9	77.2	78.0	78.2	69.8	70.2
MemNet (Tang et al., 2016)	76.8	77.7	77.2	77.6	71.4	72.8
AOA (Huang et al., 2018)	77.2	77.7	77.7	77.8	71.2	71.7
TAN (Du et al., 2017)	77.1	77.5	77.6	77.9	71.6	72.0
ASGCN (Zhang et al., 2019)	76.8	77.0	78.2	78.4	70.6	70.8
Bert_Spc (Devlin et al., 2019)	81.4	81.6	81.5	81.7	78.3	78.4
ChatGPT	**82.8**	**83.2**	**82.3**	**82.0**	**79.4**	**79.4**

Table 3: Performance comparison on P-Stance dataset with in-domain setup.

相关讨论

- （1）如何提升ChatGPT的性能： 改进提问模板 （作者没做实验）
- （2）如何获得可解释性： 直接提问chatgpt，如下图所示。



从gpt的回答来看，效果不错

d

what's the attitude of the sentence '@ReformedBroker "to da honies gittin money playing #bleep words like dummies" #POPPA '. to the target 'Hillary Clinton'? select a answer from (favor, against, none), don't explain anything, can make some guesses.

It is difficult to accurately determine the attitude of the sentence towards Hillary Clinton as the sentence does not mention or directly reference her. It is possible that the attitude of the sentence may be neutral or unrelated to Hillary Clinton.

d

why

The sentence does not mention or reference Hillary Clinton in any way, so it is difficult to determine what attitude the sentence may have towards her. Without more context or information about the intended meaning or purpose of the sentence, it is not possible to accurately determine the attitude of the sentence towards Hillary Clinton.

chatgpt对无法确定的立场给予了解释

Is ChatGPT Equipped with Emotional Dialogue Capabilities?

Weixiang Zhao, Yanyan Zhao*, Xin Lu, Shilong Wang, Yanpeng Tong, Bing Qin
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{wxzhao, yyzhao}@ir.hit.edu.cn

