Predicting Face Acts with Pre-trained Language Models

Anonymous ACL submission

Abstract

Face is an individual's public image we seek to establish in human interaction, and *face acts* are speech acts that either positively or negatively affect faces. The current study employed conventional neural networks, although the model requires training to classify face acts in a specific domain, which results in a lack of generalizability. For two reasons, we attempt to classify face acts using GPT-3, a well-known pre-trained language model (PLM) that can solve various classification tasks with few-shot learning. First, we hypothesize GPT-3 to know what face acts are, and we hope to elicit that ability for the task with few-shot learning. Second, we assume that pre-training positively impacts face act classification, and we can see the effect by comparing fine-tuned GPT-3 with the previous model. Experiments reveal that we cannot elicit GPT-3's ability for this task with few-shot learning. However, we confirm that fine-tuned GPT-3 could outperform the previous study and maintain almost the same performance as the previous study, even with a quarter of the original training data.

1 Introduction

Politeness theory explains how we care for others to facilitate human relationships. A concept called *face* is employed to define one of the reliable politeness theories (Brown and Levinson, 1978). Face is the innate human need for self-esteem or freedom from imposition, and *face acts* refer to speech acts that affect faces in conversations. For example, a desire to be recognized by others is related to self-esteem and called *positive face*, and an utterance that praises others is regarded as a face act that *raises* the other's positive face. Face and face acts have been firmly established concepts in sociolinguistics and pragmatics and have recently attracted increasing attention in persuasive dialogue system development (Dutt et al., 2021, 2020). In persuasion, there is a need to make inconvenient requests; thus, considering others more consciously than in everyday conversation is inevitable. Therefore, it is essential to consider how face acts should be used to construct a persuasive dialogue system.

Face act classification is a task for deciding whether an utterance has a face act and, if so, which type of face act it has. Face acts are the speaker's true intentions hidden behind the utterance, and face act classification is difficult because it requires reading these intentions. The previous study employed hierarchical neural networks to classify face acts (Dutt et al., 2020). However, such models require training in specific domains and lack generalizability. For instance, if the model was trained on a conversation in a persuasive situation related to a donation, it may not be helpful for face act classification in another persuasive situation, such as a hostage negotiation.

In this research, we verify the usefulness of GPT-3 (Brown et al., 2020), a pre-trained large-scale general-purpose language model, for face act classification. There are several ways to apply GPT-3 for classification tasks. Among them, we employ few-shot learning and fine-tuning. Our research hypotheses are as follows. First, we assume that GPT-3 acquires knowledge about face acts through in-context learning and can classify face acts with few-shot learning, which is less domain-dependent and more general than supervised learning. To verify this hypothesis, we provide several pairs of an utterance and a face act along with the test input and confirm whether GPT-3 can classify face acts referring to demonstrations. Second, we expect that even if GPT-3 does not know what face acts are, it has enough knowledge to classify them; thus, we can boost that knowledge with fine-tuning. To certify this assumption, we first prepare sets of pairs of an utterance and a face act label. Regarding them as a training dataset, we conduct supervised learning and apply the fine-tuned model for inference.

There are two main contributions of this study:

- We clarified that it is difficult for GPT-3 to classify face acts with few-shot learning. In other words, in-context learning of GPT-3 does not provide enough ability to solve face act classification.
- We confirmed that fine-tuning GPT-3 could outperform the previous study and found that fine-tuning with about 25% of the training data can produce results comparable to the previous one. Furthermore, we scrutinized the output of fine-tuned GPT-3 and analyzed the current bottlenecks in improving classification performance.

2 Background

2.1 Face

Face is our primary need related to human relationships with others in social life. This concept was introduced by Goffman (1967). In Brown and Levinson's politeness theory, face can be divided into two categories: *positive face* and *negative face*.

A positive face is a desire to be recognized, admired, and liked by others. On the other hand, a negative face is a desire not to let others invade one's freedom or domain. Brown and Levinson established politeness theory by applying the concept of face, and systematized the verbal behaviors that influence faces as politeness strategies.

2.2 Face Act

Face acts are speech acts that affect either oneself or others' faces. Face acts can be divided into two types. *Face Threatening Act* (FTA) is a speech act that attacks either positive or negative faces. On the other hand, *Face Saving Act* (FSA) is a speech act that saves either positive or negative faces.

According to the politeness theory, people tend to avoid attacking faces as much as possible to manage relationships. Also, even when they must attack faces, they will do it in a way that reduces the risk of attacking faces by employing politeness strategies such as implying their needs or apologizing for what they have requested.

The previous study divided face acts into eight categories based on the following three criteria (Dutt et al., 2020).

- whether it is directed toward the *speaker* or the *hearer* (**s/h**)
- whether it is directed toward a *positive* or *negative* face (**pos/neg**)
- whether the face is *saved* or *attacked* (**+/-**)

Table 1 is a generalized framework of face acts shown in Dutt et al. (2020). Suppose a persuasive situation where there are two people. One of them who makes the other mind change is called *persuader*, and the other side is called *persuadee*. When the persuader requests the persuadee to do something, the utterance is a face act categorized as **hneg-**. That is because the speaker is taking away the hearer's freedom. On the other hand, when the persuader shows the validity of his argument, the utterance has face act categorized as **spos+**, as the speaker is defending his positive face.

2.3 Face Act Classification

2.3.1 Task Formulation

This research follows the task formulation employed in Dutt et al. (2020). Given one conversation with n utterances $D = [u_1, u_2, ..., u_n]$, consider assigning them face act labels $y_1, y_2..., y_n$. The label $y_i \in Y$ can be one of the eight face acts explained in Section 2.2 or **other**, which means that the utterance does not affect faces. Those utterances categorized as **other** are greetings, fillers, and utterances unrelated to the main topic of the conversation. The model predicts one face act for each utterance in face act classification.

