Investigating the Impact of Freezing Layers in Pre-trained BERT Models for Enhanced Fake News Classification

# Abstract

The growth of fake news results in challenges for the integrity of information and trust. Artificial Intelligence and Models like BERT (Bidirectional Encoder Representations from Transformers) offer ways to detect fake news. This report explores the impact of freezing layers in pretrained BERT models on the accuracy of classifying fake news across domains like crime, health, politics, science, and social media.

The approach employs a two step system: first, classifying news articles into specific domains and then applying domain specific models to determine if the news is fake or not. Experiments were conducted to compare the performance of models with frozen layers against those with unfrozen layers by evaluating key metrics like Precision, Recall, F1-Score, Accuracy, and G-Mean. A baseline model with a 1 step structure involving a single model to classify news from all domains as fake or not fake was also trained to have baseline metrics in order to compare to domain specific metrics.

Results indicate that freezing layers can lead to improved precision in certain domains such as crime while reducing recall. In contrast, domains like politics showed less performance with frozen layers thereby emphasising the need for adaptability in complex linguistic environments. This offers insights into optimising AI models for fake news detection to mitigate misinformation.

# Introduction

### Background

Classifying fake news in 2024 is hard. Misleading or false information on social media platforms spreads rapidly. Artificial intelligence (AI), specifically deep learning and large language models (LLMs), offers a potential solution to this.

### Deep Learning and Large Language Models

Deep learning has revolutionised the field of natural language processing (NLP). Large language models like BERT (Bidirectional Encoder Representations Transformers) have revolutionised text classification, sentiment analysis, and more. These models use transformer architectures allowing processing text bidirectionally.

### Previous Solutions and Gaps

Existing solutions for detecting fake news use traditional machine learning techniques which lack the depth of important contextual meaning. Their application to fake news detection remains suboptimal due to domain variability and complexity of patterns in information contained in false news. Large language models are also prone to overfitting and can require significant resources to train thereby limiting scalability.

### Motivation

Given the limitations of current fake news detection approaches, this report explores a technique to improve the accuracy of BERT models in detecting fake news. The core motivation is to determine if freezing layers in pretrained BERT models can lead to better performance in classifying fake news.

### Research Question

Does freezing model layers in pre-trained BERT models improve accuracy for classifying fake news?

### Problem Statement

Classifying fake news involves complex language patterns, context specific language, and a lot of variability across domains. This report evaluates the impact of freezing layers in pre-trained BERT models to explore if this can improve the model's accuracy.

# Literature and Background

### Detecting Fake News

Fake news detection went from rule based to more sophisticated approaches. Early methods relied on manual fact checking and feature based techniques. But they lacked scalability and accuracy in rapidly changing environments.

A development was deep learning methods for fake news detection. For example, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) became popular due to their ability to capture textual patterns. Shu et al. (2017) used a model combining CNNs and RNNs to detect fake news.

Liu et al. (2019) also employed a Graph Convolutional Network (GCN) to detect fake news. This involved social context into the detection process thereby recognizing that fake news has different contexts than news. This showed the importance of context in fake news detection.

### Transfer Learning in Fake News Detection

Wu et al. (2019) explored using transfer learning for fake news detection and using a pre-trained BERT model. He demonstrated that fine tuning a BERT model on domain specific data improved accuracy and robustness and that transfer learning could be a technique viable solution to fake news detection.

### Domain-Specific Models in Fake News Detection

Chen et al. (2020) employed a pre-trained BERT model fine-tuned with domain specific data used in health articles with the results demonstrating a noticeable improvement in accuracy with domain specific data.

### Large Language Models and Deep Learning Models

BERT, introduced by Devlin et al. (2018), uses a transformer architecture allowing it to read text bidirectionally to capture contextual information. The original BERT model comes in these sizes: BERT-base (12 transformer layers with 768 hidden units) and BERT-large (24 transformer layers with 1024 hidden units).

Llama is another LLM which is designed to be more resource efficient while maintaining high performance. Llama comes in sizes from smaller models (7 billion parameters) to larger ones (65 billion).

Despite their versatility, BERT and Llama require tedious fine tuning for tasks. This report aims to determine whether freezing layers in these models during fine tuning can lead to improved performance in fake news detection. Transfer learning and domain specific adaptation techniques will be used also.

# Methodology and Experimental Setup

### System Overview

The core system for this project is a two-phase BERT based model. First, the text is classified into specific domains and then second it is determined whether the news is fake or not. Figure 1 shows this where the first phase is a multi-domain classifier and the second phase is domain specific models. A baseline model to be used for baseline results and comparison was also trained. This baseline model utilised a 1 stage approach with no domain classification. Figure 2 shows the architecture of the baseline model.

