Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 66–72
Marseille, 20 June 2022

Variation in the Expression and Annotation of Emotions: a Wizard of Oz Pilot Study

Sofie Labat[◇], Naomi Ackaert[◇], Thomas Demeester[♣] and Véronique Hoste[◇]
[◇]LT3, Language and Translation Technology Team, Ghent University, Belgium
[♣]T2K, Text-to-Knowledge Research Group, IDLab, Ghent University - imec, Belgium
{sofie.labat, naomi.ackaert, thomas.demeester, veronique.hoste}@ugent.be

Abstract

This pilot study employs the Wizard of Oz technique to collect a corpus of written human-computer conversations in the domain of customer service. The resulting dataset contains 192 conversations and is used to test three hypotheses related to the expression and annotation of emotions. First, we hypothesize that there is a discrepancy between the emotion annotations of the participant (the experiencer) and the annotations of our external annotator (the observer). Furthermore, we hypothesize that the personality of the participants has an influence on the emotions they expressed, and on the way they evaluated (annotated) these emotions. We found that for an external, trained annotator, not all emotion labels were equally easy to work with. We also noticed that the trained annotator had a tendency to opt for emotion labels that were more centered in the *valence-arousal* space, while participants made more 'extreme' annotations. For the second hypothesis, we discovered a positive correlation between the personality trait *extraversion* and the emotion dimensions *valence* and *dominance* in our sample. Finally, for the third premise, we observed a positive correlation between the internal-external agreement on emotion labels and the personality traits *conscientiousness* and *extraversion*. Our insights and findings will be used in future research to conduct a larger Wizard of Oz experiment.

Keywords: emotion analysis, Wizard of Oz study, conversational data collection, customer profiling, customer service

1. Introduction

Customer service (CS) delivery models are transforming due to recent technological advances (Deloitte Digital, 2021). Besides assisting human operators in their tasks, NLP techniques are increasingly implemented in autonomous conversational agents that can engage with clients on a 24/7 basis. To improve the quality of conversation, novel resources and methodologies are introduced to make human-computer interactions more personalized and empathic.

In this paper, we investigate variation in the expression and annotation of emotions during human-computer conversations. Insights in these types of variation will not only be helpful to craft more representative annotation frameworks, but they can also be used in the design of emotion detection systems. We present a pilot Wizard of Oz (WOZ) experiment that was conducted to study these variations. In a WOZ experiment, a *wizard* (the experimenter) pretends to be an autonomous conversational agent that interacts with the participants. Our experimental setup involved 16 voluntary participants that each had 12 successive conversations with the wizard. Each conversation was grounded in an event associated with a commercial sector (e-commerce, tourism, telecommunication) and was linked to a predefined sentiment trajectory along which the wizard tried to steer the conversation (e.g., *negative → positive*). The events and sentiment trajectories were kept consistent across participants, while we also tried to restrict the variation in responses of the wizard to a minimal. The conversations were afterwards annotated for emotions by both the participant and a trained annotator. Finally, we collected profiling information (age, gender, personality) on the participants.

The resulting dataset is also used to tentatively investigate three hypotheses. First, the annotation and subsequent prediction of emotions are notoriously difficult tasks due to the high degree of ambiguity that is involved. The fact that it is hard to obtain acceptable scores of inter-annotator agreement (IAA) on emotion annotations underscores this point (Schuff et al., 2017; De Bruyne et al., 2020; Troiano et al., 2021). **We thus hypothesize that not all emotions are equally easy to annotate by external annotators, as some might simply be expressed too implicitly**. Second, we regard emotions as dynamic attributes of the customer that can shift at each utterance in the conversation. Even though dialogue participants remain often in the same emotional state while exchanging turns, this can change if external stimuli are introduced (Poria et al., 2019). Emotions are therefore closely linked to (i) the event that happened prior to the conversation, and (ii) the response strategies the wizard applied. **We hypothesize that the effect of external stimuli on emotions differs across individuals depending on their personality.** We combine the two previous hypotheses in our final premise **by postulating that a participant's personality influences the annotator agreement he/she obtains with the external annotator.**

The remainder of this paper is structured as follows. Section 2 introduces the related research on emotion annotations and the Wizard of Oz technique. Section 3 describes the experimental setup of our study, while the resulting dataset is analyzed along three hypotheses in Section 4. Finally, Section 5 concludes this study with our main findings and suggestions for future research.

