Dialogue Quality and Emotion Annotations for Customer Support Conversations

John Mendonça[1,2,*†], Patrícia Pereira[1,2,*]
Miguel Menezes[1,3,4], Vera Cabarrão[4], Ana C. Farinha[4], Helena Moniz[1,3,4]
João Paulo Carvalho[1,2], Alon Lavie[5,6] and **Isabel Trancoso[1,2]**
[1] INESC-ID, Lisbon
[2] Instituto Superior Técnico, University of Lisbon
[3] Faculdade de Letras, University of Lisbon
[4] Unbabel, Lisbon
[5] Carnegie Mellon University, Pittsburgh
[6] Phrase, Pittsburgh
{john.mendonca, patricia.pereira}@inesc-id.pt

Abstract

Task-oriented conversational datasets often lack topic variability and linguistic diversity. However, with the advent of Large Language Models (LLMs) pretrained on extensive, multilingual and diverse text data, these limitations seem overcome. Nevertheless, their generalisability to different languages and domains in dialogue applications remains uncertain without benchmarking datasets. This paper presents a holistic annotation approach for emotion and conversational quality in the context of bilingual customer support conversations. By performing annotations that take into consideration the complete instances that compose a conversation, one can form a broader perspective of the dialogue as a whole. Furthermore, it provides a unique and valuable resource for the development of text classification models. To this end, we present benchmarks for Emotion Recognition and Dialogue Quality Estimation and show that further research is needed to leverage these models in a production setting.

1 Introduction

Artificial Intelligence (AI) has evolved to become a ubiquitous technology in our lives. Yet, its performance is limited by the amount of data it is trained on. Therefore, and in order to maximise the rewards of such technology, substantial research and engineering effort has been devoted to collecting and annotating data according to needs and goals.

One of the main limitations of most task-oriented conversational datasets is their lack of variability. The majority of these datasets are collected in controlled environments where annotators are

Agent: Delivery usually takes place within 1-7 working days after dispatch but this can vary depending on the couriers availability in your area.			
Cor: 2	**Tem:** 1	**Eng:** 1	**Emo:** *Neutral*
Und: 1	**Sen:** 1	**IQ:** 5	**Pol:** 1
Customer: Sorry, that doesn't satisfy me.			
Cor: 2	**Tem:** 0	**Eng:** 1	**Emo:** *Anger*
I'm already waiting for my sofa for almost 8 weeks!			
Cor: 2	**Tem:** 0	**Eng:** 1	**Emo:** *Anxiety*
Und: 1	**Sen:** 1	**IQ:** 1	**Pol:** 0
A: Regrettably as we do not have control over the speeds of the shipping processes we are unable to expedite orders, the item is still on a boat but we are doing our best to get it to you as soon as possible.			
Cor: 2	**Tem:** 0	**Eng:** 0	**Emo:** *Disappointment*
Any delays the item may encounter on its way to our distribution center are out of our hands and cannot be predicted.			
Cor: 2	**Tem:** 0	**Eng:** 0	**Emo:** *Neutral*
Und: 2	**Sen:** 1	**IQ:** 5	**Pol:** 1
C: And now? Should I sit or lie on the floor?			
Cor: 2	**Tem:** 0	**Eng:** 1	**Emo:** *Frustration*
This is not okay			
Cor: 2	**Tem:** 0	**Eng:** 1	**Emo:** *Frustration*
Und: 1	**Sen:** 1	**IQ:** 1	**Pol:** 0
A: I understand this is frustrating and disappointing.			
Cor: 2	**Tem:** 1	**Eng:** 1	**Emo:** *Neutral*
Und: 2	**Sen:** 1	**IQ:** 5	**Pol:** 1

Table 1: Adapted example of a portion of a dialogue from the MAIA DE-1 subset, from the point of view of the Agent (which receives and sends messages in English). The customer interacts with the agent in their corresponding language (in this case German). This is achieved by employing Machine Translation on both ends (DE → EN and EN → DE).

encouraged to follow specific guidelines, and are limited to a restrictive set of topics, and outcomes (El Asri et al., 2017; Budzianowski et al., 2018; Rastogi et al., 2020). This leads to highly struc-

tured dialogues that do not accurately reflect genuine conversations. In contrast, customer support conversations provide a broader range of topics and contexts, and are more linguistically diverse (Lowe et al., 2015). Furthermore, most datasets are monolingual, resulting in a lack of representation of diverse linguistic and cultural features such as tone and idiomatic expressions (Gonçalo Oliveira et al., 2022).

One approach to equip NLP models with multilingual and diverse domain knowledge capabilities is to leverage LLMs pretrained on extensive amounts of publicly available data (Conneau et al., 2020; Xue et al., 2021; OpenAI, 2023). However, lacking benchmarking dialogue datasets, it is not clear these models, applied to dialogue, are able to fully generalise to other languages and/or domains, even if other dimensions of variability remain unchanged.