2.3.2 Dataset

The representative English dataset employed in face act classification is created by Dutt et al. (2020). This study annotated face acts in persuasion dialogues about fundraising for a charity named *Save the Children* (STC)[1]. In the whole conversation, there are two people called *persuader* (**ER**) and *persuadee* (**EE**), and ER persuades EE to donate to a charitable organization. Table 2 is a part of a conversation in the dataset.

The dialogue was initially collected in Wang et al. (2019). Only one face act is attached to each utterance in Dutt et al. (2020). Although it might be possible that one utterance has two or more face acts, the previous study reported that those utterances comprise only 2% of the dataset. Therefore, they randomly selected only one face act out of possible face acts. They regarded it as a gold label and formulated face act classification as predicting only one face act for each utterance.

There are two types of conversations called "Donor conversation" and "Non-Donor conversa-

[1] https://www.savethechildren.org

Table 1: A generalized framework for situating and operationalizing face acts in conversations presented in Dutt et al. (2020).

Face Act	Description
spos+	(i) S posit that they are virtuous in some aspects or they are good. (ii) S compliment the brand or item they represent or endorse and thus project their credibility. (iii) S state their preference or want, something that they like or value.
spos-	(i) S confess or apologize for being unable to do something that is expected of them. (ii) S criticise or humiliate themselves. They damage their reputation or values by either saying they are not so virtuous or criticizes some aspect of the brand/item they endorse or support.
hpos+	(i) S compliment H either for H's virtues, efforts, likes or desires. It also extends to S acknowledging the efforts of H and showing support for H. (ii) S can also provide an implicit compliment to incentivize H to do something good. (iii) S empathize / sympathize or in general agree with H. (iv) S is willing to do the FTA as imposed by H (implying that the FTA is agreeable to S.)
hpos-	(i) S voice doubts or criticize H or the product/brand that H endorses. (ii) S disagree with H over some stance, basically contradicting their viewpoint. (iii) S is either unaware or indifferent to H's wants or preferences.
sneg+	(i) S reject or are unwilling to do the FTA. Stating the reason does not change the circumstances of noncompliance but sometimes helps to mitigate the face act.
sneg-	(i) S offer to assist H.
hneg+	(i) S seek to decrease the imposition of the FTA on H by either decreasing the inconvenience such as providing alternate, simpler ways to carry out the FTA or decrease the threat associated with the FTA. (ii) S apologize for the FTA to show that S understood the inconvenience of imposing the request but they have to request nevertheless.
hneg-	(i) S impose an FTA on the H. The FTA is some act which H would not have done on their own. (ii) S increase the threat or ranking of the FTA (iii) S ask/request H for assistance?

Table 2: An example of a part of an annotated conversation with face act labels. In this two people's conversations, they are given roles *persuader* (**ER**) and *persuadee* (**EE**). ER persuades EE to donate to a charitable organization.

Speaker	Utterance	Face act
ER	Would you be interested today in making a donation to a charity?	**hneg-**
EE	Which charity would that be?	**other**
ER	The charity we're taking donations for is save the children!	**other**
EE	I've seen a lot of commercials about them, but never did a lot of research about them.	**hpos+**
ER	They are actually really great.	**spos+**

tion". Donor conversation is a conversation where ER successfully persuades EE. On the other hand, a Non-Donor conversation is a conversation where ER fails to persuade EE. The annotated dataset has 231 Donor conversations and 65 Non-Donor conversations.

2.4 GPT-3

GPT-3 is one of the most significant publicly available transformer models developed by OpenAI (Brown et al., 2020). It is an autoregressive NLP model that can be applied to various tasks such as classification, question-answering, translation, and summarization.

We can elicit GPT-3's ability by using *prompts*, which are queries written in natural language (Radford et al., 2019; Brown et al., 2020). Prompts composed of natural language tokens are called discrete prompts, and many studies are related to designing discrete prompts. One of the commonly used prompt formats is a question-answering style (Zhao et al., 2021; Wei et al., 2022b; Lu et al., 2022). It is a format in which the sentences targeted by a task, such as summarization or sentiment analysis, are passed to the GPT-3 in the form of questions, and the generated output is considered the answer to that task. In addition, as the classification is conducted by generating the continuation of the prompt freely, the generation might be out of track. Therefore some previous studies apply a method to suppress invalid labels by including instructions in a prompt (Mishra et al., 2022; Chiu and Alexander, 2021; Wei et al., 2022a; Wang et al., 2022).

One of the most basic methods for exploiting GPT-3's inference capabilities with discrete prompts is *few-shot* learning. In few-shot learning, the prompt consists of an explanation about the task, a few input-output pairs (demonstrations), and test data. Since model parameters are not updated, it has the advantage of not having to create a new data set for each task as in fine-tuning and of not overfitting the target task's data.

In addition to the low training resource methods

mentioned above, we can take a fine-tuning approach to utilize GPT-3. When we fine-tune GPT-3, models are trained with enough training data for the specific target task in a supervised manner. After that, we provide prompts with only test data for inference.

3 Method

3.1 Experimental Overview

We let GPT-3 output a continuation of the prompt in this experiment. We consider a valid face act string in the completed token sequence as a prediction. As we mentioned in Section 2.4, we tried the following two methods: few-shot learning and fine-tuning.

