An Emotional Journey: Detecting Emotion Trajectories in Dutch Customer Service Dialogues

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

Abstract

The ability to track fine-grained emotions in customer service dialogues has many real-world applications, but has not been studied extensively. This paper measures the potential of prediction models on that task, based on a real-world dataset of Dutch Twitter conversations in the domain of customer service. We find that modeling emotion trajectories has a small, but measurable benefit compared to predictions based on isolated turns. The models used in our study are shown to generalize well to different companies and economic sectors.[1]

1 Introduction

While emotion recognition in conversations (ERC) has recently become a popular task in NLP (Poria et al., 2019b), its application potential to real-life business-related settings remains understudied. Our research focuses on applying ERC to the domain of customer service (CS), as it can be used to model customer satisfaction, reduce churns, prioritize clients, and detect emotional shifts in clients throughout CS interactions. Since the provision of customer service is gaining ground in both public and private chat channels, timely delivering high-quality assistance is crucial in mitigating the effects of negative word-of-mouth (van Noort and Willemsen, 2012) and creating relational bonds between customers and brands (Deloitte Digital, 2020).

As emotion recognition is often implemented on 'artificial', open-domain conversations (Busso et al., 2008; Li et al., 2017), we worked on real-world, domain-specific data that is more imbalanced and noisy. Moreover, we are the firsts to tackle the ERC task in Dutch dialogues. To these ends, we annotated *emotion layers* in a Dutch subset of 9,489 conversations from the Twitter corpus introduced by Hadifar et al. (2021), which we called EmoTwiCS ('Emotions in CS interactions on Twitter') (Labat et al., 2022b).[2] These emotion layers function as building blocks for *emotion trajectories*, a term emphasizing that emotions are dynamic attributes that can shift at each customer turn in the conversation.

We report classification effectiveness for six prediction tasks (focusing on cause, response strategies, subjectivity, valence, arousal, and emotion clusters). Besides subjectivity prediction which is applied to the conversation level, the five other tasks are run on isolated turns. To investigate the portability of our trained models to future data and other companies or sectors, we introduce three well-chosen train-test segmentation scenarios. We then zoom in on emotions and hypothesize that they follow a trajectory throughout conversations, whereby the operator tries to help the customer, thus deflecting negative emotions. To investigate whether knowledge about recurring emotion transitions may be useful for emotion prediction, we apply a Conditional Random Field (CRF; Lafferty et al., 2001) to the sequence of user turn encodings from a conversation, to make a joint prediction for the emotions in the conversation. We observe a weak, but consistently positive effect with respect to the isolated turn baselines in support of that premise.

2 Related work

Although emotion detection has often been applied to tweets (Mohammad et al., 2018) and chat logs (Ma et al., 2005), the context-aware detection of emotions throughout conversations is a relatively recent development in NLP. State-of-the-art results for emotion detection on isolated texts are achieved by fine-tuning large pretrained language

[*] Both authors contributed equally.

[1] Dataset and code are available at https://github.com/SofieLabat/EmoTwiCS-data and https://github.com/hadifar/DutchEmotionDetection, respectively.

[2] We refer to Labat et al. (2022b) for a detailed inter-annotator study and data analysis on EmoTwiCS.

models. For Dutch, there currently exist two such models named BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020), although cross-lingual language models such as XLM (Conneau et al., 2019) can also be applied to Dutch texts.

In contrast to these ‘vanilla’ emotion detection systems, recent work on ERC models additional information such as the conversational context, the temporal order of turns, and interlocutor-specific attributes (Poria et al., 2019b). There exist two approaches for ERC: we either view it as a sequence labeling task, or we predict emotions for a turn given the previous (and, in some variants, future) utterances. The latter approach was first addressed by recurrence-based models such as LSTMs (Poria et al., 2017), conversational memory networks (Hazarika et al., 2018), and attentive RNNs (Majumder et al., 2019). Afterwards, graph-based (Ghosal et al., 2019; Shen et al., 2021) and knowledge-enriched transformer models (Zhong et al., 2019; Zhu et al., 2021) were also investigated. The sequence labeling approach was introduced by Wang et al. (2020) who used information about the emotional consistency in conversations. His model combines a global context encoder (transformer) with an individual context encoder (LSTM) into a CRF layer to jointly predict emotions for all utterances. Guibon et al. (2021) implemented ERC in a few-shot learning sequence labeling problem. In our second experimental setup, we also tackle emotion detection as a sequence labeling task.

