IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2020, accepted May 1, 2020, date of publication May 12, 2020, date of current version May 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994147

ScenarioSA: A Dyadic Conversational Database for Interactive Sentiment Analysis

YAZHOU ZHANG[1], ZHIPENG ZHAO[2,3], PANPAN WANG[4], XIANG LI[5], LU RONG[6], AND DAWEI SONG[7]

[1] Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[2] Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Minister of Education, Zhengzhou 450001, China
[3] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China
[4] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[5] Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, China
[6] Personnel Department, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[7] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

Corresponding author: Zhipeng Zhao (zzpeng@tju.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0831704, in part by the Natural Science Foundation of China under Grant U1636203, Grant 61772363, Grant U1736103, Grant U1504608, and Grant 61802352, in part by the Major Project of Zhejiang Laboratory under Grant 2019DH0ZX01, in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie under Grant 721321, in part by the Foundation and Cutting-Edge Technologies Research Program of Henan Province under Grant 192102210294, in part by the Project of Science and Technology, Henan, under Grant 202102210178, and in part by the Industrial Science and Technology Research Project of Henan Province under Grant 202102210387.

ABSTRACT Interactive sentiment analysis is an emerging, yet challenging, subtask of the natural language processing problem. It aims to discover the affective state and sentimental change of each person in a conversation, and has attracted an increasing attention from both academia and industry. Existing sentiment analysis approaches are insufficient in modelling the interactions among people. However, the development of new approaches are critically limited by the lack of labelled interactive sentiment datasets. In this paper, we present a new conversational database that we have created and made publicly available, namely ScenarioSA, for interactive sentiment analysis. We manually label 2,214 multi-turn English conversations collected from various websites that provide online communication services. In comparison with existing sentiment datasets, ScenarioSA (1) is no longer limited to one specific domain but covers a wide range of topics and scenarios; (2) describes the interactions between two speakers of each conversation; and (3) reflects the sentimental evolution of each speaker over the course of a conversation. Finally, we propose an extension of interactive attention networks that could model the interactions, and compare various strong sentiment analysis algorithms on ScenarioSA, demonstrating the need of novel interactive sentiment analysis models and the potential of ScenarioSA to facilitate the development of such models.

INDEX TERMS Emotion recognition, natural language processing, opinion mining, sentiment analysis.

I. INTRODUCTION

Sentiment Analysis (SA) has been a core research topic in Natural Language Processing (NLP). It aims at discovering diversified subjective information implied in the given natural language text [1]. Most existing sentiment analysis approaches focus on identifying the polarity of commentaries or similar type of texts (i.e. movie reviews, product reviews and twitter posts) at the document-, sentence- or aspect- levels [2], [3]. The commentary documents used in these studies are in the form of individual narratives, without involving interactions among the writers or speakers.

Along with the rapid development of WWW and social networking services, such as WhatsAPP, Wechat and Twitter, instant messaging has been a popular means of communication among people. As a result, a large volume of interactive texts have been produced, which carry rich subjective information [4]. Recognizing the polarity of the interactive texts and its evolution with respect to people's interaction is of a great theoretical and practical significance. Hence, interactive

The associate editor coordinating the review of this manuscript and approving it for publication was Mario Luca Bernardi.

TABLE 1. An example in ScenarioSA exhibiting the interactions between A and B, the sentimental change, the affective states, and the jumpings in logic of A, B, where 1=positive, −1=negative, 0=neutral. Note that the final sentiment label of each speaker does not necessarily equal to the sentiment label of the last turn.

A : Hi B. What are you doing? [0]
B : Hi A. I'm planning a birthday party for NAME. [0]
A : How is that going? [0]
B : Not well. I can't think of anything to do. My idea is a mess. [-1]
A : How many kids do you want to invite? [0]
B : He wants to invite all of the boys in his classroom. That's 12 boys. [0]
A : No girls? [0]
B : He doesn't want to invite any girls. He doesn't like playing with girls at all. [-1]
A : How about an outdoor party? [0]
B : We did that last year and it rained. At last, we were not happy. [-1]
A : Oh! That's not fun. [-1]
B : I was thinking about taking the kids to a pizza place. [0]
A : Kids like pizza. How about taking them to a movie theater, too? [1]
B : I don't like this idea. That's too expensive. [-1]
A : Yeah. How about renting a movie and watching it at home after the pizza place? [0]
B : I love that idea! Then they can play in our backyard if it doesn't rain. [1]
A : That sounds like a great party. [1]
B : Yes. Hey, would you like to join us? [0]
A : Sounds great! I will be there. [1]
B : Ok. Thanks for your help. See you later. [1]
A : See you. [1]
The final affective state: **A** : [1] **B** : [1]

sentiment analysis is becoming a new research direction, and has attracted an increasing attention from both academia and industry.

Interactive sentiment analysis aims to detect the affective states of multiple speakers during and after an online conversation, and study the sentimental change of each speaker in the course of the interaction. Compared with the traditional sentiment analysis, which only focuses on identifying the polarity of independent individual, this goal is challenging for three reasons: (1) in the interactive activities, the attitude of each participant is influenced by other participants and changes dynamically; (2) the interactions among people hide a wealth of information, such as their social relationships, environments and cultural backgrounds; (3) there could be jumpings in speakers' logical flow in the course of the interaction, which is different from an individual personal narrative, in which each human expresses his or her opinions logically and coherently [5]. Table 1 shows an example of these phenomena.

However, the lack of publicly available interactive sentiment datasets has been a bottleneck for advancing interactive sentiment analysis models. Because an interactive sentiment dataset is an indispensable element for benchmarking systems and assessing the quality and potential of interactive sentiment analysis algorithms. Tian *et al.* [4], [6], [7] built a Chinese interactive corpus, which was collected from a student community service, aiming to solve the problem of emotional illiteracy in e-Learning services. But this corpus was not described in detail. Ojamaa *et al.* [8] used a lexicon based technology on the conversational texts to extract the speaker's attitude. Their dataset only included 23 dialogue files, which were not suitable for machine learning based assessments. Bhaskar *et al.* [9] proposed to combine both acoustic and textual features for emotion classification of audio conversations. The dataset was an audio-visual database, rather than text dataset. Due to the limited availability of sentiment-annotated interactive text dataset, Bothe *et al.* [10] had to use the Vader sentiment analysis tool [11] to auto-annotate the sentiment labels of two spoken interaction corpora for training their model. Chen *et al.* [12] created an emotion corpus of multi-party conversations. Based on their work, Poria *et al.* [13] proposed a multimodal emotionlines dataset (MELD) to show the importance of contextual and multimodal information for emotion recognition in conversations. Both of their datasets were collected from TV scripts. Their datasets only contained the sentiment label of each utterance, while our dataset annotated both the labels of each utterance and that of the overall conversation.

