Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7049–7055
Marseille, 20-25 June 2022

Opinions in Interactions : New Annotations of the SEMAINE Database

Valentin Barriere[1,*], Slim Essid[2], Chloé Clavel[2]

[1]Joint Research Center, Via E. Fermi, 2749, I - 21027 Ispra, *valentn.barriere@ec.europa.eu*
[2]Télécom Paris, 19 place Marguerite Perey F-91120 Palaiseau, France, *chloe.clavel@telecom-paristech.fr*

Abstract

In this paper, we present the process we used in order to collect new annotations of opinions over the multimodal corpus SEMAINE composed of dyadic interactions. The dataset had already been annotated continuously in two affective dimensions related to the emotions: Valence and Arousal. We annotated the part of SEMAINE called *Solid SAL* composed of 79 interactions between a user and an operator playing the role of a virtual agent designed to engage a person in a sustained, emotionally colored conversation. We aligned the audio at the word level using the available high-quality manual transcriptions. The annotated dataset contains 5627 speech turns for a total of 73,944 words, corresponding to 6 hours 20 minutes of dyadic interactions. Each interaction has been labeled by three annotators at the speech turn level following a three-step process. This method allows us to obtain a precise annotation regarding the opinion of a speaker. We obtain thus a dataset dense in opinions, with more than 48% of the annotated speech turns containing at least one opinion. We then propose a new baseline for the detection of opinions in interactions improving slightly a state of the art model with RoBERTa embeddings. The obtained results on the database are promising with a F1-score at 0.72.

Keywords: Opinion, Multimodal Machine Learning, Interactions

1. Introduction

The way a human expresses affective phenomena such as an opinion, an emotion or a sentiment is very different depending on the media he uses. Whether it is a tweet, a phone call or a real-life interaction will definitely impact the manner to exchange information with another person. Certain kinds of phenomena are easier to detect using specific modalities and specific context. For example, there are many more studies on emotion recognition in speech than in text (Schuller et al., 2009), since often we communicate the emotions through para-linguistic like tone, pitch or emphasis. Similarly, the study of opinion phenomena was initially restricted to text (Irsoy and Cardie, 2014; Balahur et al., 2010) which has a clearer syntax, due to the complexity of the annotation of the phenomenon and the volatility of the linguistic structure in oral language.

There are numerous corpora for the study of multimodal affective phenomena whether they appear in human communication without interaction (Zadeh et al., 2018; Morency et al., 2011), with dyadic interactions (Ringeval et al., 2013; Busso et al., 2017) or with multiparty interactions (Poria et al., 2019a; Lotfian and Busso, 2018). Nevertheless, the majority of the available annotations are centered either on the emotions (McKeown et al., 2012) or on the sentiment (Zadeh et al., 2018). Here we propose to use multimodal data to detect the opinions of a speaker in a dyadic interaction. Some datasets are acted (Busso et al., 2008) since it is mandatory to create artificial situations in order to obtain enough emotion examples that would be too rare in real-life data, like fear (Busso et al., 2017). This being not a problem for opinions, we decided to annotate a corpus containing non-acted interactions. We chose to use the SEMAINE corpus of McKeown et al. (2012), which is a corpus of dyadic interactions between a spontaneous human-user and a human-agent playing a predefined role. The agent is intended to trigger an emotion in the user and engage him/her in an emotionally colored discussion. It is a dataset of spontaneous speech that was crafted to contain affective phenomena, hence it would be likely to be dense in terms of opinions.

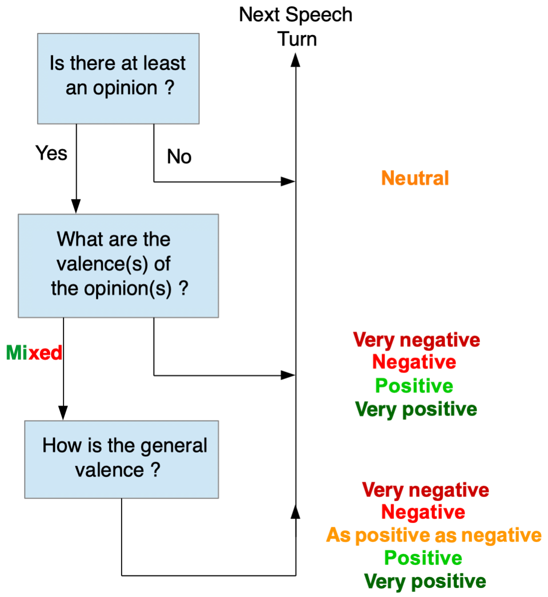


Figure 1: Block diagram of the annotation schema we used to annotate the corpus SEMAINE