This paper builds upon the original MAIA dataset release by adding extensive annotations of emotion and dialogue quality at different granularity levels, thus allowing a holistic approach at understanding the dynamics of conversations in the context of customer support. The MAIA dataset is a collection of genuine bilingual customer support conversations initially released as a challenge dataset for the WMT Chat shared-task (Farinha et al., 2022). In these conversations, which are powered by Machine Translation, the agent communicates with the customer exclusively in English, whereas the customer interacts with the agent exclusively in their native language. Our annotations cover 612 dialogues accounting for around 25k sentences, covering diverse topics, ranging from account registration issues, payment and delivery clarifications and after-sale services. Languages includes German (DE), Brazilian-Portuguese (PT_BR) and European Portuguese (PT_PT).

We argue that the MAIA dataset and the accompanying annotations have unique value in the field of customer support and conversational agents. The comprehensive annotations conducted enable the analysis of the relations between several dialogue sub-qualities and emotion. Furthermore, they can be used as a training and benchmark dataset for text classification in these distinctive settings. For instance, one could leverage this dataset for the construction of dialogue systems that support customer-agent interaction processes. Classification models trained on this data could assist customer service agents (human or machine) by measuring customer emotions and dialogue qualities in real-time and provide the agent with feedback on the fluidity and success of the dialog.

To kick-start this research, this paper provides benchmarks for Emotion Recognition and Dialogue Quality Estimation. Results show that existing models are not strong enough to perform on par with other benchmarks, indicating significant future work research will be required to reduce this performance gap.

In summary, the primary contributions of this work are as follow:

- We conduct extensive emotion and dialogue quality annotations for the MAIA dataset. More specifically, we annotate the dataset at different levels of granularity, ranging from the **sentence level**, where we perform an 8-class emotion and local conversational quality annotations; **turn-level** conversational quality annotations including IQ (Interaction Quality); up to the **dialogue level**, where annotations for task success are provided;

- We analyse these annotations and show how emotions and different aspects of conversational quality are related in the context of customer support;

- We benchmark known approaches for Emotion Recognition in Conversations and Dialogue Evaluation on this dataset;

- The corpus and accompanying benchmarks are publicly available under the Creative Commons public license Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) and can be freely used for research purposes only. [1]

The paper is structured as follows: Section 2 provides a brief literature review on task-oriented dialogues and their annotations. In Section 3, the MAIA dataset construction pipeline is presented, including the anonymization and annotation steps. The dataset is formally presented in Section 4, delving into the uniqueness of the dataset and its contributions to research. Existing AI-powered approaches for customer support chat such as Emotion Recognition in Conversations and Dialogue Evaluation are benchmarked in Section 5.

---

[1] `github.com/johndmendonca/MAIA-DQE`

2 Related Work

2.1 Task-oriented Dialogue Datasets

Perhaps the most well known open-source customer support datasets are TweetSumm (Feigenblat et al., 2021) and the Ubuntu Dialogue Corpus (Lowe et al., 2015). In both datasets, the language used is exclusively English. TweetSumm contains customer support interactions between customers and companies crawled from Twitter, whereas Ubuntu extracts its dialogues from the Ubuntu chat logs. The main difference between the Ubuntu dataset and TweetSum is the fact the former is constrained by the nature of the platform itself, typically resulting in limited turn interactions where the agent inevitably steers the customer to a dedicated costumer service chat platform. The Ubuntu dataset, similarly to MAIA, does not have this limitation and consists of live multi-turn dyadic conversations. However, unlike Ubuntu, the MAIA dataset contains customer support conversations of 4 different products and companies, where the agent is a representative of the company. This contrasts with Ubuntu, where the participant offering support is typically an experienced user without any official affiliation with Ubuntu. As such, the conversational dynamics between the two datasets are quite different, with the MAIA dataset showing more diverse emotions.

Other relevant public resources of task-oriented dialogue corpora include the MultiWoz and associated datasets (Budzianowski et al., 2018). These datasets are frequently used in the context of task-oriented dialogue, where an agent assists a customer in well defined tasks such as reservations. Unlike the MAIA dataset, the interactions are collected using English speaking crowdworkers, lacking representation of other languages. Additionally, the strict guidelines result in "sterile" and structured interactions that lack complexity known to real-world customer support interactions.

2.2 Dialogue Annotations

One of the most widely used dialogue benchmark datasets with emotion annotations is DailyDialog (Li et al., 2017), built from websites used to practice English and labelled with the six Ekman's basic emotions (Ekman, 1999). In the realm of customer support, Herzig et al. (2016) collected and annotated data in terms of emotions from two North America based customer support Twitter accounts. A particularity of this work is that a different set of emotion classes was used for the agent and customer. Furthermore, annotators were asked to indicate the intensity of each possible emotion, allowing for a multi-class setting.

With respect to quality annotations, the goal of most human annotation work is to evaluate dialogue systems or to validate proposed automated metrics. As such, two approaches are typically employed: annotators either interact with the system in a live setting and rate it, or evaluate existing responses given a context which was fed to the system. In the context of task-oriented dialogue, annotating Task Success (Walker et al., 1997), User Satisfaction and/or Emotion (Schmitt et al., 2012) are the norm. However, for open-domain dialogue, the focus has been mostly on annotating system responses on several notions of quality (See et al., 2019; Mehri and Eskenazi, 2020), since these dialogues are open in nature. To the best of our knowledge, this work is the first one to provide human judgements of customer support conversations with both task-oriented and open domain dialogue quality annotations at the turn and dialogue-level.

