Med-Abbrev Mystery: Decoding Medical Abbreviation Jargon with Transformers

Monica Napoles, Kiara Monahan, David Solow
University of California Berkeley
{mnapoles, kmonahan, david.solow}@berkeley.edu

Abstract

Electronic Health Records (EHRs) are a critical data source for Natural Language Processing (NLP) applications in healthcare. Despite their utility, the widespread use of abbreviations in EHRs can lead to misinterpretations and reduced clarity, posing challenges for clinical decision making. This study aims to improve the interpretation of medical abbreviations in clinical texts by fine-tuning Bidirectional Encoder Representations from Transformers (BERT) models using the Medical Dataset for Abbreviation Disambiguation for Natural Language Understanding (MeDAL), crafted by Wen et al. (2020) containing 5,886 abbreviations with approximately 4 expansions on average for each. The abbreviations come from 14,393,619 medical abstracts on *PubMed*. The fine-tuned BERT models were applied to two medical tasks: mortality prediction and diagnosis prediction. We hypothesize that fine-tuning on medical abbreviations will enhance the models' ability to process clinical text and improve task performance, and that performance improvements can be obtained by fine-tuning Large Language Models (LLMs) on the abbreviation disambiguation task. Results were mixed with some indication that fine-tuning BERT models on abbreviation disambiguation does offer modest performance improvements on downstream mortality and diagnosis prediction tasks in line with those observed in Wen et al. (2020) but we ultimately conclude that there is more exploration which can be done in fine-tuning BERT models on medical abbreviations to improve downstream task performance.

1 Introduction

Electronic Health Records (EHRs) are the most common data source among Natural Language Processing (NLP) applications in healthcare (Hossain et al., 2023). The use of abbreviations are standard for brief and efficient documentation in writing EHRs; however, several terms in this domain have identical abbreviations which can be misinterpreted and reduce clarity. Making these data easier to process by distinguishing shared abbreviations amongst medical text can further improve results in clinical NLP applications (Hossain et al., 2023). We use a sense inventory[1] of medical abbreviations called MeDAL, to fine tune two BERT models. The resulting LLMs are applied to two medical tasks: mortality prediction and diagnosis prediction. The resulting inclusion of domain relevant fine-tuned embedding tokens within our classification models aims to test if the disambiguation of medical abbreviations will help our models learn to process clinical text and improve performance on our tasks. While one of the goals for the MeDAL authors was to create a sense inventory for pretraining models and a secondary goal was to test this on downstream tasks, we wanted to extend the testing on downstream tasks with a different model architecture than the authors used.

2 Background

Typically, information from clinical notes is obtained manually, which is costly, time-intensive, and lacks scalability (Sheikhalishahi et al., 2019). Classification using clinical notes presents a challenge in the field of Natural Language Processing due to the presence of varying structures, domain-specific vocabulary, and use of abbreviations. Thus, there is reason to explore machine learning techniques for the purpose of classification of notes from medical records.

In a study asking physicians, doctors-in-training, and nurses to assign a corresponding long-form to abbreviations found in medical notes, long-forms were identified correctly by 0% to 87% of the volunteers, depending on the abbreviation (Jayatilake and Oyibo, 2023). This exemplifies the difficulty of

[1] A sense inventory in this context refers to a dataset that includes abbreviations and their possible meanings, also called long-forms (Moon et al., 2014).

the task of disambiguation even for human subject area experts as well as the potential for errors in interpretation by clinicians to affect medical care.

Popular model architectures in NLP for mortality and diagnosis prediction include Recurrent Neural Networks (RNNs), Long-Short Term Memory models (LSTMs), and Convolutional Neural Nets (CNNs), however, the application of BERT models on these tasks is less common. Yiyun Chen, a student at Stanford University, applied BERT models to the task of diagnoses prediction and concluded that while BERT does not seem to be the ideal model for this task, there is still potential in pre-processing and fine-tuning that may improve performance (Chen, 2020). For this reason, we were motivated to continue working with BERT models.