- 介绍

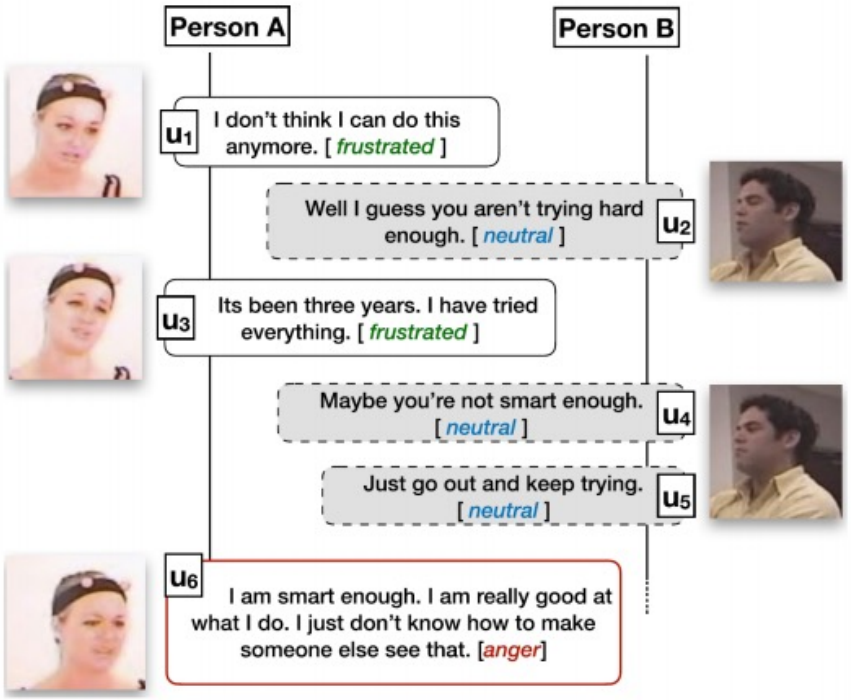
- 评估了ChatGPT在情绪对话理解和生成方面的表现；

- 测试它的两种能力：

-理解能力：ChatGPT是否能够准确地理解和解释用户的情绪？

-生成能力

理解能力



为了测试GPT的理解能力，作者进行了情感对话分类识别任务

对话情感识别要求考虑上下文

Model	IEMOCAP	MELD	EmoryNLP	DailyDialog
DialogueRNN	64.76	63.61	37.44	57.32
IEIN	64.37	60.72	-	-
COSMIC	65.28	65.21	38.11	58.48
DialogXL	65.94	62.41	34.73	54.93
DAG-ERC	68.03	63.56	39.02	59.33
DialogueCRN	66.20	58.39	-	-
CauAIN	67.61	65.46	-	58.21
CoMPM	69.46	66.52	38.93	**60.34**
MuCDN	-	65.37	40.09	-
SPCL	**69.74**	**67.25**	**40.94**	-
ChatGPT 0-shot	44.97	57.30	37.47	40.66
ChatGPT 1-shot	47.46	58.63	35.60	42.00
ChatGPT 3-shot	48.58	58.35	35.92	42.39

Table 1: Comparison of ChatGPT and other baselines on ERC task.

ChatGPT和先进基准差3~18个点

few-shot比0-shot强一点

研究发现，chatGPT的逻辑标准和实际数据集注释标准是不一致的
不能说两者谁对，只是采取了不同的标准

Speaker	Dialogue Content	Annotation	Prediction
A	Good morning. What's the matter with you?	Neutral	Neutral
B	Good morning, doctor. I have a terrible headache.	Neutral	Sadness
A	All right, young man. Tell me how it got started.	Neutral	Neutral
B	Yesterday I had a runny nose. Now my nose is stuffed up. I have a sore throat. And I'm afraid I've got a temperature. I feel terrible.	Neutral	Sadness
A	Don't worry, young man. Lat me give you an examination. First let me take a look at your throat. Open your mouth and say 'ah'.	Neutral	Neutral
B	Ah.	Neutral	Neutral
A	Your throat is inflamed. And your tongue is heavily coated. You have all the symptoms of influenza.	Neutral	Fear
B	What am I supposed to do then?	Neutral	Fear
A	A good rest is all you need, and drink more water. I'll write you a prescription.	Neutral	Happiness
B	Thank you very much.	Neutral	Happiness

作者还做了CEE任务，即哪些语句造成了目标的Happy情感；

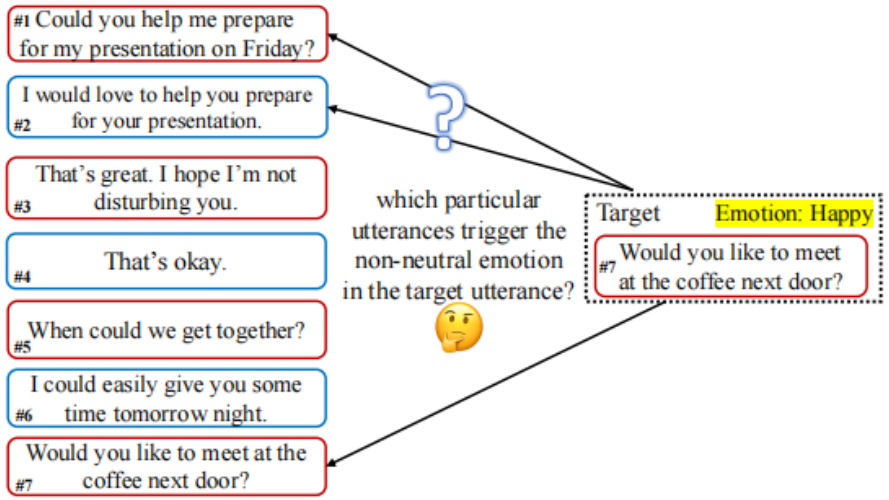


Figure 3: An example from RECCON-DD dataset (Poria et al., 2021) for identifying the causal utterances.

Model	Neg. F1	Pos. F1	macro F1
RoBERTa-Base	88.74	64.28	76.51
RoBERTa-Large	87.89	66.23	77.06
KEC	88.85	66.55	77.70
KBCIN	89.65	68.59	79.12
TSAM	**90.48**	**70.00**	**80.24**
ChatGPT 0-shot	85.25	51.33	68.29
ChatGPT 1-shot	82.10	52.84	67.47

Table 3: Comparison of ChatGPT and other baselines on CEE task.