3.1.1 Few-shot Learning

We create a prompt with input and output pairs, so-called "demonstrations." We randomly choose two demonstrations from the training data for each face act. The dataset has seven possible face acts excluding **sneg-**. Also, there are utterances labeled as **other**. Therefore, one prompt must have sixteen demonstrations. To reduce the influence of prompt variation, such as the content or order of demonstrations, we prepare three sets of prompts for each setting and average their results.

We conducted the preliminary experiment to confirm the influence of demonstrations whose output labels are **other**. From the experiment, we find that adding demonstrations labeled with **other** is beneficial; therefore, we report experimental results in that setting. See Appendix D for comparing demonstrations with or without **other**.

3.1.2 Fine-tuning

When we fine-tune GPT-3, we must prepare pairs of inputs and outputs (prompt and completion). We create the dataset as follows: first, we prepare prompts according to the basic format explained in Section 3.2, then pair them with face act labels. After that, we train GPT-3 and utilize the fine-tuned model for classification.

When creating the training data for fine-tuning, we have to consider the label bias, as the distribution of face acts in the dataset is skewed. We attempt simple oversampling to mitigate this bias. We run the experiment once since we experiment in a situation with no randomness in the output. Also, in this setting, we conduct another experiment to reduce the number of training data to see how the accuracy changes. The detailed procedure of oversampling and reducing the training data is explained in Appendix A.

3.2 Prompt

3.2.1 Basic Form

There is room for designing prompts as we adopt them written in natural language. An example of a part of a prompt we employed in this study is shown in Table 3. Based on previous studies we referred to in Section 2.4, we designed prompts that we use throughout our experiments. We can leverage the question-answering style as the task we focus on is classification, and the style is commonly employed for this significant task type. Also, the valid labels we employ in this study are not standard as options such as Yes/No, Positive/Negative, and True/False. Therefore we put the task specifications and valid labels on the top of the prompts.

In the valid label names, we do not include the face act **sneg-**, the only face act that does not appear in the dataset. The previous study trained the model based on this dataset. In other words, the model is trained not to generate **sneg-**. In order to experiment with a similar condition, we consider eight types of labels, including seven face acts and **other** that do not belong to any of them.

3.2.2 History Length

Dutt et al. (2020) claimed that it is beneficial to consider past utterances when classifying face acts. Therefore, in this experiment, we change the length of the script and examine its effect. There are three history lengths from one to three. The example shown in Table 3 is prompt with three utterances; thus, the history length is three.

3.3 Data Splitting

In this study, we employ the dataset[2] explained in Section 2.3.2. As we fine-tune GPT-3, we divide the original dataset into training and test data. Note that when we conduct few-shot learning, we use training data only for extracting demonstrations for prompts and report the classification result of the test data. When we create training data, we randomly choose 80% of all Donor conversations and all Non-Donor conversations. The remaining conversation data becomes test data. Table 4 shows the number of utterances and conversations in both training and test datasets.

[2]This data is licensed under the MIT license.

Table 3: An example of a part of a prompt with a conversation history length is three. The part highlighted in gray can be replaced, and the other part is fixed throughout the experiment. GPT-3 generates the continuation of the prompt right after "Answer:."

Prompt	Gold label
Question: Read the following script, and classify EE's last utterance of the script " Not really, please tell me anything! " based on whether its face act is spos+, spos-, hpos+, hpos-, sneg+, hneg+, or hneg-. If there is no corresponding face act, then classify it as other. Note that "STC" stands for "Save the Children." Script: ER: I'd like to tell you about an organization called Save the Children. ER: Have you heard about them? EE: Not really, please tell me anything! Answer:	hpos-

Table 4: The left eight columns represent the number of utterances for each face act in the dataset. The rightmost column shows the number of conversations. Note that an "utterance" represents one or more sentences annotated with one face act, and a "conversation" represents a set of utterances from the beginning to the end.

	spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**	**other**	#conv
Training, Donor	991	4	1928	154	110	219	656	2733	185
Training, Non-Donor	265	3	329	118	76	46	211	699	52
Test, Donor	262	3	511	40	28	33	161	682	46
Test, Non-Donor	71	2	76	22	45	7	45	186	13

In the previous study, they performed five-fold cross-validation, referring that the dataset size was relatively small. We assume that the number of utterances in the test split is enough, as we can see several datasets employed for a measure of classification tasks are the same in the order of magnitude (Socher et al., 2013; Pang and Lee, 2004; Ganapathibhotla and Liu, 2008). Therefore, we take different procedures and use the fixed training and test dataset.

4 Results

We report the classification result based on the accuracy and macro average of the F1-score calculated with scikit-learn[3]. Table 5 shows the precision, recall, and F1-score for each label in various settings. Also, it shows accuracy, a macro average of the F1-score, and the number of outputs not counted as valid predictions in each setting.

4.1 Few-shot Learning

Overall, classification performance remains lower than that of the previous study. There is a tendency for lengthening the conversation history increases the number of invalid outputs; meanwhile, there is no performance gain. The F1-score for face acts other than **other** decreases across the board, suggesting that lengthening the history may have made it harder to focus on the target utterances for face act classification. However, it seems that GPT-3 can acquire the ability to identify face act tendencies and perform classification through few-shot learning, even if the task is entirely unknown.

Looking at each face act, the F1-score of **spos+**, **hneg-**, and **other** are relatively higher than the other. **spos+** is a face act that raises the speaker's face and is frequently seen in the persuader's utterance, as the persuader needs to raise the charitable organization. On the other hand, **hneg-** is related to asking for donations or questioning about the charity. These utterances are distinct from other utterances and are stated directly; therefore, GPT-3 might be able to grasp their characteristics with a few demonstrations. Also, providing demonstrations whose labels are **other** taught GPT-3 which utterances should be classified explicitly as **other**. This result shows that GPT-3 can detect the characteristics of utterances even with no face acts.

