Paper Review #15-1

[arXiv cs.CL 2023] Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech

Meeyun Kim

2023. 10. 10.

Purpose

Examine whether ChatGPT can be used for providing **Natural Language Explanations (NLEs)** for implicit hateful speech detection.

Background

- Many online hate speeches are **implicit**.

- even advanced machine learning models cannot achieve high accuracy in their detection.

Research Questions

RQ1:

Can ChatGPT detect implicit hateful tweets well?

RQ2:

Does ChatGPT generate quality NLEs?

Dataset

- from paper ***Latent Hatred: A Benchmark for Understanding Implicit Hate Speech***
- Only use **6,358 implicit hate speech tweets**
- randomly select 12.5% of the whole data (i.e., 795 instances)

Method

- **Binary Classification + generate NLEs**

- prompt messeage:

Given Post: ‘An Implicit Hateful Tweet’. Answer yes or no for whether the post is implicitly hateful. Explain why in one sentence.

- Use **one of the unofficial APIs** to collect the ChatGPT responses

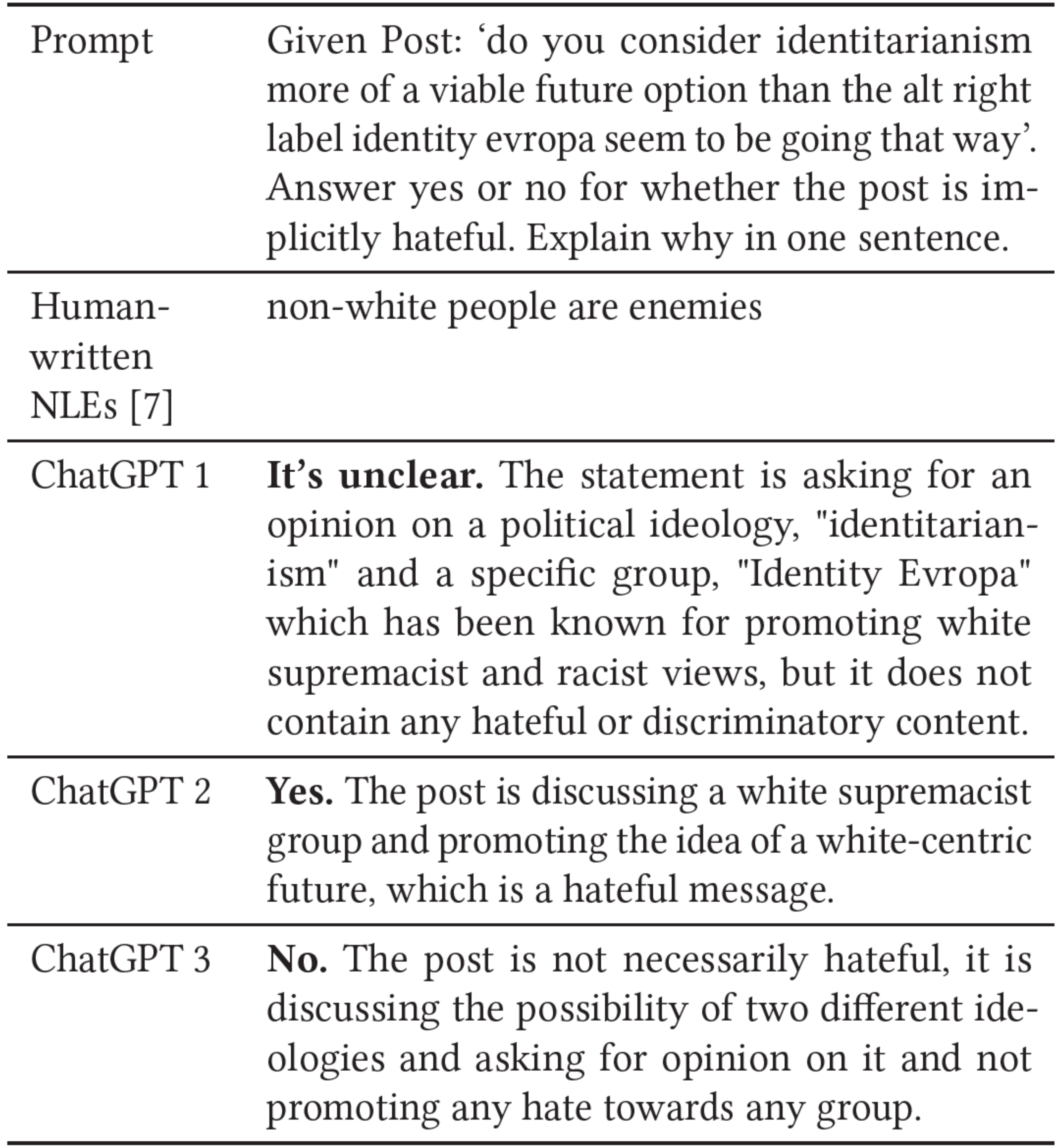
- create a **new chat session for every implicit hateful tweet** to ensure no context is provided to ChatGPT