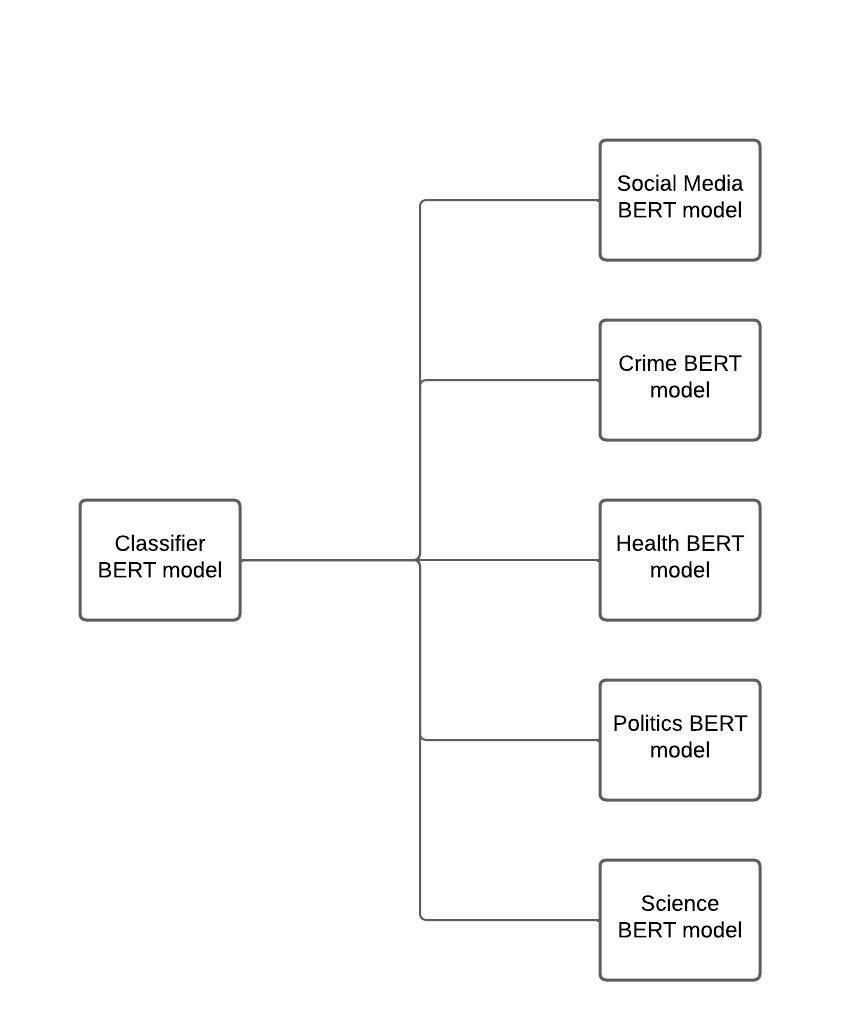


Figure 1. Two Step Architecture

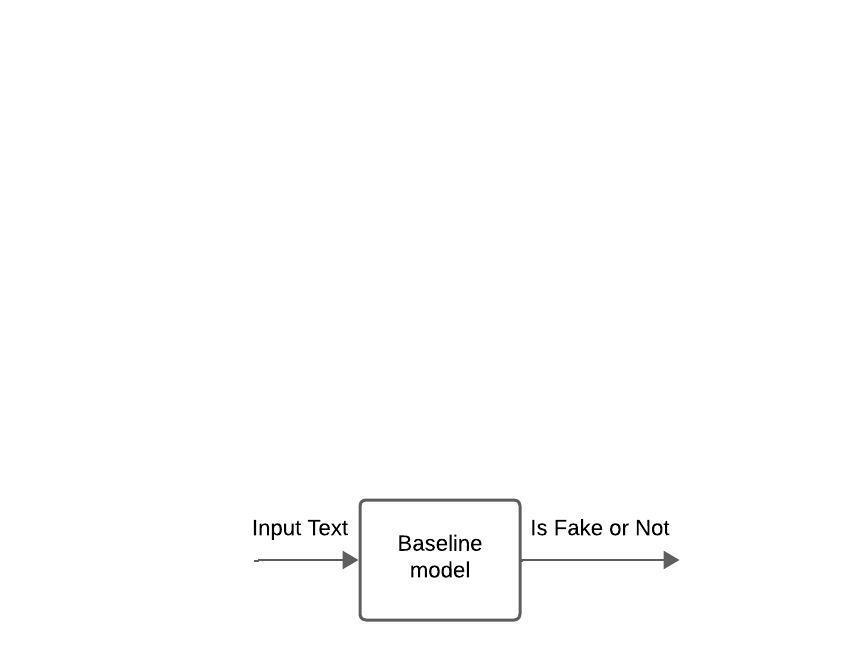


Figure 2. Baseline Model Architecture

### Dataset Details and Preprocessing

Here is a run through of the dataset sources for training, validation, and testing for each domain:

* Crime:
  + Training:FA-KES
  + Testing:snope
* Politics
  + Training:FakeNews + ISOT + LIAR
  + Testing:Pheme + Politifact
* Health:
  + Training:COVID-FN + COVID-FNIR
  + Testing:COVID\_Claims
* Science
  + Training:climate
  + Testing:isot\_science\_part
* Social Media
  + Training:gossioCob
  + Testing:isot\_social\_part
* Baseline
  + Training: All training datasets mentioned above
  + Testing: All testing datasets mentioned above

Preprocessing involved several steps:

* Text Cleaning Removing HTML tags, special characters, and redundant whitespace
* Tokenization and converting text into tokens using BERT tokenizer.
* Transforming the dataset so that the label column is binary with 0 and 1 labels representing not fake or fake news.
* Transforming data columns so that there is a centralised text column representing model input text.

### Models and Choice Rationale

Transfer learning means adapting a model developed for one task to a related task and using its pre-learned features. BERT for example is a pretrained model which is pretrained on text and understands complex language patterns which makes it suitable for data with news articles. By training our data on a pre-trained BERT model and using transfer learning, it expedited the training process and ensured that the model performs well on the specialised fake news classification task.

Figure 3. Shows the BERT models used for each domain for the second step in this process. For the first stage Classifier a DistillBert variant was used while the baseline model also used a DistillBert.

| Domain | BERT Type |
| --- | --- |
| Crime | Albert |
| Health | Albert |
| Politics | Albert |
| Science | Albert |
| Social Media | DistilBert |
| Baseline | DistilBert |

Figure 3. Domain specific BERT models

The Albert version used is BERT-base with 12 transformer layers, 768 hidden units, and 12 attention heads. BERT-base is preferred over BERT-large for this project because it offers a good balance between performance and resource requirements. Additionally, BERT-base has fewer parameters (about 110 million), making it more manageable for experimentation and deployment on smaller-scale hardware. On the other hand, the DistillBert model used in this experiment contains 6 transformer layers, 768 hidden units, and 12 attention heads.

The ‘Llama-7B’ model variant from the Python transformers library was also considered but its 7 billion parameter variant was found less effective than BERT due to increased complexity and higher demand for bidirectional context. No metrics are available for LLama comparison as not even 1 epoch could be completed.

### Training Environment

Training was conducted on Google Collab with the following specifications:

* Processor: Intel(R) Xeon(R) CPU, 2.20GHz
* GPU: NVIDIA GeForce Tesla T4
* RAM: 15 GB

The system used Python 3.8, TensorFlow, and PyTorch. The trainer used to train the models was the default Trainer module from the Transformers library in Python.

### Training

First, each domain BERT model was trained on their respective datasets for 5 epochs using Stochastic Gradient Descent with the Adam optimizer. It was found that training and validation loss stabilised before 5 epochs across all domains therefore a threshold of 5 epochs was chosen. The baseline however was trained on 1 epoch due to resource constraints from training the large merged dataset from every domain. The parameters for training were then optimised according to iterative tuning and the optimal epoch for each domain was selected based on best validation and training loss.