2. Related Research

Section 2.1 gives a concise overview of the different models used to capture emotions. It also focuses on IAA studies conducted for emotion annotation tasks, and links these studies to research on the possible causes of annotation disagreement. Section 2.2 introduces the WOZ technique and describes other studies that applied this technique.

2.1. Emotion Models and IAA

Emotions can be captured in two types of frameworks: categorical and dimensional models. Ekman (1992) introduced the most popular categorical model that consists of six emotions based on universal facial expressions: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*. This model was extended by Plutchik (1980) who added the primary emotions *anticipation* and *trust*. In recent years, researchers have realized that our ability to express and interpret emotions goes beyond a small set of basic emotions (Skerry and Saxe, 2015), which resulted in new datasets annotated along large taxonomies of categorical labels (Cowen and Keltner, 2017; Rashkin et al., 2019; Demszky et al., 2020).

Dimensional emotion models are less frequently used, even though, in contrast to categorical frameworks, they are not limited in the number of emotions they can capture (Canales and Martínez-Barco, 2014). Moreover, they can more easily be compared across different domains (Buechel and Hahn, 2016). Dimensional emotion annotations are made along two or three independent axes. The first dimension *valence* represents emotions on a displeasure-pleasure continuum; the second dimension *arousal* depicts the intensity of emotions on a passive-active continuum; the third (often omitted) dimension *dominance* portrays the degree of control over the affective state on a submissive-dominant scale (Mehrabian and Russell, 1974).

As emotion annotations are linked to a high degree of subjectivity and ambiguity, both categorical and dimensional models struggle to reach acceptable levels of inter-annotator agreement (Wood et al., 2018; De Bruyne et al., 2020). Moreover, the more fine-grained the annotation framework is, the lower the agreement amongst annotators becomes (Labat et al., under review). Some researchers have recently started to look at factors that potentially cause disagreement between annotators (Troiano et al., 2021). Our current study contributes to this line of work.

2.2. Wizard of Oz Study

The Wizard of Oz technique is mainly used to mimic human-robot interactions and to test hypotheses in that setting. Participants of a WOZ experiment interact with a wizard that pretends to be an autonomous computer system, but that in reality is partially/fully controlled by a human operator (Riek, 2012). Some WOZ studies involve prior knowledge of the participant, other studies apply a low level of deceit to elicit more natural responses. Since its introduction in the mid-80s, the technique is frequently used in interdisciplinary research on a variety of topics, such as the effect of politeness on learning outcomes (Wang et al., 2008), analysis of customer experiences (Wei and Le, 2018), diagnosis of mental health problems (Gratch et al., 2014), the successfulness of persuasion strategies (Adler et al., 2016), and the creation of data-driven dialogue systems (Budzianowski et al., 2018).

3. Experimental Design

To collect written conversational data, we designed an online interface in which participants acted in the role of customers and chatted with a wizard about events that occurred in a customer service setting. Our participants did not know that the so-called computer system they interacted with was actually fully controlled by the experimenter. As we collected profiling information, the experimental setup was submitted to and approved by the Ethics Committee of the faculty of Arts and Philosophy at Ghent University.[1]

3.1. Events and Sentiment Trajectories

All participants had 12 successive conversations with our wizard. Each of these conversations was grounded in a predefined event description. Descriptions are linked to a company that is active in Flanders, the Dutch-speaking community in Belgium, and that represents one of three economic sectors: Bol.com (e-commerce), Airbnb (tourism), and Telenet (telecom).