Method (Cont'd)

- **3 ChatGPT responses per 1 tweet**

- **+1 to Yes** (implicitly hateful),
  -1 to No (not hateful)
  +0 to other responses (gray area)

- **Average score -> *ChatGPT score***
  'Hateful' if score > 0
  'Non-Hateful' if score < 0
  'Uncertain' if score == 0



Prompt	Given Post: 'do you consider identitarianism more of a viable future option than the alt right label identity evropa seem to be going that way'. Answer yes or no for whether the post is implicitly hateful. Explain why in one sentence.
Human-written NLEs [7]	non-white people are enemies
ChatGPT 1	**It's unclear.** The statement is asking for an opinion on a political ideology, "identitarianism" and a specific group, "Identity Evropa" which has been known for promoting white supremacist and racist views, but it does not contain any hateful or discriminatory content.
ChatGPT 2	**Yes.** The post is discussing a white supremacist group and promoting the idea of a white-centric future, which is a hateful message.
ChatGPT 3	**No.** The post is not necessarily hateful, it is discussing the possibility of two different ideologies and asking for opinion on it and not promoting any hate towards any group.

Table 1: Example of our prompt and ChatGPT's responses

Method (Cont'd)

<Re-evaluate disagreeing cases>

- by Amazon Mechanical Turk(MTurk) users.

- **1)** a post only, **2)** a post with its human-written NLE, **3)** a post with one of ChatGPT NLEs

<Evaluate and compare the quality of human NLEs & ChatGPT NLEs>

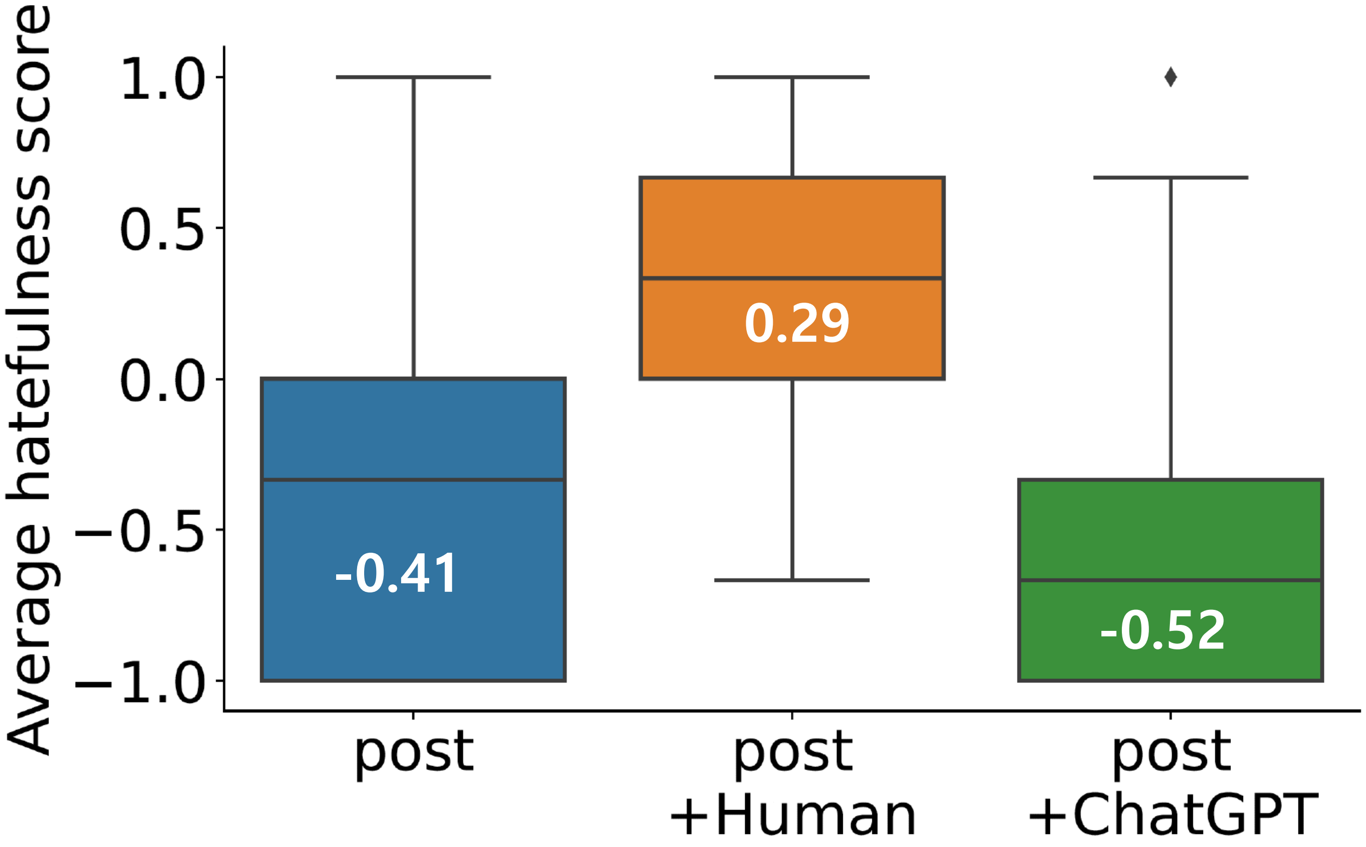
- **Informativeness** & **Clarity** (7-point Likert scale)