This process was done to maintain consistency across domains while allowing for differences in training times based on complexity and size of the datasets and domain. The final optimised parameters on which the domain specific models were trained on are outlined in Figure 4.

| Parameter | Value |
| --- | --- |
| Weight decay | 0.01 |
| Warmup Steps | 500 |
| Per Device Train Batch Size | 8 |
| Per Device Eval Batch Size | 8 |
| Learning Rate | 5E-05 |
| Batch Size | 16 |
| Weight Decay | 0.01 |
| Warmup Steps | 500 |
| Gradient Clipping | 1.0 |
| Epochs | 5 (Crime, Politics), 2 (Health), 4 (Science), 1 (Social Media, Baseline) |

Figure 4.

The largest difference was noticed in varying the learning rate. Increasing it to 1E-04 was beneficial in domains with large datasets like Politics and Social Media as it made the loss converge very fast. However, this learning rate made domains like Science not converge at epochs less than 10. Keeping the loss at 5E-05 did not negatively affect Politics and Social Media however, so the learning rate was set at this value.

With gradient clipping, lowering to as low as 0.5 resulted in stable validation loss for the Health domain but increasing it to as high as 2.0 resulted in suffering validation loss albeit improved training loss. A 1.0 gradient clipping value was found the most effective.

Additionally, batch size and warmup steps were also tuned and the optimal values of 16 and 500 were arrived at across all domains.

### Introducing Layer Freezing

Once the models were trained with the optimal parameters, the models were then trained again but this time with layer freezing to determine its impact on performance. The way layers were frozen involved locking transformer layers during training to prevent weight updates. Layers were frozen starting from the first layer.

One to three layers frozen was experimented and optimised for each model as shown below and the performance metrics were compared to the initial models where no layers were frozen which is discussed in the next section.

| Model | Optimal # of Layers to Freeze |
| --- | --- |
| Crime | 2 |
| Health | 2 |
| Science | 1 |
| Politics | 2 |
| Social Media | 1 |
| Baseline | 1 |

### Training Loss and Validation Loss

For both freezing layer models and un-frozen layer models, the data was split into training and validation sets with a 90-10 ratio. This split provided good enough data to train while at the same time maintaining validation sets for evaluation. The Training Loss and Validation Loss figures for sample models are shown below in Figure 5.

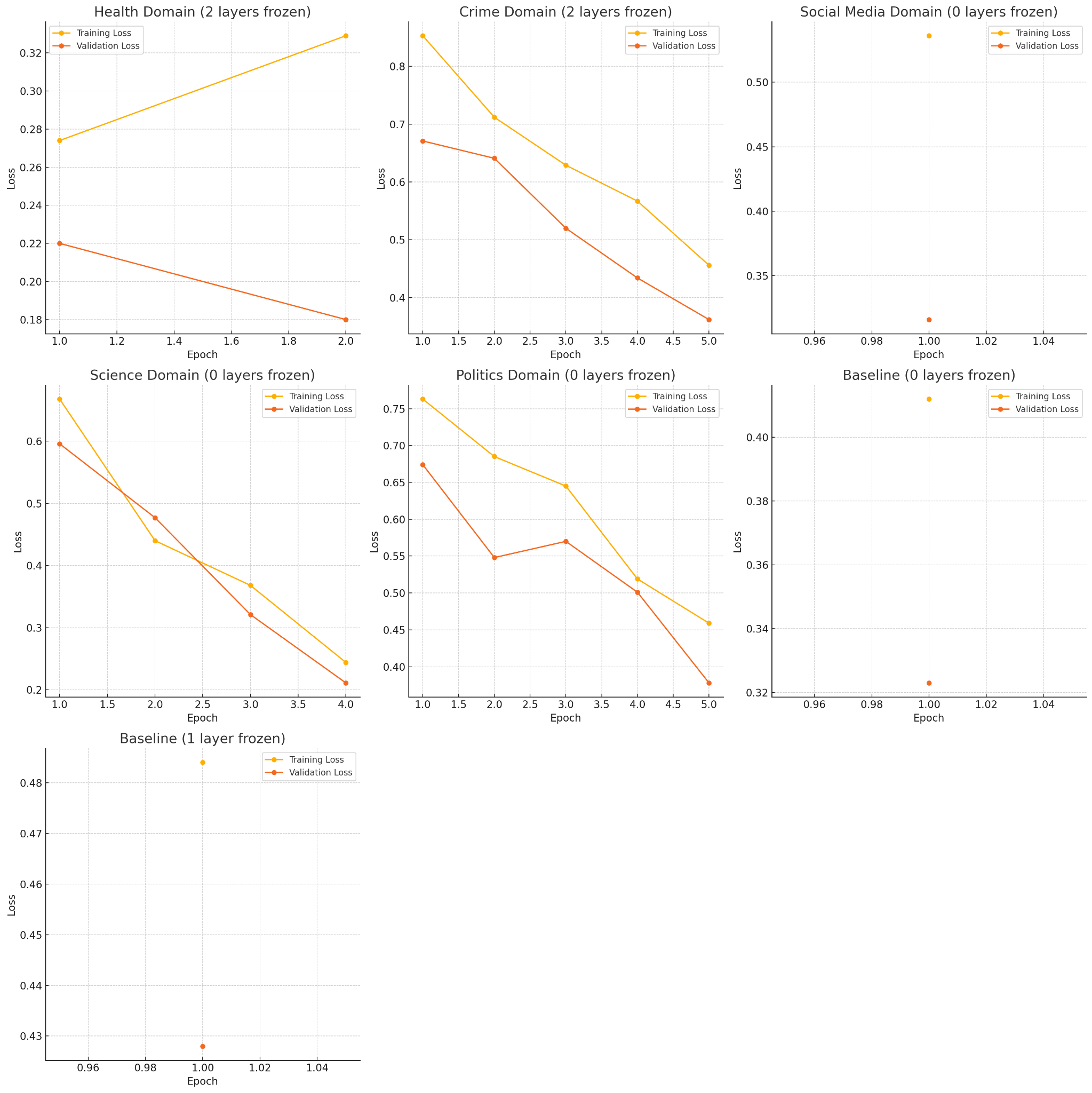


Figure 5. Sample Training/Validation Loss for some of the models trained

### 

### Testing the Final Models

Once the frozen and unfrozen models were trained they were tested using these datasets respectively:

* Crime:
  + snope
* Politics
  + Pheme + Politifact
* Health:
  + COVID\_Claims
* Science
  + isot\_science\_part
* Social Media
  + isot\_social\_part
* Baseline
  + All datasets mentioned above

