Intent classification using contextual embeddings

CHEN Yunhao *
ENSAE
yunhao.chen@ensae.fr

RONDEY Hélène *
ENSAE
helene.rondey@ensae.fr

Abstract

Sequence Labelling problems, such as Dialogue Act (DA) and Sentiment/Emotion (S/E) identification, play a large part in Natural Language Understanding tasks, like implementing Conversational Agents. In the past few years, the methods used for vector representation of words and utterances, an essential component of these problems, have been significantly improved thanks to Large Language Models, like BERT (Devlin et al., 2018), which is based on the Transformer Architecture.

In this work we experiment on the Sequence Labelling Problem by proposing our own architecture for classification and relying on the fine-tuning of a pre-trained large language model, either RoBERTa (Liu et al., 2019) or DeBERTa (He et al., 2020). We then compare the performance of the two models for our architecture. We also study the difference in performance between a total fine-tuning and a partial fine-tuning of the pre-trained models.

We conduct our analysis on the annoted spoken language datasets from the SILICONE benchmark (Chapuis et al., 2020). Our work shows that the two models are quite similar in test accuracy and seem to outperform the models from (Colombo, 2021). We also show that partially fine-tuning provides lesser performances than totally fine-tuning. We conclude by discussing some other leads of exploration to tackle the Intent classification problem, like hierarchical structures and contrasive learning.

1 Introduction

In recent years, virtual assistants like Google Home have become increasingly popular, providing users with a convenient and intuitive way to interact with technology. These devices rely on natural language processing (NLP) technologies to understand and respond to user requests, which often involve complex dialogues with multiple turns and linguistic cues.

One crucial aspect of NLP technology for virtual assistants is the identification and classification of dialog acts, which are the fundamental units of communication in a conversation. Dialog acts refer to the specific actions performed by a speaker during a dialogue, such as asking a question, giving an order, expressing an opinion, or providing information. Accurate identification of these dialog acts is essential for the virtual assistant to understand the user's intent and provide appropriate responses (Jalalzai* et al., 2020; Colombo* et al., 2019; Colombo et al., 2021).

The importance of dialog act classification in virtual assistants like Google Home is further highlighted by the fact that these devices are becoming ubiquitous in daily life, assisting users with a wide range of tasks, from setting reminders and playing music to controlling smart home devices and ordering groceries. Inaccurate classification of dialog acts can result in incorrect responses, leading to user frustration and potentially damaging the reputation of the virtual assistant.

In this work, we address the problem of DA and E/S classification using deep learning technics. From non-contextual embeddings (such as Word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), FastText (Bojanowski et al., 2017)) to contextual embeddings, generic vector representations of natural language has been proven powerful in NLP tasks. Especially the emergence of attention mechanisms (Bahdanau et al., 2014) and the Transformer architecture (Vaswani et al., 2017) has brought NLP to a large language model (LLM) era. Many transformer-based models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) are trained on large corpora of written language, and provide the contextual vector repre-

*Stands for equal contribution

sentations which capture the semantic information in natural language.

In this work, we focus on showing the generality of vector representations trained on written language and their effectiveness in DA and E/S tasks. More specifically, we employ the pre-trained RoBERTa and DeBERTa and fine-tune them using SILICONE (Chapuis et al., 2020), which is a collection of sequence labelling tasks, gathering both DA and E/S annotated datasets. Five datasets in SILICONE have been chosen for fine-tuning: `DyDA_DA`, `MapTask`, `MELD_e`, `MELD_s` and `SEM`. The source code can be found on Github.[1]

2 Problem Framing

We start by formally defining the Sequence Labelling Problem. There are three levels: dialogue level, utterance level and word level. At the dialogue level, we have a set D of conversations composed of utterances, i.e., $D = (C_1, C_2, ..., C_{|D|})$ with $Y = (Y_1, Y_2, ..., Y_{|D|})$ being the corresponding set of labels (e.g., DA or E/S). At the utterance level, each conversation C_i is composed of utterances u, i.e $C_i = (u_1, u_2, ..., u_{|C_i|})$ with $Y_i = (y_1, y_2, ..., y_{|C_i|})$ being the corresponding sequence of labels: each u_i is associated with a unique label y_i. At the word level, each utterance u_i can be seen as a sequence of words, i.e $u_i = (\omega_1^i, \omega_2^i, ..., \omega_{|u_i|}^i)$. Some concrete examples can be found in Table 1.