3 Processing and Annotations

3.1 Collection and anonymization

The conversations that compose this corpus are extracted from the original WMT22 Chat shared-task dataset (Farinha et al., 2022). It consists of dialogues obtained from companies that provide customer support and that gave written consent to use their data for research purposes [2]. This was achieved by using a mix of proprietary anonymization tools and human annotations was used to anonymize all PII (Personally Identifiable Information) from the data[3].

3.2 Annotations

The annotations were conducted by expert linguists in the given language. A single annotator for each language was used to fully annotate the dataset. Given its structure, we annotated the dataset along three dimensions: **Sentence level**: corresponding to a single message; **Turn level**: one or more sentences sent by one of the participants within a given time frame. **Dialogue level**: a succession of turns between the customer and agent denoting the full

---

[2]In accordance with the EU General Data Protection Regulation (GDPR).

[3]Additional information, including the anonymization tokens, are available in the original paper.

conversation. Considering dialogues are collaborative acts between speakers, we annotated data from both participants, customer and agent. This allowed us to evaluate the interaction as a whole and understand how one's action may impact the following response and how that affects the outcome of the conversation. A fully annotated dialogue is presented in Appendix B.

3.2.1 Sentence Level Evaluation

The metrics used to assess each sentence are as follows:

- **Correctness {0,1,2}**
- **Templated {0,1}**
- **Engagement {0,1}**

The **Correctness** metric was expressed resorting to three different scores measuring the sentence fluency. A score of 0 applies to a sentence indicated ungrammaticalities at several levels, both in terms of structure and in terms of orthography, originating a sentence that is difficult to understand. A score of 1 indicates that the analysed sentence contains minor mistakes but still remains fully understandable. A score of 2 was used when the sentence showed no mistakes and was fully understandable and coherent.

The **Templated** metric measured the type of sentence. For each sentence, a score of 0 was given for non-templated sentences, and a score of 1 for templated sentences. Note that by templated sentences we refer to predefined scripts used by customer support agents.

The **Engagement** metric was also expressed as one of two scores, measuring the level of engagement from both conversation parties. A score of 0 indicates a lack of engagement, whereas with a score 1 the participant was fully engaged in the conversation.

Besides the above-mentioned metrics, we also found to be reasonable to measure real emotions that usually go hand in hand within a customer support scenario. Following the previous strategy, the assessment was provided at a sentence-level, identifying the emotions conveyed by each sentence. The set of emotions used are as follows: **Happiness; Empathy; Neutral; Disappointment; Confusion; Frustration; Anger; and Anxiety**. We selected these emotions because upon analyzing the dataset we observed that these were the most common emotions displayed from a pool of several customer support emotions. With regards to empathy, it is a crucial emotion to analyze to measure agent performance. In terms of emotion annotation, and since a situation often triggers multiple emotions, annotators had the opportunity to select multiple emotions for a single sentence, ranking from the main emotion expressed to the others that are less evident. For example, a customer can be both disappointed and frustrated.

3.2.2 Turn Level Evaluation

The annotation process was designed to measure the interaction between participants within a dialogue. Since dialogues are a multi-tier architecture structure engineered not just around sentences but also around turns, it was necessary to account for these compositional properties. An analysis at the turn level allowed us to understand the overall mood and attitude of the turn-taker w.r.t what was previously stated by the other dialogue participant, at any given stage of the conversation. As a metric deeply dependent of the previously sentences, it is important to note that the initial turns were considered as non-evaluatable, since their function within the dialogue is to set the tone and the context that allow the newly started conversation to flow. The set of categories used for the turn taking evaluation were as follow:

- **Understanding {0,1}**
- **Sensibleness {0,1}**
- **Politeness {0,1}**
- **Interaction Quality [1,5]**

The category **Understanding** measured how well the participant was able to understand the message from the other dialogue participant, with a score of 0 meaning the understandability was somehow compromised, and the score 1 meaning understandability was reached.

Sensibleness measured the response appropriateness to what was previously stated by the other dialogue shareholder. A score of 0 means the response did not *follow* what was previously stated or requested, indicating that the current turn-taker ignored the conversation history. Conversely, a score of 1 indicates that the turn-taker acknowledged the conversation history and provided a suitable response.

Politeness measured the courtesy level of each participant towards one another. A score of 0 shows disrespect, discourtesy *inter alia* concerning the

Agreement (%)	Sentence				Turn				Dialogue	
	Emot	Corr	Temp	Enga	Unde	Sens	Poli	IQ	DC	TS
Full	72.39	81.45	76.10	71.24	88.98	92.12	98.36	51.97	90.00	30.00
Partial	23.06	17.39	23.90	28.76	11.02	7.88	1.64	41.73	10.00	60.00
None	4.56	1.16	0.00	0.00	0.00	0.00	0.00	6.30	0.00	10.00

Table 2: Observed agreement as a percentage of the total annotations per category between 3 annotators on a subset of PT_PT-3. Annotation types are abbreviated for brevity.