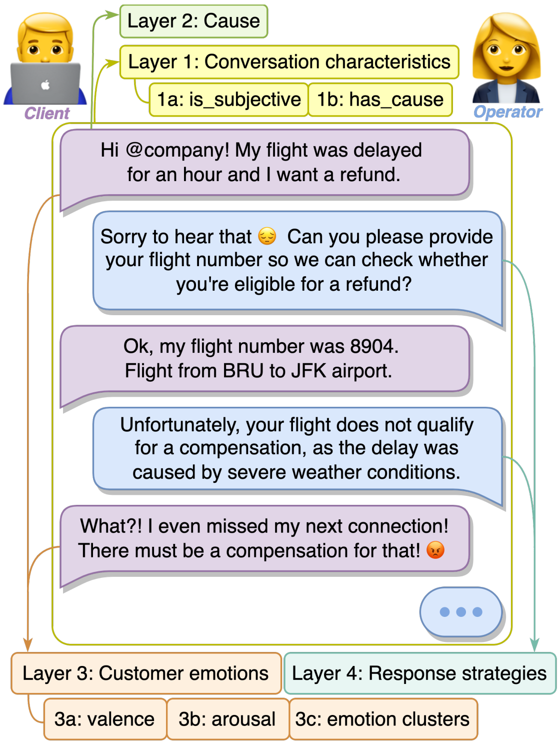
All but the two previously mentioned models are trained on publicly released datasets in English containing open-domain conversations (Busso et al., 2008; Poria et al., 2019a). There is only one small Dutch dataset (Vaassen et al., 2012) with 11 conversations and emotions rated on Leary’s Rose (Leary, 1957), a dimensional framework with two axes representing the degree of control and agreeableness. For ERC, the corpus is less suitable given its small size, low agreement, fixed events, and uncommon emotion model. Unlike standard sentiment analysis, the fine-grained task of ERC has not yet become commonplace in CS departments. To our knowledge, there exist only a few papers that apply ERC to CS (Herzig et al., 2016; Maslowski et al., 2017; Mundra et al., 2017; Guibon et al., 2021).

3 Experimental setup

After describing the EmoTwiCS corpus along with its prediction tasks (Section 3.1), our data segmentation strategies are introduced (Section 3.2), followed by the models and their implementation details (Section 3.3).

3.1 EmoTwiCS task descriptions

We rely on a newly annotated corpus of emotion layers called EmoTwiCS. The corpus contains 9,489 Dutch Twitter dialogues in the domain of customer service that were collected for three economic sectors: telecommunication, public transportation, and airline industry. The conversations were annotated for four emotion layers: conversation characteristics, cause, response strategies, and customer emotions. Figure 1 illustrates how the layers and sublayers are annotated on a conversation, while the remainder of this section provides more details about each of them.



Hi @company! My flight was delayed for an hour and I want a refund.

Sorry to hear that 😔 Can you please provide your flight number so we can check whether you're eligible for a refund?

Ok, my flight number was 8904. Flight from BRU to JFK airport.

Unfortunately, your flight does not qualify for a compensation, as the delay was caused by severe weather conditions.

What?! I even missed my next connection! There must be a compensation for that! 😡

Figure 1: An English mock-up conversation to illustrate how conversations are annotated in the EmoTwiCS corpus along four emotion layers (conversation characteristics, cause, emotions, and response strategies).

Experiments were conducted for the following classification tasks on our emotion layers:

Subjectivity – Detect whether the conversation is subjective, which is the case if at least one customer turn contains emotions. The task involves classifying the concatenation of all customer turns.

