ScenarioSA: A Large Scale Conversational Database for Interactive Sentiment Analysis

Yazhou Zhang
yzhou_zhang@tju.edu.cn
Tianjin University

Lingling Song
Tianjin University

Dawei Song
Beijing Institute of Technology

Peng Guo
Tianjin University

Junwei Zhang
Tianjin University

Peng Zhang
Tianjin University

ABSTRACT

Interactive sentiment analysis is an emerging, yet challenging, subtask of the sentiment analysis problem. It aims to discover the affective state and sentimental change of each person in a conversation. 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 emotion database that we have created and made publically available, namely ScenarioSA[1]. We manually label 2,214 multi-turn English conversations collected from natural contexts. In comparison with existing sentiment datasets, ScenarioSA (1) covers a wide range of scenarios; (2) describes the interactions between two speakers; and (3) reflects the sentimental evolution of each speaker over the course of a conversation. Finally, we evaluate various state-of-the-art algorithms on ScenarioSA, demonstrating the need of novel interactive sentiment analysis models and the potential of ScenarioSA to facilitate the development of such models.

KEYWORDS

Sentiment analysis, conversational database, human interaction

ACM Reference Format:
Yazhou Zhang, Lingling Song, Dawei Song, Peng Guo, Junwei Zhang, and Peng Zhang. 2019. ScenarioSA: A Large Scale Conversational Database for Interactive Sentiment Analysis. In *SIGIR '19: ACM SIGIR Conference, July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Sentiment Analysis (SA) has been a core research topic in natural language processing. Most existing sentiment analysis approaches focus on identifying the polarity of commentaries or similar type of texts (i.e. reviews and tweets) [6]. However, These commentary documents are in the form of individual narratives, without involving interactions among the authors. Along with the rapid development of WWW, instant messaging has been a popular means of communication among people. As a result, a large volume of interactive texts have been procuded, which carry rich subjective information [8]. 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 sentiment analysis 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 conversation, and study the sentimental change of each speaker in the course of the interaction. Compared with the traditional sentiment analysis, which only foucses on identifying the polarity of independent individual, ignoring the interactions, 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; (3) there may 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 [2]. Table 1 shows an example of these phenomena. From Table 1, we can notice that ***A*** and ***B***'s affective states change dynamically because of interactions.

Table 1: An example in ScenarioSA exhibiting the interactions between *A* and *B*, the sentimental change, the affective states, where 1=positive, -1=negative, 0=neutral.

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. My idea is a mess. [-1]
A : How many kids do you want to invite? [0]
....
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. [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]

[1]The dataset and source code are available on: https://github.com/anonymityanonymity/ScenarioSA

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

However, the lack of publicly available interactive sentiment datasets has been a bottleneck for advancing interactive sentiment analysis models. Tian et al. [8] built a Chinese interactive corpus, aiming to solve the problem of emotional illiteracy. But this corpus was not described in detail and was not publicly available. Ojamaa et al. [4] 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. Due to the limited availability of sentiment-annotated interactive text dataset, Bothe et al. [1] had to auto-annotate the sentiment labels of two spoken interaction corpora for training their model.

To fill the gap, we present ScenarioSA, an English conversational dataset with sentiment labels. 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 sentiment polarity: positive, negative or neutral. The final affective state of each participant is also labelled when the conversation ends. The advantages of ScenarioSA over existing sentiment datasets (e.g., Movie Reviews [6], the SemEval-2014 [7], etc.) can be summarized as follows: (1) broad coverage in various scenarios and conversation styles; (2) ScenarioSA depicts the interactions between two speakers, and reflects the sentimental evolution of each speaker over the course of a conversation; (3) ScenarioSA introduces a 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.

Finally, we design a comparative experiment on ScenarioSA over a number of neural network approaches, including a Convolutional Neural 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. Out results show that the IAN and AT-LSTM model only achieve an accuracy of 70.7% and 71.4% on ScenarioSA, in comparison to 78.6% and 83.1% on SemEval 2014. Through considering social influence, IAN-INF achieves better result, which is 72.1%. This indicates that the existing approaches cannot effectively model the evolution of sentiment in interactive conversations and new methodologies are required.

2 DATASET CONSTRUCTION

2.1 Data Collection & Pre-processing

Our goal is to construct a large scale emotion 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 (e.g., eslfast.com, focusenglish.com, etc.)[2]. The conversations are collected in the various daily life contexts and cover a wide range of topics, such as shopping, work, etc., which is important to ensure unbiased evaluation with this dataset.

All the conversations are then pre-processed. Some of the crawled conversations involve three or more participants. 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. 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 2.

[2]Note that the original copyright of all the conversations belongs to the source owners, and the dataset is only for research purposes.

