Reevaluating Data Partitioning for Emotion Detection in EmoWOZ

Anonymous ACL submission

Abstract

This paper focuses on the EmoWoz dataset, an extension of MultiWOZ that provides emotion labels for the dialogues. MultiWOZ was partitioned initially for another purpose, resulting in a distributional shift when considering the new purpose of emotion recognition. The emotion tags in EmoWoz are highly imbalanced and unevenly distributed across the partitions, which causes sub-optimal performance and poor comparison of models. We propose a stratified sampling scheme based on emotion tags to address this issue, improve the dataset's distribution, and reduce dataset shift. We also introduce a special technique to handle conversation (sequential) data with many emotional tags. Using our proposed sampling method, models built upon EmoWoz can perform better, making it a more reliable resource for training conversational agents with emotional intelligence. We recommend that future researchers use this new partitioning to ensure consistent and accurate performance evaluations.

1 Introduction

Emotion recognition in task-oriented conversational agents is challenging because it requires the agent to accurately interpret and respond to a user's emotional state in real-time. Emotional signals can be complex and difficult to detect accurately, especially in unstructured conversations where users may not express their emotions explicitly. Therefore, conversational agents need to interpret the contextual factors accurately to provide appropriate responses sensitive to the user's emotional state. Despite the challenges, emotion recognition in task-oriented conversational agents is important because it can improve the overall user experience (Zhang et al., 2020). By accurately detecting and responding to a user's emotional state, conversational agents can provide more personalized and empathetic interactions, increasing user satisfaction and engagement. Additionally, emotion recognition can help agents identify when a user is experiencing frustration, confusion, or other negative emotions, allowing them to intervene and provide support to prevent user dropout or dissatisfaction(Andre et al., 2004). Emotion recognition is critical to developing effective task-oriented conversational agents that can provide a human-like user experience (Polzin and Waibel, 2000).

Sentiment analysis in task-oriented conversational agents has been addressed in the literature as an essential aspect of natural language processing that can improve the overall user experience (Shi and Yu, 2018; Saha et al., 2020; Wang et al., 2020). However, the lack of publicly available data for emotion recognition is a significant limitation for task-oriented conversational agent applications.

MultiWOZ (Multi-Domain Wizard-of-Oz) is a large-scale dataset of human-human written conversations for task-oriented dialogue modeling. The dataset was initially collected for training and evaluating dialogue systems, particularly those designed to assist users with completing specific tasks such as booking a hotel or reserving a table at a restaurant (Budzianowski et al., 2018). Feng et al. (2022) extended the MultiWOZ dataset by including dialogues between humans and a machine-generated policy, which they named DialMAGE. They added emotional labels to the user message and called the resulting dataset EmoWOZ. EmoWOZ is a large-scale, manually emotion-annotated corpus of task-oriented dialogues. The corpus contains more than 11K dialogues with more than 83K emotion annotations of user utterances, which makes it the first large-scale open-source corpus of its kind. The authors propose a novel emotion labeling scheme tailored to task-oriented dialogues and demonstrate the usability of this corpus for emotion recognition and state tracking in task-oriented dialogues. The authors highlights that while emotions in chit-chat dialogues have received considerable attention (Li et al., 2017; Poria et al., 2018; Zahiri and Choi,

2017), emotions in task-oriented dialogues remain largely unaddressed. They argue that incorporating emotional intelligence can help conversational AI generate more emotionally and semantically appropriate responses, making a better user experience.

EmoWOZ authors maintained the original split of the MultiWOZ dataset and divided the DialMAGE dataset into three training, validation, and testing sets with a ratio of 8:1:1. While the original partitioning of the MultiWOZ dataset is suitable for the development of task-oriented conversational agents, it may not be optimal for detecting emotions. For emotion recognition, the EmoWoz dataset exhibits a phenomenon known as *dataset shift*, which refers to a discrepancy between the joint distribution of inputs and outputs during the training phase as opposed to the validation and test phases (Quinonero-Candela et al., 2008). This inconsistency leads to a decrease in performance when using the dataset. It is essential to address this issue by ensuring that the training, validation, and test sets are representative of the same underlying distribution to improve model performance. Creating a new partitioning that accommodates this additional purpose improves the dataset's reusability. By doing so, researchers can leverage the existing dataset for detecting emotions. This approach enables more efficient and cost-effective use of the data while maintaining high-quality results.

