ChatGPT Prompting

In the following (first) prompt we insert specific examples to train the prompting classifier in one instance of each class.

There is a total of 5 classes following this configuration:

Beneficial/Ehical	---- > class label ------ -> 1
Helpful	---- -> class label ------ -> 2
Neutral	---- -> class label ------ -> 3
Provoking	----- > class label ------ -> 4
Toxic	----- > class label ------ -> 5

Note-1: The following prompts contain the training and application of the prompting classifier, meaning that there are prompts regarding the classification of the comments.

Note-2: The repetition of the prompts is a result of the prompting limitations of a free ChatGPT user, meaning that, we were unable to classify all comments at once (single prompt).

----------FIRST PROMPT ----------:

Let's classify the following sentences based on their toxicity level {1,2,3,4,5}:

"Σε ανακάλυψα πριν 2 χρόνια όταν έψαχνα στο YouTube για κάποιο video και έπεσα πάνω στο επικο podcast (Ιλιάδα) και από τότε δεν έχω χάσει video!! Το πιο αγαπημένο όμως είναι η Οδύσσεια!!! Εξαιρετική δουλειά!! Σου εύχομαι καλή επιτυχία στο βιβλίο (το ψήφισα). Εύχομαι να συνεχίσεις και έτσι στο μέλλον!!! Συγχαρητήρια είσαι μοναδικός!!!"

+This sentence is in Greek language and expresses excitement and thankfulness. It is a highly positive comment and does not contain any level of toxicity. This is classified as beneficial/ethical: 1 .

"Μλκα ειναι αστείο αυτό το άρθρο ειδικά για εμάς που ξέρουμε τον Hayate από παλιές εποχές"

+This sentence is in Greek language and expresses unsettlement and irony in a negative way. It is a negative comment but does not contain any toxicity. This sentence is classified as provoking: 4 .

"HAYATOBASTARDO KLAPSE MOULIKO KRIPSU KOTA STH FWLIA SU 8ASU KANUN THN TRYPA SOURWTHRI STHN MPOOUZOU OSO KAINA KRYFTEIS ERXETAI PUTSAAAAAAAAA "

+This sentence is in Greeklish and it contains hate speech and is extremely toxic. It is a highly negative and highly toxic sentence. This sentence is classified as toxic: 5.

"giati tous upervarous xarakthres tous paizoun upervaroi ithopoioi"

+This sentence is in Greeklish contains a ironic way of expressing toxicity. It implies that overweight movie/video characters are also overweight in their everyday life outside the movie/video characters. This sentence is classified as toxic: 5.

"The transformation of the warrior to poet (bow and lyre) and back to warrior to save his society starting with Antinoos (names are very important). A blue print we should follow to save our country 4,000 yrs later."

+This sentence is in English and can be considered neutral in terms of toxicity and it tries to explain a historic fact which is related to the theme and channel of the respective YouTube video. This sentence is classified as neutral: 3.

"Κρίμα να τελειώσει το επικό podcast θα μπορούσε να συνεχιστεί και με άλλα Έπη όπως η Ελένη του Ευρυπίδη"

+This sentence is in Greek and appears to express remorse in a positive manner. It implies that the author of the comment experiences sadness to let this podcast end. This sentence is classified as helpful: 2

----------SECOND PROMPT ----------:

In this prompt we are utilizing the training of the first prompt to classify 80 or so comments which is a chunk of the crawled YouTube comments originating from the crawl.csv file

The actual prompt:

For our next step I am going to give you a series of comments from YouTube posts, and you should classify each one of them to at most one of the 5 classes I trained you for.

You should present your results in a Python dictionary. For example, if you want to classify the sentence with indexing 0 with the label Toxic: 5 then you should right 0: 5 inside the dictionary.

Also, the Python dictionary must be readable and have 20 classifications in each row parallel to each other.

Here are some of the comments: *### Here the user pastes the comments of interest.*

----------THIRD PROMPT ----------:

At this point our prompting classifier has been trained and the actual chat has understood our needs of classifying comments in a toxicity scale.

But it should be noted that if we do not specify manually the descriptions of the classes then ChatGPT will at some point assign different classes. This is based on my prompting experience for this task/assignment.

The actual prompt:

Now, you should classify the following YouTube comments in at most one of the five classes.

Reminder the classes are:
Beneficial/ethical: 1,

Helpful: 2,
Neutral: 3,
Provoking: 4,
Toxic: 5

You should present your results in Python dictionary. For example if you want to classify the sentence with indexing 0 with the label Toxic : 5 then you should right 0: 5 inside the dictionary.

Also the Python dictionary must be readable and have 20 classification in each row pararel to each other.

Here are some of the comments: *### Here the user pastes the comments of interest.*

----------FOURTH PROMPT ----------:

At this point there is a repetition and justifiably so, considering that the classification task remains the same throughout these prompts.

The actual prompt:

Next, classify the following YouTube comments as you did previously.

Reminder the classes are:
Beneficial/ethical: 1,
Helpful: 2,
Neutral: 3,
Provoking: 4,
Toxic: 5

You should present your results in Python dictionary. For example if you want to classify the sentence with indexing 0 with the label Toxic : 5 then you should right 0: 5 inside the dictionary.

Also the Python dictionary must be readable and have 20 classification in each row pararel to each other.

Here are some of the comments: *### Here the user pastes the comments of interest.*