To fill the gap, we present ScenarioSA,[1] a high-quality English conversational dataset with sentiment labels, for interactive sentiment analysis. The dataset contains 2,214 multi-turn conversations, altogether over 24K utterances. There are two speakers, anonymized as **A** and **B** in each conversation. Each utterance is manually labelled with its corresponding sentiment polarity: positive, negative or neutral. The final affective state of each participant is also labelled when the conversation ends. Additionally, the conversations in ScenarioSA cover more grounded and natural contexts, such as shopping, college life, work, etc. The advantages of ScenarioSA over existing sentiment datasets can be summarized as follows: (1) broad coverage in various scenarios and conversation styles; (2) unlike existing sentiment datasets, ScenarioSA depicts the interactions between two speakers of each conversation and reflects the sentimental evolution of each speaker over the course of a conversation; (3) ScenarioSA introduces a new requirement for future sentiment analysis models: They should be able to identify the sentiment polarities of each utterance and of each speaker at the end of a conversation. A comparison of ScenarioSA with several existing sentiment and dialogue datasets is shown in Table 2.

Finally, we design a comparative experiment on ScenarioSA over a number of typical sentiment analysis models, including a lexicon-based approach (which is SentiStrength), two machine learning based algorithms (which are support vector machine and joint sentiment/topic model), four popular neural network approaches: a convolutional neural

[1] The dataset is available on: https://github.com/yazhouzhang2008/InteractiveSentimentDataset.

TABLE 2. Comparison of ScenarioSA with other databases.

Dataset	Scenario	Large scale	Interactivity	Sentiment label	Publicly available
ScenarioSA	Web chat	√	√	√	√
EmotionLines [12]	TV show	√	√	√	√
MELD [13]	TV show	√	√	√	√
ICTs [4]	Community	√	√	√	×
Movie Reviews [14]	Community	√	×	√	√
the SemEval-2017 [15]	Real-world	√	×	√	√
eNTERFACE'05 emotion [9]	Inducible video	√	×	√	√
First-encounter dialogues [8]	Inducible video	×	√	√	×
CHILDES [10]	Real-world	√	√	×	√
DailyDialog [16]	Real-world	√	√	√	√
Cornell Movie-Dialogues [17]	Movie	√	√	×	√

network (CNN), two long short-term memory (LSTM) networks, an interactive attention networks (IAN), and an improved interactive attention networks with influence (IAN-INF) that incorporates three learned influence matrices into the output gate of each LSTM unit for obtaining hidden states of words. Out results show that the IAN and AT-LSTM models perform the best among all baselines, but only achieve an accuracy of 63.3% on A, 72.8% on B and 61.5% on A, 72.3% on B, in comparison to 78.6% and 83.1% on SemEval 2014 [18], due to the challenges listed above. Through considering social influence, IAN-INF achieves better classification accuracy results, which are 64.2% on A, 72.4% on B. This indicates that the existing approaches cannot effectively model the evolution of sentiment in interactive conversations and new methodologies are required.

The major contributions of the work presented in this paper are summarized as follows.

- We create a manually labelled large scale conversational dataset, ScenarioSA, for conversational sentiment analysis.
- There are more than 24,000 utterances in ScenarioSA, which makes our dataset suitable for machine/deep learning based assessments.
- ScenarioSA not only indicates the final affective state of each speaker, but also reflects the sentimental change of each speaker in the conversation.
- We evaluate several typical sentiment analysis methods over the ScenarioSA collection, showing that our dataset would facilitate the development of future sentiment analysis models.

The rest of this paper is organized as follows. Section 2 gives a brief formulation of conversational sentiment analysis problem. Section 3 gives a brief introduction to our ScenarioSA. In Section 4, we conduct a detailed analysis of the ScenarioSA dataset, and describe the interactions between speakers. In Section 5, we report and analyze the empirical experiments. Section 6 concludes the paper and points out a number of future research directions.

II. PROBLEM FORMULATION

The granularity of sentiment analysis can range from conversation, document, sentence to aspect levels. In this work, we target determining the attitude of each speaker at the utterance and conversation levels, in terms of positive (expressing positive sentiment), negative (expressing negative sentiment) and neutral (expressing unbiased sentiment or not expressing any sentiment), but not bipolar (expressing both positive and negative sentiment in one sentence). We assume that humans prevailingly express only one main sentiment polarity [19]. When both positive and negative sentiment are implied in a sentence, we believe that the author always leans towards one of the two.

Moreover, we associate the speakers' final affective states with the polarities of utterances, allowing for finer-grained sentiment analysis, such as at sentence (utterance) level. Similarly, we also assume that the results of sentence level sentiment analysis can be aggregated in effective ways to obtain high-level statistics: conversation-level sentiment analysis.

With the aforementioned assumptions, we formulate the problem as follows: *Given a multi-turn conversation between speakers written in English, how to determine whether this conversation carries subjective information and if it does, what is the sentiment polarity of the conversation, then how to depict the sentimental change of each speaker in the conversation?*

Now, referring to the example shown in Table 1, we formulate each conversation in our ScenarioSA as a 4-tuple $\left\{\left(s_i, u_{ij}, l_{ij}, p_i\right) | i = A, B; j = 1, 2, \ldots, n\right\}$. Specifically, n is the number of speaker turns (n = 11 in Table 1). s_i denotes the speakers **A**, **B** in the conversation. u_{ij} is the j-th utterance expressed by the speaker s_i, for example, u_{A2} represents the sentence: "How is that going?". l_{ij} is the corresponding sentiment label of u_{ij}, $l_{ij} \in [-1, 0, 1]$, denoting negative, neutral, positive sentiment respectively. In Table 1, for example, $l_{A2} = 0$. p_i represents the final polarity of the speaker s_i, $p_i \in [-1, 0, 1]$, e.g., in Table 1, we have $p_A = 1$, $p_B = 1$. We take the speaker information s_i and the utterance u_{ij} as input and produce its sentiment label l_{ij} as intermediate result, then infer the final polarity indication of each speaker p_i as output.

III. DATASET CONSTRUCTION

In this section, we describe the process of creating our ScenarioSA dataset and analyze its basic features.

TABLE 3. The ScenarioSA dataset statistics.

	Dataset Statistics
Total Conversations	2,214
Total Utterances	24,072
Total Words	228,047
Average Turns Per Conversation	5.9
Average Words Per Conversation	103.0
Average Words Per Utterances	9.5

A. DATA COLLECTION & PRE-PROCESSING

Our goal is to construct a large scale sentiment dataset to support the interactive sentiment analysis task. First, we crawl over 3,000 multi-turn English conversations from several websites that support online communication.[2] The conversations are collected in the various daily life contexts and cover a wide range of topics, such as shopping, work, travel, and food. More details of the topics will be introduced in Sec.IV-A. Each conversation is human written and thus is more formal than the transcribed text from a spoken corpus such as the First-encounter dialogue [8].

Since each conversation revolves around a certain topic, it usually ends after a reasonable number of turns (less than 25 turns in our ScenarioSA). The crawled conversations are clearly distinguishable from other dialogue datasets such as Cornell Movie-Dialogues Corpus [17] and The NPS Chat Corpus [20].

All the conversations are then pre-processed. Some of the crawled conversations involve three or more participants. We think the conversation among multiple speakers will exacerbate the jumpings in logic of each speaker. In this work, we prefer studying the interactions between two speakers, and thus discard those involving three or more speakers. Further, for sake of privacy protection, we replace the first speaker's name with **A**, the second with **B**, and replace others' names mentioned in the conversation with **NAME**. We also correct the spelling mistakes automatically, and check if each conversation is composed of illegible characters.