Cause – Recognize the event that triggered customers to start a conversation, as a multi-class classification problem with eight classes (see Appendix, Table 3). Since 99% of all causes reside in the first customer turn, we use that as our model's input.

Response strategies – Recognize one or more response strategies operators applied in their responses. This is a multi-label prediction task over eight response strategies (see Table 3). Response strategies have only been annotated for subjective conversations, but cannot be assumed absent in objective ones. We therefore restrict the prediction task to subjective conversations only, and we use single operator turns as input to our models.

Valence/Arousal – Given a customer turn, predict its valence/arousal score (integer from 1 to 5). While valence represents the sentiment of an emotional state ranging from very negative to very positive, arousal stands for the amount of activation an emotion elicits and ranges from calm to excited. We implement both as multi-class tasks.

Emotion clusters – Given a customer turn, predict the emotion clusters that it contains.[3] While annotators could assign multiple labels to a single turn, we find that only 6.5% of the customer turns received two or more annotations. We therefore convert the task from a multi-label to a multi-class detection task by assigning an order of importance to the labels.[4] To validate our heuristic, an external annotator extracted the most prominent emotions from 100 customer turns with multiple emotion annotations. We find that the annotator and our heuristic agree in 78% of the cases.

3.2 Data segmentation

To investigate the out-of-domain transferability of our models on the different prediction tasks, we work with three train-test segmentation strategies. The size of the different splits is given in Table 4 in the Appendix.

Temporal split – 80-20 train-test split based on the chronological order of the first tweet in each conversation, stratified over companies. This way, we want to demonstrate that prediction systems trained on past data generalize well to unseen, future data. The split is also used for the in-context classification experiments (see Section 4.2).

Company splits – As telecom is the most frequent sector in EmoTwiCS, we split the six companies within this sector into three train-test splits, with each four companies for training and two for testing. Averaging the prediction results over these splits gives an idea of the transferability of our models to new companies within the same sector.

Sector splits – Given that EmoTwiCS has data for three economic sectors, we create three corresponding train-test splits in which we train on two economic sectors and evaluate on the third one. Cross-validation over these splits will demonstrate the transfer potential of our models to new sectors.

3.3 Models and implementation details

For the experiments on isolated tweets, we select the following models: majority class baseline, Support Vector Machines (SVM; Cortes and Vapnik, 1995) with tf-idf features, BERTje (de Vries et al., 2019), RobBERT (Delobelle et al., 2020), and XLM (Conneau et al., 2019). For all pretrained transformer models, we use their publicly available 'base' versions and place a single feedforward layer on top to predict the classes. We only tune the learning rate and number of epochs on 15% of the train data for the temporal setup, and reuse the same hyperparameters for the company and sector setups. For the second set of experiments, we put a CRF layer on top of RobBERT to predict the emotion trajectories of conversations (Lample et al., 2016). Given a conversation and its sequence of turns, we first extract the turn embeddings by using the [CLS] token representations from the last layer of the pretrained language model, which are then given to a classifier to estimate emotion cluster probabilities. These probabilities are subsequently fed into a CRF layer to maximize valid emotion sequence predictions.

4 Results and Discussion

We present the results of our models for six classification tasks on isolated tweets across the different train-test setups in Section 4.1. In Section 4.2, we focus on the emotion trajectories, and cast the detection of emotion clusters as a context-aware sequence labeling task.The presented metrics are micro and weighted F1 scores (Table 1), as well as accuracy (Fig. 2) and individual class F1 (Table 2) for emotion trajectories.

4.1 Experiments on isolated tweets

The results of our experiments for the six classification tasks are shown in Table 1, while the standard

[3] We use the term *clusters* to remain consistent with the EmoTwiCS data description paper. In that paper, 28 emotion labels were grouped into 9 emotion clusters.

[4] Heuristic: Anger > Annoyance > Disappointment > Nervousness > Gratitude > Relief > Joy > Desire > Neutral.