2 Data partitioning

When building a predictive model, we typically split our data into three sets: a training set, a validation set, and a test set. The purpose of the training set is to estimate model parameters, and the purpose of the validation set is to tune the model's hyperparameters and assess its performance. The test set aims to get an unbiased estimate of the model's performance on new, unseen data. If the test set has a different distribution than the validation set, a model that performs well on the validation set may not be the best model for the test set, and vice versa. Therefore, it's essential to ensure that the distributions of the three sets are as similar as possible.

We specifically concentrate on the MultiWOZ subset of EmoWOZ in this paper as it is widely used in various applications. Nonetheless, our approach can be readily extended to the entire EmoWOZ dataset. Approximately 2.5% of the MultiWOZ subset in (Feng et al., 2022) underwent manual annotation for resolution. However, this manual annotation was not evenly distributed across all classes, and minority classes were affected more than majority classes. For instance, the *abusive* class in the test set was completely annotated manually, while in the validation set it was manually labeled in 50% of cases. Also, the *Fearful* class had a relative frequency three times higher in the training set, as shown in Table 1.

	Relative Frequency			Manual resolution		
	Fear.	Abus.	Dis.	Fear.	Abus.	Dis.
Train	0.62	0.06	1.29	28.86	87.88	16.53
Val.	0.22	0.08	1.00	62.50	50.00	21.62
Test	0.20	0.07	1.47	40.00	100.0	20.37

Table 1: Relative frequency and Manual resolution percentage for the three minority classes with less than 2% relative frequency in the MultiWOZ dataset

Data	F1 for each Emotion Label							Macro F1	
	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	w/o N.	w N.
Train	93.94	54.71	50.19	72.85	32.49	42.32	88.78	56.89	62.18
Val.	94.09	26.25	46.72	73.35	50.00	43.19	88.73	54.71	60.33
Test	94.14	29.73	52.03	71.98	32.00	34.97	88.86	51.6	57.67

Table 2: Performance of annotators based on F1 for emotion labels (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied) on MultiWOZ . Following the benchmarks in the literature, in the aggregated level, we report Macro F1 scores **w**ith**o**out **N**eutral and **w**ith **N**eutral emotion.

Three annotators labeled the utterances according to the task. The final label was determined primarily by the majority vote of the annotators. Among all utterances, 72.1% had a complete agreement among the three annotators. A partial agreement was found for 26.4% of the utterances, while for 1.5%, there was no agreement. The paper reports that these instances were resolved manually to address cases where the annotators could not reach an agreement. In a small portion of the data, a label different from the majority vote was chosen.

We use F1 scores to compare annotators across data partitions to assess inter-annotator agreement. In essence, we measure the effectiveness of annotations by the three annotators across the training, validation, and test sets using the final labels. Table 2 presents the results, indicating model performance discrepancies across the training, validation, and test sets. This phenomenon is known as *dataset shift* in which there is a difference between the joint distribution of inputs and outputs during the training stage compared to the validation and test stage (Quinonero-Candela et al., 2008), leading to a decrease in performance. Dataset Shift is a

common problem in machine learning, and it can have significant consequences, such as a decrease in accuracy. In the case of EmoWoz, the dataset Shift arises from the fact that (Feng et al., 2022) kept the original partitioning of MultiWOZ, which is not evenly distributed across partitions for this particular task.

It's worth noting that while the original partitioning of MultiWOZ data is suitable for many tasks related to the development of task-oriented conversational agents, it may not be ideal for emotion detection. Emotion detection requires consideration of different contextual aspects, which may require a new partitioning approach. For instance, in the original partitioning, 60% of messages with the *Fearful* emotion in the training set were in conversations with the police or hospital. However, in the validation and test sets, none of the conversations with *Fearful* emotions were related to the police or hospital. This contextual aspect of the conversation plays a crucial role in accurately recognizing emotions.