Metric	DE-1	DE-2	PT_BR-2	PT_PT-3	PT_BR-4	Total
# Dialogues	370	65	113	21	43	612
# Sentences	12,169	3,823	6,673	815	1,480	24,960
# Tokens	359,030	101,001	166,049	22,656	41,410	690,146
Avg. Sen/Dial	32	58	59	38	34	40
Avg. Token/Sen	29	26	24	27	28	27

Table 3: Statistical information of the MAIA dataset. The number of tokens includes tokens from Source and MT.

remaining participant; score 1 shows the participant was at worst civil and respectful.

The category **Interaction Quality (IQ)** was adapted from Schmitt and Ultes (2015) and scores the turn-taker disposition regarding the previous turn issued by the other dialogue part-taker. This category metric ranges from 1 to 5. With a score of 1, the turn-taker found the previous response to be extremely unsatisfactory; score 2, unsatisfactory; score 3, somewhat unsatisfactory; score 4, somewhat satisfactory; score 5, satisfactory.

With the above metrics we were able to have a better outlook of the different types of customers and agents, distinguishing behaviour and attitude patterns within a customer support dialogue.

3.2.3 Dialogue Level Evaluation

Lastly, we focused on the full dialogue, measuring the conversation in terms of:

- **Dropped Conversation {0,1}**
- **Task Success [1,5]**

Dropped Conversation responds to the questions: *"Was the conversation terminated without a conclusion?"* and/or *"Was the conversation dropped?"*. A score 0 means the conversation reached its end. Conversely, a score of 1 means a dropped conversation, i.e., the conversation did not reach its end, implying that the issue was not resolved.

Task Success dwells with the success of the interaction. This category responds to the following question: *"Was the agent able to fulfil the customer's request?"* The dialogue success was measured according to the following scores:

- A score of 1 means the agent failed to understand and fulfil the customer's request;
- A score of 2 means the agent understood the request but failed to satisfy it in any way;
- A score of 3 means the agent understood the customer's request and either partially satisfies the request or provided information on how the request can be fulfilled;
- A score of 4 means the agent understood and satisfied the customer request, but provided more information than what the customer requested or took unnecessary turns before meeting the request;
- A score of 5 means the agent understood and satisfied the customer request completely and efficiently.

3.3 Interannotator agreement (IAA)

Since all annotators were also fluent in European Portuguese (PT-PT), we conducted a trial annotation using 10 dialogues of the corresponding subset to gauge inter-annotator agreement between the annotations. The observed agreement is presented in Table 2 [4]. Of note, we observe that IQ and Task Success are the annotations that have the lowest agreement, which is expected given the highly subjective nature of these annotations and the fact they are annotated using a Likert Scale. By mapping these annotations to a binary decision (joining the last 2 and 3 ranks together for IQ and Task Success, respectively), the (full/partial) agreement increases to **(87.4/12.6)** and **(80.00/20.00)** for IQ and Task Success, respectively.

---

[4]Due to the class imbalance, regular IAA metrics such as Cohen's kappa (Cohen, 1960) are uninformative.

4 MAIA Dataset

4.1 Statistics

Annotation	Count
Correctness {0,1,2}	205 | 938 | 23,730
Templated {0,1}	18,174 | 6,602
Engagement {0,1}	315 | 23,712
Understanding {0,1}	136 | 9,470
Sensibleness {0,1}	127 | 9,478
Politeness {0,1}	345 | 9,390
IQ [1,5]	89 | 479 | 1,665 | 4,358 | 3,012
Dropped Conv. {0,1}	499 | 112
Task Success [1,5]	35 | 63 | 141 | 27 | 347

Table 4: Statistical information pertaining to the annotations of the MAIA dataset.
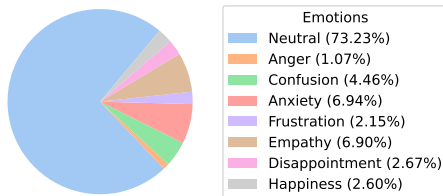

Figure 1: Emotion distribution of the MAIA dataset.

The dataset consists of a total of 612 dialogues, split into 5 subsets of different languages and/or companies (identified using a unique integer). Table 3 presents the statistical information of the dataset and corresponding subsets. Additional statistics on the quality annotations is presented in Table 4, with Figure 1 illustrating the emotion distribution.