存在差距

解锁ChatGPT性能的全部潜力需要对数据集的规范有深入的理解

生成能力

机器人在生成回复时，不仅要建模上下文，还要具有同理心并考虑用户情感，即“会说话”

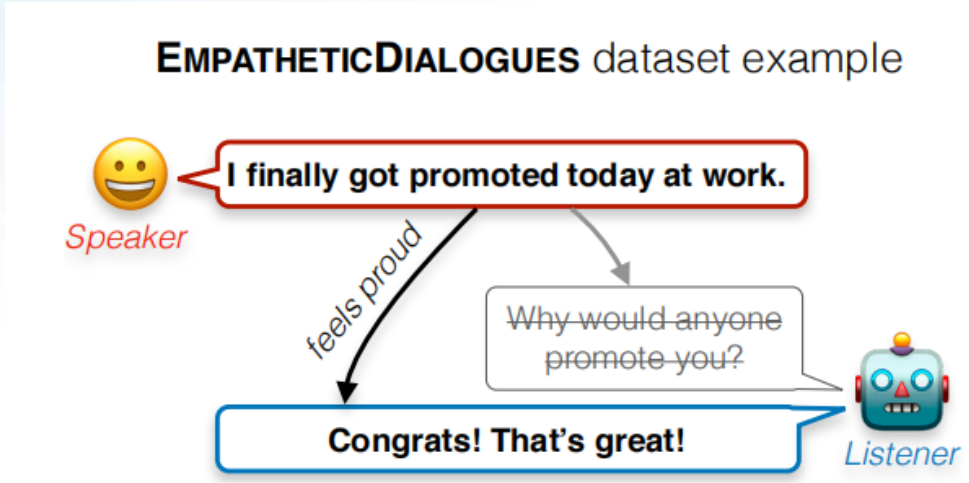


Figure 4: An example of empathetic conversation from EMPATHETICDIALOGUE dataset (Rashkin et al., 2018).

Model	D-1	D-2	B-1	B-2	B-3	B-4	R-L
Multi-TRS	0.44	1.98	21.42	7.60	3.67	2.13	21.66
MoEL	0.58	2.91	21.70	7.75	3.58	1.96	**22.08**
MIME	0.42	1.71	**22.20**	**8.07**	3.89	2.22	21.54
EmpDG	0.44	1.91	21.99	8.02	3.86	2.19	22.03
CEM	0.65	3.03	18.69	6.84	3.37	1.92	21.65
EmpSOA	0.69	3.87	21.41	8.06	**4.14**	**2.41**	21.64
SEEK	0.66	2.74	15.21	4.40	1.81	0.85	19.44
ChatGPT 2-shot	**4.17**	**23.23**	12.99	4.07	1.90	1.05	12.70

生成回复的多样性高　　词汇语义质量低

Emotion	Devastated
Context	My dog I had for five years just passed away. He was hit by a car last night.
MIME	Oh no! I am so sorry to hear that.
EmpDG	Oh no! I am so sorry to hear that.
CEM	Oh no! I am so sorry to hear that.
EmpSOA	Oh no! I am so sorry to hear that.
ChatGPT	I'm so sorry for your loss. Losing a beloved pet can be incredibly hard. Would you like to talk about it more?
Ground-Truth	Oh I am so sorry. That must be fresh on your heart, may he rest in peace.
Emotion	Content
Context	My wife made me pancakes for breakfast. I have a full belly and feel rather happy now. Yum! I love pancakes. Is it your favorite food too? One of my favorite breakfast foods. She also made some thick sliced bacon. She treats me so well.
MIME	That is a great idea. I love my girlfriend too.
EmpDG	That is a great attitude to have!
CEM	That is good, I love it!
EmpSOA	That sounds so sweet. I am glad you have a great time!
ChatGPT	It's so wonderful to have someone who cares for and treats us well. Enjoy your delicious breakfast!
Ground-Truth	That is great you have a sweet wife!

Table 8: Case study of the generated empathetic responses by ChatGPT and the baselines.

chatgpt回复更充分

但也存在词语累叠问题，让用户感到无聊

- 作者还做了一些其他测试...都大同小异...

- 局限性：
- 更先进的gpt4+还没测试；
- 未能测到chatgpt的性能上限；
- 度量方法有待改进；