Precision, Recall, F1-Score, Accuracy, and G-Mean was also calculated and data for True Positives, True Negatives, False Positives, and False Negatives was collected and analysed as discussed below

# Results and discussion

Here is a summary of the results for freezing vs not freezing model layers for domain specific models:

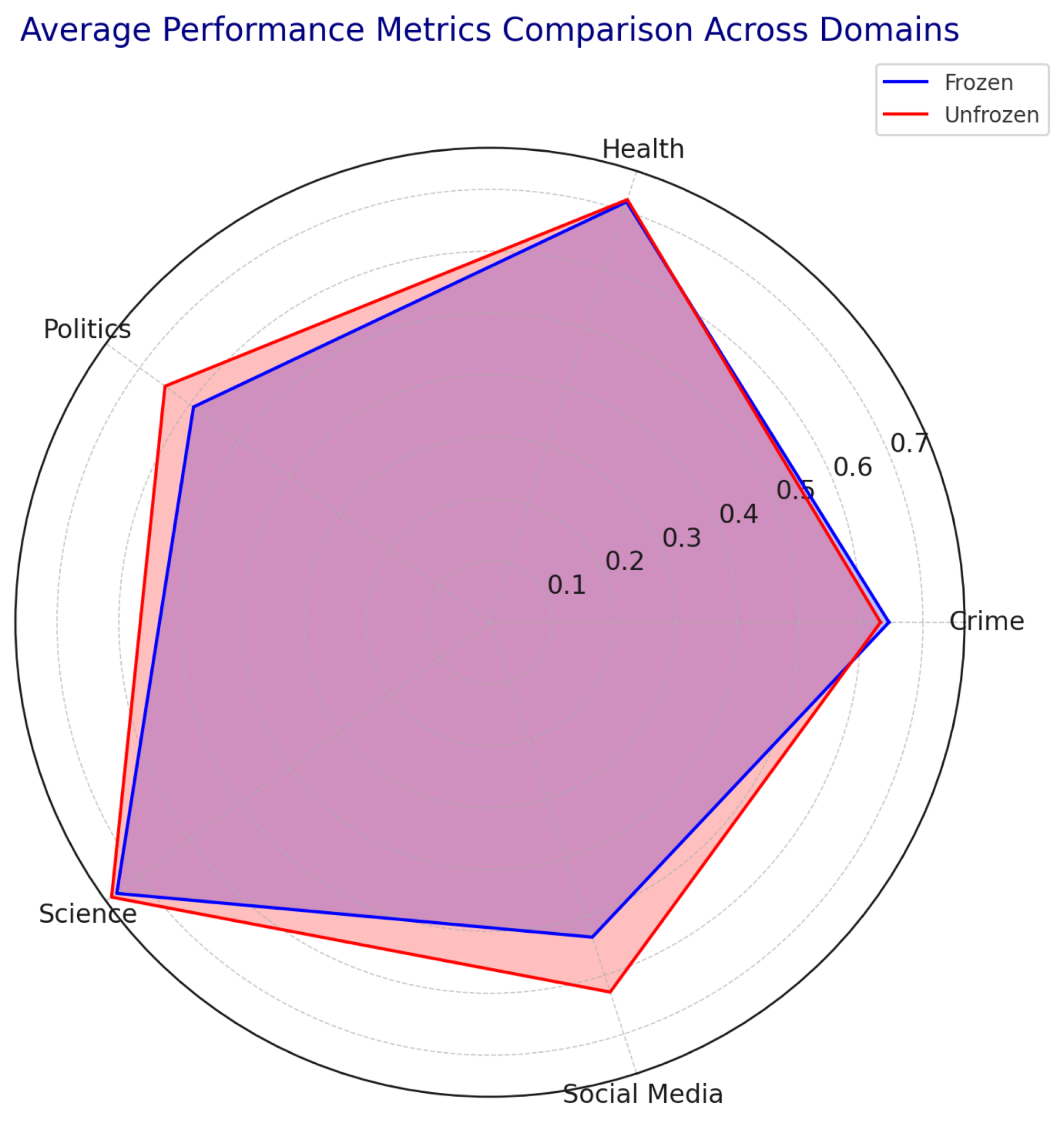


Figure 6.