4.2 Structure

Whilst the majority of dialogues follows a typical turn-taking approach, we find some instances where one of the participants breaks the flow of the conversation. This occurs when the next turn taker does not respond within an appropriate time frame (according to the other side). This is especially true at the end of the dialogues, where the customer terminates the conversation abruptly, irrespective of whether the issue was resolved. Additionally, these interactions are aided by automated system that responds on behalf of the agent: (1) when the customer doesn't reply within a given time frame, resulting in the system reminding the customer of the ongoing customer support interaction before terminating the conversation; (2) at the end of the

dialogues, requesting customer satisfaction survey and providing additional steps, if applicable.
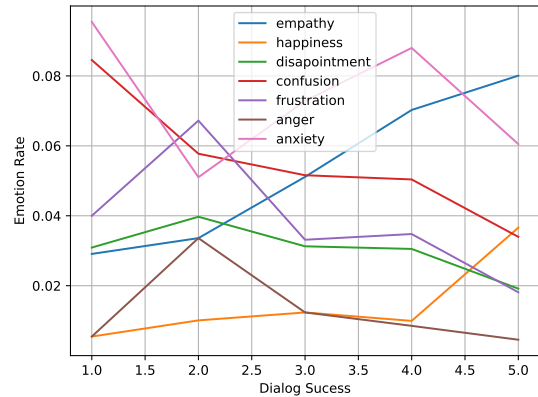
4.3 Observations and Discussion

Figure 2: Proportion of non-neutral Emotion Rates across all Dialogue Success levels

Emotion correlates with interaction quality and dialogue success. We hypothesise a positive correlation between emotion and dialogue success levels since the emotions of the interlocutors are related with the outcome of the experiment. This can be observed in Figure 2, where we note a rise in empathy and happiness, together with a decrease in negative emotions. Simultaneously, a positive correlation between emotion and Interaction Quality (IQ) should also be observed. For each turn, we mapped the emotions into a 3 class sentiment (-1,0,1) and report a Pearson and Spearman correlation of **0.4136** and **0.5494**, respectively.

Figure 3: Pairwise Pearson correlation matrix of sentence and turn level annotations.

Low dialogue subqualities result in loss of customer satisfaction. One assumes a decrease in IQ can be attributed, in part, to the occurrence of low quality responses by either participants. Looking at Figure 3, subqualities such as *Understanding*, *Sensibleness* and *Politeness* are in fact somewhat correlated with IQ. Engaging responses are uncorrelated with IQ, however. This is likely due to the
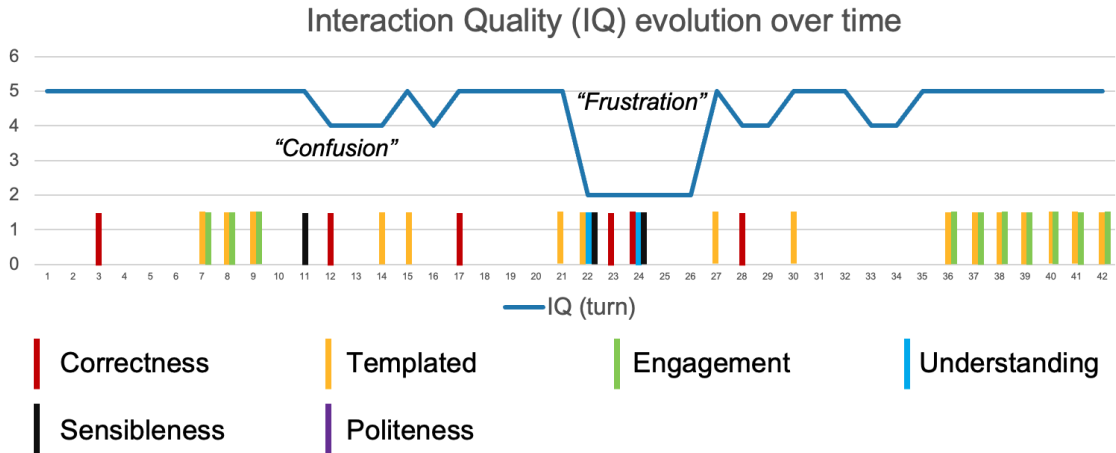
Figure 4: Evolution of the annotation *Interaction Quality* over a dialogue, together with relevant sentence and turn level annotations. Each spike in the lower portion of the figure denotes a negative annotation.

nature of the dialogue itself, which generally involves the agent dictating steps and/or terms and conditions pertaining to the product, which are verbatim of existing content.

Low quality interactions can be recovered successfully. Figure 4 presents a use case where a decrease of IQ is observed and rectified by the agent, resulting in a positive outcome: Around turn 21 we observe a large degradation in IQ which is paired with frustration. This is a result of the responses by the agent being templated and ineffective to solve the issue at hand. This is further exacerbated due to the lack of understanding between the participants, which is eventually resolved, increasing the quality of the interaction.

5 Benchmark Evaluation

Given the focus of the annotation work was on emotions and dialogue quality, in this section we evaluate existing mainstream approaches for emotion recognition and automatic dialogue evaluation.

5.1 Emotion Recognition in Conversations

State-of-the-art approaches for Emotion Recognition in Conversations (ERC) produce representations of each sentence using pretrained language models and then model the interactions between these representations with classification modules. Approaches such as leveraging conversational context or speaker specific modelling typically resort to architectures such as gated and graph neural networks (Poria et al., 2019).

5.1.1 Experiments

For our benchmark, we finetuned a pretrained Encoder model, more specifically XLM-RoBERTa (Conneau et al., 2020). We conducted train/dev/test splits at the dialogue level for each subset, employing a distribution of 70%/10%/20%, respectively, and ensuring the original distribution of emotion classes on all splits whenever possible. During training and evaluation, we used the source text while considering only the primary emotion labels, disregarding secondary emotion annotations. Performance is evaluated using Macro, Micro and individual emotion label F1 scores across all languages and the whole dataset. Additional training details are available in Appendix A.

