Intents Classification for Neural Text Generation

Idrissa Konkobo *
ENSAE Paris
idrissa.konkobo@ensae.fr

Sorelle loveline Methafe *
ENSAE Paris
sorelleloveline.methafe@ensae.fr

Abstract

The hype around OpenAI's ChatGPT has more than ever sparked interest in AI-based bots where labeling and classification of utterances are a centerpiece in order to improve user experience. Broadly, Dialogue Acts (DA) and Emotion/Sentiment (E/S) tasks are identified through sequence labeling systems that are trained in a supervised manner. In this work, we propose four encoder-decoder models to learn generic representations adapted to the spoken dialog, which we evaluate on six datasets of different sizes of a benchmark called Sequence labellIng Evaluation benChmark fOr spoken laNguagE benchmark (SILICONE). Designed models are represented with either a hierarchical encoder or non-hierarchical encoders both based on pre-trained transformers (BERT/XLNet). We notice the failure of the models to learn some datasets due to their inherent properties but in general, the BERT-GRU architecture is the best model regarding accuracy.

1 Introduction

In recent years, conversational agents and chatbots have become increasingly popular for a variety of applications, from customer service to mental health support. These systems rely heavily on natural language processing (NLP) techniques to understand user inputs and generate appropriate responses. However, to create engaging and effective interactions, it is not enough to simply understand the literal meaning of the user's input. Emotion and dialog act recognition are crucial components of effective communication that must be taken into account.

Emotion recognition enables the chatbot to identify the user's emotional state (Witon* et al., 2018) and respond accordingly, which can help establish rapport and build trust between the user and the system. For instance, a mental health chatbot may use emotion recognition to detect signs of distress and provide appropriate support. Additionally, dialog act recognition can help the chatbot understand the purpose and intent of the user's input. This can enable the system to generate more targeted and effective responses, leading to a more satisfying user experience.

In this context, this paper explores the importance of dialog act recognition for conditioning the response of chatbots and conversational agents (Colombo et al., 2019; Jalalzai* et al., 2020; Colombo et al., 2021b). We will examine different methods for incorporating these techniques into chatbot systems and the impact that they have on the effectiveness of the interaction. By doing so, we hope to provide insights into how these techniques can be leveraged to improve the performance of chatbot systems and create more engaging and effective interactions.

2 Problem Framing

A decisive step in conversational AI systems is the characterization of user's utterances since it enhances the identification of both DA and E/S on spontaneous dialogue (Dinkar et al., 2020). This can involve two levels of modeling, the utterance level to understand the subtlety of the user messages and the dialogue level to figure out the inner patterns over long sequences of conversations. In this work, we focus on an English setting and the purpose is to fine-tune a pre-trained model on discourse level and/or utterance level and then build a label decoder according to a given architecture. First of all, let's state the sequence labeling problem as in (Chapuis et al., 2020; Colombo et al., 2021a) and in (Colombo et al., 2020). At the highest level, we have a set of D conversations composed of utterances. The set can be either monolingual or multilingual conversations

while in this article, we focus on monolingual conversations. Therefore, $D = \left(C_1, C_2, \ldots, C_{|D|}\right)$ with $Y = \left(Y_1, Y_2, \ldots, Y_{|D|}\right)$ being the corresponding set of labels (e.g. DA or E/S). At the lower level, each conversation C_i is formed of utterances u, i.e $C_i = \left(u_1, u_2, \ldots, u_{|C_i|}\right)$ with $Y_i = \left(y_1, y_2, \ldots, y_{|C_i|}\right)$ being the corresponding sequence of labels. Each utterance u_i is associated with a unique label y_i. $u_i = \left(w_1^i, w_2^i, \ldots, w_{|u_i|}^i\right)$ is a sequence of words. The below table shows concrete examples with emotion and sentiment.

Utterance	E/S
Good job Joe! Well done! Top notch!	pos
You liked it? You really liked it?	pos
Oh-ho-ho, yeah!	pos
Which part exactly?	neu
The whole thing! Can we go?	neu
Oh no-no-no, give me some specifics	neg

Table 1: Example of dialogue labelled with E/Staken from MELD_s. The labels pos, neu and neg respectively stand for positive, neutral and negative

2.1 Architecture

The aforementioned definitions highlight the hierarchical relation that captures some possible multi-utterance dependencies. Besides a lack of high computation infrastructures motivates us to restrain on smaller models which leverage the fine-tuned capability of transformers. On this basis, we draw four encoder-decoder strategy architectures of sequence labeling prediction. Typically an architecture (Figure 1) is a merge of an encoder which is based on either a layer of transformers (Wolf et al., 2019) or hierarchical transformers i.e. a bloc of Transformers (Chen et al., 2017) (Li et al., 2018) and a decoder designed by either a feedforward or a recurrent neural network.