[*] Work partially done at Télécom Paris

While the links between the different affective phenomena are pretty blurred (Munezero et al., 2014), the difference between emotion and opinion is pretty clear. The emotions are physiological when the opinions are constructed and prominently based on objective and/or subjective probabilities of information about a topic, and not necessarily charged in emotion (Munezero et al., 2014). Finally, SEMAINE already contains annotation in emotions regarding four affective dimensions, namely activity, expectation, power, and valence (Schuller et al., 2012). The dataset has been used recently to compare different models of Emotion Recognition in Conversation (ERC) (Poria et al., 2019b) and can be useful to investigate the links between emotions and opinions in a multimodal dyadic context.

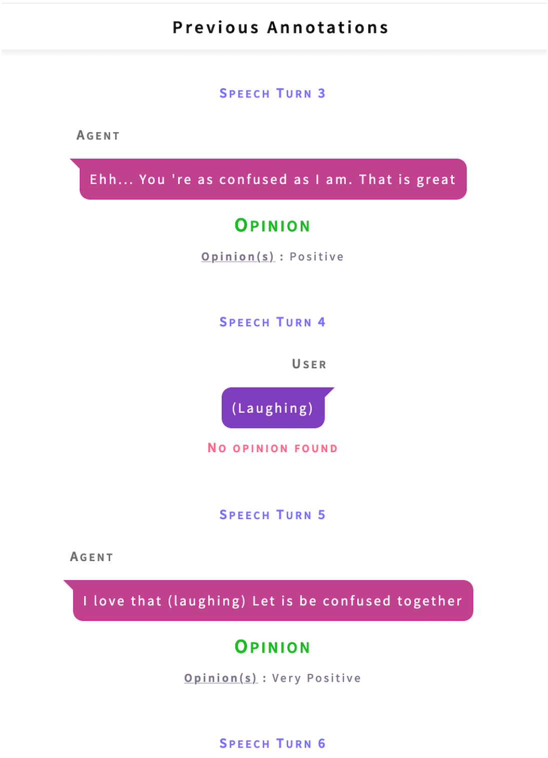
In our annotation process, we focused on different points: the distinction between an opinion and an emotion in order to get clean annotation and the use of the conversational context. In order to achieve the former goal, we helped the annotators to understand the difference by firstly, showing him/her examples of expected annotations in a preliminary learning phase before starting the annotation, following the protocol of Langlet et al. (2017) on another task and corpus, and secondly, we gave them feedback on their work after the annotation of an entire discussion. In order to achieve the latter goal, we used all the conversational context to detect opinions, which is different than most of the previous works on multimodal data that were focusing on annotating the emotion or sentiment of a speaker within its speech turn or in an utterance (Zadeh et al., 2018).

The rest of this paper is organized as follows. The schema and the web platform used to annotate long dyadic interactions in a simple manner will be presented in Section 2. In Section 3, we will present an analysis of the corpus and the annotations, whether on the inter-agreement or the ground truth obtained by majority vote. Finally, Section 4 will present the results of two baseline state-of-the-art models for Affective Phenomenon Recognition in Conversation (APReC), namely valence and arousal for the emotions and opinion.

This dataset is part of the Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE benchmark (Chapuis et al., 2020).

2. Annotation Collection

We collected the annotation using a schema and a platform firstly developed by Langlet et al. (2017) on another task and corpus that we improved in several ways: the schema we crafted is a series of questions designed to reduce the cognitive load of the annotator, we added the audio content to the textual content in order to give more information to the workers, we created a visualization platform allowing the annotator and ourselves to look at their annotations, we sent feedback to the annotators in order to help them improve the quality of their future annotations.