4.2 Fine-tuning

No matter the length of the history, the results outperform the previous study. Although the prediction of face act labels with less training data is less accurate than those with more training data, simple oversampling generally works. The F1-score for each face act improved with longer histories, possibly because fine-tuning enabled GPT-3 to classify face acts considering the conversation history.

[3] https://scikit-learn.org/stable/

Table 5: Statistics for face act classification in each setting. In the table, **Dutt** represents the result obtained in Dutt et al. (2020), "Few" represents few-shot setting with demonstrations labeled as **other**, and "Fine" represents fine-tuning. The number next to the "Few" and "Fine" indicates the length of the conversation history. "Rand" represents the situation where the face act is chosen randomly, and "Maj" represents the situation where all predictions are **other**. "Acc" represents the accuracy, "F1" represents the macro average of the F1-score for each label, and "#Inv" represents the number of outputs that were not counted as valid face acts out of all 2174 outputs. Each cell below face acts represents precision, recall, and F1-score (p/r/f1) of a face act in a particular setting. For the few-shot setting, we take the average of three trials.

	spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**	**other**	Acc	F1	#Inv
Rand	-	-	-	-	-	-	-	-	.13	-	-
Maj	-	-	-	-	-	-	-	-	.40	-	-
Dutt	-	-	-	-	-	-	-	-	.69	.60	-
Few, 1	.32/.75/.44	.07/.87/.13	.49/.25/.29	.05/.02/.03	.03/.14/.19	.04/.48/.07	.48/.57/.51	.79/.19/.30	.33	.25	0
Few, 2	.27/.76/.39	.07/.87/.12	.41/.13/.17	.04/.02/.03	.55/.07/.13	.02/.30/.05	.48/.22/.29	.82/.19/.31	.26	.19	18
Few, 3	.27/.77/.39	.07/.92/.12	.48/.12/.17	.06/.03/.04	.50/.02/.04	.02/.20/.03	.54/.22/.28	.85/.22/.34	.27	.18	74
Fine, 1	.74/.70/.72	1.0/.20/.33	.70/.77/.73	.53/.52/.52	.78/.55/.65	.38/.62/.48	.74/.79/.77	.75/.71/.73	.72	.62	2
Fine, 2	.79/.70/.74	1.0/.80/.89	.75/.77/.76	.57/.53/.55	.81/.60/.69	.38/.62/.48	.79/.76/.78	.75/.78/.77	.75	**.71**	0
Fine, 3	.79/.72/.75	.67/.80/.73	.78/.78/.78	.57/.52/.54	.71/.66/.68	.43/.60/.50	.82/.76/.79	.76/.79/.78	**.76**	.69	1

Table 6: The number of utterances which contain "Your donation will be directly deducted from your task payment."

	hneg+	**other**	**hneg-**	Total
Training	26	22	2	50
Test	3	7	0	10

Table 7: The number of utterances which contain "You can choose any amount from 0$ to all of your payment."

	hneg+	**other**	Total
Training	17	17	34
Test	3	3	6

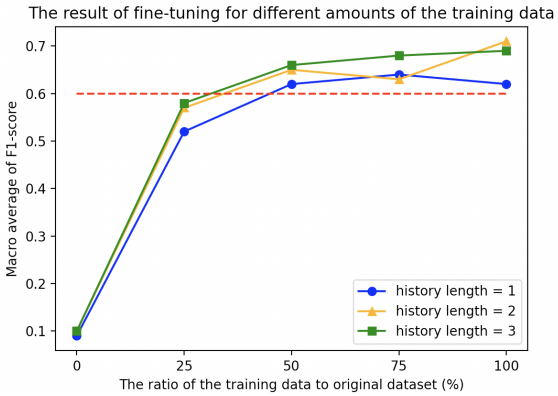

Figure 1: The result of fine-tuning for different amounts of the training data. The red dashed line shows the macro average of the F1-score obtained in Dutt et al. (2020). The leftmost points represent the result of the zero-shot learning (using no training datasets).

From Table 5, we can see that the performance of predicting **hpos-** and **hneg+** are lower than the other face acts. Looking at the result where the conversation history length is three, we can see the cause of this performance degradation.

The recall rate of **hpos-** is worse than precision, as we can see eight utterances whose labels are **hpos-** are classified as **hneg-**. The reason may be that **hpos-** is often euphemistically phrased because it is a criticism of the other and tends to be similar to other face act utterances. For example, a statement that doubts STC is classified as **hpos-**, and similar wording is sometimes used as a question about the STC, which is classified as **hneg-**.

On the other hand, 21 cases that should have been classified as **other** are mistakenly classified as **hneg+**, resulting in low precision. If we look closely at the result, most incorrectly classified utterances contain the following two patterns. Pattern 1 is "Your donation will be directly deducted from your task payment," and pattern 2 is "You can choose any amount from 0$ to all of your payment."

Table 6 and 7 shows the breakdown of utterances. The distribution of **hneg+** and **other** is nearly the same; however, GPT-3 predicts almost all of them as **hneg+** (9 out of 10 for pattern 1, and 5 out of 6 for pattern 2). Therefore if all these gold labels were assigned with **hneg+**, the accuracy would have improved. Overall, the label inconsistency that occurred at the annotation stage was a cause of the performance degradation.