2.2 Encoder

We choose two types of encoders for outcome embeddings. The first one is built on a simple layer $\mathcal{T}^d$ of transformers basis and performs encoding at the discourse level (Chapuis et al., 2020). Indeed, given a conversation C_j we have

$$\mathcal{E}_{C_j} = \mathcal{T}^d\left(w_1^i, \ldots, w_{|u_i|}^i, 1 \leq i \leq |C_j|\right) \quad (1)$$

where $\mathcal{E}_{C_j} \in \mathbb{R}^{d_d}$ is the embedding of C_j. The other encoding design is the hierarchical one aiming to capture dependencies at different granularity levels (Chen et al., 2018), (Chen et al., 2017) . Formally it is expressed with an additional function $\mathcal{T}^u$ verifying:

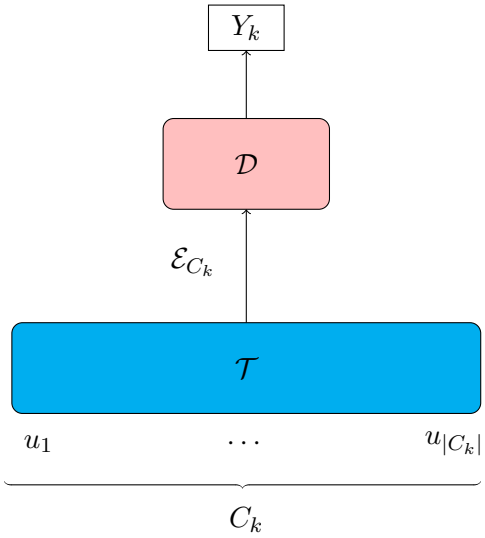


Figure 1: Architecture of a model where $\mathcal{T}$ and $\mathcal{D}$ are respectively an encoder and a decoder

$$\mathcal{E}_{u_i} = \mathcal{T}^u\left(w_1^i, \ldots, w_{|u_i|}^i\right) \quad (2)$$

$$\mathcal{E}_{C_j} = \mathcal{T}^d\left(\mathcal{E}_{u_1}, \ldots, \mathcal{E}_{u_{|C_j|}}\right) \quad (3)$$

where $\mathcal{E}_{u_i} \in \mathbb{R}^{d_u}$ is the embedding of utterance u_i.

2.3 Decoder

After encoding conversations, the following step is the building of a model which can predict labels (DA, E/S). The model is fed with the discourse-level embeddings $\mathcal{E}_{C_j}$ and in view of this, a decoder with the purpose to predict in one shot the sequence of labels is more suitable. Concretely given a conversation C_i the associated predicted labels are:

$$\hat{Y}_i = \left(\hat{y}_1 \quad \cdots \quad \hat{y}_{|C_i|}\right) = \mathcal{D}\left(\mathcal{E}_{C_i}\right) \quad (4)$$

where $\mathcal{D}$ is the decoder with either a recurrence or a forward property and its performance evaluates according to:

$$Acc\left(\mathcal{D}\right) := \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{|C_i|}\sum_{j=1}^{|C_i|}\mathbb{1}\left(y_j = \hat{y}_j\right) \quad (5)$$

3 Experiments Protocol

3.1 Datasets

The datasets used for the model evaluation are collected from SILICONE (Godfrey et al., 1992; Li

et al., 2017; Leech and Weisser, 2003; Busso et al., 2008; Passonneau and Sachar., 2014; Thompson et al., 1993; Poria et al., 2018; Shriberg et al., 2004a; Mckeown et al., 2013), a reference in the community composed of a set of sequence labeling tasks, gathering both DA and E/S annotated datasets. The corpora statistics are summarised in Table 4

From DA datasets, we pick three which are Switchboard Dialog Act Corpus (SwDA) a telephone speech corpus consisting of two-sided telephone conversations with provided topics, DailyDialog Act Corpus ($\texttt{DyDA}_\texttt{a}$) a multi-turn dialogues and daily communication corpus and ICSI MRDA Corpus (MRDA) introduced by (Shriberg et al., 2004b) and composed of transcripts of multi-party meetings hand-annotated with DA.

