Dialog Act Classification with BERT Models

Serine Mechide*
ENSAE
serine.mechide@ensae.fr

Beatriz Farah*
ENSAE
beatriz.farahnoresgoncalves@ensae.fr

Abstract

The identification of Dialog Acts (DA) through sequence labeling systems is an important part of Spoken Dialog (SD) understanding. Nowadays DA recognition has gained attention given its importance in Chatbot training, since understanding the role that a user's message plays in a message is crucial in order to respond in an appropriate and helpful way. In this work, we perform DA classification adapted to SD, which we implement using 2 BERT models (Bidirectional Encoder Representations from Transformers). We evaluate the models on 3 DA databases from the SILICONE benchmark[1], and we compare the models' performances by obtaining their accuracy, their average training time and their loss.

1 Introduction

Dialog act classification is a vital component of chatbot systems that enable them to interpret and respond appropriately to user inputs in a conversational context. Dialog act classification involves the identification of the underlying intention or purpose behind a user's utterance in a conversation. For example, in a customer service chatbot system, a user's input might be classified as a request for information, a complaint, or a question about a particular product or service.

Accurate dialog act classification is crucial for chatbot systems as it enables them to provide the most relevant and helpful response to the user's input (Colombo et al., 2021b; Jalalzai* et al., 2020; Colombo* et al., 2019). Furthermore, it facilitates a smooth and natural conversation flow that can lead to higher user satisfaction and engagement.

With the increasing popularity of chatbot systems in various domains, such as customer service, healthcare, and education, the importance of dialog act classification has only grown. As such, there has been a significant amount of research in recent years focused on developing more accurate and efficient dialog act classification methods. These advances have helped improve the overall performance of chatbot systems, leading to more effective and engaging interactions between users and chatbots.

1.1 Problem Framing

Following the notation from (Colombo* et al., 2020; Colombo, 2021), in linguistics and in natural language understanding, a dialog D is defined as a sequence of conversations C:

$$D = (C_1, C_2, C_3, ..., C_{|D|})$$

Each conversation is composed of multiple utterances U and can be defined as follows:

$$C_i = (U_1, U_2, ..., U_{|C_i|})$$

A Dialog Act (DA) represents the speaker's intention in delivering an utterance (Dopp, 1962), and each utterance U_i is associated with a unique DA label y_i.

The goal of this paper is to use Spoken Dialogue (SD) datasets to predict the DA Classification Y_i of each conversation C_i as the sequence of labels y_i corresponding to each utterance $Y_i = (y_1, y_2, ..., y_{|C_i|})$.

In order to predict the DA classification on SD data, we will use different sizes of the BERT pretrained encoder model (Devlin et al., 2018). These models will be evaluated on 3 out of the 5 DA datasets that are a part of the SILICONE benchmark(Godfrey et al., 1992; Li et al., 2017a; Leech and Weisser, 2003a; Busso et al., 2008; Passonneau and Sachar., 2014; Thompson et al., 1993;

*These authors contributed equally to this work

[1]Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE

Poria et al., 2018; Shriberg et al., 2004a; Mckeown et al., 2013)[2], each with a different size and set of labels (Chapuis et al., 2020a). Examples of extracts from a dialog from one of the DA databases used in training and their respective labels are found in Table 1.

DA recognition is largely used in training Chatbots and technologies such as ChatGPT. This is due to the fact that the analysis of the DA of a user's message allows for building a natural language system that provides an appropriate and helpful response to the user's requests, therefore avoiding the generic response problem (Kumar et al., 2018).

The models and the corresponding training were developed in Python, using the Pytorch implementation of the BERT models provided by the Hugging Face transformers library (Wolf et al., 2020). Our code is available on GitHub[3].