The events are further associated with one of four predefined sentiment trajectories: *positive* → *negative*, *negative* → *positive*, *neutral* → *negative*, *neutral* → *positive*. The sentiment trajectories were only visible to the experimenter who had to steer the conversation towards a given end sentiment. We decided to work with sentiment trajectories instead of emotion trajectories (e.g., *anger* → *admiration*) to give more conversational freedom to both the participant and the wizard.

In the Appendix, Figure 5 contains an example conversation, while Table 2 offers a detailed overview of the 12 event descriptions in which the conversations were grounded. Even though we worked with these 12 event descriptions for all participants, the order in which they were presented to the participants differed to avoid undesired sequential effects.

3.2. Response Strategies

The wizard tried to direct each conversation along a fixed sentiment trajectory. For example, positive emo-

[1] Participants could withdraw their participation up to 5 days after the experiment. The data records were anonymized in order to assure the privacy of the participants.

tions could be evoked by being helpful or showing empathy, while negative emotions were induced by being impolite, introducing repetitions, or answering beside the point. To remain as consistent as possible across different participants, we worked with standardized replies for eight response strategies that are typical in the domain of customer service: (i) *apology*, (ii) *cheerfulness*, (iii) *empathy*, (iv) *gratitude*, (v) *explanation*, (vi) *help offline*, (vii) *request information*, and (viii) *other* (Labat et al., 2020). We must, however, acknowledge that one can never fully control the participant's conversational output. As the wizard must reply at all times, its responses can slightly differ across participants. Nevertheless, responses are always in line with the given sentiment trajectory.

3.3. Emotion Annotations

Once all conversations were collected, both the participant and the external, trained annotator (the experimenter) proceeded to annotate emotions. Both parties were given a set of 15 emotions to label utterances: *admiration*, *amusement*, *anger*, *annoyance*, *approval*, *confusion*, *desire*, *disappointment*, *disapproval*, *disgust*, *fear*, *gratitude*, *joy*, *love*, *sadness*. An additional *neutral* category was introduced to label objective utterances. We composed the emotion taxonomy by combining a concise set of five emotions used for cross-domain comparisons (De Bruyne et al., 2020) with emotion labels that are frequent in the domain of customer service (Labat et al., under review). Besides categorical annotations, the experimenter also made dimensional annotations for *valence-arousal-dominance* (VAD) on three 5-point scales. While annotators were not restricted in the number of emotion labels they could assign to a given utterance, only one score per utterance could be made for each VAD dimension.

3.4. Participants and Profiling Information

This pilot study was conducted with 16 participants. Participants had to be older than 18 years, have a stable internet connection, and speak Dutch as a mother tongue. Given the small scale of our experiment, participants were recruited through word-of-mouth advertising and participated on a voluntary basis without remuneration. The experiments were conducted from mid-March to mid-April 2021.

After the WOZ session, participants were asked to fill out their customer profile. We collected three types of profiling information: year of birth, biological gender, and personality. To collect personality types, participants filled in a Dutch version of the IPIP-NEO-120 test (Johnson, 2014). The test measures personality across five dimensions: *neuroticism*, *extraversion*, *openness*, *agreeableness*, and *conscientiousness*. For each dimension, 24 questions are answered with one of five possible answers ranging from *very inaccurate* to *very accurate*.

4. Corpus Analysis along Hypotheses

The resulting dataset of our experiment consists of 192 conversations that contain 3,089 utterances in total. 1,684 of these utterances are written by the wizard, while the remaining 1,405 are written by the participants and have been annotated for emotions. In Section 4.1, we introduce the proposed hypotheses with respect to the variables of our corpus, and their interrelationship. Afterwards, we analyze our three hypotheses in chronological order in Sections 4.2, 4.3, and 4.4.

4.1. Hypotheses

We are interested in three hypotheses:

- **H1**: Not all emotions are equally easy to annotate by external annotators. Some experienced emotions might simply be expressed too implicitly.
- **H2**: The effect of external stimuli (such as event and responses) differs across individuals, depending on their personality.
- **H3**: A participant's personality influences the level of agreement he/she achieves on emotion annotations with the external annotator.