Wen et al. (2020) previously showed that model performance on clinical text classification tasks can be improved upon by pre-training ELECTRA transformer-based or LSTM models on an abbreviation disambiguation task. We build upon their work by fine-tuning two BERT models BERT Base Cased (Devlin et al., 2018) and MS BERT (NLP4H, 2024) on the abbreviation disambiguation task using the MeDAL dataset. MS-BERT is a model that has been pre-trained on several clinically relevant text sources: clinical notes from the MIMIC-III notes database, PubMed abstracts, and clinical notes from multiple sclerosis examinations (NLP4H, 2020). MS-BERT was chosen due to its strong prior documented performance in a medical abbreviation disambiguation task (Jaber and Martìnez, 2022). Base BERT cased was chosen due to the presence of capitalization in many abbreviations. We subsequently used the fine-tuned BERT models to predict patients' diagnosis and mortality using clinical notes from The Medical Information Mart for Intensive Care-III (MIMIC-III).

The MIMIC-III database is composed of deidentified clinical data from over 40,000 patients who were admitted to critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016b). This data can be accessed by completing CITI Data or Specimens Only Research training and signing a Data Use Agreement related to HIPAA protections and responsible use of the data (Johnson et al., 2016a).

3 Methods

3.1 Data Preprocessing

The number of expansions in MeDAL is 22,555 and the number of unique abbreviations is 5,798. For fine-tuning BERT on the medical abbreviation disambiguation task, the train, test, and validation splits from MeDAL were used since the authors implemented a class-balancing algorithm that iteratively selected a certain threshold of samples of each expansion and removed classes in increasing frequency (Wen et al., 2020). We observed that most of the abbreviations in the text occurred before the 200th position before tokenization and to save computing costs we decided to discard any records whose abbreviation fell outside of the first 200 words shown in Figure 1. This step filtered out about 500,000 rows in the train set, which had an initial size of 3 million rows. Tokenization of input texts was performed using the associated tokenizers for the BERT Base cased and MS-BERT LLMs . Given that location labeling provided in MeDAL is word-based, adjusted start and end locations were tracked and tested during tokenization to confirm our models identified the abbreviation positions. As previously stated, only samples where the labeled abbreviation remained in the sequence after tokenization and truncation (i.e. the adjusted end location was less than or equal to the maximum sequence length) were kept.

For our downstream tasks, only medical notes in the MIMIC-III database written by nurses or physicians at least 24 hours before hospital discharge were utilized since patients that expire tend to have discharge notes within those 24 hours mentioning their outcome. To balance classes, a maximum of 4 notes were kept per individual patient if they survived given that most do in accordance with the processing done by the authors of MeDAL. One difference with our processing of this text was that we kept casing given we utilized the BERT Base Cased model and theorized that casing may be important in regards to learning abbreviations. We also excluded "Nursing/Other" notes, a category of note we thought was less informational compared to the content of the "Nursing" and "Physician" note categories. Furthermore, we excluded any records from these notes that did not contain an abbreviation that appeared in MeDAL. In pre-processing, a head and tail truncation approach was used as many of the notes exceeded the max number of tokens (512) that can be input into the chosen BERT models. We chose the head and tail method as it resulted in the strongest performance compared to head-only or tail-only truncation for classification with BERT using long text in Sun et al. (2019).

3.2 Model Architecture

For our fine-tuned models, a layer was added to extract the hidden states specific to the abbreviations' tokens using the adjusted abbreviation start and end locations. Non-abbreviation embeddings are converted to padding as shown in Figure 2. Abbreviation token embeddings were then pooled by taking the average and ignoring the padded tokens. After extraction, a final softmax classification layer with a size of 22,555 (the total number of expansions) was used to classify the abbreviation to its long-form utilizing the pooled hidden states from the abbreviation location as the input. Sparse categorical cross-entropy was used to calculate loss. The fine-tuned BERT weights were exported for later application to the downstream task. See Figure 3 for a diagram of the complete model.

The mortality and diagnosis prediction models were constructed beginning with the chosen BERT model and feeding the pooled BERT output into a dense layer, with a final sigmoid classification layer illustrated by Figure **??**. For mortality prediction, 2 classes were used: "died in hospital" or "did not die in hospital". For diagnosis prediction, multi-label classification was used in the model used to predict diagnosis probabilities. Binary cross-entropy was used as the loss function and our optimizer was Adam. To evaluate diagnosis prediction, top-k recall was calculated as:

$$\text{Top-}k\text{ recall} = \frac{N_k}{N_{\text{total}}} \qquad (1)$$

where N_k is the number of true positives in the top K predictions of the model, and N_{total} is the total number of diagnoses listed for the patient's admission (Choi et al., 2015).