Based on the bootstrap test, the p-values for the three dataset pairs were close to zero, suggesting that the observed differences in label proportions between the train, validation, and test sets are unlikely to have occurred by chance. Therefore, we can conclude that there is evidence of dataset shift between the train, validation, and test sets, indicating that a model trained on the train set may not perform well on the test set.

To address this issue, we use stratified sampling, which is a sampling technique that ensures that each sub-group in the data is represented proportionally in the sample. In this case, we use stratified sampling to ensure that the training set has a distribution similar to the validation and test sets. Stratified sampling is particularly useful in situations where the distribution of the target variable is imbalanced or varies across sub-groups in the data. In the case of EmoWoz, the distribution of emotions in the training set has differed from that in the validation and test sets, which could have contributed to the dataset shift. We used the Algorithm 1 to get a new partitioning of the data. By using stratified sampling, we have ensured that the emotion distribution in the training set was similar to that in the validation and test sets, which helps to reduce the dataset Shift and improve the model's performance. Table 3 shows the annotator's F1-score after the new partitioning.

Data: MultiWoz dataset with emotion labels from EmoWOZ

1. Group the dataset based on their utterance ids and find a list with the emotion sequence in each utterance.
2. Determine the frequency of emotional sequences in the dataset.
3. Make a dictionary called *emotion_seq_dict* with the emotional sequence as the key and the counts of the sequence in the dataset as the value.
4. Partition the whole dataset into one set called *frequent_seq* with conversations of more than six emotional list frequencies and another set *non_frequent_seq* with the rest of the data.
5. For the *frequent_seq*, do the stratified sampling of conversation based on the emotion sequence and partition it to the training, validation, and test set, with a 80-10-10 split similar to the original split of the data.
6. Use random sampling to partition the *non_frequent_seq* to the training, validation, and test sets.
7. Find the union of the two partitions to get the partitioning of the whole dataset.

Algorithm 1: Stratified sampling for emotion recognition in the conversation

3 Case study

Feng et al. (2022) used Bert, Contextual BERT, DialogueRNN (Majumder et al., 2019), and COSMIC (Ghosal et al., 2020) for the baseline methods. Among methods used in EmoWOZ, BERT did not incorporate the sequential aspects of the conversation, yet it yielded the best Macro F1 scores in most EmoWOZ subsets. We build up the BERT model by computing the relative embedding of the message from the chatbot and agent and employ a transformer model to address the sequential aspects of the problem. Hyperparameters for this method using both the original and proposed partitioning are illustrated in Table 4. Upon examining the results, we can see that in the original partitioning, the hyperparameters corresponding to the top-

Data	F1 for each Emotion Label							Macro F1	
	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	w/o N.	w N.
Train	94.02	52.38	50.97	73.06	34.87	41.73	88.83	56.98	62.27
Val.	93.86	48.58	47.16	71.49	37.50	41.65	88.69	55.85	61.28
Test	93.79	52.08	46.13	72.20	32.26	41.54	88.49	55.45	60.93

Table 3: Performance of annotators based on F1 for emotion labels (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied) on MultiWOZ after new partitioning.

		Original splits			Stratified splits		
Batch	Epochs	Val.	Test	Dif.	Val.	Test	Dif.
8	4	**51.4**	48.92	-2.48	52.25	51.58	-0.67
8	8	49.52	52.35	2.83	53.5	52.38	-1.12
16	4	50.58	54.03	3.45	53.09	51.76	-1.33
16	8	50.31	51.77	1.46	53.54	49.68	-3.86
32	4	48.96	**54.23**	5.27	55	**53.73**	-1.27
32	8	49.86	52.27	2.41	**56.8**	53.14	-3.66

Table 4: Performance of different hyper-parameters.

performing model on the validation set produced the worst model on the test set. This discrepancy could be indicative of data drift, as we previously discussed. This implementation uses five distinct