To better explain these hypotheses in the context of our dataset, we created an interrelationship digraph in Figure 1. In this digraph and our hypotheses, we distinguish internal annotations A_i (made by the person who experienced an affective state) from external annotations A_e (made by a trained annotator who has only access to the written utterance). For the first hypothesis, we will look at the agreement between internal and external annotations to investigate which emotion labels cause disagreement when the point of view of the annotator shifts. The second hypothesis focuses on the relationship between personality (part of the customer profile P) and emotion annotations A_e. Finally, the third hypothesis studies the relationship between personality (part of P) and the internal-external annotator agreement (A_i-A_e).

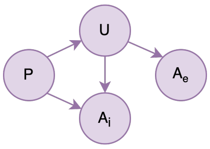


Figure 1: Interrelationship digraph of the variables customer profile (P), expressed utterances (U), internal annotations (A_i), and external annotations (A_e).

4.2. Internal versus External Annotations

For H1, we explore the extent to which the participants and the external annotator agree on the task of emotion labelling. Since we are especially interested in the agreement on each emotion label, we calculate Cohen's κ for individual labels. We also take into account the frequencies with which labels were assigned to utterances, as lower levels of agreement will usually be obtained for more infrequent labels. The results of this analysis are shown in Table 1. From this table, we extract five emotion labels that occur frequently, but that still have relatively low κ scores: *confusion*, *desire*, *anger*, *disapproval*, and *approval*. These five emotions are examined in more detail in Figure 2.

Emotion	C($\boldsymbol{A}_i$)	C($\boldsymbol{A}_e$)	**Cohen's $\boldsymbol{\kappa}$**
Gratitude	184	215	0.565
Neutral	487	502	0.480
Joy	66	47	0.401
Annoyance	281	324	0.384
Disappoint.	72	38	0.340
Confusion	102	38	0.182
Admiration	16	6	0.177
Desire	58	139	0.165
Anger	60	10	0.161
Disapproval	97	153	0.126
Amusement	17	5	0.086
Disgust	23	6	0.063
Approval	48	67	0.013
Sadness	6	2	-0.002
Fear	9	0	NA
Love	0	0	NA

Table 1: The table shows the number of internal annotations (C(A_i)), the number of external annotations (C(A_e)), and the IAA of internal and external annotators (Cohen's κ) for each emotion.

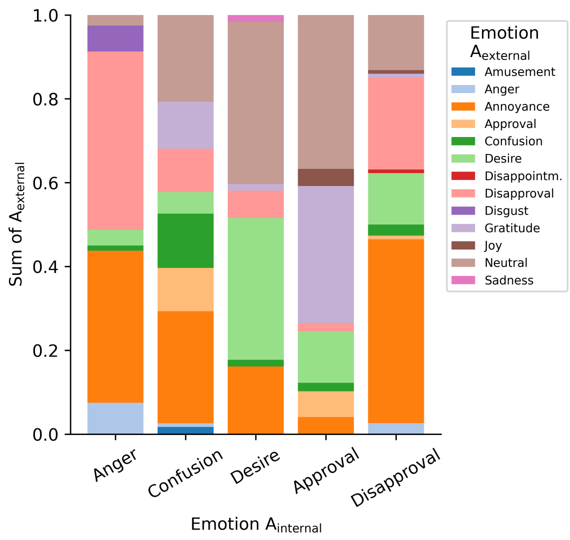


Figure 2: For all internal annotations ($A_{internal}$) with a given emotion category, this figure plots the emotions that the external annotator ($A_{external}$) picked to label the same instances.