Average Performance Metrics Across Domains

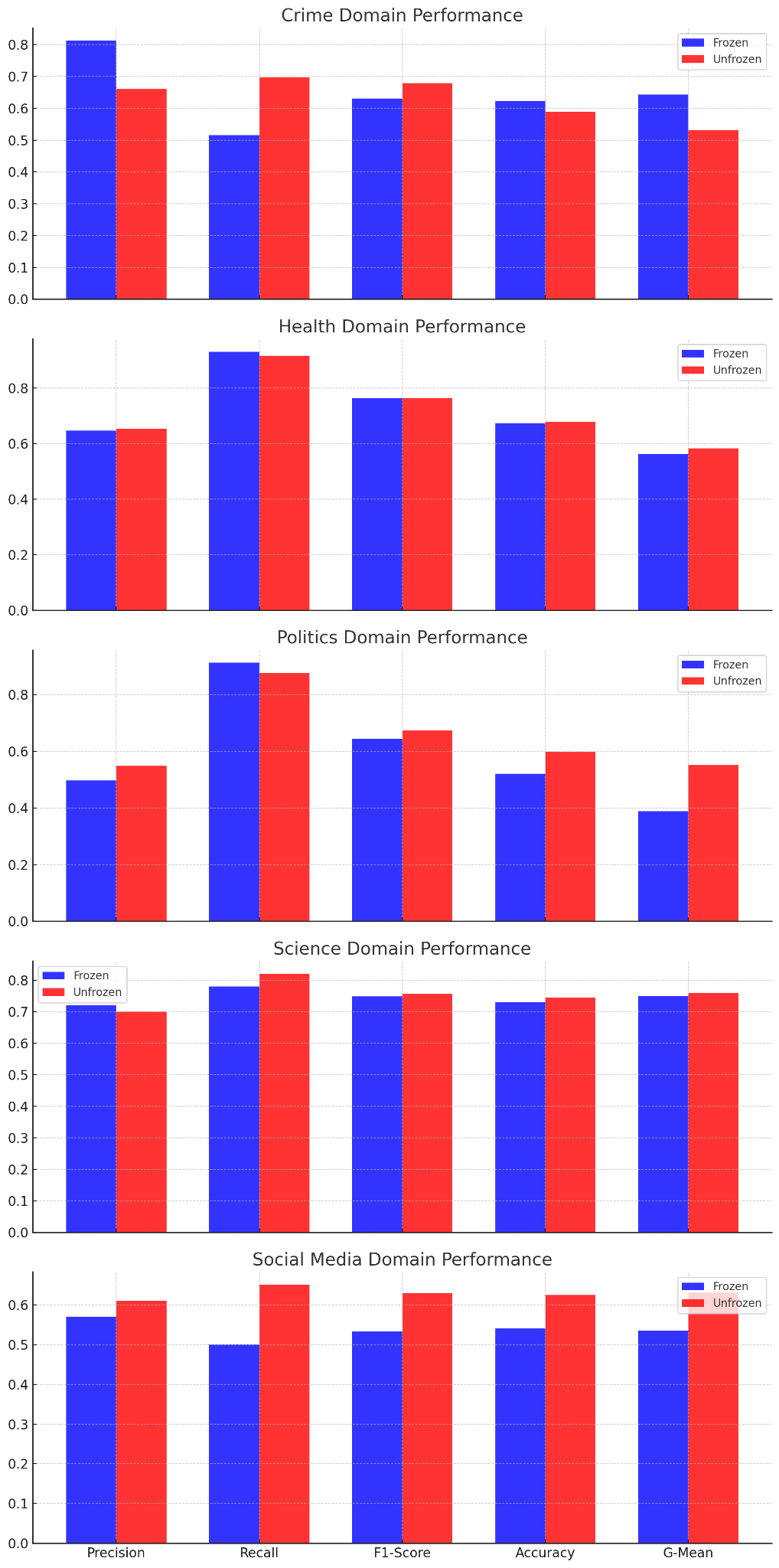
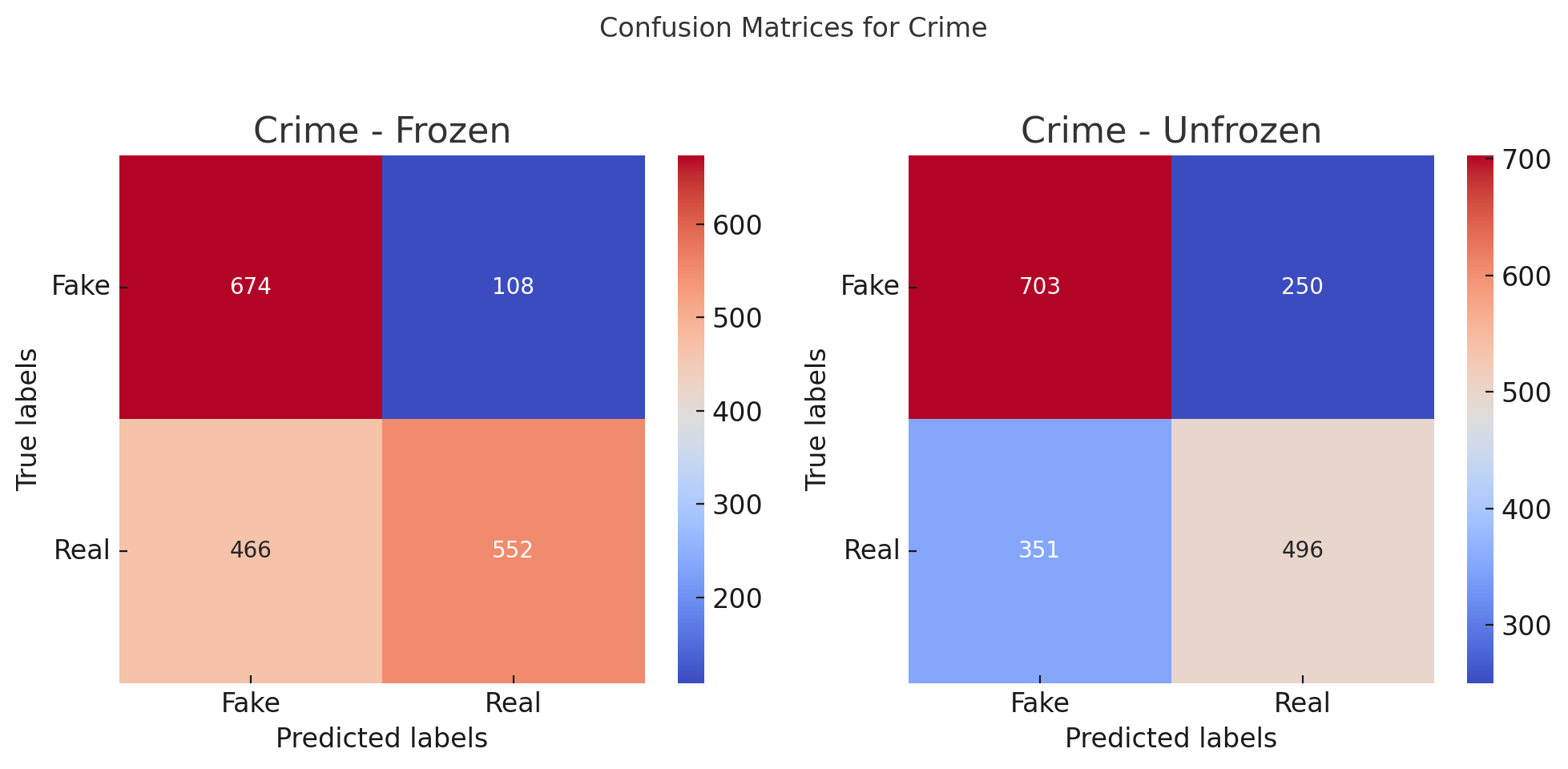


Figure 7.