Figure 1 shows the result of the setting where the training data size is reduced to 25%, 50%, and 75% of the original. The leftmost points are the result of zero-shot learning, where the basic format of the prompt is the same as a few-shot setting but provides no demonstrations. The detail of zero-shot learning is in Appendix C. We confirm that reducing the data to about 25% maintains the

Table 8: The part of the conversation whose last utterance is classified as **spos-**.

Speaker	Utterance
EE	no, I do not wish to donate at this time.
EE	there are other charities I'd like to donate to over this one.
EE	I'm sorry.

performance as much as the previous study. By experimenting with fewer data, it seems that more diverse data is needed to improve the classification performance for fewer labels to reduce label bias rather than the number of data itself.

5 Analysis

Through experiments, we found that the few-shot learning we used in this study was ineffective for face act classification. In this section, we review the merit of fine-tuning compared to few-shot learning and what is hard to classify even with fine-tuning.

5.1 Merit of Fine-tuning

From the results shown in Table 5, it can be said that GPT-3 did not sufficiently capture the essence of face acts through few-shot learning. As shown in Table 1, various patterns of utterances are classified into one face act. Therefore, two demonstrations for each face act might not be enough to teach what kind of utterances can be classified into which face act. In fine-tuning, we could provide various patterns of utterances; therefore, the typification of face acts was successful and significantly improved performance compared to the few-shot method.

Moreover, compared to the few-shot case, where there is no compulsion to keep track of the dialogue history, fine-tuning makes the classification more history-aware. We can see how it works through the classification of the rare face act **spos-**, which is directly related to the improvement in the macro average of the F1-score. Comparing the three fine-tuning settings with different lengths of the history, when conversation history is one, the F1-score of **spos-** is only around 0.3. However, if the conversation history becomes longer than one, the F1-score rises to around 0.8. We can see why considering past utterances is effective from Table 8. The EE's last utterance, "I'm sorry." is classified as **spos-**, as the utterance rejects answering ER's request. Utterances classified **spos-** have two characteristics: first, those utterances tend to be too short for classification; second, utterances are related to denying requests and tend to be said in a roundabout way. In this sense, **spos-** is a face act that requires history to be followed.

5.2 What is Hard to Classify with Fine-tuning

This section explores what types of utterances are hard to classify with fine-tuned GPT-3. Table 9 shows the four patterns of errors we observed. The two patterns we mention first are the problems that need high-level natural language understanding and are inherently difficult to solve with the current method. The latter two patterns are related to the nature of the dataset.

The first problem is that GPT-3 has a weakness against euphemisms. The first example in Table 9 implies the hesitation of "I can't donate to an organization I don't know well." The utterances that defend speakers tend to be conveyed indirectly; thus, they are hard to classify its face acts.

The second problem is that GPT-3 cannot precisely grasp the speaker's stance. In conversations about donating to charity, organizations other than STC (e.g., Make-A-Wish) could be mentioned as the second example in Table 9. When the persuadee says that he/she donates to STC, the utterance is classified as **hpos+** because it matches the persuader's desire and saves the persuader's face. However, if the persuadee shows his/her willingness to donate to another organization, the persuadee prioritizes his/her preference over the persuader's request, and the utterance is classified as **spos+**. Even if donating money itself is the same, the face act changes depending on whether or not it is fulfilling one's purpose when standing on the listener's side.

The third problem is that it is hard to classify unfinished utterances, as the following utterances cannot be referenced. The gold label of the original data may be assigned by looking at the entire conversation, and especially short utterances are assigned the same label as the preceding and following utterances. For instance, the third example in Table 9 cannot be classified **spos+** by itself. However, the following ER's utterance claims that STC is making good use of the money. Then the third example in the table is regarded as a support of the validity of STC; therefore, the utterance is classified as **spos+**.

The fourth problem is that face acts with similar characteristics are hard to discern. From examples 4-a and 4-b in Table 9, we can see that there are utterances similar in the content, but the annotated

Table 9: Examples that fine-tuned GPT-3 cannot classify. The rightmost column represents predictions of GPT-3 when the length of conversation history is one, two, and three.

	Speaker	Utterance	Gold label	Predictions
Example 1	EE	I like to learn as much as I can about an organization before donating.	**sneg+**	**hpos+/other/other**
Example 2	EE	Make a wish is really cool.	**spos+**	**other/other/other**
Example 3	ER	I'm not a fan of charities that keep a lot of their proceeds for themselves.	**spos+**	**other/other/other**
Example 4-a	ER	I think you can donate as little as one cent.	**hpos+**	**hneg+/hneg+/hneg+**
Example 4-b	ER	A little goes such a long way!	**hneg+**	**hpos+/hpos+/hpos+**

labels are inconsistent. These utterances can be interpreted in two ways: the persuader encourages the persuadee to do good (**hpos+**); the persuader wants to decrease the burden of donation (**hneg+**).

6 Related Work

Face act is a firmly established concept related to politeness theory, which is a branch of pragmatics. There are many studies about face acts in other research fields about human interaction. Indeed, there has been much research on oral communication (Naderi and Hirst, 2018; Hutchby, 2008), and the concept of face act can be employed for various types of two-party conversations. For example, in practice, there are studies about business letters (Maier, 1992; Jansen and Janssen, 2010). Business letters are basically exchanged between two persons and include contents such as requests and apologies. Therefore the politeness theory is exploited to explain the linguistic phenomenon in text messages.