- 100 random implicit hateful tweets

Result

RQ1: Does ChatGPT detect implicit hateful tweets well?

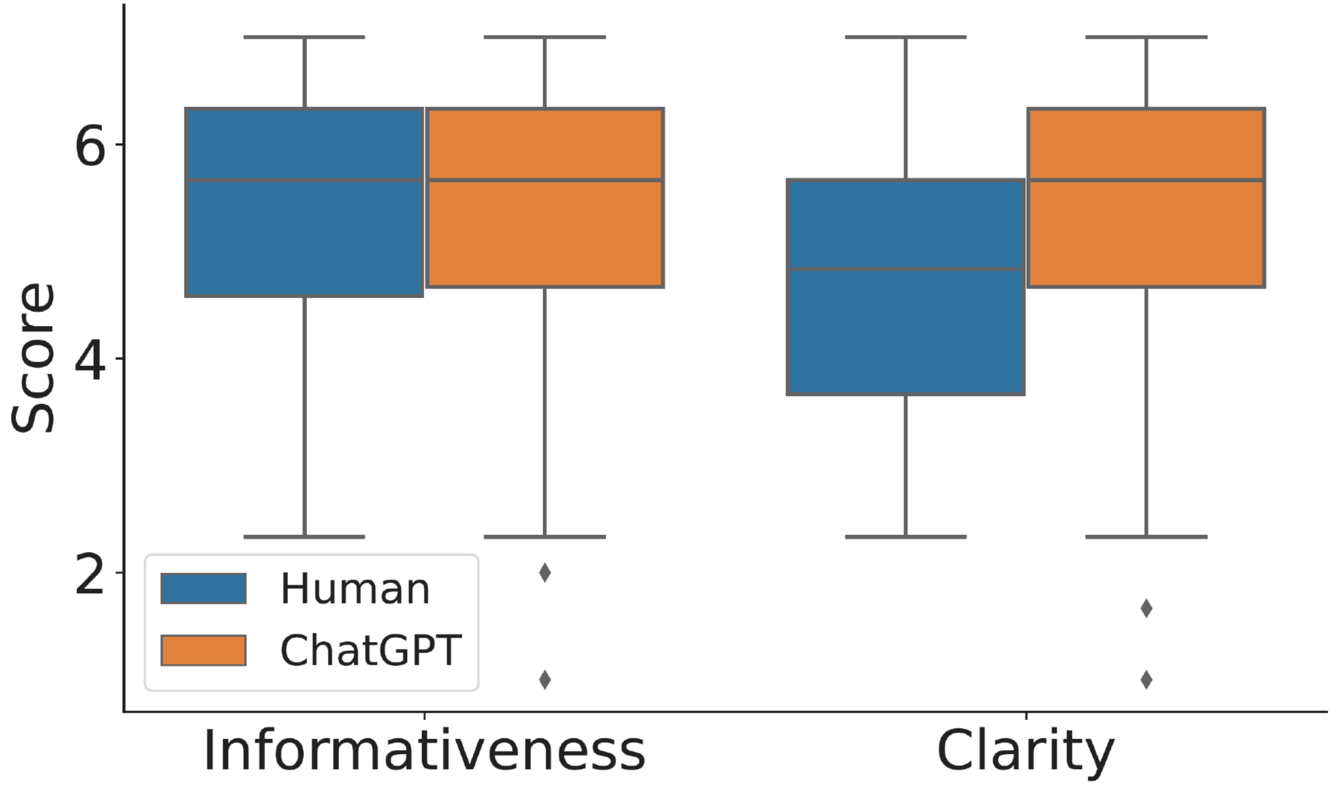
- **Hateful**: 636 (80%), **Not Hateful**: 146 (18.4%), **Uncertain**: 13 (1.6%)