Previous Annotations

Speech Turn 3

Agent

Ehh... You 're as confused as I am. That is great

Opinion

Opinion(s) : Positive

Speech Turn 4

User

(Laughing)

No opinion found

Speech Turn 5

Agent

I love that (laughing) Let is be confused together

Opinion

Opinion(s) : Very Positive

Speech Turn 6

Figure 2: Example of an annotation's visualization

2.1. Annotation Schema

Creating an annotation schema composed of simple questions to ask is difficult when one is interested in subjective phenomena such as opinions. Since we wanted one label per speech turn, the speech turns containing at least 2 different valence opinions seemed to us an important case to define. For this, we have created an annotation process based on several questions in order to make the task easier for the annotator, reduce its cognitive load and thus allow him/her to be more attentive to the detection of the desired phenomena.

We've taken up the idea of a previous conversation history and visible annotations proposed by Langlet et al. (2017) for a preference annotation task on a negotiation corpus, while completely changing the annotation schema for our task. The annotation is made on a speech-turn-by-speech-turn basis, and always showing the contextual Adjacency Pair composed of the previous and current speech turns (Zhou et al., 2013; Barriere et al., 2018). The annotation scheme used on our corpus can be summarized by the block diagram of the figure 1.

2.2. Annotation Platform

The annotation platform was built to collect a huge quantity of data from native-English people all around

Speaker	# Token per session						Time per session (second)					
	μ	σ^2	**min**	**max**	**med**	Σ	μ	σ^2	**min**	**max**	**med**	Σ
Agent	306	140	67	670	290	24,174	101	55	32	363	88	8,006
User	630	343	96	1602	640	49,770	187	90	32	447	186	14,809
All	936	407	210	2220	953	73,944	289	118	75	638	291	22,815

(a) Lengths of sessions per time and tokens

Speaker	# Token per speech turn					Time per speech turn (s)				
	μ	σ^2	**min**	**max**	**med**	μ	σ^2	**min**	**max**	**med**
Agent	8,6	8,2	1	106	6	2,7	2,8	0,1	45,2	1,9
User	15,66	18,31	1	103	9	5,3	7,1	0,1	83,0	2,9
All	13,1	19,0	1	256	7	7,9	8,2	0,1	92,9	5,5

(b) Lengths of speech turns per time and tokens

Table 1: Lengths of speech turns and sessions of our corpus (μ: mean, σ: standard deviation, Σ: sum)

the world through the website CrowdFlower.[1] All the annotators experienced annotation tasks before and had been listed as trusted annotators by Figure Eight.

2.2.1. Three-step process

In order to help the annotator acclimate to the task, we divided the work into three parts: a first part presenting the annotation schema and the User Experience of the platform, a second part where the annotator is training over a small conversation composed of eight speech turns, and a third part with the real annotation task.

We have created a preliminary learning phase where the annotator can practice on a dummy conversation composed of 8 turns of speech. During this annotation phase, he will also be able to see the annotations of an ordinary annotator called ”Robin”. This phase allows the annotator to train and compare with results that we have deemed valid before. For example, we have often seen that new annotators find it hard to tell the difference between emotion and opinion, which are two separate phenomena (Munezero et al., 2014), and they would tag a speech turn where the speaker has a cheery tone as containing a positive opinion. Similarly, many annotators often label as opinion a question of the agent intended to trigger an affective reaction of the user, but is not in itself an opinion.

2.2.2. Focused on the Audio Content

We used the aligner of Ochshorn and Hawkins (2017) following a homemade two-step process in order to roughly get the timecodes of every speech turn transcribed manually, then we used those to get the exact timecodes of every token. The rate of non-aligned words compared to the total number of words in the corpus is 1.21% which seems reasonable for our application. When a word is not aligned, we use the timecodes of neighboring words in the same turn to approximate its associated timecodes. Using those timecodes, we improved the platform of Langlet et al. (2017) by adding the audio content to the initial textual content during the annotation task. The worker had to listen to the entire audio speech turn before annotating it.