Although the number of research that analyses this language phenomena from a view of computational linguistics is limited, in the recent evolution around the realm of the persuasive dialogue system, the concept of face act is coming into focus. Several pieces of research incorporate the concept of face act into dialogue systems (Gupta et al., 2007; Zhao et al., 2018). Persuasive dialogue systems have been researched for years, and there is a recent trend in analyzing conversational tricks such as face acts with neural networks. One of the significant conversational tricks extracted from utterances is persuasion strategies. For example, Wang et al. (2019) collected conversations about fundraising for charitable organizations and annotated ten types of persuasion strategies. They also analyzed the relationships between persuasion strategies, personal background, and the persuasion outcome. Some models use dialogues collected in Wang et al. (2019) to predict the persuasion outcome (Wang et al., 2019; Dutt et al., 2021; Sinha and Dasgupta, 2021). Also, a persuasive dialogue system incorporates persuasion strategies and evaluates how effective it is in actual persuasion (Chen et al., 2022).

7 Conclusion and Future Work

In this work, we clarified that currently, it is difficult to classify face acts by utilizing GPT-3 with few-shot learning. However, we found that fine-tuned GPT-3 performed better than the previous model. We also confirmed that the performance remained comparable to the previous study, even with about 25% of the original training data.

The limitation of this work is that whether we can apply the fine-tuned models to other persuasive dialogues is unchecked. If the model is proven not applicable, a new method with high generalizability must be found. Another limitation is that we have not gone deeply into designing prompts, and there is a possibility that we cannot fully elicit GPT-3's ability. More sophisticated prompt engineering, such as utilizing sequential prompts, might be another promising way for future work. Furthermore, we were only concerned with utterances in English in this experiment. Politeness is a culturally dependent phenomenon; thus, if we deal with face acts in other languages, we have to consider whether the framework of face acts defined in English is applicable, and also, we may have to employ other languages' PLM.

We aim to create another persuasive dialogue dataset annotated with face act to analyze face act classification with PLMs further. Another future research direction is that if GPT-3 can be used to classify face acts in a broader range of situations, it is expected to be applied as a module for user-conscious speech production in developing dialogue systems using large-scale language models.

8 Ethical Considerations

This research focuses on the basic technology supporting dialogue systems; therefore, there is no risk of abuse directly. However, if this technology is applied to persuasive dialogue systems in the future, there are potential risks that it could be used to generate false claims or misused for fraud indirectly. In addition, this study utilizes GPT-3. Therefore the results we obtained may be affected by the inherent aggressive knowledge, expressions, and various biases of the pre-trained large language model.

References

Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Maximillian Chen, Weiyan Shi, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2022. Seamlessly integrating factual information and social content with persuasive dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*, pages 399–413. Association for Computational Linguistics.

Ke-Li Chiu and Rohan Alexander. 2021. Detecting hate speech with GPT-3. *CoRR*, abs/2103.12407.

Ritam Dutt, Rishabh Joshi, and Carolyn P. Rosé. 2020. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7473–7485. Association for Computational Linguistics.

Ritam Dutt, Sayan Sinha, Rishabh Joshi, Surya Shekhar Chakraborty, Meredith Riggs, Xinru Yan, Haogang Bao, and Carolyn P. Rosé. 2021. Resper: Computationally modelling resisting strategies in persuasive conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 78–90. Association for Computational Linguistics.

Murthy Ganapathibhotla and Bing Liu. 2008. Mining opinions in comparative sentences. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 241–248.

Erving Goffman. 1967. *Interaction Ritual: Essays in Face to Face Behavior*. AldineTransaction.

Swati Gupta, Marilyn A. Walker, and Daniela M. Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, volume 4738 of *Lecture Notes in Computer Science*, pages 203–217. Springer.

Ian Hutchby. 2008. Participants' orientations to interruptions, rudeness and other impolite acts in talk-in-interaction. *Journal of Politeness Research-language Behaviour Culture - J POLITENESS RES-LANG BEH CUL*, 4:221–241.

Frank Jansen and D.M.L. Janssen. 2010. Effects of positive politeness strategies in business letters. *Journal of Pragmatics*, 42:2531–2548.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Paula Maier. 1992. Politeness strategies in business letters by native and non-native english speakers. *English for Specific Purposes*, 11:189–205.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2018. Using context to identify the language of face-saving. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 111–120. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278. ACL.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Manjira Sinha and Tirthankar Dasgupta. 2021. Predicting success of a persuasion through joint modeling of utterance categorization. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3423–3427. ACM.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5635–5649. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, and Daniel Khashabi. 2022. Benchmarking generalization via in-context instructions on 1, 600+ language tasks. *CoRR*, abs/2204.07705.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Ran Zhao, Oscar J. Romero, and Alex Rudnicky. 2018. SOGO: A social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018, Sydney, NSW, Australia, November 05-08, 2018*, pages 239–246. ACM.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Appendix

A Sampling the Training Data

A.1 How to Oversample Training Data

When we oversample the training data, we copy pairs of an utterance and a face act label. The most frequent label in the dataset is **other**, and 3432 utterances out of 8542 are labeled as **other**. Therefore, we duplicate utterances labeled with other valid face acts until the number reaches nearly 3432. The procedure for how many times we multiply the number of utterances is described below.

Given k original data labeled with a face act, find n satisfying

$$nk \leq 3432 \leq (n+1)k.$$

After that, we select either nk or $(n+1)k$ closer to 3432 and copy the original data for n or $n+1$ times.

For example, in the training data, there are 1256 utterances whose label is **spos+**. The n which satisfies

$$n \times 1256 \leq 3432 \leq (n+1) \times 1256$$

is 2. Compared to $2 \times 1256 = 2512$ and $3 \times 1256 = 3768$, the closer one to 3432 is 3768. Therefore we copy each utterance with the label **spos+** three times. Following the procedure above, we create the oversampled data. Table 10 details the number of oversampled training data.