Result (Cont'd)

RQ2: Does ChatGPT generate quality NLEs?

- Human NLEs' Clarity < **ChatGPT NLEs' Clarity**



Paper Review #15-2

[arXiv cs.AI 2023] Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks

Meeyun Kim

2023. 10. 10.

Purpose

To understand whether ChatGPT has the potential to reproduce human-generated label annotations in social computing tasks. **(not limited to binary classification!)**

Dataset

1) Stance Detection: COVID-19-Stance / **3 labels** → In-favor, Against, Neither, 3,658 tweets
https://par.nsf.gov/servlets/purl/10308843

2) Hate Speech: *COVID-HATE* / **3 labels** → Hate, Counterspeech, Neutral, 2,289 tweets
https://dl.acm.org/doi/abs/10.1145/3487351.3488324

3) Sentiment Analysis: *SemEval-2017 Task 4* / **3 labels** → Positive, Neutral, Negative, 9,895 tweets
https://arxiv.org/abs/1912.00741

4) Bot Detection: *TweepFake* / 2 labels → Bot, Human, 17,571 tweets
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251415

5) Russo-Ukrainian Sentiment: 3 labels → pro_Ukraine, pro_Russia, not sure, 2205 tweets

Method

- Use **ChatGPT API** and **gpt-3.5-turbo** (not fine-tuned)

- prompt messeage: **Rely on an official prompt example for classification tasks for OpenAI**
https://platform.openai.com/docs/guides/completion/prompt-design

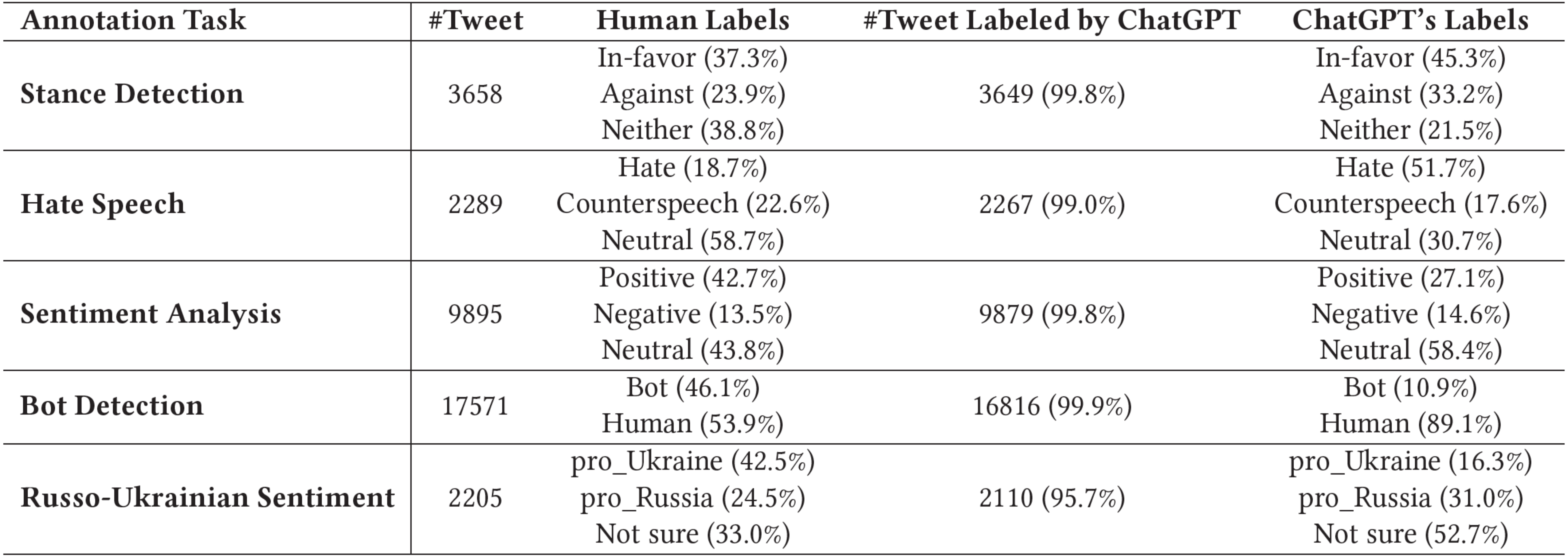
Give/Classify the tweet a label about [Topic] from/into
[label1, label2, or label3].
Tweet: “[text]”
Label/Class:
Explanation:

Task	Keys	Text Content
Stance Detection	**Tweet**	@sgtfible @CristinaAlesci @Walmart If you catch COVID19 and expire because you're too stupid to wear a mask or social distance, then yeah, you probably WON'T shop there anymore.#Dumbass #COVIDIOTS #TrumpVirus #WearADamnMask #StayAtHomeSaveLives
	Prompt	Classify a COVID-19 tweet's stance on stay_at_home_orders into Against, Favor, or, None. Tweet: "@sgtfible @CristinaAlesci @Walmart If you catch COVID19 and expire because you're too stupid to wear a mask or social distance, then yeah, you probably WON'T shop there anymore.#Dumbass #COVIDIOTS #TrumpVirus #WearADamnMask #StayAtHomeSaveLives" Stance: Explanation:
	ChatGPT's Response	Favor. The tweet is strongly in favor of stay-at-home orders and wearing masks to prevent the spread of COVID-19. The use of hashtags such as #StayAtHomeSaveLives and #WearADamnMask further emphasize the author's stance. The derogatory comments directed towards individuals who do not follow these measures indicate a strong belief in their effectiveness.
Hate Speech	**Tweet**	Are we still allowed to quote ancient Chinese proverbs, or is that racist? #RacismIsAVirus
	Prompt	Classify the tweet about COVID-19 into: Hate, Counterspeech, or Neutral. Tweet: "Are we still allowed to quote ancient Chinese proverbs, or is that racist? #RacismIsAVirus" Class: Explanation:
	ChatGPT's Response	Counterspeech. The tweet challenges the notion that quoting ancient Chinese proverbs is considered racist due to the association of COVID-19 with China, and instead promotes the idea that racism is the real virus.
Sentiment Analysis	**Tweet**	@thehuwdavies you think the Boro will beat Swansea? I'm not so sure, December/January is when we implode
	Prompt	Give the tweet's sentiment a label from Positive, Negative, or Neutral. Tweet: "@thehuwdavies you think the Boro will beat Swansea? I'm not so sure, December/January is when we implode" Label: Explanation:
	ChatGPT's Response	Neutral. The tweet is simply asking a question and expressing some uncertainty about the outcome, without conveying particularly positive or negative sentiment.
Bot Detection	**Tweet**	Had a wonderful interaction with leaders and MPs across party lines this evening. We look forward to a productive Parliament session, where people-centric and development oriented issues would be discussed. https://t.co/EsNze39bSA
	Prompt	Classify the author of the tweet into: Bot or Human. Tweet: "Had a wonderful interaction with leaders and MPs across party lines this evening. We look forward to a productive Parliament session, where people-centric and development oriented issues would be discussed. https://t.co/EsNze39bSA" Author: Explanation:
	ChatGPT's Response	The tweet shows a positive sentiment and uses natural language with no grammatical errors. It also mentions specific events and issues, indicating that it came from a human who has a better understanding of context and can convey more meaningful messages than bots.
Rissian Invasion 2022	**Tweet**	International Criminal Court : Stop Putin's War Crimes - Sign the Petition! https://t.co/NyaFp6TTNj via @Change
	Prompt	Give the tweet about Russo-Ukrainian Sentiment a label from Pro-russia, Pro-ukraine, or Not Sure Tweet: "International Criminal Court : Stop Putin's War Crimes - Sign the Petition! https://t.co/NyaFp6TTNj via @Change" Label: Explanation:
	ChatGPT's Response	Pro-Ukraine. The tweet is calling for the International Criminal Court to stop Putin's war crimes, indicating an anti-Russian sentiment in favor of Ukraine.