		Subjectivity	Cause		Response strat.		Valence		Arousal		Emotion clusters	
Setup	**Model**	**F1**$_{micro}$	**F1**$_w$	**F1**$_{micro}$	**F1**$_w$	**F1**$_{micro}$	**F1**$_w$	**F1**$_{micro}$	**F1**$_w$	**F1**$_{micro}$	**F1**$_w$	**F1**$_{micro}$
Temporal	Majority-class	55.4	29.6	46.6	23.2	41.1	38.2	54.3	48.7	63.0	34.9	51.4
	SVM (tf-idf)	75.4	62.7	64.5	79.8	80.5	58.7	61.8	67.4	69.4	66.5	68.3
	BERTje	82.0	70.0	70.4	**87.6**	**87.7**	65.6	65.2	74.2	74.9	71.6	72.0
	RobBERT	**83.4**	**71.1**	**71.7**	86.9	87.1	67.8	67.7	74.0	74.6	**72.8**	**73.7**
	XLM	83.4	70.9	71.6	87.5	87.6	**68.1**	**68.0**	**74.4**	**75.2**	72.7	73.4
Company	RobBERT	**83.0**	71.1	71.4	**84.4**	**84.8**	**66.7**	**67.0**	**73.1**	**74.5**	**71.2**	**72.7**
	XLM	76.3	**71.3**	**71.5**	80.9	81.9	65.7	66.4	72.8	74.1	68.4	71.4
Sector	RobBERT	**83.0**	61.6	**64.0**	**84.6**	**85.4**	65.6	65.7	**73.9**	**74.7**	**71.6**	**72.9**
	XLM	72.90	**63.3**	63.4	81.5	83.5	**65.8**	**66.0**	72.8	73.7	70.6	72.0

Table 1: Results for subjectivity, cause, response strategies, valence, arousal, and emotion clusters classification.

deviations on the results of the company and sector setups are reported in Table 5. In the temporal setup of Table 1, we see that the fine-tuned language models outperform the majority class and SVM baselines by a large margin. Upon comparing the two Dutch language models RobBERT and BERTje, we find that RobBERT outperforms BERTje on four tasks (subjectivity, cause, valence, and emotion clusters). Moreover, the multi-lingual XLM model also achieves good results: it is the best baseline for valence and arousal prediction, but achieves second-to-best scores on all other tasks. As for the company and sector setups, we report scores for the two best-performing systems from the temporal setup. We observe that the results for two latter setups are less than, but still very comparable to the temporal experiments. Our models thus generalize well to other companies within the same domain, and to other economic sectors. This generalizability across sectors is significantly less outspoken for cause detection, which illustrates that cause classes are often linked to a specific domain (e.g., *delay* for public transportation vs. *breakdown* for telecom).

4.2 Modeling emotion trajectories

We hypothesize that emotions follow recurring trajectories that reflect the attempts of the CS operator to mitigate negative customer emotions. This motivated our reformulation of the emotion clustering task as a sequence labeling task (see also Wang et al., 2020; Guibon et al., 2021), modeled with a CRF to make joint predictions for emotion clusters in the conversation. As we work with joint predictions, we test our hypothesis on the subset of subjective conversations with at least two customer turns. We focus on subjective conversations, as these contain a varied distribution of emotion clusters. Figure 2 plots the results of our experiment across the conversations with a given number of customer turns. We notice a weak, yet consistent trend in which the CRF model slightly outperforms the isolated turn predictions. There is no clear indication that this effect is stronger for longer conversations, although that is hard to measure due to the low number of longer conversations. The improved results of the CRF model are thus an indication that there is some signal in modelling the sequence of emotions, although not statistically significant, given the size of the test set.