As E/S annotated, we choose DailyDialog Emotion Corpus ($\texttt{DyDA}_\texttt{e}$) with eleven emotional labels, and Multimodal EmotionLines Datasets (Chen et al., 2018) $\texttt{MELD}_\texttt{s}$ and $\texttt{MELD}_\texttt{e}$ with respectively three sentiments and seven emotions created by enhancing and extending the EmotionLines dataset where multiple speakers participated in the dialogues.

3.2 Data prepocessing

Processing the data to make it suitable for a chosen architecture is the first task of our experiment[1]. An important point to notice is the non-uniform length of discourse for a dataset. We then decide to split a dialog to T utterances where T is determined by analyzing the average number of utterances per dialogue given a dataset. The table 5 in the appendix gives values of T taken for each dataset. We then outline two data formats to match our architectures. The concatenate format where for a conversation C_i the T utterances and their associated labels are concatenated to make the model inputs and target features. The second format is to group utterances and labels by dialogue to form a batch. Furthermore each y represents a class and we proceed to one-hot encode it to a vector of dimension the number of distinct labels in the datasets. We finally denote by Y the target variable after preprocessing which is represented in Table 2. The below table shows the shape of inputs and outputs i/o tensors after processing:

[1]Our experiments and plots can be reproduced thanks to our code available on GitHub repository link: github.com/intent_classification

	Concatenate	Separate[2]
i shape	$(?, d_d)$	$(?, T, d_d)$
o shape	$(?, T \times \lvert Labels \rvert)$	$(?, T, \lvert Labels \rvert)$
Y_i	$\begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} 1 & 0 \end{bmatrix} \end{bmatrix}$

Table 2: Shape of inputs and outputs tensor after preprocessing where ? stands for the batch size

3.3 Pre-trained transformers used for encoding

We build encoders based on pre-trained transformers through the PyTorch implementation provided by the Hugging Face transformers library. We use mainly two models: BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019) in their base-cased architecture for the non-hierarchic encoding. Regarding hierarchical encoding, we try to bring up two BERT transformers at their tiny variant. Since a base BERT does not allow input with a dimension larger than 512. However, we do not succeed to find an appropriate output from the hierarchical decoder to conduct decoding and consequently we focus on a non-hierarchic decoder.

3.4 Models for prediction

The use of a neural network enables us to show the dependencies between utterances. As mentioned in Problem Statement two methods are mainly rolled out. Firstly we set up a multilayer perceptron that fits with the concatenated format in order to produce a one-shot prediction. We implement an architecture of 2 hidden layers since a neural net with 2 layers can model any relationship. At each layer, we utilize a number of neurons according to Equation (6), 20 % dropout, and a ReLU activation function.

$$\#neurons = \frac{2}{3} d_d + T \left| Labels \right| \tag{6}$$

Several activation functions are suitable for classification tasks. In our problem, we choose a sigmoîd to output a value in $(0, 1)$. After the model is fitted a label prediction is made by voting per batch of $|Labels|$ on a argmax strategy. Concretely the predicted labels for the dialogue C_i are obtained by:

$$\hat{y}_j \equiv \underset{(j-1)T+1 \leq k \leq jT}{\operatorname{argmax}} \mathrm{Y}_{ik} \pmod{T} \tag{7}$$

	SwDA	DyDA_a	MRDA	DyDA_e	MELD_e	MELD_s
BERT + MLP	37.4	63.5	69.1	86.1	52.0	57.8
BERT + GRU	44.0	**81.9**	**69.3**	**86.7**	**60.5**	**70.3**
XLNet + MLP	39.1	61.7	69.3	85.7	52.3	53.7
XLNet + GRU	**58.7**	78.3	69.3	85.3	51.2	63.9

Table 3: Performances of all mentioned models with different decoders such as MLP, GRU. The datasets are grouped by label type (DA vs E/S) and order by decreasing size

Afterward, we implement a Gated Recurrent Unit to perform sequence prediction. The architecture is composed of a GRU layer. Since the model outputs a tensor of dimension[3] $(?, T, |Labels|)$, the prediction of each label is done similarly to the MLP case. We select the binary cross-entropy loss (8) as a loss function for both decoder architectures.

$$\mathcal{L} = -\sum_{i,j} \mathrm{Y}_{ij} \log p\left(\mathrm{Y}_{ij}\right) \tag{8}$$

Finally to accelerate computational time we train models on Onyxia - SSP Cloud Datalab using a service of 1 GPU.