A.2 How to Reduce Training Data

We divide the training data into 25%, 50%, and 75% of the original dataset, based on the success

Table 10: The number of utterances in the oversampled training data. A number inside a parenthesis represents how many times the labeled data is duplicated.

Face Acts	Donor	Non-Donor	Total
spos+	2973	795	3768 (3x)
spos-	1960	1470	3430 (490x)
hpos+	1928	329	2257 (1x)
hpos-	2002	1534	3536 (13x)
sneg+	1980	1368	3348 (18x)
hneg+	2847	598	3445 (13x)
hneg-	2624	844	3468 (4x)
other	2733	699	3432 (1x)
Total	19047	7637	26684

or failure of the persuasion. For example, if we sample 25% of the total dataset, we randomly select 25% of the 185 Donor conversations in the training data (46 conversations) and 25% of the 52 Non-Donor conversations (13 conversations). This way, we collect 59 conversations for 25% of the training dataset. Note that to ensure that at least one of every face act is selected, the first 25% is randomly re-selected to include utterances with all face acts.

When we sample 50% of the training dataset (93 Donor conversations and 26 Non-Donor conversations), we first select the same conversations we selected for the 25% of the dataset to ensure including at least one utterance from all face acts. After that, we select another 47 Donor conversations and 13 Non-Donor conversations.

Likewise, when we sample 75% of the training dataset (139 Donor conversations and 39 Non-Donor conversations), we first select the same conversations we selected for the 50% of the dataset. After that, we select another 46 Donor conversations and 13 Non-Donor conversations.

B Model and Decode Settings

There are four variations of GPT-3, which are all available. Among them, we employed two models; Curie and Davinci. In few-shot settings, we employed the Davinci model (text-davinci-002), which has 175 billion parameters and is the most powerful. On the other hand, in the fine-tuning setting, we employed the Curie model (text-curie-001), which has 13 billion parameters and is the second best-performing model. It also requires less money and time for training and inference than Davinci. The entire experiment cost approximately $2,100. Each experiment takes about 20 minutes for inference and 40 minutes for fine-tuning. When we train the models and let them infer, we adopt the OpenAI API[4]. We checked OpenAI's usage policies and experimented with following them.

We can set the model's parameter related to decoding. Those are "temperature" and "output token number". We set the temperature to 0 to eliminate randomness in the output. Also, we limit the output token number to 3. According to GPT-3's tokenization scheme, **other** is one token, and the other valid face acts are all three tokens. We plan to avoid outputting extra tokens other than the face act as much as possible. Therefore we set the output token number to 3.

C Zero-shot Learning

We conducted a preliminary experiment and confirmed that zero-shot learning is ineffective for face act classification. The example of the prompt we employed in the zero-shot setting is shown in Table 3. The prompt includes only the task description and the script. We provide it to GPT-3, and GPT-3 fills the face act of the last utterance of the script. The experimental settings are the same as the few-shot learning described in Appendix B. In zero-shot learning, we run the experiment once since we experiment in a situation with no randomness in the output.

The middle three rows in Table 11 shows the result of zero-shot learning. The accuracy is slightly better than what would be obtained if the prediction were made randomly among the possible eight choices. However, the overall performance is far lower than the previous study's. Since the face act label used in this study is defined in Dutt et al. (2020), it does not exist in the GPT-3's pre-training data; thus, GPT-3 cannot utilize in-context learning, and we confirm that face act classification is hard through zero-shot learning.

The fact that there are few invalid outputs suggests that at least GPT-3 understood this task as a classification task and what the valid label options are. Including task instructions in the prompt for reducing invalid outputs is also effective in this study. Table 12 shows the list of generated tokens in zero-shot learning.

D Few-shot Learning

We experimented to test the influence of demonstrations labeled as **other**. We prepared prompts with fourteen demonstrations. In other words, we removed demonstrations labeled as **other** from what

[4] https://openai.com/api/

Table 11: Statistics for face act classification in each setting. In the table, **Dutt** represents the result obtained in Dutt et al. (2020), "Rand" represents the situation where the face act is chosen randomly, and "Maj" represents the situation where all predictions are **other**. "Zero" represents Zero-shot learning and "Few" represents few-shot setting without demonstrations labeled as **other**. The number next to the symbol indicates the length of the conversation history. "Acc" represents the accuracy, "F1" represents the macro average of the F1-score, and "#Inv" represents the number of outputs that were not counted as valid face acts out of all 2174 outputs. Each cell below face acts represents precision, recall, and F1-score (p/r/f1) of a face act in a particular setting. For the few-shot setting, we take the average of three trials.

	spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**	**other**	Acc	F1	#Inv
Rand	-	-	-	-	-	-	-	-	.13	-	-
Maj	-	-	-	-	-	-	-	-	.40	-	-
Dutt	-	-	-	-	-	-	-	-	.69	.60	-
Zero, 1	.19/.75/.31	.02/.60/.04	.22/.17/.19	.18/.15/.16	.00/.00/.00	.03/.10/.04	.00/.00/.00	.00/.00/.00	.17	.09	11
Zero, 2	.19/.40/.25	.00/.00/.00	.26/.48/.34	.12/.21/.15	.00/.00/.00	.03/.13/.05	.04/.01/.02	.00/.00/.00	.20	.10	13
Zero, 3	.18/.37/.24	.00/.00/.00	.27/.51/.36	.12/.19/.15	.00/.00/.00	.02/.10/.04	.03/.01/.02	.00/.00/.00	.20	.10	26
Few, 1	.29/.79/.43	.07/.87/.13	.36/.21/.25	.02/.03/.02	.37/.11/.17	.03/.48/.06	.49/.48/.48	.95/.04/.08	.24	.20	0
Few, 2	.26/.78/.39	.06/.93/.10	.31/.14/.17	.02/.04/.03	.55/.04/.08	.02/.32/.04	.51/.23/.31	1.0/.05/.09	.21	.16	9
Few, 3	.25/.80/.38	.05/.85/.10	.37/.11/.15	.03/.05/.04	.56/.01/.03	.01/.17/.02	.53/.24/.30	1.0/.08/.14	.22	.15	64