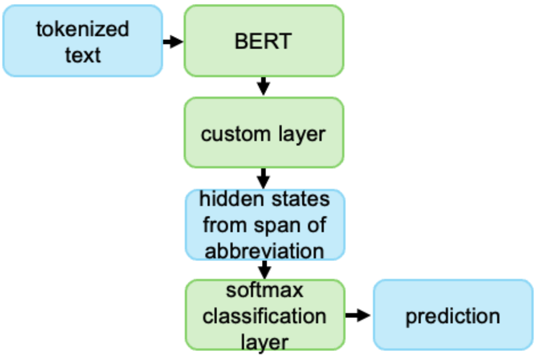

tokenized text
BERT
custom layer
hidden states from span of abbreviation
softmax classification layer
prediction

Figure 3: Structure of model used for the abbreviation disambiguation fine-tuning task.

We present the following eight models: baseline Base BERT cased for mortality prediction, baseline MS-BERT for mortality prediction, Base BERT cased fine-tuned on abbreviation disambiguation then used for mortality prediction, MS-BERT fine-tuned on abbreviation disambiguation then used for mortality prediction, baseline Base BERT cased for diagnosis prediction, baseline MS-BERT for diagnosis prediction, Base BERT cased fine-tuned on abbreviation disambiguation then used for diagnosis prediction, and MS-BERT fine-tuned on abbreviation disambiguation then used for mortality prediction.

4 Results and Discussion

4.1 Results

Training and validation results for our fine-tuned models are reported in Figure 9 in Appendix 1.

In keeping with the authors of MeDAL we report validation metrics for our downstream tasks. Figures 5 and 6 show the loss and validation scores for our baseline and fine-tuned models on mortality prediction.

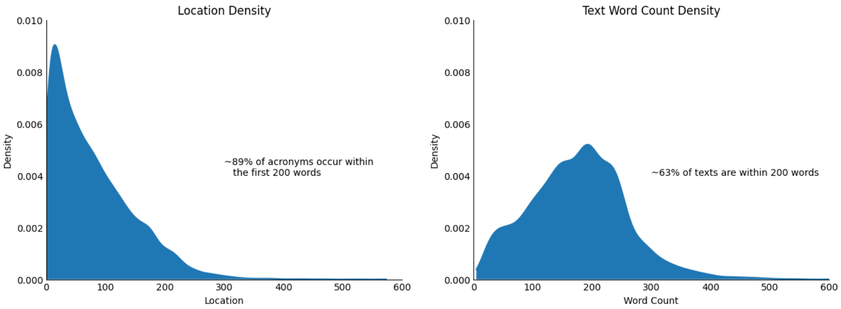
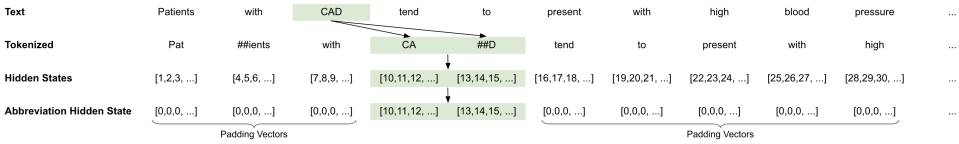


Figure 1: Abbreviation Location and Text Word Count Densities in MeDAL train subset.



Text	Patients	with	CAD	tend	to	present	with	high	blood	pressure	...
Tokenized	Pat	##ients	with	CA	##D	tend	to	present	with	high	...
Hidden States	[1,2,3, ...]	[4,5,6, ...]	[7,8,9, ...]	[10,11,12, ...]	[13,14,15, ...]	[16,17,18, ...]	[19,20,21, ...]	[22,23,24, ...]	[25,26,27, ...]	[28,29,30, ...]	...
Abbreviation Hidden State	[0,0,0, ...]	[0,0,0, ...]	[0,0,0, ...]	[10,11,12, ...]	[13,14,15, ...]	[0,0,0, ...]	[0,0,0, ...]	[0,0,0, ...]	[0,0,0, ...]	[0,0,0, ...]	...