Utterance	Label
Are we supposed to read digits at the same time?	q
No	s
Oh okay	s

Table 1: Extract of a dialog from the MRDA database from SILICONE. Labels are: "s" (Statement), "d" (Declarative Question), "b" (Backchannel), "f" (Follow-me) or "q" (Question)

2 BERT Model

In order to solve our classification problem, we use a variant of transformers: the Bidirectional Encoder Representations from Transformers (BERT), introduced by (Devlin et al., 2018). In contrast to the context-free models like word2vec (Mikolov et al., 2013), it is a contextual model that will use other words in the sentence to give a representation of each word so that it takes into account the bidirectional relationship between words in the same sentence during the training phase.

2.1 The architecture

Regarding its architecture, BERT is a trained Transformer Encoder stack with originally 12 encoder layers in the base version and 24 in the large version. The particularity of BERT is that it was pre-trained on a large set of unlabelled texts including Wikipedia and Book Corpus which contain more than 10,000 books of different genres. This allows us to use this pre-trained model for various language-based tasks and to fine-tune these pre-trained models on smaller task-specific datasets. To understand how this model works, it is necessary to have a great understanding of how Transformers work.

A Transformer consists of an encoder to read the text input and a decoder to predict for the task. A transformer is a model architecture that relies entirely on an attention mechanism to learn relationship between input and output. Attention refers to a neural network structure. It was introduced in 2015 (Luong et al., 2015) to be used for translation tasks. Attention mechanism allows the model to focus on every single word in the sentence when choosing how to translate words in the ouput sentence. The model, thanks to training data, learns what words are connected to each other and therefore on which words to focus its attention. The attention mechanism is important for translation (Vaswani et al., 2017).

Self-attention is also an important concept in Transformers because it allows neural networks to really understand a word based on the context of the other words in the sentence. Self attention is necessary to to remove all ambiguity and accomplish tasks like semantic roles or part of speech tagging for instance.

In this current paper, we have chosen to compare different BERT models for each DA database chosen. Following (Devlin et al., 2018) we let L denote the number of layers (ie transformer blocks), H the number of hidden states, A the number of states and P the number of trainable parameters. We will report results on the following models:

Model	L	H	A	P (in millions)
BERT_{MINI}	4	256	4	11
BERT_{SMALL}	4	512	8	28

Table 2: BERT models we use in the current paper

These models were firstly introduced in (Turc

[2]SILICONE benchmark can be found in the dataset library from HuggingFace (Chapuis et al., 2020a) at https://huggingface.co/datasets/silicone

[3]https://github.com/beafarah/NLP_Intent_Classification

et al., 2019)[4], and are available on GitHub.[5]. They were built by reducing the size of the standard BERT model (BERT BASE). These models are ideal for environments with restricted computational resources as they are lighter than the BERT BASE model (that has $L = 12, H = 768, A = 12, P = 110$ millions).

Our main objective in running these different BERT models is to be able to compare their performance, as well as the time that each one takes to run, for each one of the 3 selected DA datasets.

We chose these two BERT models out of all of the other possible BERT models because they are more lightweight than the other usual BERT models (MEDIUM, BASE and LARGE) which need more computational time.

2.2 Text Processing

There are some processing steps to make before building the model, given that BERT expects input text in a specific format. The input to the encoder is a sequence of tokens, which are first transformed into vectors and then processed in the neural network.

Every input embedding contains 3 embeddings: the position, the segment and the token embeddings. For each token, the input representation is constructed by summing the 3 embeddings. The position embeddings allows to learn the position of words in a sentence. The segment embeddings allows BERT to take sentence pairs as inputs in order to distinguish between them and accomplish tasks like Question-Answering. Token embeddings are the embedddings learned from the specific token from the WordPiece token vocabulary.

2.3 Pre-training tasks

BERT is pre trained on two Nature Language Processing tasks: Masked Language Modelling and Next Sentence Prediction (Devlin et al., 2018). Instead of predicting the next word in a sequence like models that were trained on a left-to-right context, the BERT model "masks" words in a sentence and then tries to predict them. Masking means taking into account both the previous and next tokens at the same time to consider the full context of the sentence to predict the masked word.