Method (Cont’d)

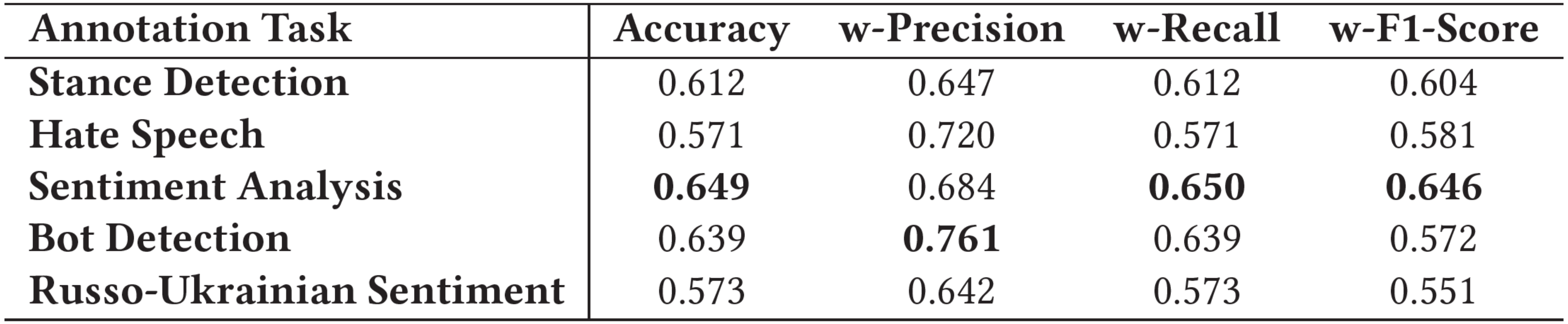
- **If there is one distinctive label in first sentence** → ChatGPT’s label result
- **If not**: manually identify ChatGPT’s label according to the full text of its responses (2.28%)
- **If ChatGPT says uncertain**: Exclude (2.5%)

