Annotation Guidelines

Evaluate each dialogue response given its dialogue context (i.e., dialogue history) and the following criteria, distinguished in two categories.

Please, consider that the responses have been tokenized, lowercased and that punctuation and articles have been removed, therefore, these are not mistakes made by the models.

CATEGORY I

- the properties are evaluated on a Likert scale *(1-5).*
- if score equals *0*, put *1* instead
- if score has first decimal < *.5* round down
- if score has first decimal >= *.5* round up

Soundness

The response should contain conceptually logical information that is likely to be true based on common sense and factual knowledge. If you are not certain about easy-to-retrieve information, as in (b), please search online. You don't have to search for information that is challenging to find, as in (c).

5 all statements are true

4 the truthfulness of one statement is obscure

3 one statement is not true, but there are other true statements

2 two or more statements are not true, but there are other true statements

1 none of the statements are true

eg.

a)	RESPONSE:	"*Michael Jackson is a country.*"	***(1)***
b)	RESPONSE:	*"Michael Jackson is a performer and a politician."*	***(3)***
c)	RESPONSE:	*"Michael Jackson went to France 30 years ago."*	***(4)***
d)	RESPONSE:	"*I like Michael Jackson!*"	***(5)***

Conciseness

The response should not provide more content than necessary for the communicative goal to be addressed and the meaning to be conveyed.

NOTE: The communicative goal does not have to be achieved (see d).

5 no redundant statements

4 one redundant statement

3 two redundant statements

2 three redundant statements

1 four or more redundant statements

eg.

HISTORY: *“When did Michael Jackson die?”*

a) RESPONSE: “*Michael Jackson died in 2009.”* ***(5)***

b) RESPONSE: *“Michael Jackson died from intoxication.”* ***(4)***

c) RESPONSE: *“Michael Jackson died from intoxication in 2009.”* ***(4)***

d) RESPONSE: *“I don’t know”* ***(5)***

Completeness

The response should not provide less information than necessary for the communicative goal to be addressed and its meaning to be conveyed.

.

NOTE: A complete response does not guarantee that the communicative goal is achieved. In (c) the response is complete and scores 5, even though the communicative goal (i.e., knowledge acquisition) is not reached.

Completeness = (the amount of necessary info stated / the amount of info we judge is necessary) * 5

eg.

HISTORY: *“Would you recommend this movie? Who is starring?”*

a) RESPONSE: *“Yes, I could totally recommend it.”* ***(3)*** ½ * 5 = 2.5 = 3
(It doesn’t mention who is starring)

b) RESPONSE: *“I don’t know.”* ***(1)*** 0/2 * 5 = 0 = 1
(It doesn’t mention what the speaker doesn’t know)

c) RESPONSE *“I would totally recommend it. I don’t know who is starring.”* ***(5)*** 2/2 * 5 = 5

Relevance

The response should relate to the conversation history and the communicative goal, as you perceive it.

NOTE: The communicative goal does not have to be achieved, but it does need to be considered for the generation of the response.

5 relevant to most recent turn and communicative goal of the entire history with specific details

4 relevant only to the most recent turn with moderate specificity and likely containing a rather generic queue *(eg. Do you like X?)*

3 a very generic response, but still applicable *(eg. I don’t know).*

2 only thematically (i.e., topic-based) relevant to the most recent turn, but not the communicative goal

1 not relevant at all

eg.

HISTORY: – *"Do you know Jordan Smith?"*
– *"I think I saw him on TV. He's a golfer born in Dallas"*
– *"Do you know any other athletes?"*
– *"What about Rohit Sharma? He plays for the Mumbai Indians."*

a) RESPONSE: *"He's a great player! He also plays in the national team."* ***(5)***
(Relevant to most recent turn. Relevant to the communicative goal of the entire history i.e., sports chit chat and knowledge exchange. The second sentence adds specificity).

b) RESPONSE: *"I haven't heard of him. Do you like the Mumbai Indians?"* ***(4)***

c) RESPONSE: "Rohit Sharma is an athlete." ***(2)***

Clarity

The response should NOT be:

- semantically ambiguous (its meaning allows more than one interpretation)
- syntactically ambiguous (its syntax allows more than one interpretation)
- semantically obscure (the concepts or words do not convey a clear meaning, sound likely unnatural and are hard to understand)
- syntactically obscure (the structure and grammar are complex and/or unnatural and require careful parsing to interpret the response.)