After pre-processing, the ScenarioSA dataset contains 2,214 multi-turn conversations, altogether 24,072 utterances and 228,047 word occurences. The average speaker turns and average number of words per conversation is about 6 (turns) and 103 (words), respectively. The detailed statistics are shown in Table 3.

B. ANNOTATION CRITERIA AND PROCEDURE

The pre-processed dataset is manually annotated with three labels: -1, 0, 1. In order to guarantee the annotation quality, we recruited five volunteers. They are all fluent in English, and have a good knowledge in sentiment analysis. Before labelling the whole dataset, they were instructed to independently annotated 100 examples first, with the aim to minimise ambiguity while strengthen the inter-annotator agreement. We define the gold standard of a utterance or conversation in terms of the label that receives the majority votes. The annotation procedure consists of two steps:

1) SENTENCE (UTTERANCE)-LEVEL ANNOTATION

As we are interested in detecting sentimental change of each speaker, the annotators were first asked to mark up each sentence with one of the following three sentiment labels: -1, 0, 1.

2) CONVERSATION-LEVEL ANNOTATION

In the second step, the annotators were instructed to tag whether each speaker expresses positive, negative or neutral opinion at the end of the conversation. The motivation of adding this tag comes from our interest in developing a classification model to detect the affective state of each speaker of the conversation.

Note that the final sentiment label of each speaker does not necessarily equal to the sentiment label of the last turn. Because in a few conversations, such as seeing a doctor, the patient always shows thanks to the doctor in last turn, but he still feels bad in the whole conversation. After calculation, there are 578 (26.11%) conversations whose final sentiment labels are different from the sentiment labels of the last turn.

C. AGREEMENT STUDY

After annotating the whole dataset, we assess the reliability of our sentiment annotation procedure, through an agreement study.

1) AGREEMENT ASSESSMENT

we first use the percent agreement calculation method to calculate the average agreement. At the conversation level, the average agreement among five annotators on three sentiment labels is about 78.6%. At the utterance level, the average agreement is about 73.2%. Specifically, for the task of determining whether a conversation is subjective (i.e., positive and negative) or objective (neutral), the average agreement is 85.6%. For the task of determining whether a conversation is positive or negative, the average agreement is 81.9%.

Moreover, we also introduce Kappa metric [21], which is generally thought to be a another robust measure except the aforementioned percent agreement calculation, to verify inter-rater agreement for the dataset. The definition of κ is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_0 is the relative observed agreement among annotators, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the annotators are in complete agreement then $\kappa = 1$. If there is no agreement among the annotators, then $\kappa = 0$. Applying our formula for Kappa metric, we get $\kappa = 0.57$.

[2] Note that the original copyright of all the conversations belongs to the source owners, and the dataset is only for research purposes and cannot be used for any commercial purposes.

TABLE 4. The probability distribution of the annotators' judgments on sentiment polarity.

%	Positive	Neutral	Negative
Positive	87.0	6.7	6.3
Neutral	20.2	61.6	18.2
Negative	8.9	17.9	73.2

To examine which pair of labels is the most difficult to distinguish, Table 4 summarizes the probability distribution of all annotators' judgments. Each row describes the probabilities of a finally assigned label being annotated as other labels (including itself). For example, we finally assign a −1 label to one utterance as the gold standard. However, when annotating this utterance, different annotators may give different labels (−1,0,1 are likely candidates). We check and calculate the disagreement among annotators. We can observe that it is most difficult to determine whether a conversation is "neutral" or not. Further, a "negative" conversation can sometimes be confused with "neutral".

2) ANNOTATOR-LEVEL NOISE

given the disagreement among annotators, we aim to evaluate the accumulated noise level introduced by each of five annotators. Similar to the work [22], we also associate the noise level with the deviation from the gold standard labels. The noise level of each annotators $noise\,(anno_i)$ is estimated through accumulating the deviation frequency of the annotations received from this annotator. Statistical results show that there does exist one annotator (i.e., his noise level is 31.9%) who yields more noisy annotations than the others (whose noise levels are 10.6%, 13.4%, 16.1% and 18.3%). The noise level reflects the reliability of annotators. When performing annotation, we give higher weights to the opinions of the annotators who have lower noise levels.

IV. DATASET ANALYSIS

In this section, a detailed analysis of the ScenarioSA dataset is conducted, which shows that our dataset stands out in the following aspects.

- ScenarioSA covers a wide range of topics.
- ScenarioSA describes the interactions between two speakers, and exhibits how one speaker influences another.
- ScenarioSA not only indicates the final affective state of each speaker, but also reflects the sentimental change of each speaker in the conversation.

A. TOPIC ANALYSIS

The topic information is critical for identifying the sentiment polarity of the conversation. For example, when it comes to accidents or crime, the speakers usually express negative attitude. When it comes to romance or travel, the speakers usually feel good. The conversations in ScenarioSA, as described above, are collected from several websites. When we were crawling the raw data, we observed a fact that some conversations have been labelled with their topics, while the others have not. Therefore, we investigated our dataset carefully, and empirically estimated the number of topics as around 15, which in our belief, are enough to illustrate the diversity of ScenarioSA.

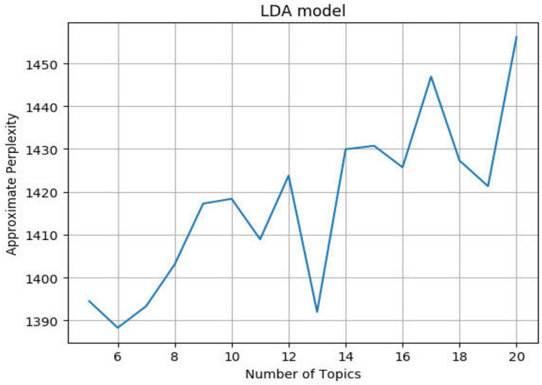


FIGURE 1. The perplexity results on ScenarioSA.

Then we determine the topics based on those existing labeled topics and newly extracted ones using topic modelling. As one of the most popular topic models, latent dirichlet allocation (LDA) is a probabilistic model with interpretable topics. Hence, in this work, we choose LDA model [23], which allows sets of observations to be explained by unobserved sets, to mine the hidden topics in a collection of conversations. In LDA, each document is viewed as a mixture of topics, and each topic is viewed as a mixture of words. Based on the assumption, the LDA model formulates a joint probability distribution to describe the generative process:

$$p\left(z_m, w_m, \theta_m \Phi|\alpha, \beta\right) = \prod_{n}^{N_m} p\left(w_{m,n}|\phi_{z_{m,n}}\right) \cdot p\left(z_{m,n}|\theta_m\right) p\left(\theta_m|\alpha\right) p\left(\Phi|\beta\right) \quad (2)$$

where $z_{m,n}$ is the topic for the n-th word in a document m, $w_{m,n}$ is the specific word, θ_m is the topic distribution for the document m, Φ is the word distribution for topic, α, β are the parameters of the Dirichlet prior.