Table 1: Examples of dialogues labelled with DA and E/S taken respectively from `DyDA` and `MELD`.

Utterances	DA
Can you study with the radio on?	`question`
No, I listen to background music.	`inform`
What is the difference?	`question`
The radio has too many comerials.	`inform`
That's true, but then you have to buy a record player.	`inform`
Utterances	**E/S**
You had no right to tell me you ever had feelings for me.	`anger`
What?	`surprise`
I was doing great with Julie before I found out about you.	`anger`

2.1 Related Work

Many approaches have been proposed to tackle the DA and E/S classification problem. Early work relies on the independent classification of each utterance using various techniques, such as Hidden Markov Models and SVM (Stolcke et al., 2000), (Surendran and Levow, 2006). More recently, deep learning approaches (Kalchbrenner and Blunsom, 2013) have been widely applied to capture contextual dependencies between input sentences. (Li et al., 2018) proposes a dual attention mechanism to capture information about both DAs and topics. Another refinement is to consider the inter-tag dependencies (Kumar et al., 2018). By making use of the perfect alignment between the utterance sequence and DA sequence, a guided attention mechanism is used to force the decoder to focus more on some specific utterances (Colombo et al., 2020).

The hierarchical structure has been considered to tackle the multiple levels of DA and E/S tasks (Kumar et al., 2018), (Garcia et al., 2019), (Colombo, 2021). Various pieces of information have been proven useful to improve the accuracy of sequence labelling tasks, such as the fillers in spoken language (Dinkar et al., 2020) and the information in other modalities (e.g. video and audio) (Garcia et al., 2019). Some other works focus on the code-swithed phenomenon by proposing the DA classification models in multi-language settings (Chapuis et al., 2021).

2.2 Our proposals

This work mainly focuses on making use of generic representations of natural language provided by pre-trained LLM, to tackle the sequence labelling problem.

We propose to use the vector representation of each utterance to perform an independent classification (without considering the intra-utterance dependencies), as shown in Fig. 1. Following the idea in BERT (Devlin et al., 2018), we extract the vector representation of the first special token in the utterance, as the representation of the whole utterance.

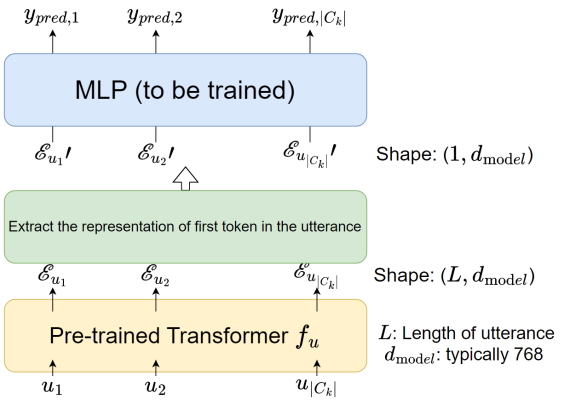

Figure 1: Independent classification

[1] Our repository on Github

2.3 The SILICONE datasets and tasks

As described in (Chapuis et al., 2020), SILICONE[2] gathers nine labelling tasks datasets (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., 2004; Mckeown et al., 2013), including five DA datasets and four S/E datasets. We have selected five datasets of various sizes for our tuning tasks, although we favoured in our choices some of the smaller-sized datasets for the sake of shorter computation times and saving ressources. Among the chosen DA datasets, `DyDA_DA` is the third biggest set of SILICONE (102k utterances in total), and `MapTask`, a small set of 27k utterances. Then, among the S/E datasets, `SEM` is the smallest SILICONE dataset with 5,6k utterances, while `MELD_e` and `MELD_s` share the same 13k utterances.

The two `MELD` datasets are representative of the diversity of tasks found in SILICONE: for a given utterance, `MELD_s` identifies the sentiment among three labels "negative", "neutral", or "positive" while `MELD_e` assigns an emotion among seven labels ("anger", "disgust", "fear", "joy", "neutral", "sadness" and "surprise").