5 zero undesired properties present

4 one undesired property present

3 two undesired properties present

2 three undesired properties present

1 all undesired properties present

eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: "*I don't know!"* ***(4)***
(Semantically ambiguous)

b) RESPONSE: "*Recommend the stars!"* ***(3)***
(Syntactically ambiguous, semantically obscure)

c) RESPONSE: "A performance to like." ***(3)***

(Semantically ambiguous, syntactically obscure)

Brevity

The response should not contain any unnecessary verbalizations, such as word or phrase repetitions. The response should demonstrate ability to use anaphoric expressions either within itself or in relation to the previous context

NOTE: Brevity should be distinguished from Conciseness. Conciseness refers to the conceptual content of the response, while Brevity to the lexical content of the response.

5 no unnecessary verbosity

4 one unnecessary, but grammatical verbosity

3 two unnecessary, but grammatical, verbosities

2 ungrammatical unnecessary verbosities, but the meaning can still be discerned

1 random ungrammatical and unnecessary verbosities that hinder interpretation

eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: *"I don't know who is starring in this movie."* ***(3)***

(Two grammatical verbosities: "who is starring" and "in this movie").

b) RESPONSE: *"I would recommend this movie and this movie."* ***(2)***

(The second "this movie" is an ungrammatical verbosity, given that there is no context suggesting the existence of a second movie)

c) RESPONSE: *"I would recommend this movie and on".* ***(1)***

("on" is a random, ungrammatical verbosity that hinders interpretation).

Coherence

The information presented in the response should be semantically and syntactically connected to each other and the most recent history turn in a logical order.

5 strong coherence both within the response and in relation to the previous turn

4 weak coherence either within the response OR in relation to the previous turn

3 coherence is lacking within the response OR in relation to the previous turn, but the meaning is still conveyed

2 coherence is lacking within the response OR in relation to the previous turn, and difficult to interpret the meaning

1 no coherence

eg.

HISTORY: *"Do you know Selena Gomez?"*

a) RESPONSE: *"Yes, she's an American singer. Do you like rock music?"* ***(4)***

(The second sentence displays weak semantic coherence in relation to the first sentence since Selena Gomez belongs in the pop genre).

b) RESPONSE: *"I've never heard of Katy Perry! What genre of music does she sing?"* ***(2)***

(The first sentence lacks coherence in relation to the last history turn, causing confusion in interpretation).

c) RESPONSE: *"Tell me a song of hers! I like "Liar."* ***(3)***

(The logical connection between the two sentences and between the response and the last history turn is lacking but the meaning can still be conveyed. The first sentence in the response suggests that the speaker is not sure if they know the singer and ask for details. However, the second sentence suggests they already know the singer).

CATEGORY II

- The properties are evaluated using the following fixed categorical values:

Y (yes)
N (no)
P (part)

Perspective: Dialogue Act

Please find the dialogue act classes used in this work on Appendix A1.

NOTE: -Always take into account only the given context, and not other factors that might influence the dialogue act in the real world, such as previous conversations or the physical context. For instance, the dialogue act in (c) might fit in a real-world setting (eg. if the speaker has been asked the same question repeatedly), but it does not match the given dialogue context.
-The most neutral dialogue act type is *statement-non opinion* and is likely to fit in most contexts, but without being the perfect candidate (see e). It is up to your judgment to decide whether this type is appropriate, sufficient and natural given the context.

Y (yes) the response displays an appropriate dialogue act, given the dialogue history

N (no) the response displays an inappropriate dialogue act, given the dialogue history

P(part) only part of the response displays an appropriate dialogue act

eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: *"I am so sorry!"* ***(N)***

(Apology)

b) RESPONSE: *"You're going to love it. Leo is starring."* ***(Y)***

(General Opinion + Statement-non opinion)

c) RESPONSE: *"Not again! Really?"* ***(N)***

(Complaint)

d) RESPONSE: *"This movie is directed by Martin Scorsese starring Leonardo DiCaprio."* ***(P)***

(The *statement-non-opinion* dialogue act addressing the second sentence of the history turn, is appropriate. There is no dialogue act directed to the first sentence of the history turn).

e) RESPONSE: *"This movie has received good reviews."* ***(N)***

(The *statement-non-opinion* dialogue act addressing the first sentence of the history turn is not appropriate. There is no dialogue act directed to the first sentence of the history turn.)

Perspective: Emotion

Please, find the emotion classes used in this work on Appendix A2.