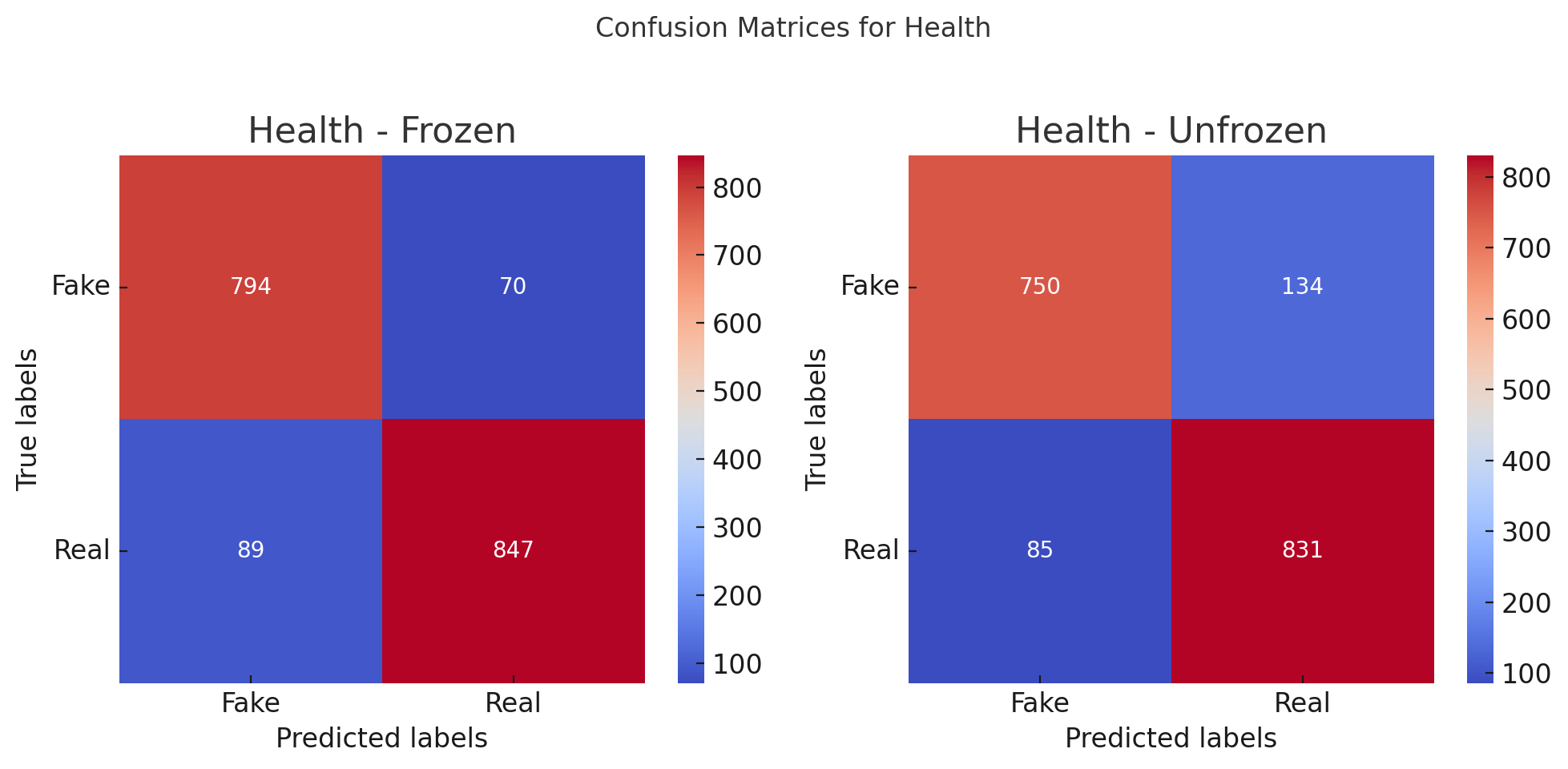
Precision, Recall, F1-Score, Accuracy, G-Mean Metrics Per Domain Frozen vs Unfrozen

### Scenario 1:Crime Domain with Frozen Layers Vs Unfrozen Layers



In the Crime, frozen layers show high precision but low recall which suggests that a conservative model with less false positives but more false negatives. This is probably due to the structured nature of crime related content which allows precise classification but limits the flexibility for the actual model. In contrast, the unfrozen model exhibits a better balance between precision and recall which indicates increased adaptability and lesser rate of false negatives thereby contributing to better recall.

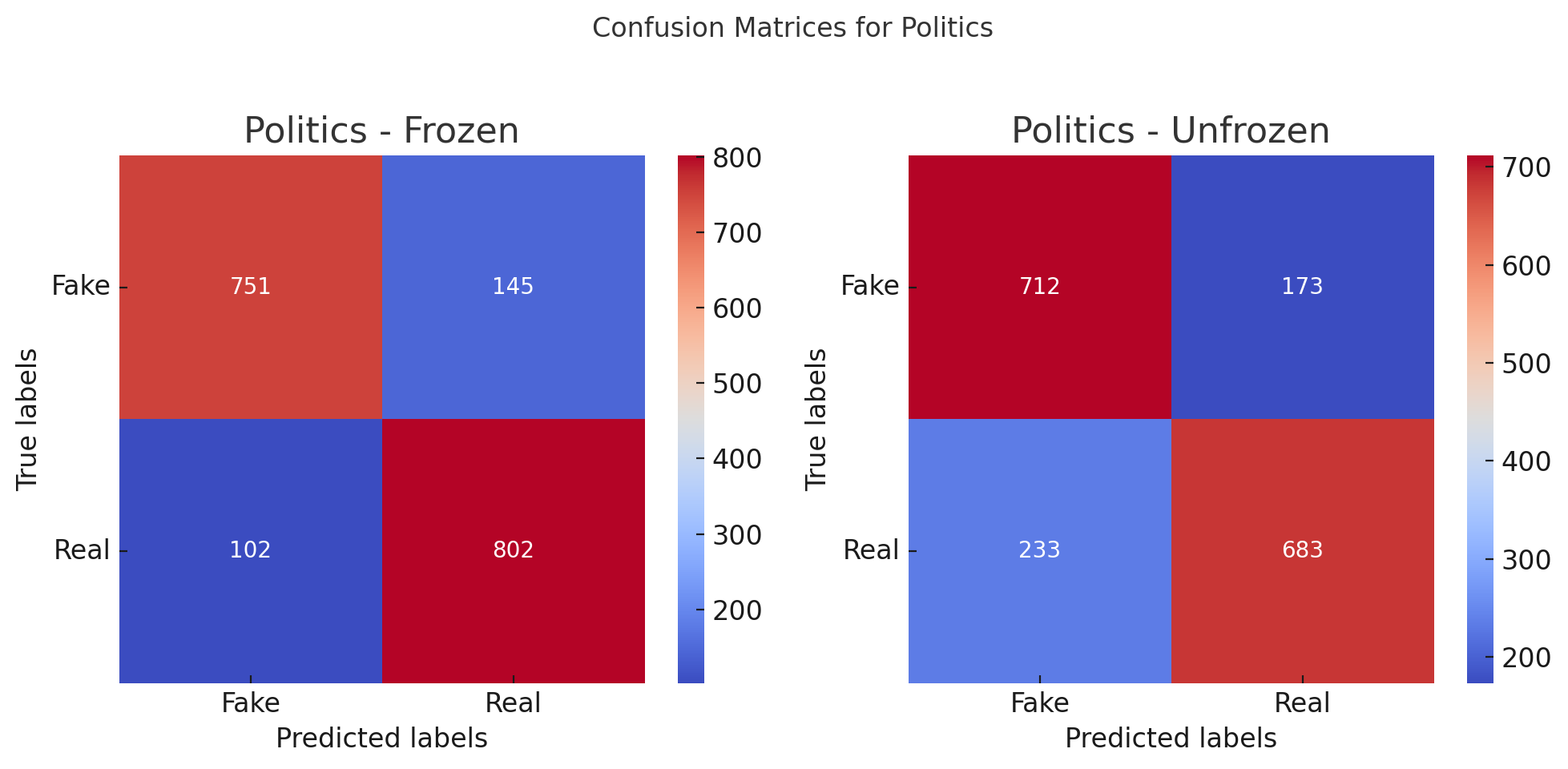
### Scenario 2:Health Domain with Frozen Layers Vs Unfrozen Layers