Figure 2 investigates the extent to which the external annotator agreed with the internal annotators. If disagreement occurred, we explore which other emotion labels the external annotator selected. Although labels selected by the external annotator do not always correspond to the internal annotations, we find that the two groups of annotations are often semantically related. For the more extreme emotion *anger*, we see that the external annotator prefers similar labels that are, however, more centered in the *valence-arousal* (VA) space (see Labat et al. (under review) for a detailed overview). Similarly, internal annotations with more 'moderate' labels (e.g., *approval*, *confusion*, *desire*) are often confused with *neutral*. Finally, the internal emotion *confusion* seems particularly daunting to label, as it is often labelled with both negative and positive emotions by the external annotator.

4.3. Personality and Emotion Expressions

For our second hypothesis, we explore how variation in expressed emotions A_e can be linked to personality (part of P). As our study was conducted with a small group of 16 participants, we only aim to tentatively investigate whether some correlations can be found. We decided to work with the external annotations A_e for this hypothesis, as (i) they are consistently made by the same annotator across different experimental trials, and (ii) they contain VAD scores.

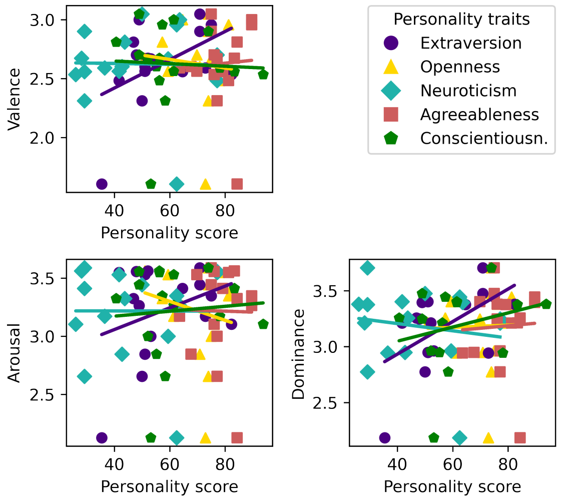


Figure 3: Three scatterplots with regression lines that plot the correlations between each personality dimension and the variables *valence*, *arousal*, *dominance*.

We plotted three scatterplots in Figure 3. Each plot shows the relation between the independent variable personality and one of the VAD dimensions. Since personality traits were captured on five dimensions, the colour and form of the markers distinguish between these five traits. For each personality dimension, we plotted one point per participant. To this end, we used a single score per VAD dimension, which was obtained by averaging all scores for a given dimension across the different utterances of a participant. For each personality dimension, we also plotted a linear regression line to better visualize possible correlations. In most cases, there seems to be no correlation between personality and emotional dimensions. There are, however, two exceptions to this trend, as the personality trait *extraversion* correlates positively with valence (Pearson's correlation coefficient r-value = 0.480, p-value = 0.060) and dominance (r-value = 0.551, p-value = 0.027^{*}). This implies that in our small sample, extraverted participants were more positive and dominant in the emotions they expressed than their introverted counterparts.

4.4. Personality and Emotion Annotations

The third and final hypothesis states that participants' personality (part of P) not only influences the emotions they express, but also the way in which they evaluate their own emotional states through annotations. To study variation in annotations across participants, we looked at the level of agreement between their emotion annotations (A_i) and the annotations of our external, trained annotator (A_e), since the latter made consistent annotations across the different participants. We used Krippendorff's α (Krippendorff, 2004) with Jaccard distance to calculate internal-external annotator agreement on the emotion labelling task.

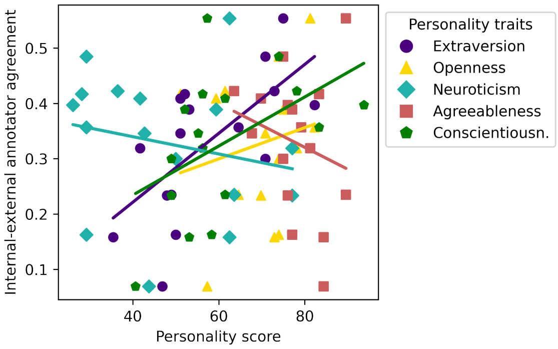


Figure 4: Scatterplot with regression lines showing the correlation between each personality dimension and the internal-external annotator agreement.