NOTE: - The most neutral emotion type is *neutral* and is likely to fit in most contexts, but without being the perfect candidate. It is up to your judgment to decide whether this type is appropriate, sufficient and natural given the context.
- Always take into account only the given context, and not other factors that might influence the emotion in the real world, such as previous conversations or the physical context.

Y (yes) the response displays a relevant emotion

N (no) the response displays an irrelevant emotion

P(part) only part of the response displays a relevant emotion

eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: *"I didn't really like it."* ***(P)***

(The emotion addressing the first sentence of the dialogue history is appropriate, but the second sentence is not addressed emotionally).

Communicative Goal

Knowledge exchange, knowledge acquisition and chit-chat on a specific topic are the most frequent communicative goals in the data.

NOTE: - For a response to be labeled with Y, the response needs to achieve the communicative goal, not just be relevant to it.

Y (yes) the communicative goal is achieved

N (no) the communicative goal is not achieved

P(part) the communicative goal is partially achieved or not all communicative goals are achieved.

eg.

HISTORY: *“Would you recommend this movie? Who is starring?”*

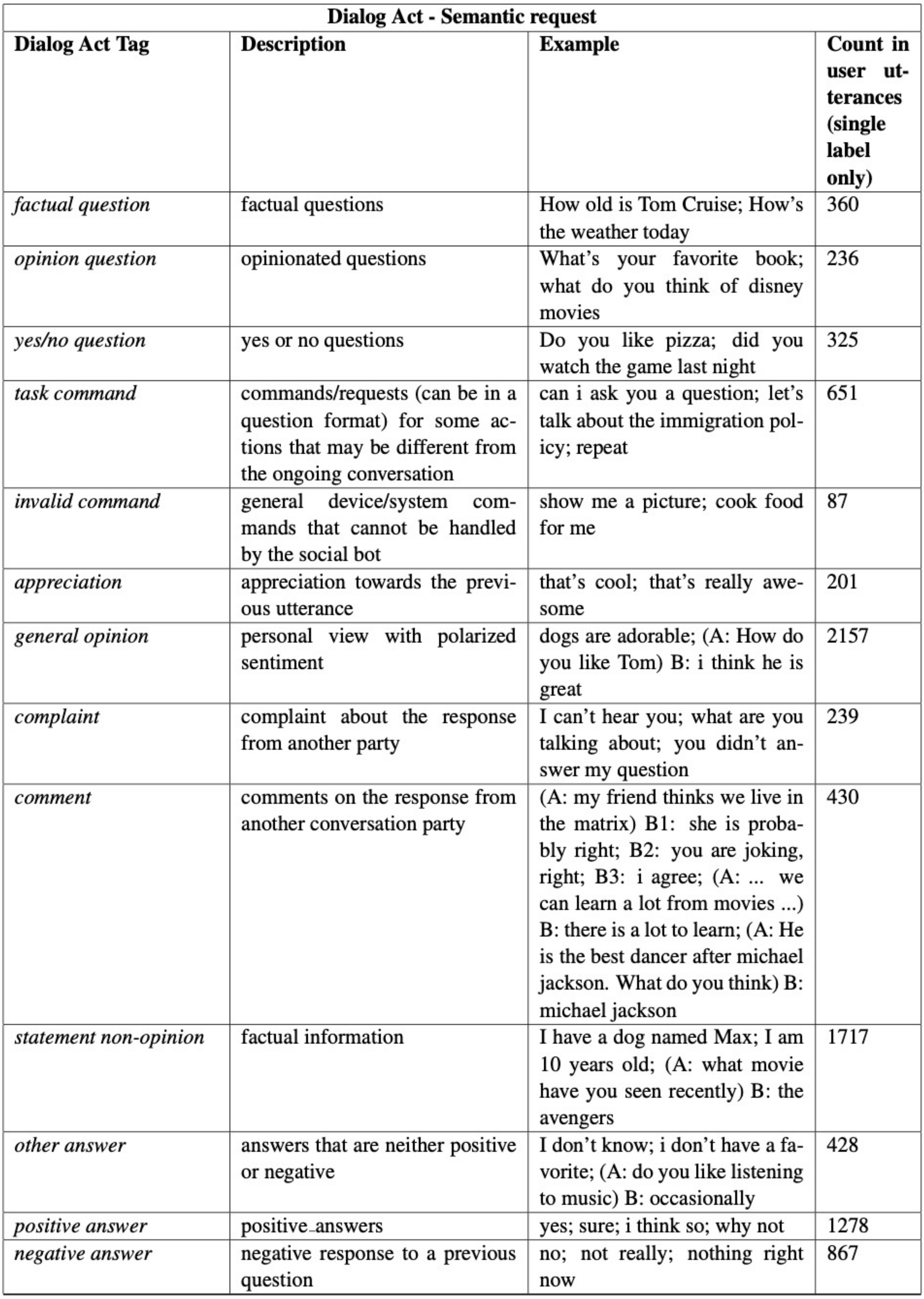
a) RESPONSE: “*Johnny Depp is starring in the Pirates of the Caribbean. I like him.”* ***(P)***