We consider each conversation as a document, and select the top 10 words associated with each topic. We pick 5 to 20 topics to train the corresponding models, respectively, and compute the perplexity of each model. The perplexity results are shown in Figure 1. We can determine the number of topics in ScenarioSA as 13, since it has a second lowest perplexity. The derived topics from LDA are displayed in Table 5, from which we can uncover a range of topics, such as vote, apartment, school life, food, driving, and work. Incorporating the topics that have been labelled, we cluster all conversations into 13 categories, which are Apartment, Financial transaction, Crime, Daily life, Romance, Traffic, Food, Health, School life, Shopping, Travel, Vote and Work. Figure 2 shows diverse topics and their statistics. Among

TABLE 5. List of the top 10 words associated with each topic.

Topics	Top 10 words									
Topic 0	job	house	home	offer	price	thousand	office	make	good	help
Topic 1	really	vote	late	won	computer	clothes	problem	right	sure	new
Topic 2	help	apartment	right	problem	thank	great	today	good	fine	okay
Topic 3	flight	want	room	airport	pay	ticket	airline	money	bring	carry
Topic 4	car	bus	sure	good	really	pass	buy	right	look	drive
Topic 5	school	hi	class	good	time	great	year	really	sure	today
Topic 6	happened	key	got	stay	use	books	news	happen	doctor	come
Topic 7	phone	right	let	help	really	stop	today	number	oh	pcc
Topic 8	good	month	people	rent	news	company	money	business	time	market
Topic 9	classes	parking	time	sir	want	room	park	sure	good	tell
Topic 10	look	new	good	let	buy	looks	want	home	looking	house
Topic 11	want	really	work	make	good	time	food	great	eat	let
Topic 12	problem	today	money	help	thank	sale	want	account	right	good

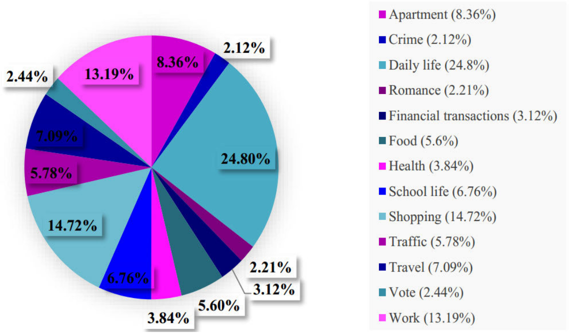


FIGURE 2. Topic distributions in ScenarioSA.

the topics, Daily life (24.8%), Shopping (14.7%) and Work (13.2%) are what people talk about the most, while Crime and Vote are the least. Obviously, this is also in line with our actual life.

B. ANALYSIS OF INTERACTION EFFECT BETWEEN TWO SPEAKERS

Different from commentary documents, in which the affective states of the authors keep constant, the sentiment polarity implied in the conversation is dynamically changing with the conversation going on. The existence of interaction effect requires a change of the sentiment evaluation methodology.

In ScenarioSA, the interaction effect is defined as the combined influence of one speaker on the others. When an interaction effect is present, what one speaker says will lead to a specific type of the other speaker's response. For example, in adjacency utterances, the first utterance actually determine how the second utterance is constructed. The construction of the third utterance will be influenced by both the first and the second utterances. Example interaction patterns include question-answer (service scenario), offering-response (party scenario), apology-minimization (work scenario), greeting-greeting (daily life scenario), etc [24]. We summarize three main interaction patterns in ScenarioSA as follows:

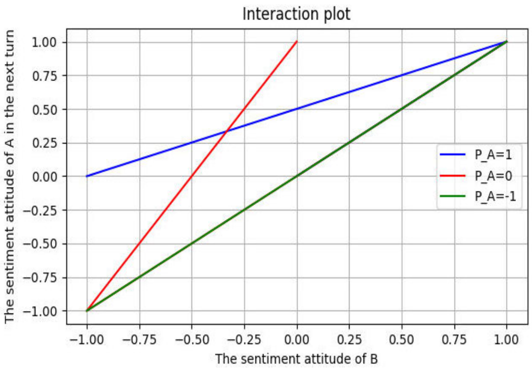


FIGURE 3. Interaction of A and B and their effect on the sentiment polarity of A in the next turn.

1) QUESTION&ANSWER

In the service related scenarios, one speaker usually acts as a questioner aiming to acquire some information. S/He is the leader of a conversation, who will raise one question after another until her/his need is met. The other speaker acts as a service provider, who will answer the questions. About 347 (15.67%) conversations contain this pattern.

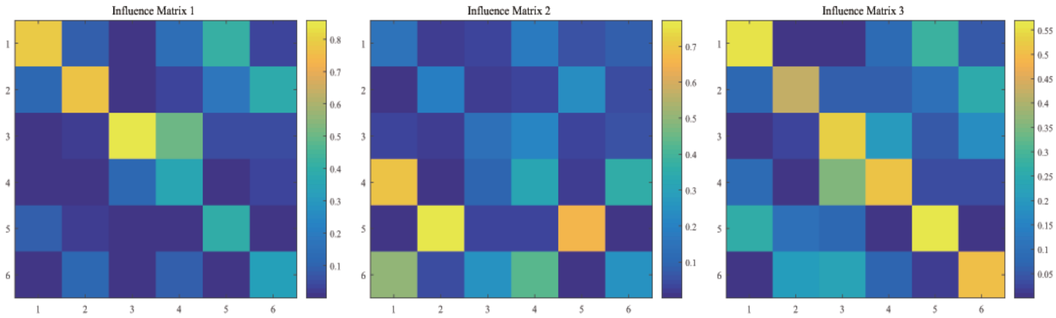
2) OFFERING&RESPONSE

In the party and similar scenarios, one speaker often throws an invitation or gives some advice to the other. The other speaker chooses to accept or reject it in response. It is an active-passive relationship between two speakers. About 281 (12.69%) conversations contain this pattern.

3) GREETING&GREETING

In daily life scenarios, any speaker can initiate a conversation through talking about the weather, the news or the work, etc. The other speaker usually expresses her/his opinions for exchanging information. They generally focus on a common topic, and the role of them are equal. About 923 (41.68%) conversations contain this pattern.

Then, we employ a statistics method, namely the interaction plot [25], to check the interaction effect between two speakers, A and B. If the lines are parallel, then there is no



(a) Influence matrix 1 (b) Influence matrix 2 (c) Influence matrix 3

FIGURE 4. Three learned influence matrices. Different colors denote different influences. Yellow denotes great influence, and deep blue represents small influence.

interaction effect. Conversely, the more non-parallel the lines are, the greater the strength of the interaction. We consider the sentiment polarities of A, B in the current turn as two ternary variables $\mathbf{P_A}$ and $\mathbf{P_B}$, and consider the sentiment polarity of A in the next turn as a third variable $\mathbf{P_{NextA}}$. The interaction plot is shown in Figure 3, which shows a clear interaction effect.