One can notice that the label repartition can be quite imbalanced in the datasets (see Table 2). It can be explained by the construction of the datasets, since they are a collection of consecutive lines (utterances) of natural speech and dialog.

Table 2: Label repartition in some of the selected SILICONE datasets

`SEM` - Training set	Negative	Neutral	Positive	
	844	2291	1129	
`MELD_s` - Training set	Negative	Neutral	Positive	
	2945	4710	2334	
`DyDA_DA` - Training set	Commissive	Directive	Inform	Question
	8081	14242	39873	24974

3 Experiments and Analysis

As the implementation of our first proposal, we have fine-tuned the pre-trained "RoBERTa-base" [3] and pre-trained "DeBERTa-base" [4], on NVIDIA RTX A2000 with 12GB RAM. On top of the pre-trained model, we added three linear layers, each with 768 neurons (the last layer serves as classification layer). Other hyper-parameters can be found in Table 3, which use the fine-tuning setting in (He et al., 2020). Figures 6, 7, 8 and 9 in the Appendix show respectively the evolution of validation accuracy and loss in DeBERTa and RoBERTa. Generally speaking, after a few (2-4) epochs, the validation accuracy reaches a peak. This could be explained by the small amount of fine-tuning data, compared to the huge pre-training corpus.

Table 3: Hyper-parameters in fine-tuning

Hyper-parameter	Value
Dropout in classification layer	0.15
Optimizer	Adam
Learning rate	2e-5
Adam ϵ	1e-6
Gradient Clipping	1.0
Maximum fine-tuning epochs	10
Patience	2
Batch size	32 (64 for Dyda_da)

Information on several datasets from SILICONE and test accuracy on them is shown in Table 4. RoBERTa and DeBERTa have both 120 million trainable parameters. The execution time per epoch is the same and is reasonably proportional to the dataset size. There is no obvious difference between these two models in terms of test accuracy. We also notice that on several datasets such as `DyDA_DA`, `MELD_e` and `MELD_s`, RoBERTa and DeBERTa achieve better performance than (Colombo, 2021).

Table 4: For each tasks: Number of samples, Execution time per epoch and Test accuracy (fine-tuning whole model)

Task	#Samples	Exec. time	DeBERTa	RoBERTa
`DyDA_DA`	87170	2h26mins	-	82.7
`MapTask`	20905	35mins	63.6	64.0
`MELD_e`	9989	17mins	63.7	63.0
`MELD_s`	9989	17mins	70.5	70.0
`SEM`	4264	7mins	62.7	63.2

We then compared total fine-tuning and partial fine-tuning. The latter means that the backbone model is frozen and only the added linear layers are trained. Fig. 2 shows the result which is expected: there is an obvious drop of performance if we just update the parameters in added linear layers.

As a final illustration of the importance of fine-tuning and an assessement of our architecture's performance, we take a look in Fig. 3 at the confusion matrix obtained on the `SEM` test set with RoBERTa. On this dataset, the accuracy from Table 4 was 63.2%. The confusion matrix shows that

[2] SILICONE on HuggingFace
[3] Roberta Model on HuggingFace
[4] Deberta Model on HuggingFace

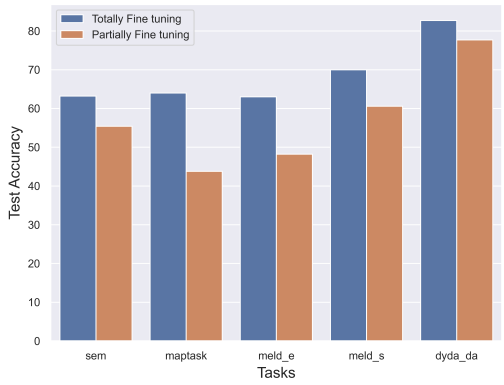

Figure 2: Total and partial fine-tuning on RoBERTa

for the Negative label, 56% of the instances are matched with a correct prediction, for the Neutral label 64%, and for the Positive label 63% of the predictions are correct. The fact that the predictions for the Negative label are weaker than for the other two can be explained by the lower proportion of this label over the training set, as we showed in Table 2. However, these predictions are significantly improved by the total fine-tuning: the initial RoBERTa (without being fine-tuned) predicts mostly a Negative label regardless of the true label (see Figure 10).