(The communicative goal expressed by the first sentence in the dialogue history is not achieved).

b) RESPONSE: *“I wouldn’t recommend. I don’t remember the actor’s name.”* ***(P)***

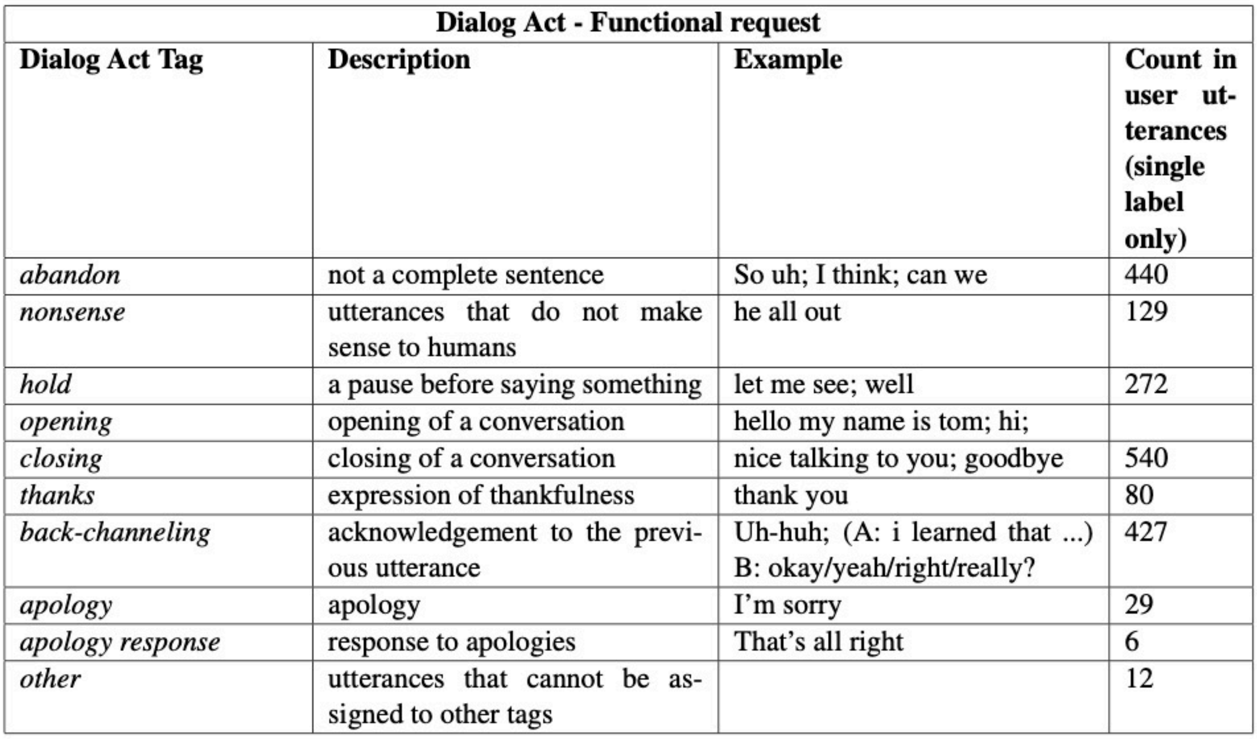
(The second part of the response is relevant to the second question of the dialogue history, but the goal i.e., acquisition of knowledge is not achieved).