4 Results

This section gathers experiences carried out on the SILICONE benchmark. Globally, models do not issue outstanding performances certainly due to the decision to split each dialogue into T utterances. The sequential GRU decoder outperforms slightly the MLP with an average accuracy of 7.56% of over. This awaited difference could be explained by the awareness of GRU to leverage past labels in a sequence. Moreover, The designed architectures have higher prediction accuracy on E/S tasks rather than DA ones.

On the dataset level, models produce good performance since the preprocessing method and the corpus structure generate sufficient data for training. For example, designed architectures reach roughly 80 % on the DyDA based datasets. The corpus SwDA stands out itself with an average performance of 44.8% and this fact is explainable by its atypical ratio of the number of utterances over the number of labels after data management. While the models applied to MRDA and DyDA_ecorpOra show acceptable performances, on average respectively 69% and 85%, they output confusion matrices which highlight a problem of imbalanced classes.

[3] ? refers to the batch size

5 Conclusion

In this work, we have proposed a building of an intent classifier whose purpose is to predict the sequence of labels in a dialogue based on different works. We have proposed four encoder-decoder models either non-hierarchical or hierarchical. We have implemented those models on different DA/E/S datasets of SILICONE: SwDA, DyDA_a, MRDA for DA task and DyDA_e, MELD_e, MELD_s which are E/S datasets. performance of each model depends on the dataset on which it is applied. Therefore, the best model depends on the dataset considered. However, in general[4], the best model is the one obtained by applying a GRU-based sequential decoder on the dialogues encoded with the BERT transformer.

Futur work: In addition to addressing label imbalance, future work in this area should also consider fairness concerns. Bias in machine learning models can lead to unfair and discriminatory outcomes (Pichler et al., 2022; Colombo et al., 2022). Therefore, it is important to ensure that the models developed for emotion and dialog act recognition are fair and unbiased. By incorporating fairness concerns into future research on emotion and dialog act recognition, we can help to ensure that these models are not only effective but also ethical and equitable. This can help to build trust in these systems and increase their adoption in various domains, leading to more positive outcomes for users and society as a whole.

[4] According to Table 4

References

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hcrc map task corpus: natural dialogue for speech recognition.

Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004a. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004b. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

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

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

R. Passonneau and E. Sachar. 2014. Loqui human-human dialogue corpus (transcriptions and annotations).

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

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2017. Dialogue act recognition via crf-attentive structured network. *CoRR*, abs/1711.05568.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Wojciech Witon*, Pierre Colombo*, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at iest 2018: Predicting emotions using an ensemble. In *Wassa @EMNP2018*.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2018. A dual-attention hierarchical recurrent neural network for dialogue act classification. *CoRR*, abs/1810.09154.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. *CoRR*, abs/1802.08379.

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

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. 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 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. *CoRR*, abs/2002.08801.

Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for

text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7985–7993, Online. Association for Computational Linguistics.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021a. Code-switched inspired losses for spoken dialog representations. In *EMNLP 2021*.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *() ACL 2021*.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *ICML 2022*.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Learning disentangled textual representations via statistical measures of similarity. *ACL 2022*.

Corpus	$\|Train\|$	$\|Val\|$	$\|Test\|$	Utt.	$\|Labels\|$	Task
SwDA	$1k$	100	11	$200k$	42	DA
DyDA_a	$11k$	$1k$	$1k$	$102k$	4	DA
MRDA	56	6	12	$110k$	5	DA
DyDA_e	$11k$	$1k$	$1k$	$102k$	7	E/S
MELD_e	934	104	280	$13k$	7	E/S
MELD_s	934	104	280	$13k$	3	E/S

Table 4: Statistics of used datasets part of SILICONE where sizes of Train, Val and Test are given in number of conversations.