Figure 4 plots the relation between the independent variables personality traits and the dependent variable internal-external annotator agreement. As in Figure 3, the colour and form of the markers represent the different personality traits. For each personality trait, we also plotted a linear regression line to visualize possible correlations. We find that there exists a strong positive correlation between the personality trait *extraversion* and the annotator agreement (r-value = 0.655, p-value = 0.006^{**}). Moreover, we notice a moderate positive correlation between the personality trait *conscientiousness* and the annotator agreement (r-value = 0.484, p-value = 0.058). This means that in our sample, participants who are more outgoing or conscientious achieve higher agreement with the standard emotion annotations of our external annotator. We are unknown to the exact causes of this correlation, as multiple other variables may also play a role. The positive effect on annotator agreement could, for example, also be caused by the fact that (i) these participants lexicalize their emotions more strongly, or that (ii) their personality corresponds better to the personality of our external annotator. More research is needed to see whether these findings hold for a larger sample size and for other external annotators with different personalities.

5. Conclusion

In this paper, we presented a WOZ experiment that was conducted to investigate variation in both the annotation and expression of emotions during human-machine conversations in the domain of customer service. We found that some emotion classes are more easy to label in written chat conversations than others. Moreover, in contrast to the internal annotations, our external annotator often opted for emotion labels that were less extreme in their *valence* and *arousal*. This finding is interesting for the design of annotation guidelines in the domain of CS, as it is crucial to detect negative emotions in time before they become too extreme. For the link between personality and the expression of emotions, we discovered that the personality trait *extraversion* correlated positively with both *valence* and *dominance* in our sample. Finally, as for the relation between personality and internal-external annotator agreement, we observed that the personality traits *extraversion* and *conscientiousness* correlated positively with annotator agreement. Given the promising results of this study, we will apply our insights and findings to conduct a similar Wizard of Oz experiment on a larger group of participants.

6. Acknowledgements

This research received funding from the Flemish Government under the Research Program Artificial Intelligence (174U01222) and from the Research Foundation Flanders (FWO-Vlaanderen) with grant number 1S96322N. We would also like to thank the anonymous reviewers for their valuable and constructive feedback.

7. Bibliographical References

Adler, R. F., Iacobelli, F., and Gutstein, Y. (2016). Are you convinced? A Wizard of Oz study to test emotional vs. rational persuasion strategies in dialogues. *Computers in Human Behavior*, 57:75–81.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (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. Association for Computational Linguistics.

Buechel, S. and Hahn, U. (2016). Emotion Analysis as a Regression Problem — Dimensional Models and Their Implications on Emotion Representation and

Metrical Evaluation. In *Proceedings of the Twenty-Second European Conference on Artificial Intelligence*, ECAI'16, page 1114–1122, NLD. IOS Press.

Canales, L. and Martínez-Barco, P. (2014). Emotion Detection from text: A Survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43. Association for Computational Linguistics.

Cowen, A. S. and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

De Bruyne, L., De Clercq, O., and Hoste, V. (2020). An Emotional Mess! Deciding on a Framework for Building a Dutch Emotion-Annotated Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1643–1651. European Language Resources Association.

Deloitte Digital. (2021). From cost center to experience hub. Tapping the potential of customer service to help drive business growth.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., and Morency, L.-P. (2014). The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128. European Language Resources Association (ELRA).

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51:78–89.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 2nd edition.

Labat, S., Demeester, T., and Hoste, V. (2020). Guidelines for annotating fine-grained emotion trajectories in customer service dialogues (version 1.0). Technical report, Ghent University.

Labat, S., Demeester, T., and Hoste, V. (under review). EmoTwiCS: a corpus for modelling emotion trajectories in Dutch customer service dialogues on Twitter. *Language Resources and Evaluation*.

Mehrabian, A. and Russell, J. A. (1974). *An Approach to Environmental Psychology*. The MIT Press.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Robert Plutchik et al., editors, *Theories of Emotion*, pages 3–33. Academic Press.