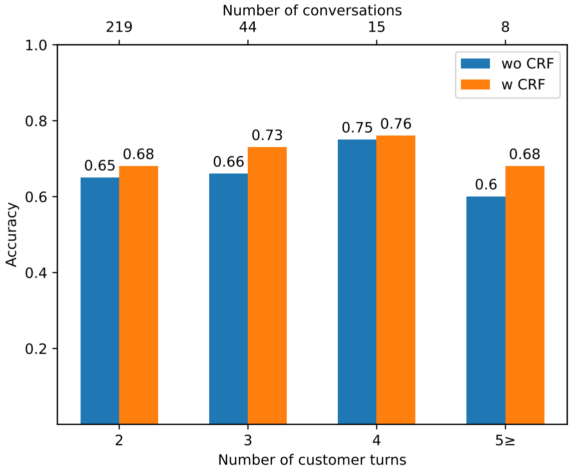


Figure 2: Emotion accuracy for all test conversations with at least two customer turns, calculated on the temporal setup, for the RobBERT baseline without CRF (wo CRF) vs. the one with the CRF (w CRF).

We further investigate the models' performance on individual emotion clusters in Table 2. We find that for some classes there is too little support leading to very low scores (e.g., Relief, Nervousness, and Desire). The F1 scores of both systems are generally higher for classes with more support. Nevertheless, the CRF model outperforms the baseline by a large margin on classes with lesser support (e.g., Anger, Disappointment, and Joy). Note that the high scores for Gratitude may be due to the rather standard lexicalization of it in the corpus.

Classes	w CRF	wo CRF	Support
Anger	0.40	0.04	45
Annoyance	0.53	0.58	182
Desire	0.11	0.0	17
Disappointment	0.45	0.0	36
Gratitude	0.92	0.90	123
Joy	0.51	0.32	35
Nervousness	0.00	0.00	11
Neutral	0.73	0.73	230
Relief	0.00	0.00	8

Table 2: Results (F1) for individual emotion clusters.

5 Conclusion

We presented the first experiments on a newly collected corpus of Dutch Twitter conversations annotated along four emotion layers. For our experiments on isolated tweets, we find that the best performance is obtained by fine-tuning pretrained language models such as RobBERT and XLM. We show that these two models transfer well across (i) time, (ii) companies within the same sectors, and (iii) across sectors. We also demonstrate that the detection of emotion clusters slightly benefits from knowledge about frequently occurring emotion trajectories, especially for classes with lower levels of support. In future research, we will extend our approach to model emotion trajectories for the purpose of real-time prediction (e.g., in chatbots), thus having access to past utterances only. We will also investigate emotion trajectories in longer conversations (e.g., on data collected through Wizard of Oz experiments (Labat et al., 2022a)) and focus on joint prediction tasks such as emotion-cause or emotion-response strategy extraction.

6 Acknowledgements

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

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Proceedings of EMNLP 2020*.

Deloitte Digital. 2020. Creating human connection at enterprise scale: What our research suggests about turning brands into bonds.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of EMNLP-IJCNLP 2019*.

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, and Chloé Clavel. 2021. Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks. In *Proceedings of EMNLP 2021*.

Amir Hadifar, Sofie Labat, Véronique Hoste, Chris Develder, and Thomas Demeester. 2021. A Million Tweets Are Worth a Few Points: Tuning Transformers for Customer Service Tasks. In *Proceedings of NAACL 2021*.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of NAACL*.

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 SIGDIAL 2016*.

Sofie Labat, Naomi Ackaert, Thomas Demeester, and Véronique Hoste. 2022a. Variation in the Expression and Annotation of Emotions: a Wizard of Oz Pilot Study. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC 2022*.

Sofie Labat, Thomas Demeester, and Véronique Hoste. 2022b. EmoTwiCS: a corpus for modelling emotion trajectories in Dutch customer service dialogues on Twitter. *Language Resources and Evaluation*. Accepted.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML 2001*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. *arXiv preprint arXiv:1603.01360*.

Timothy Leary. 1957. *Interpersonal Diagnosis of Personality: A Functional Theory and Methodology for Personality Evaluation*.

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 IJCNLP 2017*.

Chunling Ma, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion Estimation and Reasoning Based on Affective Textual Interaction. In *Proceedings of ACII 2005*.