	SwDA	DyDA_a	MRDA	DyDA_e	MELD_e	MELD_s
T	50	5	50	5	5	5

Table 5: Value of T per corpus

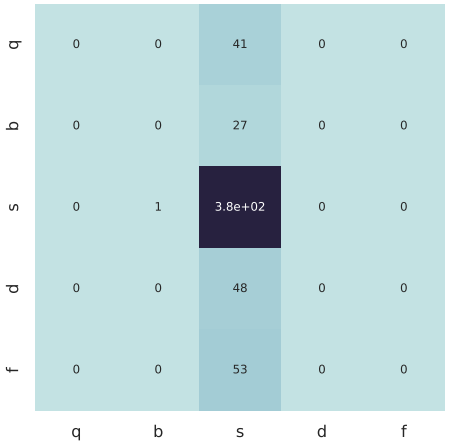
(a) BERT+MLP

(b) BERT+GRU

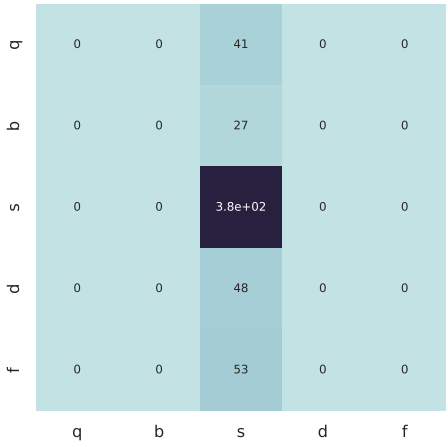
(c) XLNet+MLP

(d) XLNet+GRU

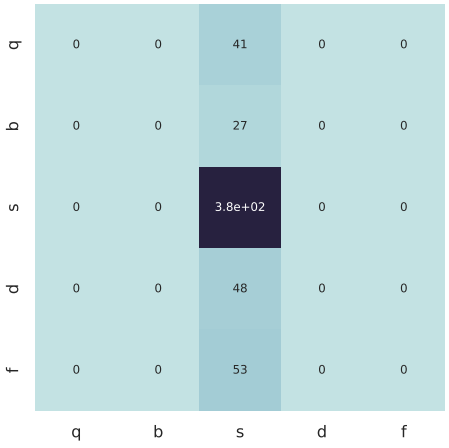
Figure 2: Confusion matrices on DyDA_a corpus