In addition to Masked Language Models, BERT is also trained on the task of Next Sentence Prediction which is useful for tasks like question answering. The idea is that the model gets as input pairs of sentence and it learns if the second sentence comes after the first one in the original text as well or if it is random. To improve accuracy, the model is trained on those both masked LM and Next Sentence Prediction together.

3 Experiments Protocol

3.1 Datasets

Following (Chapuis et al., 2020b), we chose to evaluate our models on the SILICONE[6] benchmark, which gathers datasets designed for training and evaluating natural language understanding systems designed for spoken language. SILICONE is a collection of Dialogue-Act (DA) and Emotion/Sentiment (E/S) annotated datasets, each with a different size and set of labels.

In this current paper, we have performed labelling tasks for DA datasets from the SILICONE benchmark. The DA databases in which we performed classification were the following:

- **BT OASIS Corpus:** This DA database contains 636 dialogues and 15067 utterances in total, and gathers the transcripts of live calls made to the BT and operator services. Each utterance can be classified into one out of 42 possible labels, and the distribution of the data in each category can be seen in Figure 1 in the Annex. Following (Chapuis et al., 2020b), we use a random train/dev/test split [7] where the training accounts for 80% of the dialogues (508), test and validation correspond to 10% each (64 dialogues). This corpus was introduced by (Leech and Weisser, 2003b).

- **The Daily Dialog Dataset (DyDA) :** Produced by (Li et al., 2017b), this database is a multiturn dialogue between people in their everyday life which contains 102 979 utterances. Each utterance is classified as one of the 4 following labels : "commissive" (0), "directive" (1), "inform" (2) or "question" (3). As shown in graph 2, most of the utterances are labelled as information and ques-

[4]They are meant for uncased data in English, and were trained with WordPiece masking

[5]https://github.com/google-research/bert

[6]SILICONE stands for Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE

[7]Split made originally in https://github.com/NathanDuran/BT-Oasis-Corpus

tion. We have used the train/validation/test split introduced by the authors in (Li et al., 2017b).

- **ICSI MRDA Corpus (MRDA):** Introduced in (Shriberg et al., 2004b), the MRDA dataset gathers scripts from multi-party meetings. It contains a total of 109229 utterances, which are separated into train, validation and test sets according to the official split introduced by the authors. The utterances are classified according to 5 possible labels: "s" (0: Statement/Subjective statement), "d" (1: Declarative Question), "b" (2: Backchannel), "f" (3: Follow-me) or "q" (4: Question). We can see in Figure 3 (in the Annex) the distribution of the labels in the training set. We notice that the dataset is unbalanced, as approximately 60% of the training data is classified in the category "s".

For each one of the DA datasets, we use the available official split into a respective train, test and validation dataset. Table 3 summarizes the number of utterances in each of them.

Dataset	Intention	Train	Test	Validation
Oasis	42	12076	1478	1513
MRDA	5	83944	15470	9815
Dyda	4	87170	7740	8069

Table 3: Summary of the DA databases

3.2 Loss function

Following Tensorflow's tutorial for training a BERT model[8], we decided to use the Cross Entropy loss function for training. This function measures the performance of a classification model as follows:

$$\mathcal{L}(X, \hat{Y}) = \sum_{i=1}^{n} -y_i log(p_i) \qquad (1)$$

Where p_i is the Softmax probability for class i, y_i is the true label of input X, $\hat{Y} = (p_1, ..., p_n)$ and n is the number of classes.

[8] https://www.tensorflow.org/text/tutorials/bert_glue

4 Results

In order to evaluate each of the BERT models on the test data and compare them, we use a metric called Accuracy. Accuracy is one of the most commonly used metrics in multi-class classification problems and it gives us the proportion of correct predictions of our model out of the set of data. It is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP and TN correspond to the correctly classified elements by the model, while FP and FN are the observations that have been incorrectly classified.

