**Reviewer #1:**

R1Q1: We split each conversation into utterances and provided four annotators with one utterance at a time when performing sentence-level annotation.

R1Q2: We have introduced the Kappa metric, and obtained the average κ = 0.586 (the highest score is 0.74, the lowest score is 0.48). We will report it in the paper.

R1Q3: Figure 4 illustrates that there exist different types of influences between speakers in three different interaction patterns, and different affective states have different influences. Specifically, Influence Matrixes 1, 2 and 3 describe the influences existing in the “Question&Answer”, “Offering&Response” and “Greeting&Greeting” scenarios respectively. The learned influence matrices can be used to simulate the interactions between speakers. We will add the above explanation in the paper.

R1Q4: Since we have obtained the predicted and annotated labels of all utterances, we use them to show the sentimental change of two speakers during the conversation, as shown in Figure 6.

R1Q5: In order to achieve more convincing results by using the dataset, we do not provide a standardized training set/testing set in current version, and we encourage the dataset users to perform cross validation to assessing their models.

R1Q6: Yes, all utterances in one conversation share the same topic.

R1Q7: ScenarioSA does not contain explicitly defined aspects, so that we have to choose aspects by ourselves. Since tf-idf could reflect how important a word is to a sentence, it is natural to take the words that have the largest tf-idf values as aspects.

R1Q8: Yes, we plotted the sentimental change of speaker B in one single conversation (referring to Table 1).

R1Q9: Yes, we calculate the proportion of the sentiment labels based on the final affective states of speakers A, B. We’ll replace Figure 5 with a clearer picture in the revised version.

**Reviewer #2**

R2Q1: Bhaskhar’s work involved emotion classification of audio conversations, which aimed at estimating the speaker’s emotions. In this regard, it is related to our task, even though it used multimodal features.

R2Q2: In this paper, the notion of “interactivity” is equivalent to “interaction effect”. When revolving around a certain topic, we think that two speakers could strongly influence each other, producing a clear interaction effect (we have shown this inference in Sec 4.2). As for multi-topic conversations, we admit that they also imply interaction effect. However, exploring which type of conversation produces more interaction effect is beyond the scope of this work. We just consider that our conversations could produce significant interaction effect. Compared with NPS Chat Corpus, we focus on showing interactions between speakers via their sentiments.

R2Q3: That means a speaker might not always expresses his or her opinions logically and coherently. One speaker’s opinions may be largely based on another speaker’s responses/answers.

R2Q4: Our conversations were collected from a series of sites, including eslfast.com, focusenglish.com, dailyenglishconversation.com. The conversations might take place among people of different types of relations, such as strangers, friends, workmates, families, etc.

R2Q5: No, they are not TV show, they are real conversation. The source owners have declared its truthfulness. I have replaced “Connie” with “NAME”.

R2Q6: In such case, we always discuss and determine the label together with the annotators. Moreover, we give higher weights to the opinions of the annotators who have lower noise levels.

R2Q7: The final sentiment polarity of each speaker does not necessarily equal to the sentiment label of the last turn. This is because in some conversations, such as seeing a doctor, the patient always says thanks to the doctor in last turn, but s/he may still feel bad in the whole conversation. After a careful calculation, there are 578 (26.11%) conversations whose final sentiment labels are different from the sentiment labels of the last turn.

R2Q8: We have introduced Kappa metric, and obtained the average κ = 0.586 (the highest score is 0.74, the lowest score is 0.48).

R2Q9: The noise level reflects the reliability of annotators. Thus when we were determining the gold standard label, we would value the opinions of annotators who have lower noise levels more than the ones who have higher noise levels.

R2Q10: Figure 4 illustrates that there exist different types of influences between speakers in three different interaction patterns, and different affective states have different influences. Specifically, Influence Matrixes 1, 2 and 3 describe the influences existing in the “Question&Answer”, “Offering&Response” and “Greeting&Greeting” scenarios respectively. The learned influence matrices can be used to simulate the interactions between speakers. We will add the above explanation in the paper.