Table 12: The list of generated tokens where the setting is zero-shot.

history length=1	spos+, hpos+, spos-, hpos-, hneg-, hneg+, Hi! is, \n\nHello, \n\ns, \n\nThere
history length=2	spos+, hpos+, spos-, hpos-, hneg-, hneg+, Hi! is, \n\nHello, \n\ns
history length=3	spos+, hpos+, spos-, hpos-, hneg-, hneg+, Hi! is, \n\nHello, \n\ns

we employed in the few-shot setting we reported in this paper. We employed the same procedure as the few-shot setting with the demonstrations labeled as **other**.

The bottom three rows in Table 11 shows the result of the experiment. The performance is lower than few-shot learning with demonstrations labeled as **other**. The difference in performance is mainly due to the F1-score of the **other** label. If we provide which utterance should be classified as **other**, it improves the recall of **other**; thus, the F1-score improve. Therefore, we provide demonstrations with **other** in the actual experiment.

Table 13 shows the list of generated tokens in few-shot learning without demonstrations labeled as **other**, and Table 14 shows the list of generated tokens in few-shot learning with demonstrations labeled as **other**. Note that these tables combine the results of three trials.

E Fine-tuning

E.1 Parameters

We select "text-curie-001" as a model for fine-tuning. We fine-tune the model with OpenAI API. We unchange hyperparameters from the default setting. In detail, the batch size is 32. The learning rate multiplier is set to 0.1, and the number of epochs is 4. The prompt loss weight is 0.01.

E.2 Outputs

We observe that fine-tuning causes generating extra tokens to follow when predicting the label **other**. All prediction of **other** follows extra tokens such as "other\nAnswer", "other other other", "other+\n", and so on. We judge those outputs as **other** since other face acts consist of three tokens and are not predicted with **other** in a complete form. Replacing the **other** with another sequence of tokens with three tokens would probably eliminate this problem, so it can be considered a less fundamental problem. Nonetheless, in this experiment, **other** was never predicted by itself in fine-tuning settings.

Table 13: The list of generated tokens where the setting is few-shot. The demonstrations exclude examples with **other** labels. We show all outputs we obtained in three trials for each setting.

history length=1	spos+, hneg+, hneg-, spos-, hpos+, hpos-, other, sneg+, \n\nother
history length=2	hpos+, hneg+, spos-, spos+, hneg-, other, \n\ns, \n \nh, \n\nOther, \n\nother, hpos-, sneg+
history length=3	hneg+, hneg-, spos+, spos-, hpos+, \n\ns, \n\nh, other, \n\nother, \n\nHello, hpos-, sneg+, \n\nOther, \n\nHi

Table 14: The list of generated tokens where the setting is few-shot. The demonstrations include examples with **other** labels. We show all outputs we obtained in three trials for each setting.

history length=1	spos+, hneg+, hneg-, other, hpos+, sneg+, spos-, hpos-
history length=2	spos+, hneg+, hneg-, spos-, hpos+, hpos-, other, sneg+, \n\nother, \n\ns
history length=3	other, hneg-, hneg+, spos+, \n\ns, spos-, hpos+, \n\nother, \n\nh, sneg+, \n\nQuestion, hpos-

Table 15: The list of generated tokens where the setting is fine-tuning.

history length=1	other\nAnswer, hpos+, other+\n, spos+, hpos-, hneg-, other other other, other\n\n, sneg+, hneg+, other\nNote, other hpos, other- sp, other hneg, monthly\nAnswer, other spos, one time donation, otherAnswer:, other good-, other: other, spos-, other\nThus
history length=2	other\nAnswer, hpos-, other other other, spos+, hneg-, hpos+, other\nQuestion, other\nER, other\nScript, other\nNote, other\n, hneg+, sneg+, other+ have, other countries\n, other hpos, other Other:, other negative other, other 10+, other\nNotes, other+ sent, other than that, other+\n\n, otherAnswer\n, spos-, other: other, other+!, other - other, other hneg, other 2018 other, otherAnswer:, other\n\n, other difference other, other celebrities other, otherpos+, other_chat, other than being, other priority\n, other ER:, other future sp, other problems\n, other than not, other-ing, other:other, other wait\n, other $0, other\nEE, other future h, other organizations\n, other initial easiest, otherOther other, other+ h, other+$, otherAnswer other, other people other, spos+, other URL\n, other+ER, other Stc, other hopes\n, other+ Internet, other wrong\n
history length=3	other\nAnswer, hpos-, other\n\n, spos+, hneg-, hpos+, other: other, other\nEE, sneg+, other_other, other word choice, other\nER, other+\n, other other other, hneg+, other-in, other+:, other-other, other\nIf, other+ sp, other+ other, other+other, other\nNote, laughing\nAnswer, otherAnswer:, other+ directly, spos-, other hneg, other 0., other+ bye, other+ h, other-st, otherAnswer other, other source\n, other 0%, other than being, other+!, other+ Read, other than to, other $0, other- other, other:No, other URL\n, other+$, other+h, other-ing, other+ wonderful, other hpos, other+link, otherURL\n