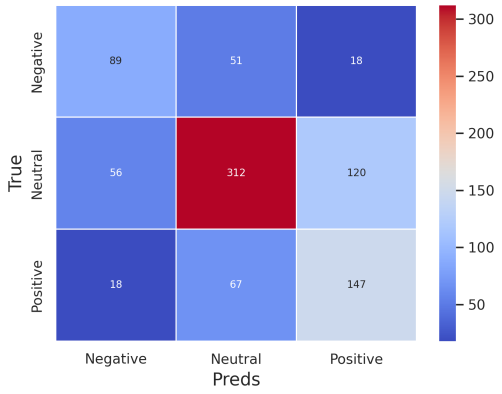

Figure 3: Confusion matrix of the fine-tuned RoBERTa on the SEM test set

4 Discussions

In this work, we have illustrated one use of Large Language Models in the Sequence Labelling problem with a quite simple approach. To do so, we added linear layers on top of pre-trained models, RoBERTa and DeBERTa, and then fine-tuned the weights accordingly to each of the datasets we selected from the SILICONE benchmark. We have observed that the two models are both quite satisfying and similar in performance when we consider the accuracy on the test sets, and seem to outperform one of our reference papers, (Colombo, 2021). The experiment we made by comparing total and partial fine-tuning demonstrated that fully tuning the model and not only the additional layers allows a higher classification accuracy. Also, observing the confusion matrices reminded us that the balance in the training samples has a part to play in the quality of the predictions, in addition to tuning. This work was also an opportunity to appreciate the diversity of tasks (Dialogue Act and Sentiment/Emotion labelling) that can be achieved thanks to the variety of available training sets from SILICONE and the flexible vector representation of the Transformer architecture.

However, we only proposed a simple architecture that performs an independent classification, which means that the model doesn't consider the hierarchy between utterances. To obtain a more complete model, as shown in Fig. 4 and Fig. 5, we suggest that the structure LSTM (Hochreiter and Schmidhuber, 1997) or Transformer (Vaswani et al., 2017) could be applied on top of the utterance representation, in order to capture the intra-utterance dependencies in one dialogue.

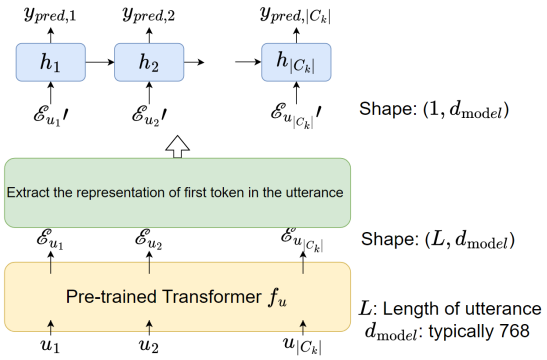

Figure 4: Hierarchical structure with LSTM

Another idea we have explored comes from the fact that each of the five datasets we selected from SILICONE contains a different number of classes, which forces us to fine-tune one model for one dataset. Ideally, we would like to use one model for all downstream tasks. The idea of contrasive learning in CLIP (Radford et al., 2021) has been borrowed here: the utterances and the corresponding labels are both encoded to vectors in the same latent space. The correct utterance-label pairs are

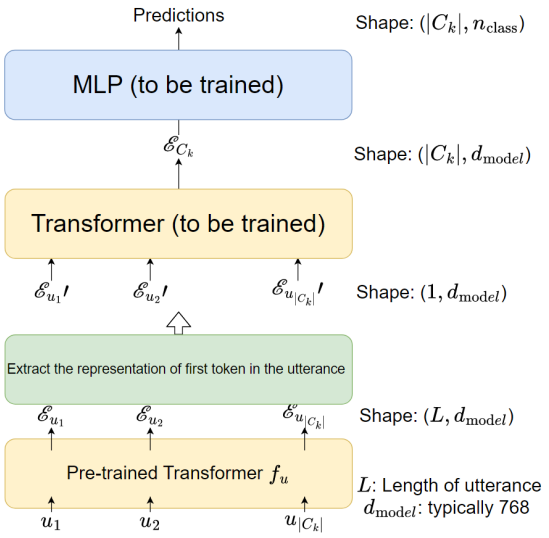

Figure 5: Hierarchical structure with Transformer

expected to have a higher cosine similarity similarity. In this way, the classification layer is not needed anymore. However, the test performance of this method is not satisfactory: the accuracy is just slightly higher than random choice. The test code can be found in our Github repository. We guess that the poor performance may be due to the difference of utterance length (more than 20 tokens) and label length (3-4 tokens). Thus, the utterance vector and label vector are probably not aligned in latent space, which makes the cosine similarity no sense. How to solve this alignment problem could be a research direction for future works.