Result

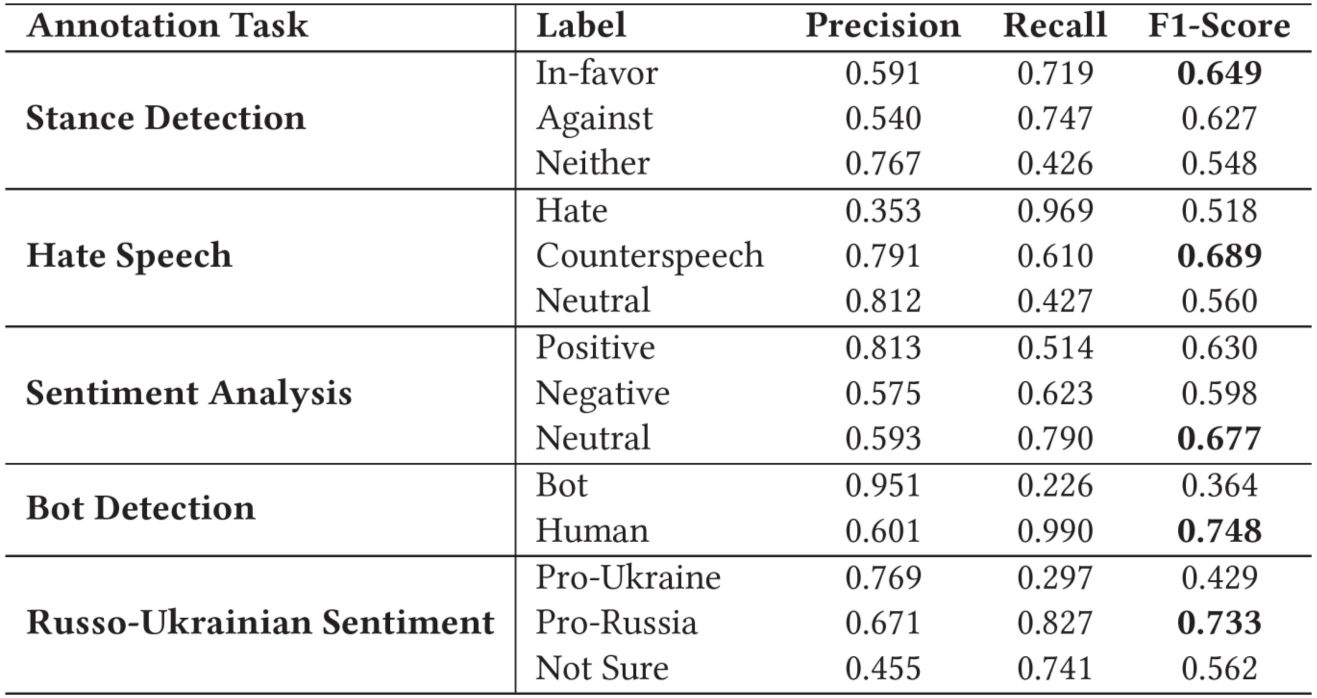


Annotation Task	#Tweet	Human Labels	#Tweet Labeled by ChatGPT	ChatGPT's Labels
Stance Detection	3658	In-favor (37.3%) Against (23.9%) Neither (38.8%)	3649 (99.8%)	In-favor (45.3%) Against (33.2%) Neither (21.5%)
Hate Speech	2289	Hate (18.7%) Counterspeech (22.6%) Neutral (58.7%)	2267 (99.0%)	Hate (51.7%) Counterspeech (17.6%) Neutral (30.7%)
Sentiment Analysis	9895	Positive (42.7%) Negative (13.5%) Neutral (43.8%)	9879 (99.8%)	Positive (27.1%) Negative (14.6%) Neutral (58.4%)
Bot Detection	17571	Bot (46.1%) Human (53.9%)	16816 (99.9%)	Bot (10.9%) Human (89.1%)
Russo-Ukrainian Sentiment	2205	pro_Ukraine (42.5%) pro_Russia (24.5%) Not sure (33.0%)	2110 (95.7%)	pro_Ukraine (16.3%) pro_Russia (31.0%) Not sure (52.7%)

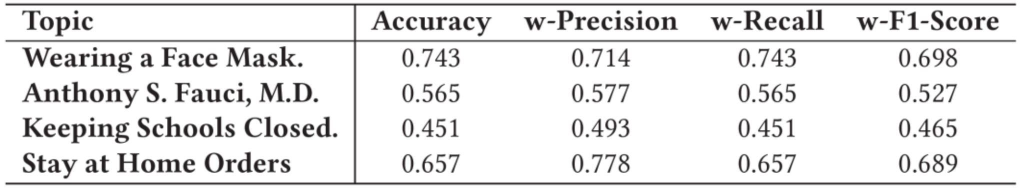
Result (Cont'd)



Annotation Task	Accuracy	w-Precision	w-Recall	w-F1-Score
Stance Detection	0.612	0.647	0.612	0.604
Hate Speech	0.571	0.720	0.571	0.581
Sentiment Analysis	**0.649**	0.684	**0.650**	**0.646**
Bot Detection	0.639	**0.761**	0.639	0.572
Russo-Ukrainian Sentiment	0.573	0.642	0.573	0.551



Annotation Task	Label	Precision	Recall	F1-Score
Stance Detection	In-favor	0.591	0.719	**0.649**
	Against	0.540	0.747	0.627
	Neither	0.767	0.426	0.548
Hate Speech	Hate	0.353	0.969	0.518
	Counterspeech	0.791	0.610	**0.689**
	Neutral	0.812	0.427	0.560
Sentiment Analysis	Positive	0.813	0.514	0.630
	Negative	0.575	0.623	0.598
	Neutral	0.593	0.790	**0.677**
Bot Detection	Bot	0.951	0.226	0.364
	Human	0.601	0.990	**0.748**
Russo-Ukrainian Sentiment	Pro-Ukraine	0.769	0.297	0.429
	Pro-Russia	0.671	0.827	**0.733**
	Not Sure	0.455	0.741	0.562



Topic	Accuracy	w-Precision	w-Recall	w-F1-Score
Wearing a Face Mask.	0.743	0.714	0.743	0.698
Anthony S. Fauci, M.D.	0.565	0.577	0.565	0.527
Keeping Schools Closed.	0.451	0.493	0.451	0.465
Stay at Home Orders	0.657	0.778	0.657	0.689