The corpus contains a wide variety of para-linguistic annotations written in an unnormalized way (left to the discretion of the transcriber). As we only study oral and textual modalities, we have chosen to exclude all visual annotations (such as head movement, smiling, etc.). To do this, we only keep para-linguistic annotations provided by the manual transcript containing strings[2] referring to an oral phenomenon: ”*voice*”, ”*laugh*”, ”*whisper*”, ”*tone*”, ”*breath*”, ”*swallow*”, ”*sigh*”, ”*say*”, ”*said*”, ”*noise*”, ”*sniff*”, ”*giggl*”, ”*cackl*”, ”*speak*”, ”*tut*”. In this way, the preserved para-linguistic phenomena are the result of intonation, pronunciation, laughter and sound volume.

2.2.3. With feedback

The platform used for the annotation process allowed us to give feedback to the workers, helping them to improve the quality of their annotation. A web-page was created especially for the workers to visualize their annotations. Thus, it was possible to give feedback to the annotators by pointing out to them several of their badly annotated speech turns with the help of the visualization platform (see Figure 2).

3. Data Analysis

3.1. Database

This subsection presents the database we have annotated. In order to build our annotated database, we chose to use the subset of SEMAINE with manual transcripts. The SEMAINE database has 20 recordings with associated transcripts that are available: this corresponds to 80 sessions. One of the sessions being unusable because the user preferred to stop the experiment, **the final corpus is composed of 79 sessions**.

[1] now appen: *https://appen.com/solutions/platform-overview/*

[2] Words or stems

Speaker	α 4 classes						α 3 classes (Mixed → General)					
	μ	σ^2	min	max	med	Total	μ	σ^2	min	max	med	Total
Agent	53.5	18.6	15.1	95.0	55.7	∅	60.4	16.1	34.0	100	60.5	∅
User	45.4	17.7	-7.1 x	89.4	48.0	∅	54.9	15.0	13.3	85.9	56.0	∅
All	52.8	13.1	25.0	91.7	50.4	57.8	60.9	11.0	39.1	90.6	58.1	66.3

Table 2: Krippendorff's α per discussion using the label Mixed *versus* General (μ: mean, σ: standard deviation, **Total** is the value obtained by calculating α over all the discussions)

The reader can find in the table 1 statistics over the number of words and the speaking time of each speaker per speech turn and discussion.
We can observe in Table 1 that the user's speech turns are longer than the agent's ones. This can be explained by the fact that the agent is supposed to follow a script and is supposed to have a simple personality. Its role is to trigger emotions in the user, who is the interactor to express himself the most.

3.2. Inter-Annotator Agreement

For each discussion, we have 3 annotators per session. The use of an inter-annotator agreement measure is conventional in the validation of data that has been coded by different individuals. In the case of opinions where subjectivity is important, it is necessary to choose a representative measure of the agreement between the different annotators taking into account the variations in labeling due to randomness, to the type of data and to the different annotators.
We chose to use the α from Krippendorff (2013), a coefficient that is suitable for a comparison of more than 2 coders, unlike Cohen's κ and Scott's π (Lombard et al., 2002). Moreover, Krippendorff's α was designed for studies on coders analyzing content (Hallgren, 2012) (and not just a value on the Likert scale). The α is also the coefficient used by Zadeh et al. (2017) to calculate the inter-annotator agreement of sentiment annotations of the CMU-MOSI database. According to Krippendorff (2013), a α measurement of 66.7 is necessary to have an acceptable agreement.
We can see in Table 2 the differences in inter-annotator agreements taking into account the mixed label as a 4th class or using the general opinion annotations as a label. The columns of the aggregate values (mean, median, standard deviation, minimum and maximum) are calculated from the α of each discussion and the "Total" column is calculated over the entire corpus.
Given the low inter-annotator agreement using 4 classes, it is unreliable to use the *Mixed* label as a fourth class. The use of 4 classes gives a Krippendorff's α of 57.8, which is low compared to the value recommended by Krippendorff. Nevertheless, this is relative because we work on opinion in speech which is a subjective phenomenon, difficult to quantify even for expert annotators (Read and Carroll, 2012). For this reason, we reduce the annotation space to 3 classes using the prominent opinion label to obtain: Negative, Positive and Other. The Other category regroups the speech turns without any opinion or with a general valence as positive as negative. We check the coherence of this process using Krippendorff's α and then use the aggregated annotations. It allows us to increase our α to a reasonable value of 66.3 over the entire corpus. In Table 2, the reader can find the values of the α coefficients per discussion, computed using 4 classes (Positive, Negative, Mixed and Neutral) or 3 classes (Positive, Negative and Other). In the case of 3 classes, the Mixed label is replaced by it's associated General Valence label.
The value of the α is higher for the agent's speech turns than for the user's speech turns: since the agent is played by an actor with a specific role, one can expect to have clearer opinions that are more easily identifiable by annotators.