We trained each model given in section 2.1 with a number of epochs equal to 5 and batch size equal to 32. We used the Adam optimizer with a learning rate of 10^{-5}. We also considered a Dropout layer for regularization, with a dropout rate equal to 0.1[9]. This layer helps to prevent the model from overfitting.

The test accuracy for each model evaluated on the test set of each dataset is given in Table 4.

Dataset	BERT_{MINI}	BERT_{SMALL}
Oasis	0.6502	0.6860
MRDA	0.8992	0.8999
Dyda	0.8232	0.8266
Average	0.7908	0.8041

Table 4: Test Accuracy for the BERT models on each datasets and the average accuracy for each model

To compare the models, we use accuracy and the average time spent training the model in each epoch in seconds, which can be seen in Table 5.

Dataset	BERT_{MINI}	BERT_{SMALL}
Oasis	1120.4	3515.6
MRDA	7668.2	10767
Dyda	3367.4	16212.6

Table 5: Average duration of the training by epoch in seconds

[9] We followed the Tensorflow tutorial available in https://www.tensorflow.org/text/tutorials/bert_glue

5 Discussion/Conclusion

We implement Small BERT and Mini BERT models on 3 DA datasets: OASIS, MRDA and Dyda. Results are summarized in Tables 4 and 5. As expected given the size of the datasets, the average duration of the training when both models are implemented is the shortest on the OASIS dataset, as it is the smallest of the 3 datasets. Moreover, the training duration of the Small BERT is superior to the MINI one as it contains less hidden states. However, Table 4 shows that the OASIS Dataset is associated with the lowest accuracy values for both models.

When comparing the models, Table 4 highlights that the model that maximizes the accuracy metric is Small BERT model for every dataset. Indeed, the average accuracy is 0.8041 and it's larger for Small BERT. Given our resources and the fact that the Base BERT model is computationally expensive, we decided to use smaller pre-trained BERT variants which are known for reducing model size while maintaining a good accuracy. The Small BERT gives the best accuracy values. On the other hand, the Mini BERT is associated with a much lower average training time and slightly lower values for the accuracy. Regarding computational time, the Mini BERT is the most suitable when we want to execute DA recognition on resource-restricted devices. It can achieve competitive performances when accomplishing natural language tasks compared to more complex and larger architectures.

To better understand the performance of those smaller architectures of the BERT model, it would be interesting to extent our analysis by implementing the BERT BASE model on the 3 DA datasets to compare its accuracy with the ones we obtained with smaller BERT models. However, estimating this model implies a large computational cost. An alternative would be the Medium BERT model which is still more complex than the models we used but is more lightweight than BERT BASE.

.

References

Joseph Dopp. 1962. Jl austin, how to do things with words. *Revue philosophique de Louvain*, 60(68):704–705.

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.

Daniel Salber and Joëlle Coutaz. 1993. A wizard of oz platform for the study of multimodal systems. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 95–96.

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

Geoffrey Leech and Martin Weisser. 2003b. Generic speech act annotation for task-oriented dialogues. In *Proceedings of the corpus linguistics 2003 conference*, volume 16, pages 441–446. Lancaster: Lancaster University.

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.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

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).

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

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. 2017a. Dailydialog: A manually labelled multi-turn dialogue dataset.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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.

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

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv: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. *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. *EMNLP 2019*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

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. *() AAAI 2020*.

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. 2020a. 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.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020b. Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Pierre Colombo. 2021. *Apprendre à représenter et à générer du texte en utilisant des mesures d'information*. Ph.D. thesis, (PhD thesis) Institut Polytechnique de Paris.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021a. Improving multimodal fusion via mutual dependency maximisation. *() EMNLP 2021*.