Last, since we have validated the existence of interaction effect, we try to model the interactions between speakers via the influence model [26], which is a generalization of HMM for describing the influence each entity's state has on the others. Suppose there are C entities in the system, and each entity e is associated with a finite set of possible states *1,2,...,S*. At different time t, each entity e is in one of the states, denoted by $q_t^e \in \{1, 2, \ldots, S\}$. Each entity omits an observable o_t^e following the emission probability $P\left(o_t^e|q_t^e\right)$. Interaction effect is treated as the conditional dependence between each entity's current state q_t^e at time t and the previous states of all entities $q_{t-1}^1, q_{t-1}^2, \ldots, q_{t-1}^C$ at time *t-1*. The conditional probability can be formulated as:

$$P\left(q_t^e \middle| q_{t-1}^1, q_{t-1}^2, \ldots, q_{t-1}^e, \ldots, q_{t-1}^C\right) = \sum_{c\in\{1,2,\ldots,C\}} R_{e,c} \times Infl\left(q_t^e|q_{t-1}^c\right) \quad (3)$$

where R is a $C \times C$ influence matrix ($R_{e,c}$ represents the element at the eth row and the cth column), $t = 1, \ldots, T$. $Infl\left(q_t^e|q_{t-1}^c\right)$ is modeled using a $S \times S$ matrix $M^{c,e}$, namely $Infl\left(q_t^e|q_{t-1}^c\right) = M^{c,e}_{q_{t-1}^c, q_t^e}$, where $M^{c,e}_{q_{t-1}^c, q_t^e}$ represents the element at the q_{t-1}^cth row and q_t^eth column of matrix $M^{c,e}$. The matrix $M^{c,e}$ is very similar to the transition matrix, which can be simplified by two $S \times S$ matrices: E^c and F^c. E^c captures the self-state transition i.e., $E^c = M^{c,c}$, and F^c represents adjacent state transition, i.e., $F^c = M^{c,e}, \forall e \neq c$. Therefore, the dynamical influence model can be defined by parameters R, E, F and $P\left(o_t^e|q_t^e\right)$.

The detailed inference procedure for learning all parameters refers to [26]. In this work, we can treat each speaker as an entity, the conversation as the system. Each speaker also has three hidden states (which are -1,0,1), representing positive, negative and neutral. Hence, interaction effect here is set to capture how each speaker's affective states dynamically change another speaker's affective states. All parameters are initialized randomly. After training, three influence matrices are learned based on the above-mentioned three interaction patterns, as shown in Figure 4. We can observe that there exist different types of influences in different interaction patterns, and different affective states have different influences. Influence matri 1 describes influences existing in the "Question&Answer" scenarios. We can see that the questioner has great influence on himself or herself, and is moderately affected by another participant. This indicates that s/he is the leader of a conversation. Influence matrix 2 describes influences existing in the "Offering&Response" scenarios. We can see that the yellow part is positioned in the lower left portion, which illustrates that the second speaker is greatly influenced by the first speaker, before s/he responds. Influence matrix 3 describes influences existing in the "Greeting&Greeting" scenarios. We can see that each speaker is greatly influenced by himself or herself, and also has a moderate influence on the other speaker. The learned influence matrices could simulate the interactions between speakers.

C. SENTIMENT ANALYSIS

Since our conversations are collected from various scenarios, there are many kinds of emotions in different conversations, such as happiness, respect, fear, sadness, anger, etc. In this work, happiness, admiration, respect, romance, etc., are seen as positive emotions (1) while sadness, seeking help, fear, anger, etc., are regarded as negative emotions (-1). We do not underscore the distinction between objectivity and neutrality, and prefer using the same label (0) to annotate both polarities.

In order to depict the sentimental change of each speaker, we manually label the sentiment polarity of each utterance and the final affective state of each speaker when the conversation ends. Finally, we obtain nine (3×3) possible combinations of labels. We count the distribution of sentiment labels, as shown in Figure 5.

Take the sentiment labels of $\mathbf{A}$ as an example, we notice that the proportion of the sentiment polarity of $\mathbf{A}$ being positive, namely, $\mathbf{A} = 1$, is 43.2%, the proportion of $\mathbf{A} = -1$

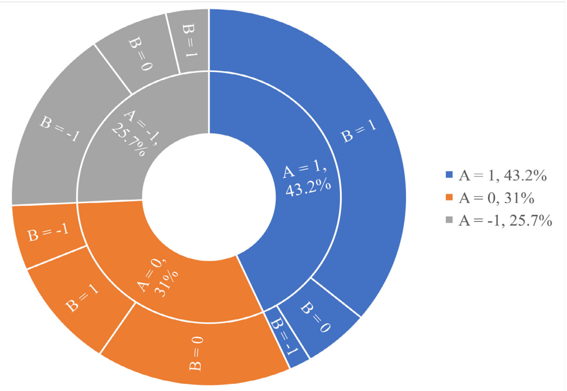


FIGURE 5. The distribution of sentiment labels.

(negative) and **A** = 0 (neutral) are 25.7% and 31.1% respectively. This indicates that our ScenarioSA is well-proportioned on sentiment information. The proportion of (**A** = 1, **B** = 1), (**A** = 0, **B** = 0) and (**A** = −1, **B** = −1) are 36.0%, 16.1%, and 15.8% respectively. This shows that two speakers could achieve consensus after communicating in most scenarios. We find about 1615 (72.9%) conversations, in which the final affective states of two speakers change, comparing with their initial states, because of the interaction effect.

V. EVALUATION WITH ScenarioSA

Note that this paper focuses on presenting an interactive sentiment dataset, demonstrating the need of novel interactive sentiment analysis models and the potential of the dataset to facilitate the development of such models, rather than model designing. Hence, in this section, we propose an extension of interactive attention networks, and evaluate several strong sentiment analysis methods over the ScenarioSA collection, checking whether existing sentiment analysis models could effectively solve interactive sentiment analysis problem or not.

A. EXPERIMENTAL SETTINGS

1) OUR MAIN RESEARCH QUESTIONS

(1) Is interactive sentiment analysis problem a challenging task? (2) How to identify the final affective state of each speaker? (3) How to depict the sentimental change of each speaker in a conversation?

To address (1), we introduce a few comparative models, including lexicon-based and machine/deep learning based methods, and evaluate their performance, demonstrating that these existing models perform poorly on ScenarioSA. It is necessary to design novel interactive sentiment models. To answer (2), we run all baselines on a ternary sentiment classification task, and predict the sentiment label of each utterance. We employ four strategies to obtain the final label of each speaker. Since the sentiment label of the last turn is very correlated to the overall sentiment, then the first strategy is only checking the sentiment label of the last utterance and compare that against the final label, called Simple Baseline here. The second is summing up the labels of each utterance belonging to each speaker, if the sum is greater than 0, the final label is seen as positive; if the sum is less than 0, the final label is seen as negativel if the sum equals to 0, the label is neutral. The third one is assigning different weights to different utterances, where the weights are learned from the dataset. Generally, we assume that the utterances which are nearer to the end of the conversation would have the greater weights. The last is a quantum Interference inspired multimodal decision fusion (QIMF) strategy, which was proposed in the work [27]. Compared with other fusion methods, the QIMF strategy considers the correlation among data at the decision level. To answer (3), based on the labels obtained in the previous step, we compare the predicted label with the true label, and check whether the sentiment of each speaker changes.

2) PRE-PROCESSING, EVALUATION METRICS

We first remove the stop words using a standard stopword list from the Python's NLTK package [28]. However, we do not filter out the punctuations, since some punctuations such as question mark, exclamation point tend to carry subjective information. We adopt **Precision**, **Recall**, **F1 score**, **Accuracy** as the evaluation metrics.