Figure 2: Abbreviation Hidden State Extraction.

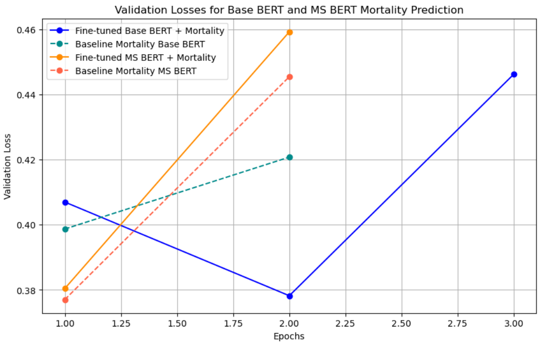


Figure 5: Validation Losses for Baseline and Fine-tuned Base BERT and MS BERT models on Mortality Prediction.

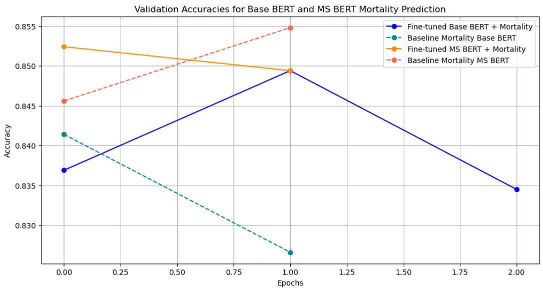


Figure 6: Validation Accuracies for Baseline and Fine-tuned Base BERT and MS BERT models on Mortality Prediction.

The results of the various models for the mortality prediction task are shown in Table 1. For Base BERT, fine-tuning on the abbreviation disambiguation task slightly improved both accuracy, recall and F1 score. Recall is included over precision given the asymmetrical impact of missing high mortality risk or high risk diagnoses in patients.

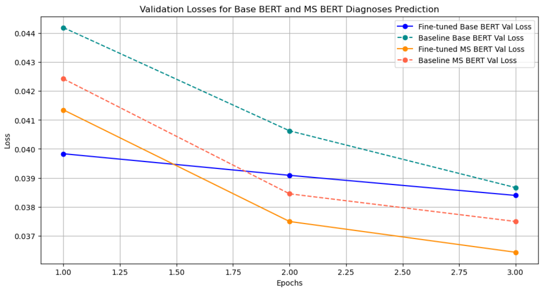


Figure 7: Validation Losses for Baseline and Fine-tuned Base BERT and MS BERT models on Diagnoses Prediction.

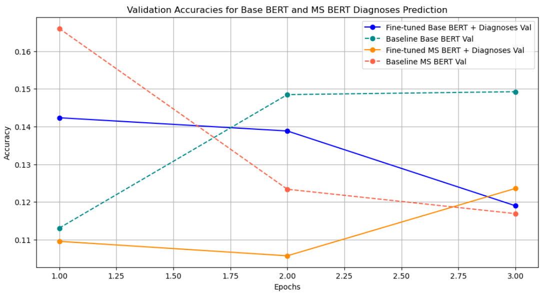


Figure 8: Validation Accuracies for Baseline and Fine-tuned Base BERT and MS BERT models on Diagnoses Prediction.

The results of the various models for the diagnoses prediction task are shown in Table 2. We saw minimal improvement with some diagnoses recall scores but overall it is difficult to tell whether these are significant. Notably, a recall of 100 % is not

BERT model	Accuracy	Recall	F1 Score
Baseline Base BERT cased	0.77	0.50	0.76
Baseline MS-BERT	0.77	0.73	0.76
Fine-tuned Base BERT cased	0.79	0.80	0.78
Fine-tuned MS-BERT	0.78	0.74	0.82

Table 1: Results of models on the mortality prediction task. *Fine-tuned* means BERT was previously fine-tuned on abbreviation disambiguation before being used in the mortality prediction model. Metrics shown are calculated from the test set.

feasible due to the varying amount of diagnoses for each patient.