Table 2: 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

2.2 Annotation Criteria and Procedure

Given that distinguishing positive/neutral/negative is more realistic and reliable than distinguishing finer graded multi-dimensional emotions. The pre-processed dataset is manually annotated with three labels: -1, 0, 1. In order to guarantee the annotation quality, we recruited four volunteers including two master students and two PhD students. They all have a good knowledge in sentiment analysis. Before labelling the whole dataset, they were asked to independently annotated 50 examples first, with the aim to minimise ambiguity while strengthen the inter-annotator agreement. We define the gold standard of an utterance or conversation in terms of the label that receives the majority votes. The annotation procedure consists of two steps:

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 three sentiment labels: -1, 0, 1. We split each conversation into utterances and provided four annotators with one turn of dialogues at a time.

Conversation-level annotation: The annotators were asked 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 after the conversation.

Note that the final sentiment label of each speaker does not necessarily equal to the sentiment label of the last turn. After calculation, there are 700 (31.6%) conversations whose final sentiment labels are different from the labels of the last turn.

2.3 Agreement Study

After annotating the whole dataset, we assess the reliability of our sentiment annotation procedure, through an agreement study using 100 sampled conversations randomly.

Agreement assessment: we first use the percent agreement calculation method to calculate the average agreement. At the conversation level, the average agreement among four 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%. Moreover, we have introduced the Kappa metric, and obtained the average $\kappa \approx 0.67$ (the highest score is 0.78, the lowest score is 0.55).

Annotator-level noise: given the disagreement among four annotators, we evaluate the accumulated noise level introduced by each of four annotators. The noise level of each annotator $noise\,(anno_i)$ is estimated through computing the deviation frequency of the labels 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 others (whose noise levels are 10.6%, 12.4% and 16.3%). The noise level reflects the reliability of annotators. We would value the opinions of annotators who have lower noise levels.

3 DATASET ANALYSIS

3.1 Sentiment Analysis

Examining the database, we calculate the proportion of the sentiment labels based on the final affective states of speakers ***A***, ***B***.

We calculate that the proportion of the sentiment polarity of **A**=1 is 43.2%, the proportion of **A**=-1 and **A**=0 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 1557 (70.3%) conversations, in which the final affective states of two speakers change, comparing with their initial states, because of the interaction effect.

3.2 Analysis of Interactions between Speakers

In ScenarioSA, the interaction effect is defined as the combined influence of one speaker on the others. We summarize three main interaction patterns in ScenarioSA as follows:

(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 524 (23.67%) conversations contain this pattern.

(2) Offering&Response: 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 305 (13.77%) conversations contain this pattern.

(3) Greeting&Greeting: Any speaker can initiate a conversation through talking about any topics. 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 1186 (53.55%) conversations contain this pattern.

Then, we employ a statistics method, namely the interaction plot, to check the interaction effect between ***A*** and ***B***. If the lines are parallel, then there is no 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 1, which shows a clear interaction effect.

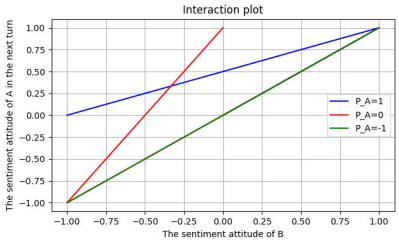

Figure 1: Interaction of *A* and *B* and their effect on the sentiment polarity of *A* in the next turn.

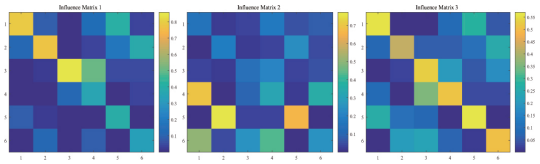

Figure 2: Three learned influence matrices. (a) Influence matrix 1; (b) influence matrix 2; (3) influence matrix 3. Different colors denote different influences.

Since we have validated the existence of interaction effect, we try to model the interactions via the influence model [5], which describes the influence each entity's state has on the others. After training, three influence matrices are learned based on the above-mentioned three interaction patterns, as shown in Figure 2. From Figure 2, we observe that there exist different types of influences in different interaction patterns. Influence matrix 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 will be incorporated into the output gate of each LSTM unit.

4 EVALUATION WITH SCENARIOSA

This paper focus on presenting an emotion database, demonstrating its potential to facilitate the development of interactive models.

4.1 Experimental Settings

We run all baselines on a ternary classification task, and predict the sentiment label of each utterance. In order to obtain the final label

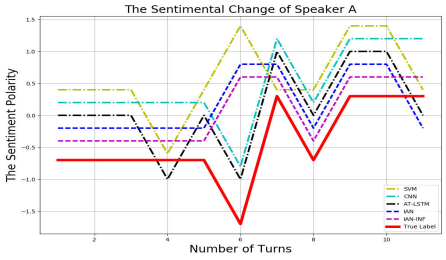

Figure 3: The sentimental change of the speaker *A*. We make small vertical shifts for illustration.

of each speaker, we employ two strategies. One is summing up the labels of each utterance belonging to each speaker, e.g., if the sum is greater than 0, the final label is seen as positive; if the sum equals to 0, the label is seen as neutral. The other is assigning different weights to different utterances, where the weights are learned from the dataset.

