A Details on Preparation for Style Datasets in XSLUE

Formality. Appropriately choosing the right formality in the situation is the key aspect for effective communication (Heylighen and Dewaele, 1999). We use GYAFC dataset (Rao and Tetreault, 2018) that includes both formal and informal text collected from the web.[7]

Humor. Humor (or joke) is a social style to make the conversation more smooth or make a break (Rodrigo and de Oliveira; Kiddon and Brun, 2011). We use the two datasets widely used in humor detection: ShortHumor (CrowdTruth, 2016) and ShortJoke (Moudgil, 2017), where both are scraped from several websites. We randomly sample negative (i.e., non-humorous) sentences from random text from Reddit corpus (Jung et al., 2019) and literal text from Reddit corpus (Khodak et al., 2017).

Politeness. Encoding (im)politeness in conversation often plays different roles of social interactions such as for power dynamics at workplaces, decisive factor, and strategic use of it in social context (Chilton, 1990; Clark and Schunk, 1980). We use Stanford's politeness dataset StanfPolite (Danescu et al., 2013) that includes request types of polite and impolite text scraped from Stack Exchange question-answer community.

Sarcasm. Sarcasm acts by using words that mean something other than what you want to say, to insult someone, show irritation, or simply be funny. We choose two of them for XSLUE: SarcGhosh (Ghosh and Veale, 2016) and SARC v2.0 (Khodak et al., 2017)[8]. For SARC, we use the same preprocessing scheme in Ilić et al. (2018).

Metaphor. Metaphor is a figurative language that describes an object or an action by applying it to which is not applicable. We use two benchmark datasets:[9] Trope Finder (TroFi) (Birke and Sarkar, 2006) and VU Amsterdam VUA Corpus (Steen, 2010) where metaphoric text is annotated by human annotators.

Offense. Hate speech is a speech that targets disadvantaged social groups based on group characteristics (e.g., race, gender, sexual orientation) in a manner that is potentially harmful to them (Jacobs et al., 1998; Walker, 1994). We use the HateOffenssive dataset (Davidson et al., 2017) which includes hate text (7%), offensive text (76%), and none of them (17%).

Romance. Since we could not find any dataset containing romantic texts, we crawl and preprocess them from eleven different web sites, then make a new dataset ShortRomance. We make the same number of negative samples from the literal Reddit sentences (Khodak et al., 2017) as the romantic text. ShortRomance text are crawled from the following websites. The copyright of the messages are owned by the original writer of the websites.

- http://www.goodmorningtextmessages.com/2013/06/romantic-text-messages-for-her.html
- https://www.travelandleisure.com/travel-tips/romantic-love-messages-for-him-and-her
- https://www.amoramargo.com/en/sweet-text-messages-for-her/
- https://www.techjunkie.com/best-romantic-text-messages-for-girlfriend/
- https://liveboldandbloom.com/10/relationships/love-messages-for-wife
- https://www.marriagefamilystrong.com/sweet-love-text-messages/
- https://pairedlife.com/love/love-messages-for-him-and-her
- https://truelovewords.com/sweet-love-text-messages-for-him/
- https://www.serenataflowers.com/pollennation/love-text-messages/
- https://www.greetingcardpoet.com/73-love-text-messages/
- https://www.wishesmsg.com/heart-touching-love-messages/

Sentiment. Identifying sentiment polarity of opinion is challenging because of its implicit and explicit presence in text (Kim and Hovy, 2004; Pang et al., 2008). We use the annotated sentiment corpus on movie reviews; Sentiment Tree Bank (Socher et al., 2013) (SentiBank).

Emotion. Emotion is more fine-grained modeling of sentiment. (degree of control). We use two datasets: DailyDialog (Li et al., 2017) from the Ekman (1992)'s six categorical model, and EmoBank (Buechel and Hahn, 2017) from (Warriner et al., 2013)'s three-dimensional VAD

[7] The dataset requires an individual authorization for access, so we only provide a script for preprocessing.

[8] SARC_{pol} is a sub-task for the text from politics subreddit.

[9] we did not include Mohler et al. (2016)'s dataset because the labels are not obtained from human annotators.

model. We also include a large but noisy emotion-annotated corpus `CrowdFlower` (CrowdFlower, 2016), including in addition to Ekman's categories, additional seven categories: enthusiasm, worry, love, fun, hate, relief, and boredom.

Persona. Persona is a pragmatics style in group characteristics of the speaker. We use the stylistic language dataset written in parallel called `PASTEL` (Kang et al., 2019) where multiple types of the author's personas are given in conjunction. `PASTEL` has six different persona styles (i.e., age, gender, political view, ethnicity, country, education) where each has multiple attributes.