R2Q11: These four metrics are widely used in sentiment analysis tasks. They reflect different aspects of the classification performance. We keep the results for all of them, in order to provide a more comprehensive and balanced evaluation. This is important to demonstrate the usefulness and applicability of the proposed new dataset.

R2Q12: We used GloVe word vectors.

**Reviewer #3:**

R3Q1: We’ll add some references in the revised version of the paper.

R3Q2: In the current version of the dataset, we focus on on-line conversations and all conversations were collected from online sites. However, we agree that interactive sentiment analysis should apply to the sentiment polarity of any kind of conversation.

R3Q3: We simply meant that our dataset covers a range of real-world scenarios and conversation styles, which is important to ensure unbiased evaluation with this dataset. Our dataset covers about 13 scenarios. We think this is a reasonable number.

R3Q4: We’ll add comments in the revised version of the paper.

R3Q5: Because LSTM and IAN have already been applied to many traditional sentiment analysis tasks, such as, product reviews, twitter posts, etc., and produced impressive results. However, our experiments show that they do not get good results in interactive sentiment analysis task. So, it is sensible to assume that these challenges imposed by the interactive sentiment analysis affected the performance.

R3Q6: This paper focus on presenting an interactive sentiment dataset, demonstrating the need of novel sentiment analysis models and the potential of the dataset to facilitate the development of such models, rather than solving all the questions.

R3Q7: There are two reasons. First, almost all sentiment datasets used 2 labels (positive, negative) or 3 labels (positive, neutral, negative). Second, sentiment is subtle and abstract, so we think distinguishing positive/neutral/negative is more realistic than distinguishing finer graded multi-dimensional emotions or simply using binary sentiment labels (positive/negative).

R3Q8: In the current version, we obtained the final labels through fusing all labels of utterances by simple summation and weighted combination approach. We will add a clarification after this line. Actually we are working on a novel mathematical method for this purpose.

R3Q9: The selected sites are all about daily English conversation. We use three rules to select sites, which have clear design, and are convenient to crawl conversations.

R3Q10: we think that 50+ turns are too disperse to capture the main topic. 20-30 turns are reasonable number.

R3Q11: In sociology, interaction is an exchange between two or more individuals, and is the processes by which we act and react to those around us. We will add the explanation in the revised version.

R3Q12: Because in multi-party conversations, each participator’s opinions may be largely based on all other speakers’ responses/answers.

R3Q15: Each utterance is labelled with its sentiment label. We split each conversation into utterances and provided four annotators with one utterance at a time when performing sentence-level annotation.

R3Q16: We’ll make a table to show data that supports these percentages, and add a reference.

R3Q17: Top 10 words are adequate to help us infer the topics; 5 to 20 topics are a reasonable interval to cluster 2214 conversations; perplexity is a measurement of how well a probability model predicts a sample, which is helpful (when using LDA) to get the appropriate number of topics in a corpus (please refers to Blei’s work). Generally, the lower the perplexity, the better.

R3Q18: HMM means Hidden Markov model.

R3Q25: Because different speakers and different affective states have different influences. We want to learn influences via our data. Due to limitation of space, we could not detail “the influence model”, so we provide the reference, which explained how it works and how to use it.

R3Q26: the states 1,2...S denote the affective states -1, 0, 1; the o^e\_t denotes the annotated labels.

R3Q27: We’ll replace Figure 5 with a clearer picture, and rewrite section 4.3 in the revised version of the paper.

R3Q31: Figure 4 illustrates that there exist different types of influences between speakers in three different interaction patterns, and different affective states have different influences. Specifically, Influence Matrixes 1, 2 and 3 describe the influences existing in the “Question&Answer”, “Offering&Response” and “Greeting&Greeting” scenarios respectively. The learned influence matrices can be used to simulate the interactions between speakers. We will add the above explanation in the paper.

R3Q32: We’ll add relevant references and correct grammatical errors in the revised version of the paper.