Speaker	# Opinions by Disc.		Opinions by ST (%)		
	μ	σ^2	Neg.	Pos.	Oth.
Agent	15.22	7.71	22.84	21.14	56.02
User	17.90	8.98	18.47	33.75	47.78
All	33.12	15.14	20,66	27.42	51.92

Table 3: Opinion per speech turn obtained after aggregation using majority vote (Oth. groups Neutral and As positive as negative)

3.3. Aggregated Opinions

While some works preferred to keep the information of great variance between the annotators (Dang et al., 2018) and incorporate it into their models, we remained on a traditional approach. A majority vote is one way to obtain values that are close to reality. To do this, we decided to aggregate our annotations using a majority vote like Wöllmer et al. (2013). Majority voting consists of taking the value given by the majority of annotators. In the case where we have 3 annotators, the majority is obtained when there are at least 2 similar answers. In the particular case where each of the annotators gave a different answer and because this case happens rarely, we used the class "*neutral*" to be consistent with the whole conversation .
We calculated the α from Krippendorff (2013) to obtain an inter-annotator agreement of 66.3, for 3 annotators per discussion. This α is relatively small but quite acceptable for a task as subjective and difficult as the annotation of opinion. By comparison, (Zadeh et al., 2017) got a value of 77 on CMU-MOSI, but on non-

Agent role	Opinions in Agent's ST (%) Neg.	Pos.	Oth.	Opinions in User's ST (%) Neg.	Pos.	Oth.
Poppy	3.73	34.66	61.1	7.54	42.04	50.42
Prudence	7.70	29.33	62.96	9.16	38.85	51.99
Obadiah	47.41	8.25	44.34	27.70	25.93	46.38
Spike	32.16	6.73	61.11	29.30	19.68	51.02

Table 4: Opinion labels of the speech turns regarding the agent's identity (ST = speech turn, Oth. groups Neutral and As positive as negative)

interactional data so without the ambiguity brought by the dialogical context. In addition, in CMU-MOSI speech was already segmented while our segmentation in speech turns does not allow one to cut long tirades, which increases the possibilities of confusion in the annotation. Finally, our annotators use only audio and not video. These points greatly facilitate the annotation task of Zadeh et al. (2017).

Opinions are a frequent phenomenon in this corpus even though it was created for the study of emotions, most of the annotations obtained with the majority vote could have been neutral opinions but it is quite the opposite. It can be seen that, on the whole corpus, about 48% of speech turns contain opinions: 20.7% in negative opinions and 27.4% in positive opinions. The statistical analysis of the ground-truth annotations obtained by majority vote is visible in Table 3.

One particularity of SEMAINE is that the agents are intended to trigger an emotion in the user, it is therefore interesting to look at the opinions of both the speakers according to the type of agent. We can see in Table 4 that the user and the agent align the valence of their opinions. Poppy and Prudence are more positive, then Obadiah is gloomy and Spike is aggressive, there is a clear pattern when ranking the positive and negative percentages of opinions in the speech turns of the agents. It is interesting to note that Obadiah has slightly more speech turns with negative opinions than Spike. This is due to the fact that Obadiah is complaining a lot, while Spike is trying to trigger the user's emotion by being mean. This singularity is not visible in the user's behavior, where the percentage of positive is higher when talking to Obadiah than when talking with Spike (25.93 vs 19.68).

Emotions and Opinions Due to the nature of the SEMAINE database and its already existing emotion-related annotations, we thought that it would be interesting to look for cross dimension relations. We reported on Table 5 the average means and standard deviations of the emotion-related annotations for each opinion class, as well as the Pearson Correlation between the emotion-related annotations and the opinion annotations. Finally, we also computed a 2/3D visualization of the annotations in Valence and Arousal that are emotion-related, with their associated opinion annotations (Figure 3).