B. COMPARATIVE MODELS AND PARAMETERS SETTINGS

1) SentiStrength

SentiStrength [29] is a lexicon based method. It assigns to each utterance three sentiment strengths: a negative strength between −1 to −5, a positive strength between +1 to +5, and a neutral strength with 0.

2) SVM

We use the bag of words method to generate histograms of word frequencies, and train an SVM classifier to analyze the polarity of each utterance. We set the kernel function to "linear". Other weights are set as the default values.

3) JST

In order to validate the importance of topics, we use the joint sentiment/topic (JST) model [30] to detect sentiments and topics simultaneously. Similar to the work [30], we also use the prior information. In the JST model implementation, we set the symmetric prior $\beta = 0.01$, the symmetric prior $\gamma = 2$.

4) CNN

We employ a CNN [31] including a convolutional layer, a pooling layer, a fully connected layer. It is trained on top of word embeddings for utterance-level classification tasks. We set the learning rate to 0.01, batch size to 60 and the dimensionality of word embeddings to 300.

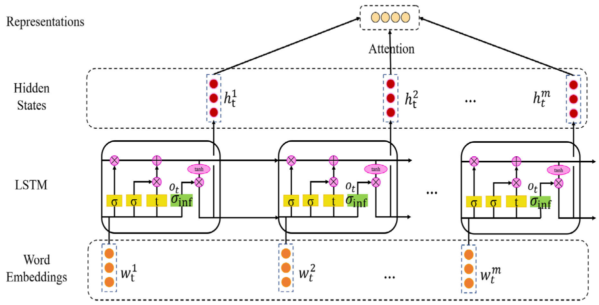


FIGURE 6. The core component of the IAN-INF model.

5) LSTM & AT-LSTM

We implement a standard LSTM [32] and an attention based LSTM [33]. We append the topic words into the input word vectors for computing attention weights. The epoch is set to 30 and the batch size to 60. The dimensionalities of word embeddings and attention vectors are set to 300.

6) IAN

Taking the words that have the largest tf-idf values as the aspects, the IAN [34] is used to generate the representations for utterances. The dimensionalities of word embeddings, attention vectors and LSTM hidden states are set to 300, the dropout rate is set to 0.5.

During the conversation, influence would control the affected speaker's response. That is, influence affects what information one speaker is going to flow out, which is similar to the role of the output gate in the LSTM network. Hence, we extend IAN through incorporating influence into the output gate of each LSTM unit. We call this generalization interactive attention networks with influence (IAN-INF), described as follows. The core component of the IAN-INF is shown in Figure 6.

7) IAN-INF

We combine the output gate of each LSTM unit in IAN with the learned influence to constitute new output gate. This new output gate has considered the previous speaker's influence, when producing the target representation and context representation. This procedure could be written as:

$$o_t = \sigma\left(W_{wo} \cdot w_t + W_{ho} \cdot h_{t-1} + b_o\right) + \sigma\left(R_k \cdot w_t\right) \quad (4)$$

where w_t is word embeddings, W and b denote weight matrices and biases respectively. $R_k \in \{R_1, R_2, R_3\}$ denotes three learned influence matrices. The dimensionalities of word embeddings, attention vectors and LSTM hidden states are set to 300, the dropout rate is set to 0.5.

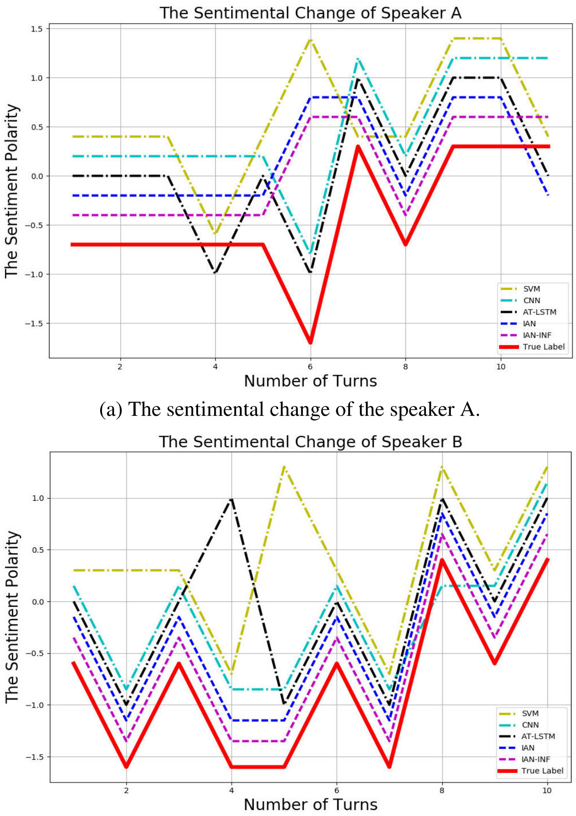
C. RESULTS ON ScenarioSA

The performance of the comparative models is summarized in Table 6. We can observe: (1) almost all the models using the QIMF and weighted combination strategies achieve a better performance. As we expected, this indicates that the weighted combination and QIMF strategies are superior to the summation strategy and Simple Baseline. Moreover, through adding an interference term, the QIMF strategy outperforms the weighted combination strategy. We believe that the QIMF strategy could more effectively incorporate some complementary decision information when data fusing. This shows that the QIMF strategy is an effective data fusion strategy, which has its mathematical principle. (2) For the QIMF and weighted combination strategy, all the models get better classification results on the speaker B than those on A. (3) For the summation strategy and Simple Baseline, the results are in the other way around. As each conversation is initiated by the speaker A, we think that A is the goal-setting one who releases some information. B often acts as the passive information consumer, whose final affective state changes more intensively in comparison with her/his initial emotional state. Hence, the models that use the summation strategy and the label of the last turn perform poorly on the speaker B. (4) Since the label of the last turn is correlated to the overall sentiment, we only check the label of the last utterance and compare it against the final label, namely, Simple Baseline. For IAN-INF model, the accuracy results of A, B are 0.592, 0.546, while the original results (the QIMF strategy) are 0.642, 0.724, declining by 10.0% and 32.6% respectively. For IAN model, the accuracy results of A, B are 0.588, 0.533, while the original results (the QIMF strategy) are 0.633, 0.728, declining by 7.7% and 36.6% respectively. For AT-LSTM model, the accuracy results of A, B are 0.561, 0.514, while the original results (the QIMF strategy) are 0.615, 0.723, declining by 9.6% and 40.7% respectively. From all these results, we can draw a conclusion that it is necessary to develop better fusion method for interactive sentiment analysis. Only checking the sentiment label of the last turn or summing up the labels is not enough.