Cross validation, evaluation metrics: In order to achieve more convincing results by using the dataset, we run the experiments using 5-fold cross-validation for all comparative models. We adopt **F1 score**, **Accuracy** as the evaluation metrics.

Comparative models: (1) **CNN:** we employ a CNN including a convolutional layer, a pooling layer, a fully connected layer. (2) **LSTM & AT-LSTM:** we implement a standard LSTM and an attention based LSTM [9]. (3) **IAN:** taking the words that have the largest tf-idf values as the aspects, the IAN [3] is used to generate the representations for utterances. (4) **Last turn**: since the label of the last turn is correlated to the overall sentiment, we would only check the label of the last utterance, which is predicted by IAN, i.e., **IAN (last turn)**. (5) **IAN-INF:** we combine the output gate of each LSTM unit in IAN with the learned influence scores to constitute new output gate for considering the previous speaker's influence.

4.2 Results on ScenarioSA

The performance of the comparative models is summarized in Table 3. We can observe: (1) since the weights could measure the mutual importance relationship between predictions, almost all the models using the weighted combination strategy achieve a better performance. (2) For the weighted combination strategy, all the models get better classification results on the speaker ***B*** than those on ***A***. For the summation strategy, the results are in the other way around. As each conversation is initiated by the speaker ***A***, ***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 his initial emotional state. Hence, simple summation strategy performs poorly on ***B***. (3) For Last turn, the accuracy results of ***A***, ***B*** are 0.588 and 0.533. It performs poorly, compared with the models using the weighted combination strategy, while it performs well, compared with the models using the summation strategy. Because ***B***'s affective states change greatly in the whole interactive activities, the label of the last turn could reflect more accurate results than the summation strategy. However, only checking the sentiment label of the last turn is not enough.

Table 3: Performance of all baselines on ScenarioSA. The best performing system is indicated in bold.

Speaker	Method	Summation		Weighting	
		F1	Accuracy	F1	Accuracy
A	CNN	0.557	0.558	0.586	0.584
	LSTM	0.553	0.556	0.585	0.582
	AT-LSTM	0.542	0.556	0.608	0.610
	IAN	0.621	0.624	0.646	0.644
	IAN (last turn)	0.580	0.588	0.580	0.588
	IAN-INF	**0.630**	**0.632**	**0.664**	**0.658**
B	CNN	0.539	0.435	0.672	0.676
	LSTM	0.491	0.406	0.678	0.667
	AT-LSTM	0.526	0.437	0.717	0.714
	IAN	**0.558**	0.462	0.719	0.707
	IAN (last turn)	0.532	**0.533**	0.532	0.533
	IAN-INF	0.553	0.510	**0.724**	**0.721**

For the weighted combination strategy, CNN and LSTM's accuracy results do not exceed 60% on ***A*** and 70% on ***B***. Through incorporating the attention mechanism, AT-LSTM and IAN 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, they drop by about 22% on ***A*** and 10% on ***B***. This shows that interactive sentiment analysis is a challenging task. Through making a simple extension, i.e., incorporating the influence scores into the original output gate of each LSTM unit, IAN-INF almost achieves the best classification results. This proves that considering social interactions is necessary for improving classification performance.

Figure 3 shows the comparison of the predicted sentimental change using SVM, CNN, AT-LSTM, IAN, IAN-INF and the actual sentimental change of ***A*** in one conversation. We observe that none of them accurately capture the sentimental change of *A*. New methodologies are required.

5 CONCLUSIONS

We present ScenarioSA, a manually labelled conversational emotion database. 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 several state-of-the-art sentiment analysis models demonstrate that interactive sentiment analysis is a challenging task, and ScenarioSA could benefit the development of new methodologies.

REFERENCES

[1] Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter. 2017. Dialogue-Based Neural Learning to Estimate the Sentiment of a Next Upcoming Utterance. In *International Conference on Artificial Neural Networks*. Springer, 477–485.

[2] Lin Deng and Emotibot. 2017. AI: How hard it is to understand the emotions in the text? Website. http://www.sohu.com/a/146212775_491255/.

[3] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. *arXiv preprint arXiv:1709.00893* (2017).

[4] Birgitta Ojamaa, Päivi Kristiina Jokinen, and Kadri Muischenk. 2015. Sentiment analysis on conversational texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*. Linköping University Electronic Press, 233–237.

[5] Wei Pan, Wen Dong, Manuel Cebrian, Taemie Kim, and A Pentland. 2012. Modeling dynamical influence in human interaction. *IEEE Signal Processing Magazine* 29, 2 (2012), 77–86.

[6] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[7] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval '17)*. Association for Computational Linguistics, Vancouver, Canada.

[8] Feng Tian, Huijun Liang, Longzhuang Li, and Qinhua Zheng. 2012. Sentiment Classification in Turn-Level Interactive Chinese Texts of E-learning Applications. In *2012 IEEE 12th International Conference on Advanced Learning Technologies (ICALT)*. 480–484.

[9] Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 606–615.