In the Health Domain, the frozen layers give high recall but low precision which indicates that it identifies fake news but at the expense of false positives. This could be due to overlapping features in health related articles. When switching to the unfrozen model, it maintains its high recall but with a slight improvement in precision. This could be due to the adaptability of unfrozen layers which helps reduce false positives.

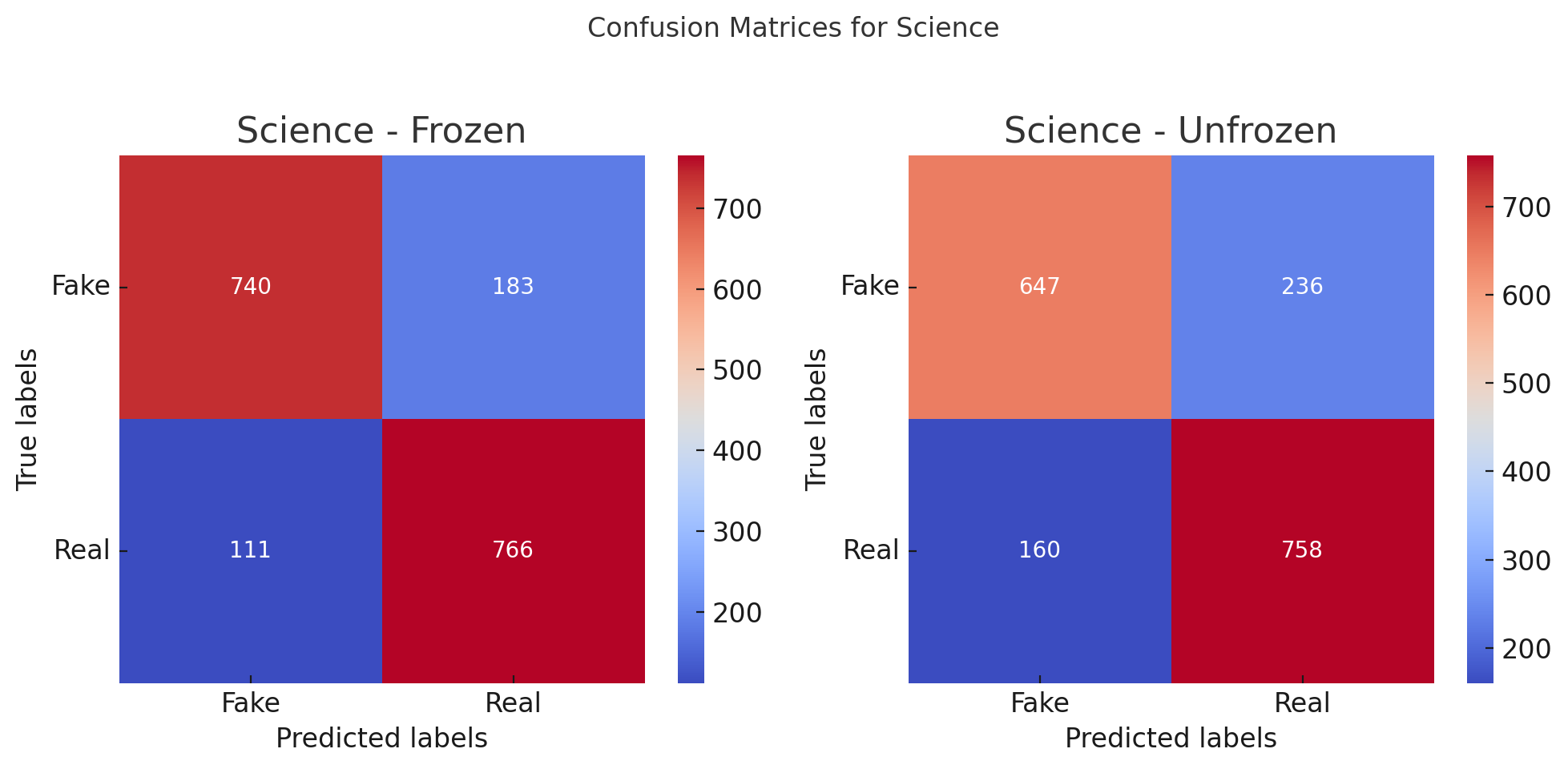
​

### Scenario 3:Politics Domain with Frozen Layers Vs Unfrozen Layers



The Politics Domain with frozen layers indicates high recall but low precision reflecting a more broad classification which is less accurate as a cost. This could be because the complexity of political language requires good flexibility which the frozen model lacks. The unfrozen model shows improved precision thereby indicating a better balance with lesser false positives. This suggests that the adaptability of unfrozen layers is valuable for the complex language of the politics domain.

### Scenario 4:Science Domain with Frozen Vs Unfrozen Layers



For the Science Domain, the frozen layers result in high precision but less recall. This outcome might stem from the fact that the structured terminology common in a scientific context allows for higher precision but reduces flexibility. The unfrozen model on the other hand achieves better recall while maintaining precision. This indicates a more balanced approach is ideal due to the unique language in science.