Specifically, for the weighted combination strategy, the SentiStrength has the worst performance, showing that the lexicon-based approach is not an effective way. JST outperforms SentiStrength, which verifies the significance of topics. SVM, CNN and LSTM outperform SentiStrength on all metrics, but their accuracy results do not exceed 60% on A and 70% on B. Through incorporating the attention mechanism, AT-LSTM and IAN achieve the best performance among all baselines. Their accuracy results achieve 61%, 64.4% on A and 71.4%, 70.7% on B. However, compared with their accuracy results (which are 78.6% and 83.1%) on SemEval 2014 (a standard sentiment analysis collection), they drop by about 22% on A and 10% on B. Through making a simple extension, i.e., incorporating the influence score into the original output gate of each LSTM unit, IAN-INF achieves the best classification results on all metrics. Compare with IAN, the accuracy results of A, B have increased by 2.2% and 1.6%. This proves that considering social interactions is important for improving the classification performance. It is necessary to model the interaction information when designing future interactive sentiment analysis models.

TABLE 6. Performance of all baselines on ScenarioSA. The best performing system is indicated in bold.

Strategy	Speaker	Method	Evaluation metrics			
			Precision	Recall	F1	Accuracy
Summation	**A**	SentiStrength	0.561	0.568	0.560	0.568
		SVM	0.601	0.587	0.588	0.584
		JST	0.573	0.562	0.568	0.569
		CNN	0.581	0.559	0.557	0.558
		LSTM	0.603	0.556	0.553	0.556
		AT-LSTM	0.571	0.555	0.542	0.556
		IAN	0.606	0.584	0597	0.589
		IAN-INF	**0.611**	**0.603**	**0.601**	**0.602**
	B	SentiStrength	0.522	**0.536**	0.528	**0.533**
		SVM	**0.734**	0.505	0.582	0.504
		JST	0.502	0.431	0.419	0.453
		CNN	0.681	0.447	0.539	0.435
		LSTM	0.646	0.404	0.491	0.406
		AT-LSTM	0.692	0.423	0.526	0.437
		IAN	0.723	0.469	0.558	0.462
		IAN-INF	0.715	0.508	**0.593**	0.510
Simple Baseline	**A**	SentiStrength	0.586	0.522	0.511	0.523
		SVM	0.587	0.527	0.515	0.538
		JST	0.567	0.544	0.558	0.551
		CNN	0.590	0.555	0.560	0.557
		LSTM	0.551	0.548	0.549	0.545
		AT-LSTM	0.541	0.564	0.545	0.561
		IAN	**0.625**	0.584	0.580	0.588
		IAN-INF	0.623	**0.586**	**0.587**	**0.592**
	B	SentiStrength	0.508	0.514	0.505	0.501
		SVM	0.533	0.530	0.528	0.529
		JST	0.508	0.510	0.510	0.507
		CNN	0.542	0.500	0.511	0.505
		LSTM	0.511	0.504	0.508	0.506
		AT-LSTM	0.509	0.514	0.512	0.514
		IAN	0.520	**0.544**	0.532	0.533
		IAN-INF	**0.553**	0.543	**0.549**	**0.546**
Weighted Combination	**A**	SentiStrength	0.583	0.562	0.554	0.560
		SVM	0.634	0.601	0.603	0.599
		JST	0.585	0.572	0.579	0.577
		CNN	0.600	0.584	0.586	0.584
		LSTM	0.600	0.582	0.585	0.582
		AT-LSTM	0.611	0.606	0.608	0.610
		IAN	**0.663**	0.645	0.646	0.644
		IAN-INF	0.662	**0.660**	**0.661**	**0.658**
	B	SentiStrength	0.672	0.675	0.673	0.662
		SVM	0.708	0.670	0.686	0.696
		JST	0.702	0.668	0.686	0.675
		CNN	0.667	0.679	0.672	0.676
		LSTM	0.681	0.676	0.678	0.667
		AT-LSTM	0.696	**0.734**	0.717	0.714
		IAN	0.735	0.706	0.719	0.707
		IAN-INF	**0.741**	0.723	**0.729**	**0.718**
QIMF	**A**	SentiStrength	0.588	0.547	0.562	0.581
		SVM	0.602	0.578	0.583	0.574
		JST	0.585	0.579	0.580	0.582
		CNN	0.605	0.587	0.594	0.589
		LSTM	0.610	0.604	0.594	0.597
		AT-LSTM	0.605	**0.625**	0.618	0.615
		IAN	0.645	0.616	0.618	0.633
		IAN-INF	**0.647**	0.614	**0.620**	**0.642**
	B	SentiStrength	0.672	0.665	0.669	0.664
		SVM	0.723	0.716	0.718	0.719
		JST	0.676	0.669	0.672	0.670
		CNN	0.701	0.691	0.702	0.700
		LSTM	0.706	0.714	0.711	0.708
		AT-LSTM	0.729	0.708	0.720	0.723
		IAN	0.728	**0.734**	**0.731**	**0.728**
		IAN-INF	**0.734**	0.720	0.725	0.724



(a) The sentimental change of the speaker A.

(b) The sentimental change of the speaker B.

FIGURE 7. The sentimental change of the speaker A and B. We make small vertical shifts for illustration.

For the QIMF strategy, we can get similar observations. In this work, we relax the constraint on the coefficients so that α^2 plus β^2 does not necessarily equal to one. We tune free parameters α, β to make $\alpha^2 = 0.7, \beta^2 = 0.7$. When $cos\theta = -0.2$ on A and $cos\theta = 0.8$ on B, most models achieve their highest classification scores. This shows that there exists slightly negative interference effect for A's labels, and strongly positive interference effect for B's labels. Finally, SentiStrength and SVM perform the worst, while CNN and LSTM perform better. AT-LSTM, IAN and IAN-INF achieve a noticeable improvement over the above models. IAN-INF achieves best performance in terms of precision, f1 and accuracy on A, which shows that IAN-INF is an effective extension of IAN. Combining the interaction information does boost performance. However, we notice that it gets the second best performance on B, which shows that only simple modifying the output gate is not enough. We need to develop more refined interactive sentiment analysis models in the future.

Figure 7 shows the comparison of the predicted sentimental change using SVM, CNN, AT-LSTM, IAN, IAN-INF and the actual sentimental change of A, B. We see that only the IAN and IAN-INF models accurately capture the sentimental change of B, but none of them could accurately capture the sentimental change of A.

In summary, we can draw a conclusion that interactive sentiment analysis is a challenging task. ScenarioSA could describe the interactions between speakers, which would facilitate the development of future sentiment analysis models.

VI. CONCLUSIONS AND FUTURE WORK

We present ScenarioSA, a manually labelled conversational dataset for interactive sentiment analysis. Compared with prior sentiment datasets, ScenarioSA covers 13 scenarios ranging from daily life, work to politics, exhibits the sentiment interactions between two speakers, and reflects the sentimental change of each speaker. Experimental results from a proposed extension of IAN model and several state-of-the-art sentiment analysis models demonstrate that interactive sentiment analysis is a challenging task, and ScenarioSA can benefit the development of new methodologies.

In the future, on the one hand, we could improve annotation instructions, annotation guidelines and introduce more expert annotators and native speakers for increasing annotator agreement. Moreover, we may improve the manner of selecting and combining annotations from different annotators, e.g., only using conversations where full agreement between annotators exists. Some conversations will not be used in case only conversations with full agreement are used. Hence, more annotated conversations would be required. We may recruit more annotators per conversation, and discard annotators that annotate significantly different from other annotators, etc.