References

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

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. 2004. 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.

Dinoj Surendran and Gina-Anne Levow. 2006. Dialog act tagging with support vector machines and hidden markov models. In *Interspeech*. Citeseer.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

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

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. *NAACL 2019*.

Alexandre Garcia, Pierre Colombo, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. *arXiv preprint arXiv:1908.11216*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7594–7601.

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 Chloe Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. *arXiv preprint arXiv:2009.11340*.

Emile Chapuis, Pierre Colombo, Matthieu Labeau, and Chloe Clavel. 2021. Code-switched inspired losses for generic spoken dialog representations. *arXiv preprint arXiv:2108.12465*.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

A Figures

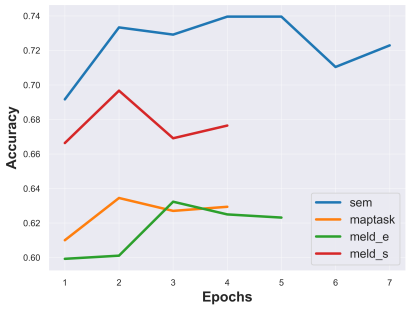

Figure 6: DeBERTa: Evolution of validation accuracy on different tasks

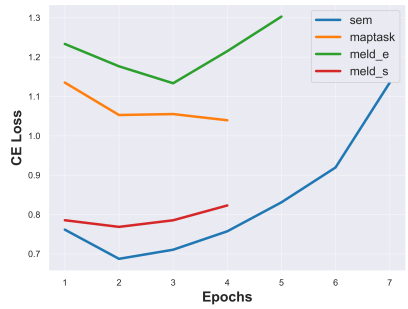

Figure 7: DeBERTa: Evolution of validation loss on different tasks

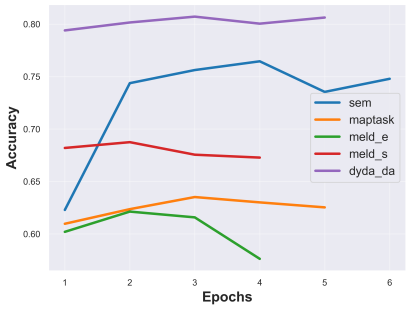

Figure 8: RoBERTa: Evolution of validation accuracy on different tasks

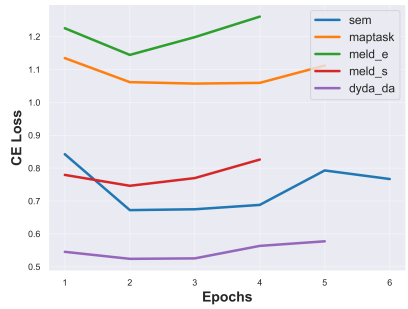

Figure 9: RoBERTa: Evolution of validation loss on different tasks

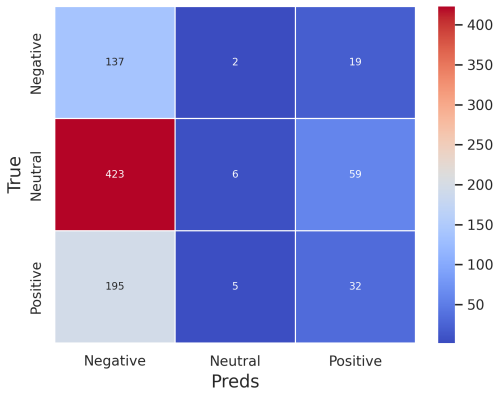

Figure 10: Confusion matrix of the initial RoBERTa (no fine-tuning) on the SEM test set