5.1.2 Results

Results for this benchmark are presented in Table 5. We report a Macro-f1 score of 47.98 for the whole MAIA dataset. This result is within the performance of typical ERC models for other datasets that also have an imbalanced class distribution. The most represented *Neutral* class has a high F1 score across all subsets, heavily influencing the Micro-F1 score. Other well represented classes such as *Empathy* and *Anxiety* also have high F-scores, whereas minority classes have lower scores. In some subsets, individual emotion labels present very low to null F1 scores, again, a result of the class imbalance issues. In fact, due to the limited number of examples for these emotions in some subsets, a handful of missclassifications yield single digit F1 scores.

15

Subset	Macro-F1	Micro-F1	Emp	Hap	Disa	Conf	Frus	Ang	Anx	Neu
All	47.98	83.49	67.16	45.74	34.37	48.59	22.22	16.96	58.05	90.71
DE-1	48.26	82.55	72.70	39.60	41.13	42.00	10.50	31.22	59.31	89.60
DE-2	44.59	88.29	63.28	37.49	26.38	53.87	16.68	8.00	57.98	93.02
PT_BR-2	39.84	83.51	42.61	53.93	31.83	53.64	25.34	0	20.13	91.27
PT_PT-3	39.27	81.94	85.33	21.33	0	0.5917	31.34	0	0	91.05
PT-BR-4	30.50	77.51	47.14	0	23.07	52.78	26.63	0	6.67	87.74

Table 5: Emotion Recognition results for each subset and the full dataset. Results are an average of 5 runs.

	Model	Correctness	Templated	Engagement	Understanding	Sensibleness	Politeness	IQ
CST	VSP	**0.6361**	**0.6541**	0.4667	0.5112	0.4943	0.5091	**0.5307**
	NSP	0.5444	0.4645	**0.5083**	**0.5734**	**0.5831**	**0.5603**	0.4842
	ENG	0.5205	0.5795	0.4545	0.5374	0.5484	0.5510	0.4740
AGT	VSP	**0.7061**	**0.6073**	0.4601	0.4648	0.4973	0.5165	**0.5083**
	NSP	0.5850	0.4888	**0.5182**	**0.5657**	**0.5864**	**0.5821**	0.5029
	ENG	0.5443	0.5794	0.4503	0.5514	0.5548	0.5756	0.4742

Table 6: Balanced Accuracy Score of the binary subquality prediction for the MAIA dataset, from the point of view of the CST (customer-LANG) and AGT (agent-EN). Best results for each of them per subquality in **bold**.

5.2 Automatic Dialogue Evaluation

Most competitive metrics for turn-level dialogue evaluation leverage pretrained Encoder models that are finetuned using well-defined self-supervised tasks (Yeh et al., 2021; Zhang et al., 2021). These approaches generate synthetically negative samples from the original dialogue data, thereby circumventing limitations w.r.t the lack of quality annotated dialogues. However, it isn't clear these approaches extend to task-oriented dialogues and/or Multilingual models, since dialogue data is exclusively open-domain and in English. As such, the MAIA dataset can be used as a benchmark to study these characteristics.

5.2.1 Experiments

Similar to approaches mentioned above, we finetuned XLM-RoBERTa for **ENG** (Engagement) using the ENDEX data (Xu et al., 2022); and **VSP** (Valid Sentence Prediction) and **NSP** (Next Sentence Prediction) using self-supervised data generated from DailyDialog (Li et al., 2017). **VSP** is mostly concerned with the syntactic fluency of the response, which maps to *Correctness* and *Templated*; **NSP** evaluates textual entailment, which maps to *Understanding* and *Sensibleness*; Finally, since we have Engagement annotations, the evaluation of the ENG submetric is straightforward. The mapping between these submetrics and the remaining annotations is less obvious, but most evaluation frameworks that leverage these submetrics have shown positive correlations with quality aspects that do not map to the submetrics (Yeh et al., 2021).

For this task, we mapped existing sentence-level annotations to turn-level by selecting the minimum of the given turn. For simplicity, we report the Balanced Accuracy Score (BAS), which in this case corresponds to the average recall obtained on the positive (1) and negative (0) classes. The BAS for outputting a single class is 0.5. As such, we consider always outputting the majority class as the baseline. For Correctness, we considered a turn to be positive when all sentences have a score higher than 0; for IQ, only turns with a score of 4 or 5 are labelled positive. We indicate results for both languages, i.e, the context-response pairs from the point of view of the Customer (CST) (original language, with agent text translated) and the Agent (AGT) (in English, customer text translated). Note that, in this case, we conducted zero-shot inference on customer languages using models finetuned only on English data. Additional details available in A.