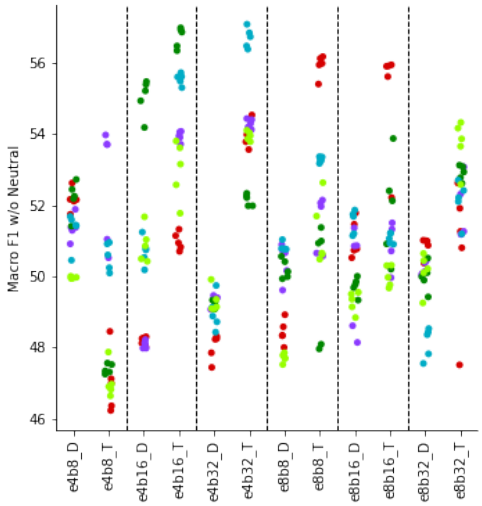

Figure 1: This figure depicts the Macro F1 score for each of the seeds utilized in implementing the sequential extension of the BERT model on the **origial** Multiwoz partitioning. For each hyperparameter, both the Macro F1 score in the validation and test sets are plotted in close proximity to one another. Additionally, the colors within the figure represent the five distinct seeds utilized in the embedding step.

seeds to generate five unique embeddings. We then employed five seeds to construct the transformer model on top of these embeddings. Consequently, we had 25 different models for each parameter in Table 4. To visualize the Macro F1 score for each of these models, we included Figures 1 and 2. The

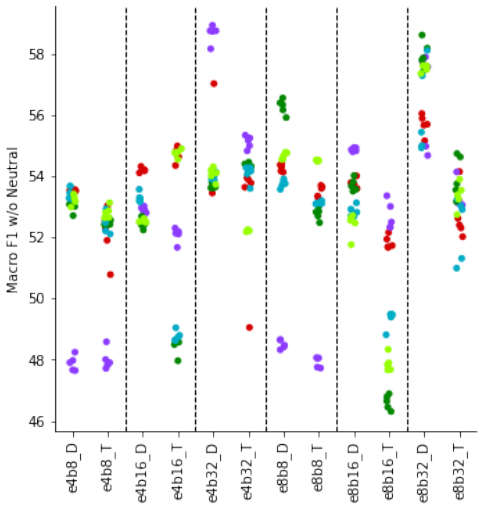

Figure 2: This figure depicts the Macro F1 score for each of the seeds utilized in implementing the sequential extension of the BERT model on the **proposed** Multiwoz partitioning. For each hyperparameter, both the Macro F1 score in the validation and test sets are plotted in close proximity to one another. Additionally, the colors within the figure represent the five distinct seeds utilized in the embedding step.

colors within these figures correspond to the five distinct seeds utilized in the embedding step. Notably, we can observe that only in the proposed partitioning of the data the change in averaged Macro F1 score is similar across the validation and test sets for various hyperparameters. Furthermore, we can observe that for models with equivalent hyperparameters, the change in score across different initial seeds is comparable in both the validation and test sets. These observations suggest that no data drift is present in the new partitioning.

4 Concluding remarks

After analyzing the original data partitioning, we have identified potential data drift and suggested an alternative approach to address this issue. Our evaluation of the results indicates that the new partitioning approach effectively reduces data drift, as demonstrated by the consistency of Macro F1 scores in both the validation and test sets across different hyperparameters and initial seeds. These findings suggest that the proposed partitioning method is a suitable alternative for researchers working on emotion detection using MultiWOZ data. This work emphasizes the significance of meticulously choosing and employing partitioning methods in the training and assessment of machine learning models.

References

Elisabeth Andre, Matthias Rehm, Wolfgang Minker, and Dirk Bühler. 2004. Endowing spoken language dialogue systems with emotional intelligence. In *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings*, pages 178–187. Springer.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašić. 2022. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Thomas S Polzin and Alexander Waibel. 2000. Emotion-sensitive human-computer interfaces. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PloS one*, 15(7):e0235367.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. *arXiv preprint arXiv:1804.10731*.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multitask learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9177–9184.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

Rui Zhang, Kai Yin, and Li Li. 2020. Towards emotion-aware user simulator for task-oriented dialogue. *arXiv preprint arXiv:2011.09696*.