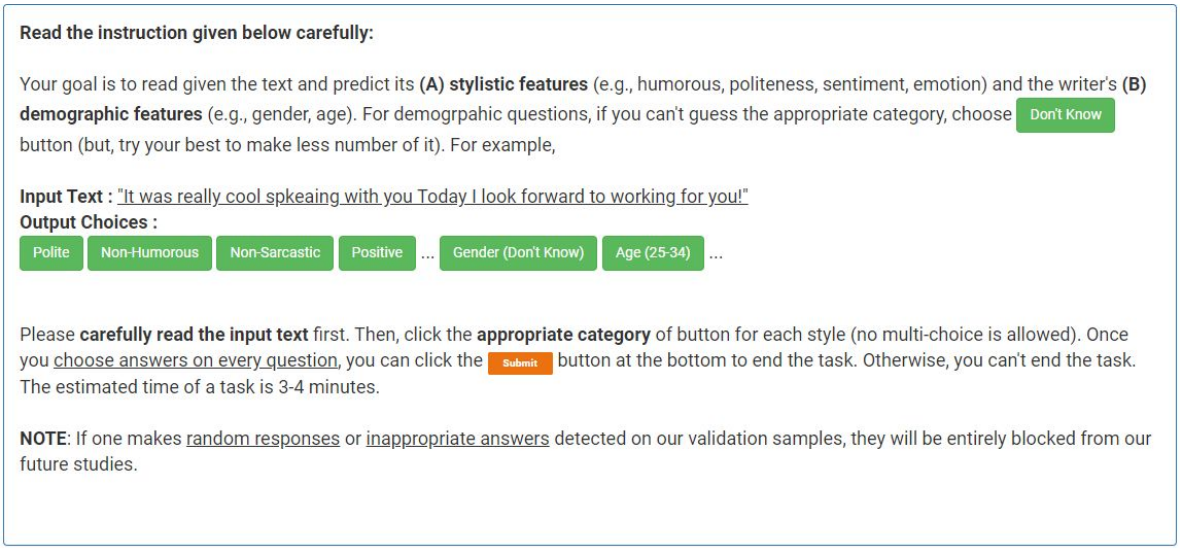
B Details on Annotation Schemes

Figure 6 shows snapshots of our annotation platform with the detailed instructions. We estimate the execution time of a task as 4 minutes, so paying $\$9/(60 minutes/3 minutes) = \$0.4$ per task. We make 10 size of batches multiple times and incrementally increase the size of batches up to 400 samples. For each batch, we manually checked the quality of outputs and blocked some bad users who abusively answered the questions.

C Details on Hyper-Parameters

For our BERT classifier, we use the uncased BERT English model. We did parameter sweeping on learning rates [2,5e-5,4] and batch sizes [32,16,8] on the validation set. For the BiLSTM baseline, we use 32[10] sizes of batching for both training and testing and 256 hidden size for LSTM layer with 300 sizes of word embedding from GloVe (Pennington et al., 2014). The vocabulary size of BiLSTM is the same as the maximum vocabulary of the BERT model; 30522. For both BERT and BiLSTM models, we use the same maximum input length 128. Both training use $2e-5$ learning rate and 1.0 maximum gradient clipping with Adam optimizer with $1e-8$ epsilon. Also, we use early stopping until the maximum training epochs of 5.

[10] 32 batch size shows slightly better performance than smaller sizes like 8 or 16.



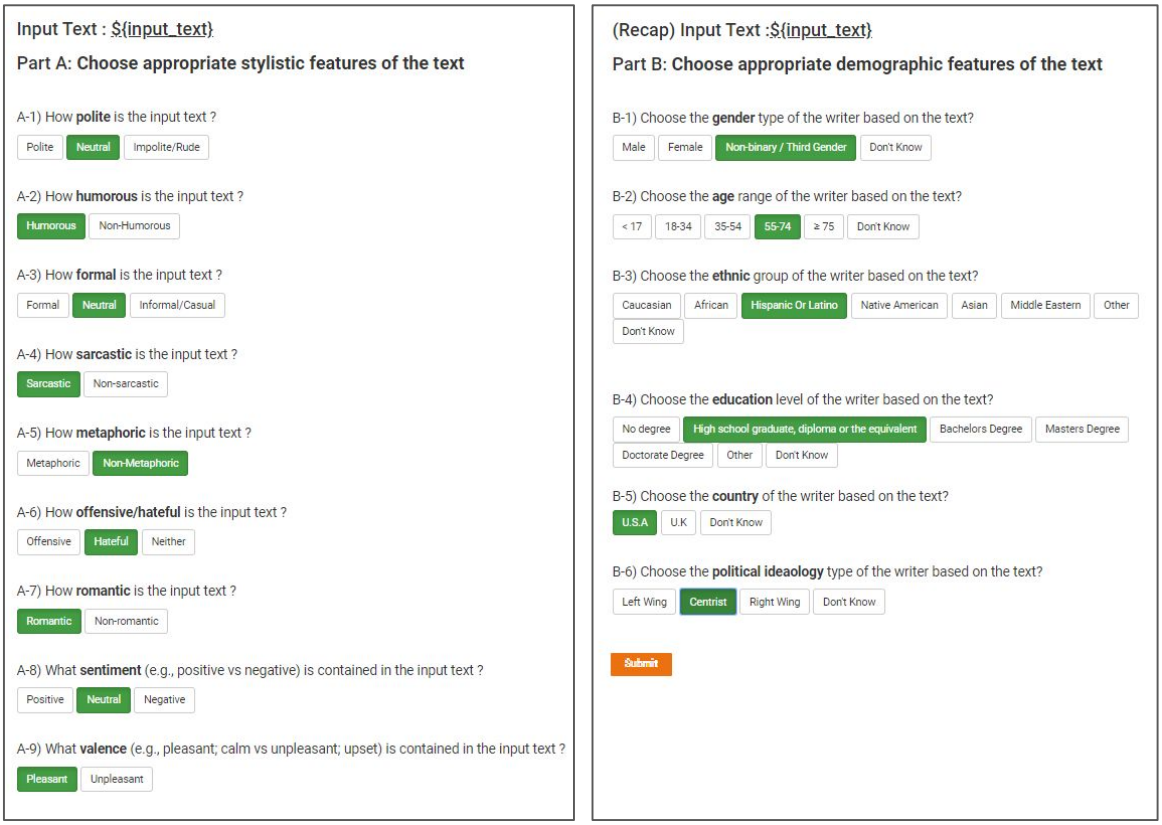


Figure 6: Snapshots of our annotation tasks: general instruction (top) and annotation tasks on each style (bottom).