5.2.2 Results

For ease of reading, we aggregate the results of all subsets and report the BAS in Table 6. It is clear some models are best suited to predict only some subqualities. However, despite **ENG** being trained on engagement data, it underperforms NSP on the *Engagement* annotation. This may be related to the training data itself: *Engagement* in the context of open-domain dialogue is different than in customer support. Further, we observe that most models only slightly outperform just predicting the positive class. This means typical approaches for automatic subquality prediction are insufficient to adequately predict low quality responses on the MAIA dataset.

Comparing the results for AGT against CST we note that the trained models do not consistently outperform on a given language. This may indicate

finetuning a multilingual encoder with English dialogue data only achieves reasonable results in a multilingual setting. However, it is important to point out (1) that the agent converses in English; (2) the result that is most sensible to linguistic differences is **VSP** for *Correctness* (since it looks at the syntax), and here we see that the model underperforms for the other languages.

6 Conclusions

This paper presents a comprehensive emotion and dialogue quality annotation for the MAIA dataset, a collection of genuine bilingual customer support conversations. All in all, we annotate 612 dialogues amounting to over 24k sentences. Besides allowing for an opportunity to study the dynamics of Machine Translation aided customer support conversations, it also provides a novel opportunity to benchmark and explore applications of existing and future NLP models applied to dialogue.

Results on the different benchmarks indicate there is still room for improving existing models. LLMs such as GPT-4 (OpenAI, 2023) show impressive classification and generation capabilities, and may prove useful in augmenting existing customer support datasets to new languages and tasks. These in turn can be used to build data-driven classifiers or end-to-end conversational agents that are robust to new languages and domains.

Limitations

Perhaps the main limitation of this work concerns the lack of several annotators for each subset. Even with well defined guidelines, individual biases may affect the annotations, especially for dialogue quality as it is highly subjective (Smith et al., 2022). By having several annotators evaluate the conversations, one could've leveraged "the wisdom of the crowd", but this approach also comes with its own limitations (Jain, 2010). Ideally we would've employed several expert annotators, but were only able to recruit a single expert for each language. In any case, we conducted a trial annotation where all annotators participated and report moderate to strong agreement on a subset of the dataset.

Another limitation pertains to the dataset itself. Despite being structured and evaluated as a dyadic interaction, the actual conversations may not follow this structure. For instance, whenever one of the participants takes to long to respond, the other may follow-up on its original turn with a reminder.

Given we do not have access to this temporal information, these sentences were lumped together into a single turn. Also pertaining to metadata information is the lack of the original customer support guidelines. This makes the *Templated* annotation a subjective observation from the point of view of the customer. However, since we are framing this annotation from a quality perspective, we believe our annotation accurately reflects the perception of quality from the P.O.V of the customer.

Ethics Statement

This work leverages real world dialogues. A comprehensive anonymization process was conducted to ensure all PII were removed, in accordance with EU's GDPR. The annotations were conducted exclusively by highly-educated European Portuguese, which were paid a fair wage according to local costs of living. Despite being native speakers of the languages they evaluated, one might argue notions of quality are strongly tied to the culture and not the language. As such, they may not accurately represent other groups.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Gonçalo Oliveira, Patrícia Ferreira, Daniel Martins, Catarina Silva, and Ana Alves. 2022. A brief survey of textual dialogue corpora. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1264–1274, Marseille, France. European Language Resources Association.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. 2016. Classifying emotions in customer support dialogues in social media. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 64–73, Los Angeles. Association for Computational Linguistics.

Radhika Jain. 2010. Investigation of governance mechanisms for crowdsourcing initiatives. In *Sustainable IT Collaboration Around the Globe. 16th Americas Conference on Information Systems, AMCIS 2010, Lima, Peru, August 12-15, 2010*, page 557. Association for Information Systems.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

John Mendonça, Alon Lavie, and Isabel Trancoso. 2023. Towards multilingual automatic open-domain dialogue evaluation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czechia. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: Assessing the quality of ongoing sporiaken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, 74:12–36.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3369–3373, Istanbul, Turkey. European Language Resources Association (ELRA).

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nischal Reddy Chandra, and Nanyun Peng. 2022. EnDex: Evaluation of dialogue engagingness at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4884–4893, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, João Sedoc, L. F. D'Haro, Rafael E. Banchs, and Alexander I. Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *ArXiv*, abs/2111.02110.

A   Experimental Setup

All experiments used XLM-RoBERTa-large downloaded from the Transformers library by Hugging Face [5]. All parameters were trained/finetuned using Adam optimizer (Kingma and Ba, 2015) and a single Quadro RTX 6000 24GB GPU for all experiments was used.

A.1   Emotion Recognition in Conversations

Training and Hyperparameters   We trained XLM-R with the cross-entropy loss with logits. An initial learning rate of 1e-5 and 5e-5 was used for the encoder and the classification head, respectively, with a layer-wise decay rate of 0.95 after each training epoch for the encoder, which was frozen for the first epoch. The batch size was set to 4 and Gradient clipping to 1.0. Early stopping was used to terminate training if there was no improvement after 5 consecutive epochs on the validation set over macro-F1, for a maximum of 10 epochs. The best performing model on the validation set was selected for testing.