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 AAAI 2019*.

Irina Maslowski, Delphine Lagarde, and Chloé Clavel. 2017. In-the-wild chatbot corpus: from opinion analysis to interaction problem detection. In *Proceedings of ICNLSSP 2017*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of SemEval 2018*.

Shreshtha Mundra, Anirban Sen, Manjira Sinha, Sandya Mannarswamy, Sandipan Dandapat, and Shourya Roy. 2017. Fine-Grained Emotion Detection in Contact Center Chat Utterances. In *Proceedings of PAKDD 2017*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of ACL*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of ACL 2019*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and E. Hovy. 2019b. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, 7:100943–100953.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of ACL-IJCNLP 2021*.

Frederik Vaassen, Jeroen Wauters, Frederik Van Broeckhoven, Maarten Van Overveldt, Walter Daelemans, and Koen Eneman. 2012. deLearyous: Training Interpersonal Communication Skills Using Unconstrained Text Input. In *Proceedings of ECGBL 2012*.

Guda van Noort and Lotte M. Willemsen. 2012. Online Damage Control: The Effects of Proactive Versus Reactive Webcare Interventions in Consumer-generated and Brand-generated Platforms. *Journal of Interactive Marketing*, 26(3):131–140.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized Emotion Recognition in Conversation as Sequence Tagging. In *Proceedings of SIGDIAL 2020*.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *Proceedings of EMNLP-IJCNLP 2019*.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *Proceedings of ACL 2021*.

Appendix

Task	Label set
Cause	employee service; product quality; delays and cancellations; breakdowns; product information; digital design inadequacies; environmental and consumer health; no cause / other.
Resp.	apology; cheerfulness; empathy; gratitude; explanation; help offline; request information; other
Emotion	anger; annoyance; desire; disappointment; gratitude; joy; nervousness; neutral; relief

Table 3: Label sets for the tasks cause, response strategies (Resp.), and emotion clusters.

	Subj-Cause		Response strat.		Cust. emotions	
Setup	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Temporal	7,587	1,902	6,477	1,489	10,272	2,443
Comp. 1	3,795	1,852	3,002	1,739	4,970	2,571
Comp. 2	3,670	1,977	2,962	1,779	4,802	2,739
Comp. 3	3,829	1,818	3,518	1,223	5,310	2,231
Sector 1	3,842	5,647	3,225	4,741	5,174	7,541
Sector 2	6,727	2,762	5,650	2,316	8,962	3,753
Sector 3	8,409	1,080	7,057	909	11,294	1,421

Table 4: Number of train-test instances for the classification tasks across the different segmentation strategies (temporal, company, sector). Subjectivity and cause are grouped together as they have the same number of train-test instances. The tag 'customer emotions' stands for valence, arousal and emotion clusters which are also grouped together for the same reason.

		Subjectivity	Cause		Response strat.		Valence		Arousal		Emotion clusters	
Setup	**Model**	s **F1**$_{micro}$	s **F1**$_{w}$	s **F1**$_{micro}$	s **F1**$_{w}$	s **F1**$_{micro}$	s **F1**$_{w}$	s **F1**$_{micro}$	s **F1**$_{w}$	s **F1**$_{micro}$	s **F1**$_{w}$	s **F1**$_{micro}$
Company	RobBERT	0.6	1.5	1.8	0.4	0.3	1.1	1.3	1.8	1.8	1.9	1.8
	XLM	9.7	2.0	1.8	1.2	1.2	1.4	1.4	2.1	2.1	2.5	2.2
Sector	RobBERT	1.4	3.2	2.1	4.4	4.3	1.1	1.2	1.6	1.8	1.9	1.9
	XLM	16.5	4.5	4.9	9.7	7.6	1.0	1.0	1.6	1.9	2.5	2.2

Table 5: Standard deviation (s) on the average performance reported in Table 1. Standard deviation is reported for those setups that have several train-test splits (viz., company and sector setups).