Emotion	μ (σ^2) Neg.	Pos.	Oth.	*Corr*
Valence	-3.6 (7.0)	18.7 (4.3)	11.9 (5.9)	28.9
Arousal	-11.5 (6.4)	-6.1 (5.8)	-7.9 (5.5)	7.4
Dominance	39.0 (5.2)	46.7 (3.2)	39.0 (5.2)	12.4
Surprise	33.8 (0.9)	31.4 (0.9)	32.7 (1.1)	-8.7

Table 5: Statistics between the emotion-related and the opinion annotations (*Corr* is the Pearson Correlation)

4. Baseline

In this section, we introduce a baseline based on an ERC (Emotion Recognition in Conversation) state-of-the-art model. We use a variant of the DialogueRNN of Majumder et al. (2018) since it aims to detect an affective phenomenon in a conversation and is state-of-the-art on the SEMAINE dataset.

In order to use the emotions and opinion labels, we used a subset of SEMAINE of 72 sessions, 5210 speech turns and 68,043 tokens, which has the annotations from AVEC and our annotations. We kept the same documents Majumder et al. (2018) used in the training and testing sets.

We further improve that state-of-the-art model, using the high-quality manual transcripts we were able to extract new textual features with the RoBERTa model (Liu et al., 2019) at the granularity of a speech turn. We chose to improve the quality of the embeddings for our task because opinions phenomena are more centered on the text than the emotions are. Furthermore, since our task is atypical, we validated this new model on ERC, out-performing the state-of-the-art in ERC on the SEMAINE dataset (see Table 6), before using the new model on our task. The version of the DialogueRNN we used is the one that worked the best for Majumder et al. (2018), namely the bidirectional version with attention and listener state update. The network was trained for 600 epochs, with a dropout of 0.3 and a ℓ_2 regularization weight of 1e-4. We trained the model 5 times, reporting average performance.

Our model using the RoBERTa embeddings improves greatly the results on the emotion recognition task on Valence and Arousal, either regarding the MAE or the Pearson coefficient. The best results on opinion were obtained using a bimodal model using the textual and the audio modalities.[3] Although the task is different,

[3] We did not use the video since our dataset was annotated

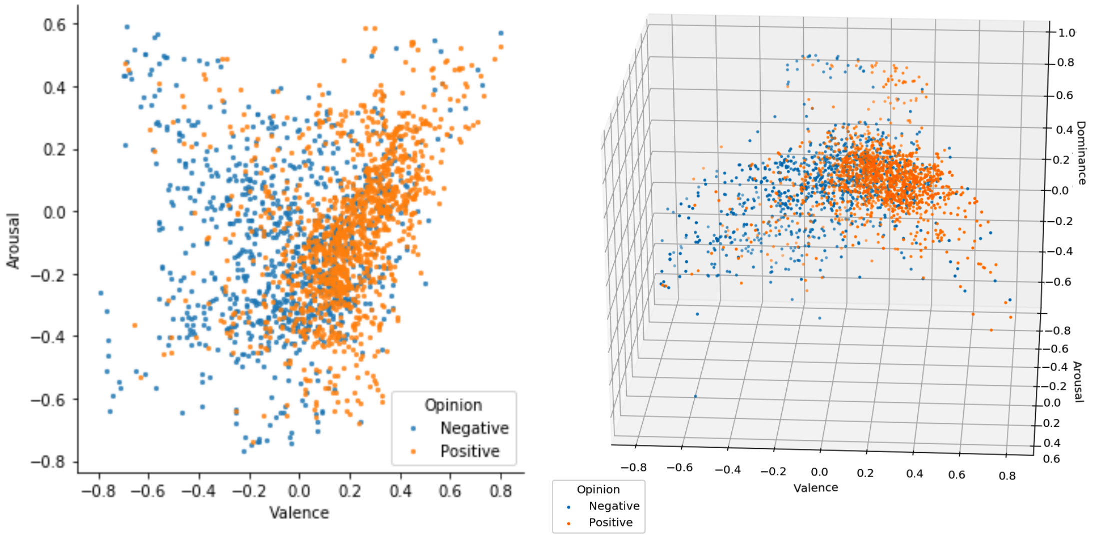


Figure 3: Visualizations in 2D and 3D of the Valence, Arousal and Dominance values of the samples annotated with positive and negative opinions. Each point represents a speech turn.