4.2 Discussion

Our 2.5% accuracy improvements for the mortality prediction task falls within the range of 0.2%-5% test accuracy improvements gained in Wen et al. (2020) from pre-training on the abbreviation disambiguation task. Additionally, in Wen et al. (2020) there was over a 70% improvement in top k recall for their diagnosis prediction models. While our improvements on diagnosis prediction were unimpressive compared to that in previous literature, they do not rule out the use of our approach. For example, better performance may have been obtained on diagnosis prediction using our approach with additional pre-processing steps, adjustments to fine-tuning, or using alternative data (i.e. using Discharge Summaries in place of physician and nursing notes for diagnosis prediction, which may be more appropriate diagnosis but less so for mortality given that they will include mortality outcomes) may help this method work more consistently.

A possible reason for the performance improvements being so modest is that the abbreviations included in the MeDAL dataset, which pulls from biomedical article abstracts, may not include all of the abbreviations present in clinical notes. Additionally, the writing styles and contents of formal paper abstracts may contrast with that of clinical notes, and thus fine-tuning on paper abstracts may not transfer well to clinical notes in MIMIC-III. To this end, we did note very significant differences in vocabulary and writing style between the two datasets. This is evident as nearly half of the abbreviations in MeDAL did not appear in our MIMIC-III subsets.

In the future, we would like to use a dataset containing text more relevant to clinical notes for the abbreviation disambiguation task as this may allow for larger improvements on model performance on downstream tasks or continue to explore different model architectures.

Limitations

There are several limitations that have impacted our work. While BERT language models are powerful and convenient due to being pre-trained, the input size restriction is limiting in the realm of clinical notes, which can sometimes be upwards of 2,000 words long. While the diagnosis and mortality classification models performed relatively well, the head and tail truncation method may have cut out important information relevant to the task. In future work we would like to explore methods of combining multiple BERT outputs or other transformer-based models that have more flexibility in terms of text length. The difficulty of utilizing BERT with clinical notes was also observed by Chen (2020).

Computational limitations also affected the fine-tuning on the abbreviation-disambiguation task, causing us to limit records and max input length. For most of our models we tried running at least three epochs but due to increasing computational and time restrictions, we could only afford two epochs for some. We decided to run two since our models began to overfit or plateau around this amount.

Ethics Statement

Ethical considerations were made when planning on the use of data from the MIMIC-III dataset due to the sensitive nature of clinical information from patients. Authors of this study completed CITI Data or Specimens Only Research training and read and acknowledged a Data Use Agreement for the database, which contained important restrictions on use of the data and Health Insurance Portability and Accountability Act (HIPAA) information. No attempts were made to identify individual patients. All clinical information in MIMIC-III is thoroughly de-identified which allows for public use of the dataset (Johnson et al., 2016a). More

BERT model	Top 5 Recall	Top 10 Recall	Top 30 Recall
Baseline Base BERT cased	0.20499	0.33434	0.59601
Baseline MS-BERT	0.22148	0.35702	0.62408
Fine-tuned Base BERT cased	0.20275	0.33586	0.60572
Fine-tuned MS-BERT	0.21064	0.34624	0.61733

Table 2: Results of models on the diagnoses prediction task rounded to 5 significant figures. *Fine-tuned* means BERT was previously fine-tuned on abbreviation disambiguation before being used in the diagnoses prediction model. Metrics shown are calculated from the test set.

information regarding the ethical use of MIMIC-III can be found on the site *PhysioNet* (Johnson et al., 2000).

Furthermore, the ethical implications of predictive modeling from clinical notes must be considered. It is very important to note that diagnosis and mortality predictions resulting from our models should not be interpreted as medical advice or information. Any use of predictive models in a healthcare setting must be thoroughly evaluated for that use case and for accuracy, usefulness, and potential biases.

Acknowledgements

We would like to acknowledge the authors of Wen et al. (2020) for their work pre-training models on medical abbreviation disambiguation before use on downstream classification tasks on clinical notes, which is what our study builds on, as well as for the creation of the publicly available MeDAL dataset. We would especially like to thank Zhi Wen for their advice on the direction of our abbreviation disambiguation model.We would also like to thank Jennifer Zhu, W266 Natural Language Processing with Deep Learning Lecturer at UC Berkeley for signing off on our application to access the MIMIC-III data.