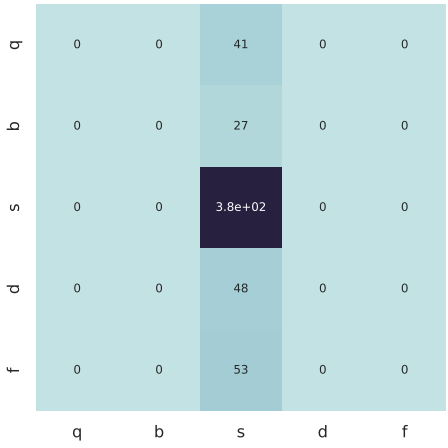
(a) BERT+MLP



(b) BERT+GRU

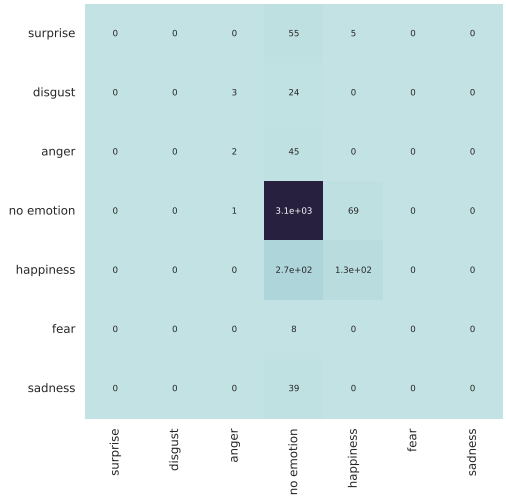


(c) XLNet+MLP on MRDA

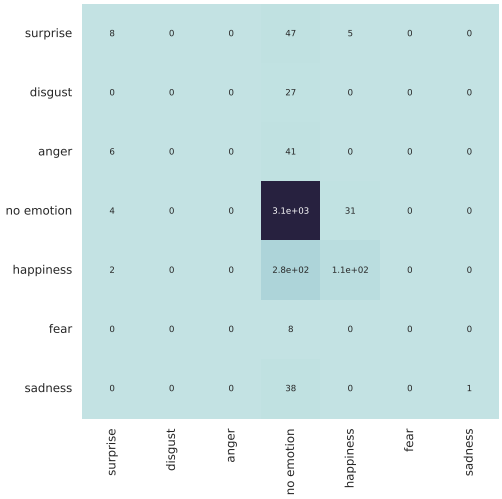


(d) XLNet+GRU on MRDA

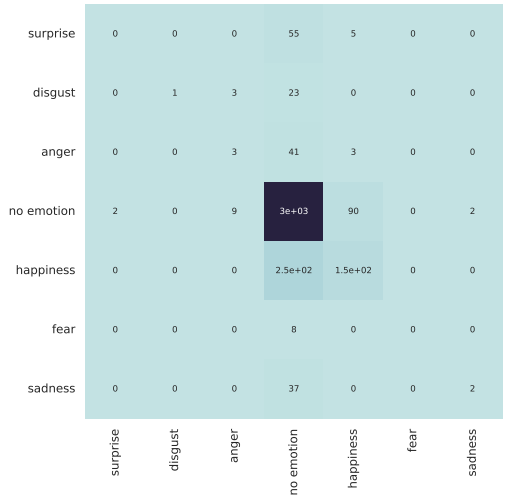
Figure 3: Confusion matrices on MRDA



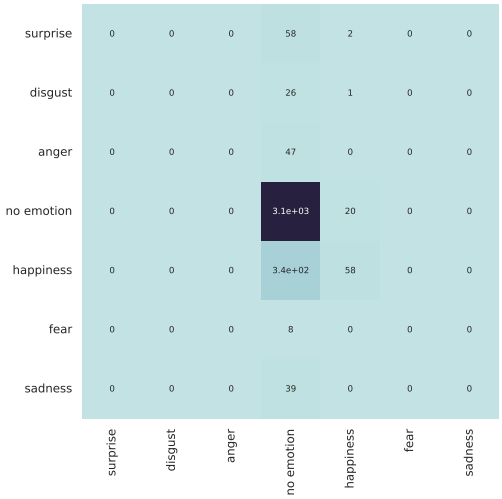
(a) BERT+MLP



(b) BERT+GRU

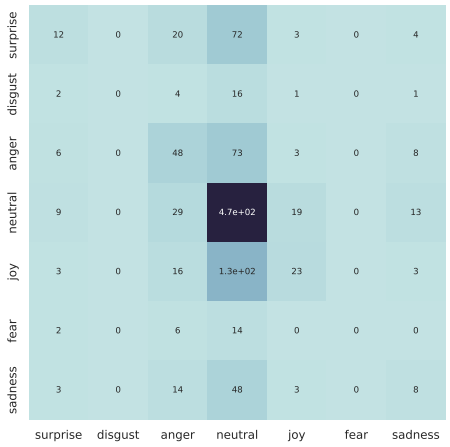


(c) XLNet+MLP

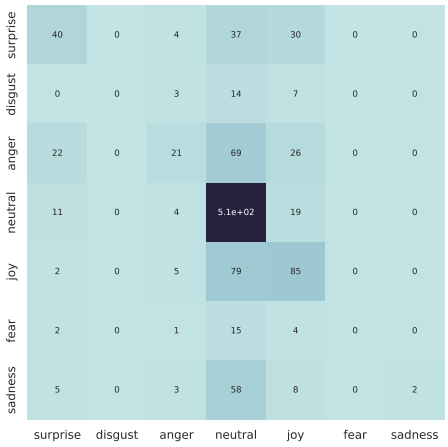


(d) XLNet+GRU

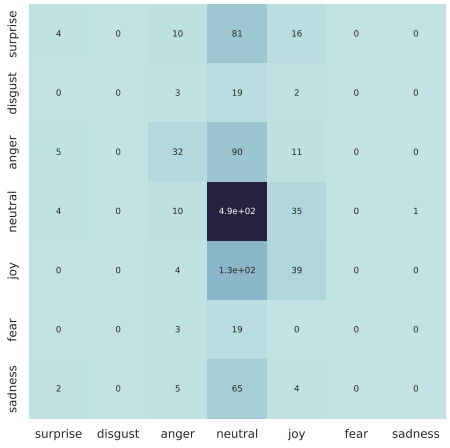
Figure 4: Confusion matrices on $\texttt{DyDA}_{\texttt{e}}$