Poria, S., Majumder, N., Mihalcea, R., and Hovy, E. (2019). Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, 7:100943–100953.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics.

Riek, L. D. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *J. Hum.-Robot Interact.*, 1(1):119–136, jul.

Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (2017). Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23. Association for Computational Linguistics.

Skerry, A. E. and Saxe, R. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology*, 25(15):1945–1954.

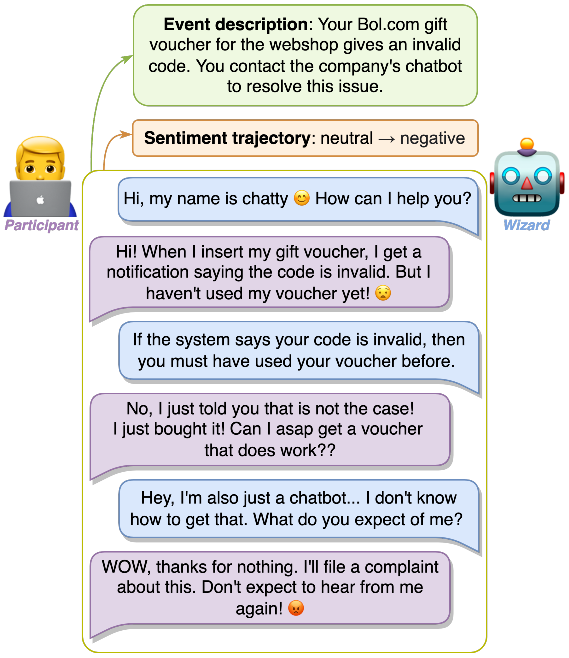
Troiano, E., Padó, S., and Klinger, R. (2021). Emotion Ratings: How Intensity, Annotation Confidence and Agreements are Entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49. Association for Computational Linguistics.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., and Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, 66(2):98–112.

Wei, Y. and Le, T. (2018). Using the Wizard-of-Oz Method for Exploring Deep Customer Experience Preferences. In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pages 1–8.

Wood, I., McCrae, J. P., Andryushechkin, V., and Buitelaar, P. (2018). A Comparison Of Emotion Annotation Schemes And A New Annotated Data Set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1197–1202. European Language Resources Association (ELRA).

Appendix



Event description: Your Bol.com gift voucher for the webshop gives an invalid code. You contact the company's chatbot to resolve this issue.

Sentiment trajectory: neutral → negative

Hi, my name is chatty 😊 How can I help you?

Hi! When I insert my gift voucher, I get a notification saying the code is invalid. But I haven't used my voucher yet! 😟

If the system says your code is invalid, then you must have used your voucher before.

No, I just told you that is not the case! I just bought it! Can I asap get a voucher that does work??

Hey, I'm also just a chatbot... I don't know how to get that. What do you expect of me?

WOW, thanks for nothing. I'll file a complaint about this. Don't expect to hear from me again! 😡

Figure 5: Example conversation to illustrate our experimental setup. Although the conversations in our dataset are in Dutch, this example is in English so that non-native Dutch speakers can also understand it.

	positive → negative	*negative → positive*	*neutral → negative*	*neutral → positive*
Bol.com	C thanks B for speedy package delivery.	C complains about undelivered product and bad service.	Gift voucher gives an invalid code.	C wants to return headphones that arrived damaged.
Airbnb	C thanks B for great service.	Host cancelled stay, C asks for sanctions.	C forgot phone charger in the accommodation.	C needs to cancel stay due to quarantine.
Telenet	C thanks B for listening to suggestion.	C missed promotion due to bad client service.	Digicorder records wrong show.	C wants to change subscription due to lack of mobile data.

Table 2: Event descriptions and corresponding sentiment trajectories in which the 12 conversations are grounded. *C* stands for customer, while *B* stands for the (chat)bot that is in reality operated by a human experimenter.