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

6 Appendix

6.1 DA Datasets labels distributions

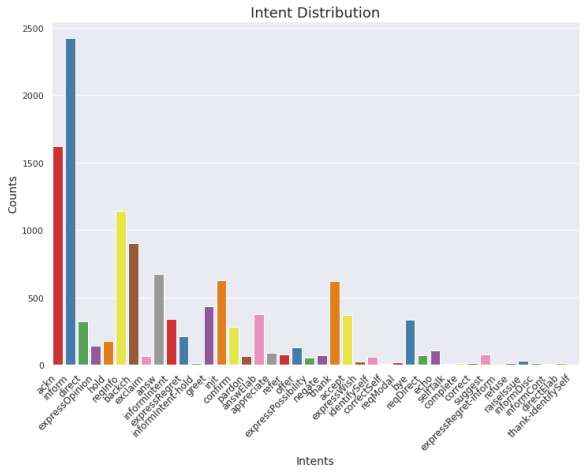

Figure 1: OASIS Dataset labels repartition

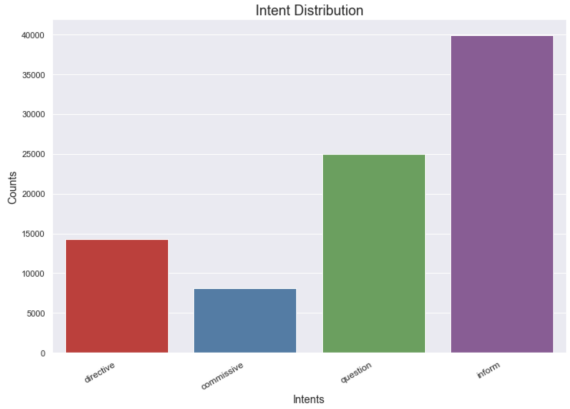

Figure 2: Dyda Dataset labels repartition

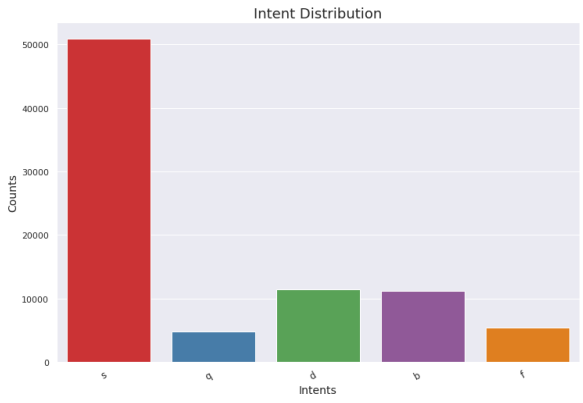

Figure 3: MRDA Dataset labels repartition

6.2 Accuracy of Mini BERT on DA datasets

Figure 4: Accuracy of the Mini BERT on dyda dataset

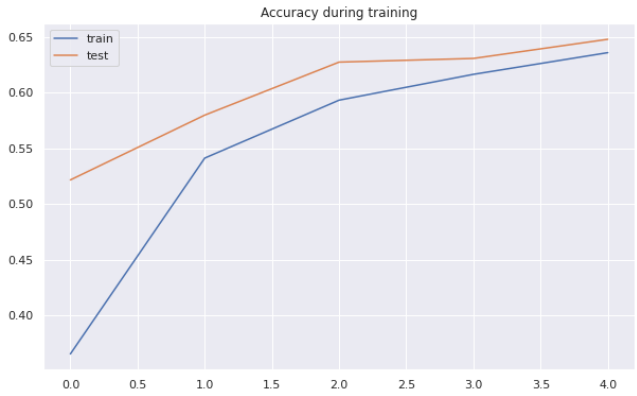

Figure 5: Accuracy of the Mini BERT on Oasis dataset

Figure 6: Accuracy of the Mini BERT on mrda dataset

6.3 Accuracy of Small BERT on DA datasets

Figure 7: Accuracy of the Small BERT on dyda dataset

Figure 8: Accuracy of the Small BERT on OASIS dataset

Figure 9: Accuracy of the Small BERT on mrda dataset