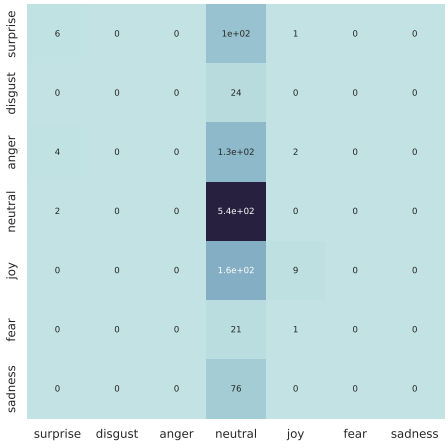
(a) BERT+MLP



(b) BERT+GRU

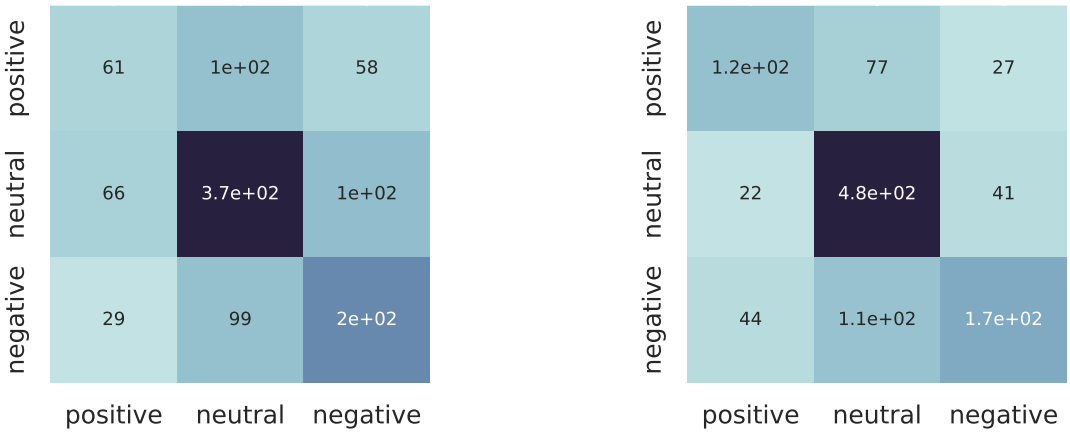


(c) XLNet+MLP



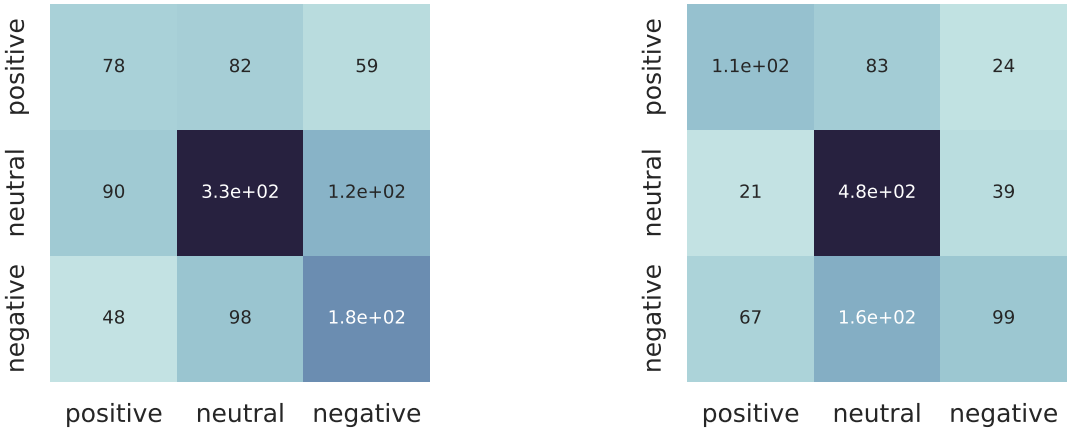
(d) XLNet+GRU

Figure 5: Confusion matrices on $\texttt{MELD}_{\texttt{e}}$



(a) BERT+MLP

(b) BERT+GRU



(c) XLNet+MLP

(d) XLNet+GRU

Figure 6: Confusion matrices on $\texttt{MELD}_\text{s}$