A.2   Dialogue Evaluation

Processing   For the dialogue data preprocessing we used spaCy [6]. In this paper, we followed the approach used by Phy et al. (2020) and initially proposed by Sinha et al. (2020). In detail, we train models trained to differentiate between positive samples and synthetic negative samples from DailyDialog (Li et al., 2017): For the **VSP** model, **Positive** samples are perturbed by randomly applying one of the following: (1) no perturbation, (2) punctuation removal, (3) stop-word removal. **Negative** samples are generated by randomly applying one of the following rules: (1) word reorder (shuffling the ordering of the words); (2) word-drop; and (3) word-repeat (randomly repeating words). For the **NSP** model, **positive** responses are drawn directly from the dialog; **negative** responses are randomly selected and a token coverage test discards semantically similar sentences. All responses are processed using the positive-sample heuristic used by VSP. The **ENG** model was trained directly on the 80k split with negative sampled data of the ENDEX dataset (Xu et al., 2022).

Training and Hyperparameters   All models were obtained following the recipe from Mendonça

---

et al. (2023). In detail, a token representing the speaker was added for each turn, and a history length of 3 turns was used. We applied a regression head consisting of a 2-layer MLP with a hidden size of 1024 and a hyperbolic tangent function as activation for prediction. A learning rate of 3e-6 for 3 epochs using a batch size of 16 was used. Evaluation was conducted every 10,000 steps. The best performing model on the evaluation set was selected for testing.

B   Example Dialogue

Text		Emo	Sentence			Turn				Dialogue	
SRC	**MT**		**Cor**	**Tem**	**Eng**	**Und**	**Sen**	**IQ**	**Pol**	**DC**	**TS**
Oii, fui cobrada por um plano q n estou usando, e estou solicitando retorno	Oii, I was charged for a plan q n am using, and I am requesting return	Frustratiom	0	0	NA	NA	NA	NA	1		
Good day :)	Tenha um bom dia!	Neutral	2	1	1						
Thanks for reaching out and I'm sorry for any confusion here!	Obrigado por entrar em contato e desculpe-me por qualquer confusão aqui!	Neutral	2	1	1						
#PRS_ORG# memberships renew each month unless you request a cancellation via one of our agents.	As assinaturas #PRS_ORG# são renovadas todos os meses, a menos que você solicite um cancelamento através de um de nossos agentes.	Neutral	2	1	1						
It seems that our system did not process a request from you before your cycle renewed which is why you were charged.	Parece que nosso sistema não processou uma solicitação de você antes da renovação do seu ciclo, e é por isso que você foi cobrado.	Neutral	2	0	1						
Unfortunately we are not able to refund #PRS_ORG# memberships retroactively.	Infelizmente, não podemos reembolsar as assinaturas da #PRS_ORG# retroativamente.	Neutral	2	0	1	1	1	5	1		
Our refund policy is stated in our Help Center here for further reference #URL#	Nossa política de reembolso está descrita em nosso centro de ajuda aqui para mais referências: #URL#	Neutral	2	1	1						
And how do I get the signature for someone else???	E tem como eu pssar a assinatura pra outra pessoa???	Frustration	1	0	1	1	1	2	0		
Pq meu dinheiro n é de graça pra ser gasto a toa	Pq meu dinheiro n é de graça pra ser gasto a toa	Frustration	0	0	1						
We apologize for the frustration here.	Lamentamos pelo inconveniente.	Neutral	2	1	1						
As an exception given the circumstances, I refunded your most recent membership charged, and I cancelled your membership immediately.	Como exceção, dadas as circunstâncias, reembolsei a sua assinatura mais recente cobrada e cancelei sua assinatura imediatamente.	Neutral	2	0	1						
You can expect this refund to arrive in 5-7 days depending on your bank/carrier, and you won't be charged again moving forward.	O reembolso estará disponível em 5 a 7 dias, dependendo do seu banco, e não haverá mais cobranças.	Neutral	2	0	1	1	1	5	1		
In the meantime, you can view the refunded charge on the billing page in your Account Settings.	Enquanto isso, você pode visualizar a cobrança reembolsada na página de faturamento nas suas Configurações da conta.	Neutral	2	0	1						
Please let me know if you have any other questions or if there is anything else that I can help with.	Se tiver outra dúvida ou se precisar de ajuda, é só avisar!	Neutral	2	1	1					1	
Muito obrigada!!	Thank you very much!	Neutral	2	0	1	1	1	5	1		
My pleasure.	O prazer é meu.	Neutral	2	0	1						
Is there anything else I can help with?	Posso ajudar com mais alguma coisa?	Neutral	2	1	1						
It seems like you're busy right now, so I'm going to close out the chat.	Parece que você está ocupado agora, então eu vou fechar o chat.	Neutral	2	1	1						5
If you have any other questions or want to get back in contact with us, you can do so here: #URL#	Se você tiver outras perguntas ou quiser entrar em contato conosco, pode fazê-lo aqui: #URL#	Neutral	2	1	1	1	1	5	1		
Have a great day!	Tenha um ótimo dia :)	Neutral	2	1	1						

Table 7: Example of a full dialogue extracted from PT_PT-3. The blue and red shaded rows correspond to turns belonging to the Customer and Agent, respectively.