### Scenario 5:Social Media Domain with Frozen Vs Unfrozen Layers

### 

Lastly, in the Social Media Domain the frozen layers show lower precision and recall suggesting that the model struggles with the dynamic nature of social media language and content. The high language ambiguity, variability, and frequent use of slang in this domain necessitates higher flexibility. Furthermore, the unfrozen model gives a slight improvement in precision and recall which indicates that the flexibility of unfrozen layers helps in addressing this.

### 

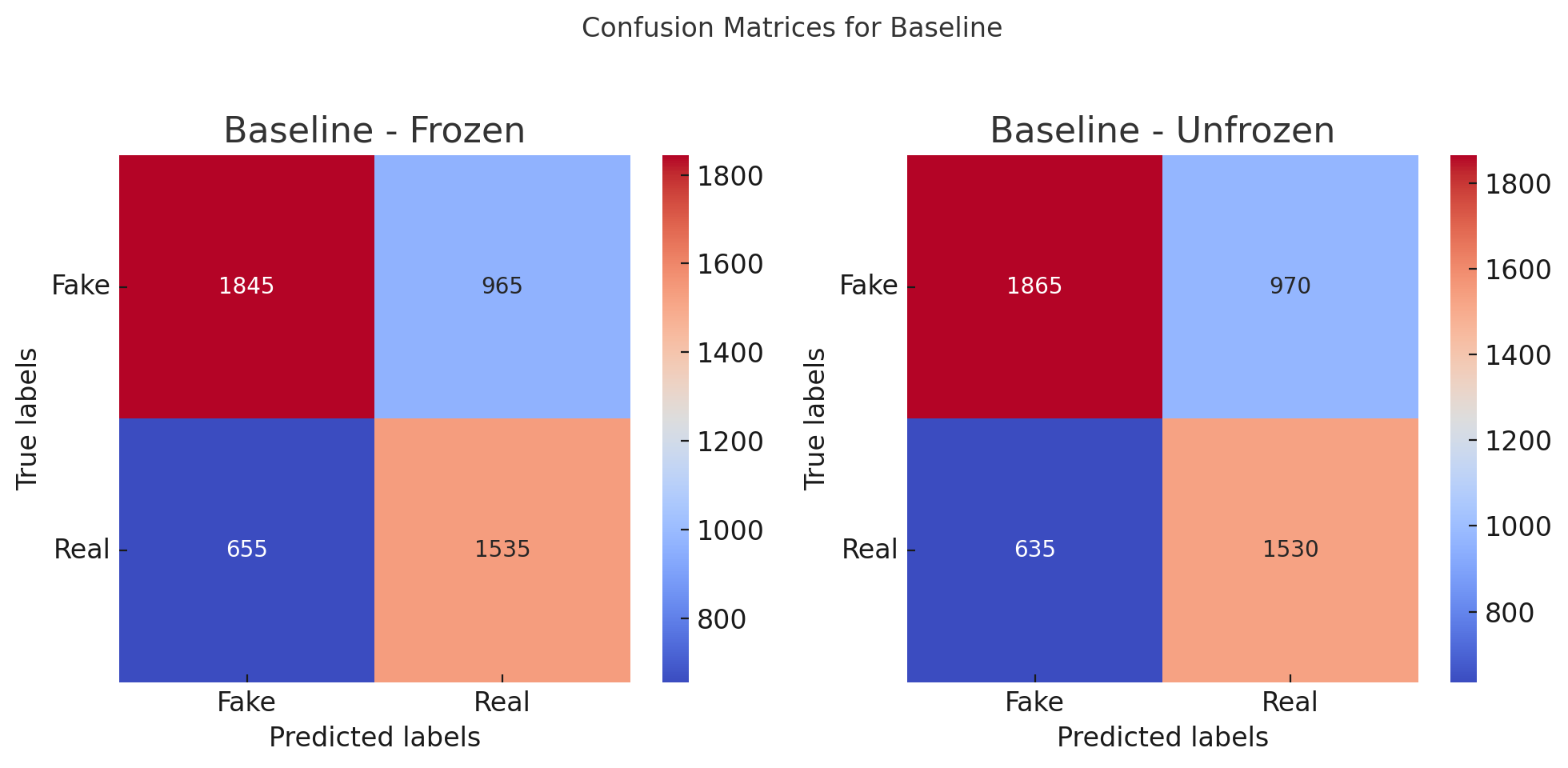
### 

### 

### 

### 

### Scenario 6:Baseline with Frozen Vs Unfrozen Layers



In the baseline model freezing the layers gave a bit more stability but at the cost of false positives. This shows there is a trade off between the model flexibility and overfitting. The frozen model showed less true positives and true negatives compared to when layers were not frozen. This suggests that freezing layers could limit the model ability to learn new and specific details of all of the domains thereby affecting its accuracy.

# Conclusion and future work

The results revealed that freezing layers and its effects on accuracy vary significantly depending on the domain.

Freezing layers led to higher precision but lower recall in the crime domain which suggests that a conservative approach could not accurately classify some fake articles. In contrast, the health domain showed minimal impact from freezing probably due to the stability in domain specific language. In politics furthermore, freezing layers reduced performance vastly. This reflects the need for leeway and adaptability due to the intricate and complex nature of political language and communication.

Using these judgments, freezing layers can be beneficial in some contexts where language patterns are stable. But it may not be ideal in dynamic and quickly changing environments where a flexible solution is more ideal as in the case for politics and social media.

In terms of future work, the following approaches could be considered:

* Hybrid Models: Hybrid approaches that combine freezing with normal learning
* Expanded Datasets: Use larger and more diverse datasets
* Different Models: Experiment other language models like Llama

# References:

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. Retrieved from https://deepai.org/publication/fake-news-detection-on-social-media-a-data-mining-perspective

Liu, Y., et al. (2019). Fake News Detection via Graph Convolutional Networks. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM). Retrieved from https://dl.acm.org/doi/10.1145/3357384.3357872

Wu, L., et al. (2019). Detecting Fake News with Transfer Learning and BERT. In IEEE International Conference on Data Mining (ICDM). Retrieved from https://ieeexplore.ieee.org/document/8970880

Chen, Y., et al. (2020). Domain Adaptation for Fake News Detection in Health-related Articles. ACM Transactions on Information Systems (TOIS). Retrieved from https://dl.acm.org/doi/10.1145/3423760

Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

Meta AI Research Lab (2023). Llama: A New Large Language Model. Meta Research. Retrieved from https://ai.meta.com/llama/