On the other hand, we will develop an elaborate interactive sentiment analysis model that considers the complex interactions. We would conduct detailed analysis of human interactions. Since the interactive sentiment analysis is not just a classification task, but involves a subjective and complex cognition process, considering this problem from a cognitive perspective is a fascinating new direction.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.

[2] A. Tripathy, A. Anand, and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 805–831, Dec. 2017.

[3] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, and O. De Clercq, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 19–30.

[4] F. Tian, H. Liang, L. Li, and Q. Zheng, "Sentiment classification in turn-level interactive chinese texts of E-learning applications," in *Proc. IEEE 12th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2012, pp. 480–484.

[5] L. Deng and Emotibot. (2017). Ai: How Hard it is to Understand the Emotions in the text? Website. [Online]. Available: http://www.sohu.com/a/146212775_491255/

[6] F. Tian, Q. Zheng, R. Zhao, T. Chen, and X. Jia, "Can e-learner's emotion be recognized from interactive Chinese texts?" in *Proc. 13th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, 2009, pp. 546–551.

[7] F. Tian, P. Gao, L. Li, W. Zhang, H. Liang, Y. Qian, and R. Zhao, "Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems," *Knowl.-Based Syst.*, vol. 55, pp. 148–164, Jan. 2014.

[8] B. Ojamaa, P. K. Jokinen, and K. Muischenk, "Sentiment analysis on conversational texts," in *Proc. 20th Nordic Conf. Comput. Linguistics (NODALIDA)*, no. 109. Vilnius, Lithuania: Linköping Univ. Electronic Press, May 2015, pp. 233–237.

[9] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid approach for emotion classification of audio conversation based on text and speech mining," *Procedia Comput. Sci.*, vol. 46, pp. 635–643, 2015.

[10] C. Bothe, S. Magg, C. Weber, and S. Wermter, "Dialogue-based neural learning to estimate the sentiment of a next upcoming utterance," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2017, pp. 477–485.

[11] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 216–225.

[12] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, Ting-Hao, Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," 2018, *arXiv:1802.08379*. [Online]. Available: http://arxiv.org/abs/1802.08379

[13] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. [Online]. Available: https://www.aclweb.org/anthology/P19-1050

[14] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Processing (ACL)*, vol. 10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.

[15] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–17.

[16] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*. [Online]. Available: http://arxiv.org/abs/1710.03957

[17] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proc. 2nd Workshop Cognit. Modeling Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 76–87.

[18] Y. Zhang, J. Wang, B. Tang, Y. Wu, M. Jiang, Y. Chen, and H. Xu, "UTH_CCB: A report for SemEval 2014—Task 7 analysis of clinical text," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 802–806.

[19] E. Tromp, *Multilingual Sentiment Analysis on Social Media*. North Brabant, The Netherlands: Department of Mathematics and Computer Science, Eindhoven Univ. of Technology, 2012.

[20] E. N. Forsythand and C. H. Martell, "Lexical and discourse analysis of online chat dialog," in *Proc. Int. Conf. Semantic Comput. (ICSC)*, Sep. 2007, pp. 19–26.

[21] K. J. Berry and P. W. Mielke, "A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters," *Educ. Psychol. Meas.*, vol. 48, no. 4, pp. 921–933, Dec. 1988.

[22] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: A study of annotation selection criteria," in *Proc. NAACL HLT Workshop Act. Learn. Natural Lang. Process. (HLT)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 27–35.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[24] J. Van Bramer, "Conversation as a model of instructional interaction," *Literacy, Teach. Learn.*, vol. 8, no. 1, p. 19, 2003.

[25] Stevens, Website. (1999). *Interaction Effects in Anova*. [Online]. Available: http://pages.uoregon.edu/stevensj/interaction.pdf/

[26] W. Pan, W. Dong, M. Cebrian, T. Kim, and A. Pentland, "Modeling dynamical influence in human interaction," *IEEE Signal Process. Mag.*, vol. 29, no. 2, pp. 77–86, 2012.

[27] Y. Zhang, D. Song, P. Zhang, P. Wang, J. Li, X. Li, and B. Wang, "A quantum-inspired multimodal sentiment analysis framework," *Theor. Comput. Sci.*, vol. 752, pp. 21–40, Dec. 2018.

[28] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Newton, MA, USA: O'Reilly Media, Inc., 2009.

[29] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010.

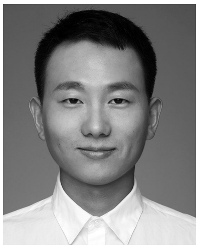
[30] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*. New York, NY, USA: ACM, 2009, pp. 375–384.

[31] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

[32] K. Kawakami, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Dept. Comput. Sci., Tech. Univ. Munich, Munich, Germany, 2008.

[33] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[34] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," 2017, *arXiv:1709.00893*. [Online]. Available: http://arxiv.org/abs/1709.00893



YAZHOU ZHANG received the Ph.D. degree from the College of Intelligence and Computing, Tianjin University, Tianjin, China, in 2020. He is currently a Lecturer with the Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include opinion mining (or sentiment analysis), data fusion, and quantum cognition. He is currently working on developing quantum inspired sentiment analysis models and their application to problems like conversational sentiment analysis, information fusion, and evolution of user emotional state.

ZHIPENG ZHAO received the Ph.D. degree from the College of Intelligence and Computing, Tianjin University, Tianjin, China, in 2019. She is currently a Lecturer with the College of Information Science and Engineering, Henan University of Technology, Zhengzhou, China. Her research interests include computer systems and networking, data center networks, high-performance switch and router, and software defined networks.

PANPAN WANG is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Tianjin University, China. Her research interests include information retrieval, sentiment analysis, and quantum cognition.

XIANG LI received the Ph.D. degree from the College of Intelligence and Computing, Tianjin University, Tianjin, China, in 2019. He is currently a Lecturer with the Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. His research interests include emotion recognition and affective computing.

LU RONG received the master's degree from the School of Foreign Languages and Literature, Tianjin University, Tianjin, China, in 2019. She is currently a Lecturer with the Personnel Department, Zhengzhou University of Light Industry, Zhengzhou, China. Her research interests include natural language understanding, machine translation, and computational language.

DAWEI SONG received the B.Eng. degree in computer science from The Chinese University of Hong Kong, in 1993, the M.Eng. degree in computer science from Tianjin University, in 1996, and the Ph.D. degree in information systems from Jilin University, in 2000. He joined Tianjin University under the Tianjin 1000-Talent Scheme, in 2012. Prior to this appointment, he has been working as the Chair of computing with Robert Gordon University, U.K., since 2008, where he has been an Honorary Professor, since June 2012. He has also worked as a Senior Lecturer and has been a Research Director, since 2007, with the Knowledge Media Institute, The Open University, U.K., from September 2005 to October 2008. He has been a Research Scientist, since September 2000, and has been a Senior Research Scientist, since May 2002, with the Cooperative Research Centre in Enterprise Distributed Systems Technology, Australia. He is currently working with the Beijing Institute of Technology. His research interests include theory and formal models for context-sensitive information retrieval, multimedia and social media information retrieval, domain-specific information retrieval, user behavior, interaction and cognition in information seeking, text mining, and knowledge discovery.