References

Yiyun Chen. 2020. Predicting icd-9 codes from medical notes – does the magic of bert applies here?

Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. 2015. Doctor AI: predicting clinical events via recurrent neural networks. *CoRR*, abs/1511.05942.

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.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. In *Computers in Biology and Medicine*, volume 155.

Areej Jaber and Paloma Martìnez. 2022. Disambiguating clinical abbreviations using a one-fits-all classifier based on deep learning techniques. In *Methods of information in medicine*, volume 61, pages e28–e34.

Dineth Jayatilake and Samson Oyibo. 2023. Interpretation and misinterpretation of medical abbreviations found in patient medical records: A cross-sectional survey. In *Cureus*, volume 15.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. Mimic-iii clinical database (version 1.4).

Alistair Johnson, Tom Pollard, and Mark Roger. 2000. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. In *Circulation*, volume 101, pages e215–e220, Online.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Anthony Celi Leo, and Roger Mark. 2016b. Mimic-iii, a freely accessible critical care database. In *Scientific Data*, volume 3. Springer Nature.

Sungrim Moon, Serguei Pakhomov, Nathan Liu, James Ryan, and Genevieve Melton. 2014. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. In *Journal of the American Medical Informatics Association: JAMIA*, volume 21, pages 299–307.

NLP4H. 2020. Ms-bert: Using neurological examination notes for multiple sclerosis severity classification.

NLP4H. 2024. ms_bert. `https://huggingface.co/NLP4H/ms_bert`. Accessed: 2024-08-05.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural language processing of clinical notes on chronic diseases: Systematic review. In *JMIR medical informatics*, volume 7. JMIR Publications Inc.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Zhi Wen, Xing Han Lu, and Siva Reddy. 2020. MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135, Online. Association for Computational Linguistics.

Appendix

Appendix 1: Abbreviation Disambiguation Performance

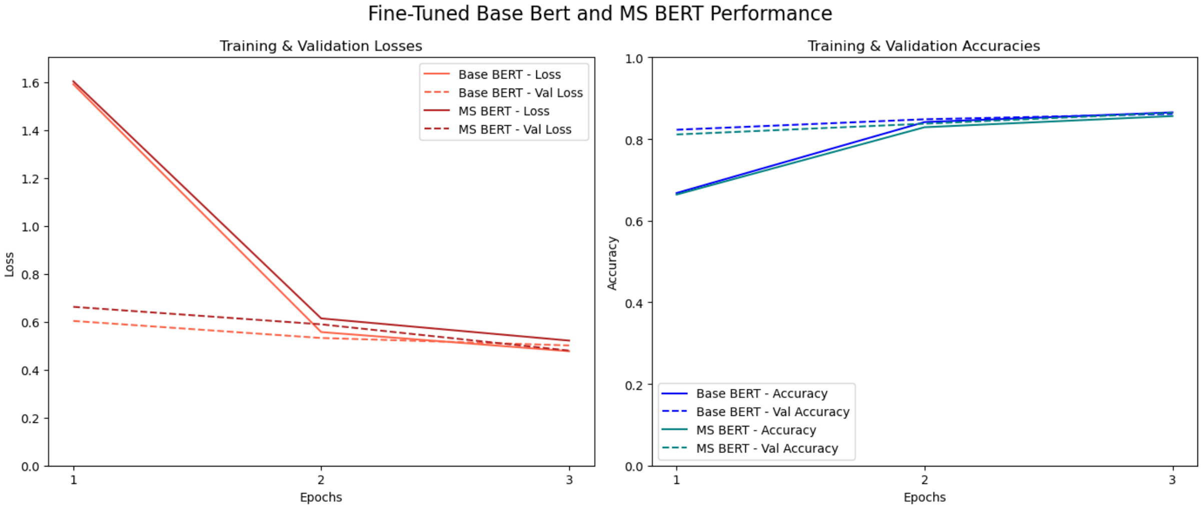


Figure 9: Losses and accuracies for the abbreviation disambiguation model.