A1. Dialogue-act classes



Dialog Act - Semantic request			
Dialog Act Tag	**Description**	**Example**	**Count in user utterances (single label only)**
factual question	factual questions	How old is Tom Cruise; How's the weather today	360
opinion question	opinionated questions	What's your favorite book; what do you think of disney movies	236
yes/no question	yes or no questions	Do you like pizza; did you watch the game last night	325
task command	commands/requests (can be in a question format) for some actions that may be different from the ongoing conversation	can i ask you a question; let's talk about the immigration policy; repeat	651
invalid command	general device/system commands that cannot be handled by the social bot	show me a picture; cook food for me	87
appreciation	appreciation towards the previous utterance	that's cool; that's really awesome	201
general opinion	personal view with polarized sentiment	dogs are adorable; (A: How do you like Tom) B: i think he is great	2157
complaint	complaint about the response from another party	I can't hear you; what are you talking about; you didn't answer my question	239
comment	comments on the response from another conversation party	(A: my friend thinks we live in the matrix) B1: she is probably right; B2: you are joking, right; B3: i agree; (A: ... we can learn a lot from movies ...) B: there is a lot to learn; (A: He is the best dancer after michael jackson. What do you think) B: michael jackson	430
statement non-opinion	factual information	I have a dog named Max; I am 10 years old; (A: what movie have you seen recently) B: the avengers	1717
other answer	answers that are neither positive or negative	I don't know; i don't have a favorite; (A: do you like listening to music) B: occasionally	428
positive answer	positive_answers	yes; sure; i think so; why not	1278
negative answer	negative response to a previous question	no; not really; nothing right now	867

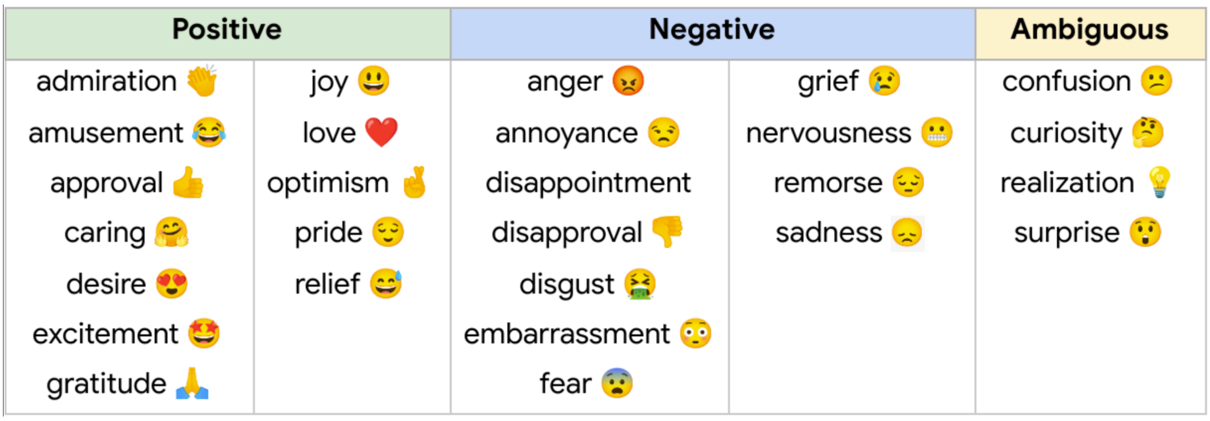
Figure 1. Yu, D., & Yu, Z. (2019). Midas: A dialog act annotation scheme for open domain human machine spoken conversations. arXiv preprint arXiv:1908.10023.



Dialog Act - Functional request			
Dialog Act Tag	**Description**	**Example**	**Count in user utterances (single label only)**
abandon	not a complete sentence	So uh; I think; can we	440
nonsense	utterances that do not make sense to humans	he all out	129
hold	a pause before saying something	let me see; well	272
opening	opening of a conversation	hello my name is tom; hi;	
closing	closing of a conversation	nice talking to you; goodbye	540
thanks	expression of thankfulness	thank you	80
back-channeling	acknowledgement to the previous utterance	Uh-huh; (A: i learned that ...) B: okay/yeah/right/really?	427
apology	apology	I'm sorry	29
apology response	response to apologies	That's all right	6
other	utterances that cannot be assigned to other tags		12

Figure 2. Yu, D., & Yu, Z. (2019). Midas: A dialog act annotation scheme for open domain human machine spoken conversations. arXiv preprint arXiv:1908.10023.

A2. Emotion classes



Positive		Negative		Ambiguous
admiration 👏	joy 😃	anger 😡	grief 😢	confusion 😕
amusement 😂	love ❤️	annoyance 😒	nervousness 😬	curiosity 🤔
approval 👍	optimism 🤞	disappointment	remorse 😔	realization 💡
caring 🤗	pride 😌	disapproval 👎	sadness 😞	surprise 😲
desire 😍	relief 😅	disgust 🤮		
excitement 🤩		embarrassment 😳		
gratitude 🙏		fear 😨		

Figure 3. Extracted from https://blog.research.google/2021/10/goemotions-dataset-for-fine-grained.html?m=1