Method	Valence		Arousal		Opinion	
	MAE	r	MAE	r	F1	Acc
DialogueRNN	0.171	0.37	0.164	0.60	49.56	51.31
Our model	**0.132**	**0.76**	**0.154**	**0.71**	**72.08**	**72.20**

Table 6: Baseline results on different tasks (MAE = Mean Absolute Error, r = Pearson correlation coefficient)

this improvement over the state-of-the-art for Valence and Arousal can let us think that the DialogueRNN with RoBERTa embeddings is a strong baseline for an Affective Phenomenon Recognition in Conversation (APReC) task.

Nevertheless, the contribution of this paper remains in a new dataset annotation, not in using different embeddings with an existing model, hence we will not analyze the results of the baseline model.

5. Conclusion

In conclusion, we collected annotations in opinions per speech turn on the SEMAINE corpus. We aligned the audio with the text word by word using handmade transcriptions containing a lot of information such as punctuation and para-linguistic phenomena. The corpus is non-symmetrical because the agent and the user have different roles, the agent is used to trigger an emotion in the user. In addition, the corpus is rich in opinion since there is at least one opinion on 48.08% of the speech turns. Finally, this corpus can be used with the continuous emotional annotations of the AVEC-2012 challenge of Schuller et al. (2012).

6. Bibliographical References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010). Sentiment analysis in the news. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pages 2216–2220.

Barriere, V., Clavel, C., and Essid, S. (2018). Attitude Classification in Adjacency Pairs of a Human-Agent Interaction with Hidden Conditional Random Fields. In *ICASSP*.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Busso, C., Parthasarathy, S., Burmania, A., Abdelwahab, M., Sadoughi, N., and Provost, E. M. (2017). MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Chapuis, E., Colombo, P., Manica, M., Labeau, M., and Clavel, C. (2020). Hierarchical Pre-training for Sequence Labelling in Spoken Dialog. In *EMNLP*, pages 2636–2648.

Dang, T., Sethu, V., and Ambikairajah, E. (2018). Dynamic Multi-Rater Gaussian Mixture Regression Incorporating Temporal Dependencies of Emotion Uncertainty Using Kalman Filters. In *ICASSP*, pages 4929–4933.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1):23–34.

Irsoy, O. and Cardie, C. (2014). Opinion mining with deep recurrent neural networks. In *Proceedings of*

using audio and text

the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 720–728.

Krippendorff, K. (2013). Content Analysis: An Introduction to Its Methodology. In *Content Analysis: An Introduction to Its Methodology*.

Langlet, C., Duplessis, G. D., and Clavel, C. (2017). A web-based platform for annotating sentiment-related phenomena in human-agent conversations. In *IVA*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. (1).

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability.

Lotfian, R. and Busso, C. (2018). Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings. *Transactions on Affective Computing, IEEE*, XX(X):1–14.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2018). DialogueRNN: An Attentive RNN for Emotion Detection in Conversations.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schröder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Morency, L.-p., Mihalcea, R., and Doshi, P. (2011). Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI-11)*, pages 169–176.

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111.

Ochshorn, R. M. and Hawkins, M. (2017). Gentle forced aligner.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019a). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*, pages 527–536.

Poria, S., Majumder, N., Mihalcea, R., and Hovy, E. (2019b). Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. pages 1–10.

Read, J. and Carroll, J. (2012). Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, 46(3):421–447.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 312–315.

Schuller, B., Valstar, M., Eyben, F., Cowie, R., and Pantic, M. (2012). AVEC 2012 - The continuous audio/visual emotion challenge. In *ICMI'12 - Proceedings of the ACM International Conference on Multimodal Interaction*, pages 449–456.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2013). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. *Proceedings of ACL*, pages 2236–2246.

Zhou, Y., Scherer, S., Devault, D., Gratch, J., Stratou, G., Morency, L.-p., and Cassell, J. (2013). Multimodal Prediction of Psychological Disorders: Learning Verbal and Nonverbal Commonalities in Adjacency Pairs. *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, pages 160–169.