Phoenix@SMM4H Task-8: Adversities Make Ordinary Models Do Extraordinary Things

Adarsh Kumar*, **Ojasv Kamal*** and **Susmita Mazumdar***
Indian Institute of Technology Kharagpur
{adarshkumar712, kamalojasv47, susmita10}@iitkgp.ac.in

Abstract

In this paper, we describe our system entry for Shared Task 8 at SMM4H-2021 , which is on automatic classification of self-reported breast cancer posts on Twitter. In our system, we use a transformer-based language model fine-tuning approach to automatically identify tweets in the self-reports category. Furthermore, we involve a **Gradient-based Adversarial fine-tuning** to improve the overall model's robustness. Our system achieved an F1-score of **0.8625** on the development set and **0.8501** on the test set in Shared Task-8 of SMM4H-2021.

1 Introduction

With increased cases of discontinuation of Breast Cancer Treatment, which often leads to cancer recurrence, there is a need to explore complementary sources of information for patient-centered-outcomes(PCOs) associated with breast cancer treatments. Social media is a promising resource but extracting true PCOs from it first requires the accurate detection of self-reported breast cancer patients. (Al-Garadi et al., 2020) presented an NLP architecture along with a dataset for automatically categorising self-reported breast cancer posts. Their dataset was released as Shared Task-8 in SMM4H 2021.

In this paper, we describe our system to automatically distinguish self-reports of breast cancer from non-relevant tweets, which we used in the final submission for the Shared Task-8 of SMM4H-2021 (our best submission).

2 Methodology

2.1 Task and Dataset Overview

The task 8 of SMM4H consists of automatic classification of tweets into self-reports of breast cancer or non-relevant categories. The dataset comprises tweets, each associated with a label. The label indicates whether the corresponding tweet is a self-report of breast cancer or not (1 for yes, 0 for no). The training set is an unbalanced dataset of 3815 labelled tweets, around 26% of which are self-reports of breast cancer. The test set comprises 1204 unlabelled tweets, and our objective is to categorise them as self-reports or non-relevant posts.

2.2 Data Preprocessing

Before feeding into the model for training, we remove the tweet ID, username, URLs and all Non-ASCII characters associated with each tweet. Furthermore, we replaced emoticons from tweets using Ekhprasis Package (Baziotis et al., 2017), with their respective dict labels present in ekphrasis.dicts.emoticons.

2.3 Our Proposed Approach

Fig 1 illustrates our proposed approach used for final submission in the Shared Task at SMM4H. It is an amalgamation of two approaches: Domain-Specific Pre-trained model fine-tuning for binary classification and Adversarial fine-tuning for model's robustness. Below, we define each module in detail.

Domain Specific Pre-Trained Model Fine-tuning: In order to classify tweets as self-reports or not, we try to leverage the information from large pre-trained models like BERT. We further try to improve the performance by using models pre-trained on Medical Dataset to leverage domain-specific information. We performed our experiments with BERT and BlueBERT (Peng et al., 2019) models from huggingface(Wolf et al., 2019) library. While fine-tuning, we use the output from the first token of the transformer model as contextualized embedding, which is then fed into a single feed-forward classifier layer, trained using Binary Cross-entropy Loss which can be mathematically formulated as:

* Equal Contribution

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 112–114

$$L(\hat{y}, y) = -\sum_{j=1}^{c} y_j log\hat{y_j} + (1 - y_j) log(1 - \hat{y_j})$$

where, c is the total number of training examples

Gradient-based Adversarial Fine-tuning: Though fine-tuned Language Models perform well on downstream tasks like Text-classification, these models are often vulnerable to Adversarial attack (Li et al., 2020). Adversarial Fine-tuning (Goodfellow et al., 2015) has proved very efficient in improved generalization by Neural Network based models in Computer Vision Tasks.(Vernikos et al., 2020) and (Chen et al., 2021) showcase a gradient-based adversarial fine-tuning approach in the text-domain. In our system, we employ a similar technique to improve the robustness of our model. The key idea is to modify the training objective by applying small gradient-based perturbations to input text that maximize the adversarial loss. These perturbations (r1, r2, ... in Fig 1) can be easily computed using backpropagation in neural networks. The loss function we used in adversarial fine-tuning can be formulated as:

$$L = -log\, p(y|x + r_{adv})$$

where

$$r_{adv} = -\epsilon \frac{g}{||g||} \quad \text{where} \quad g = \nabla_x log(y|x;\theta)$$

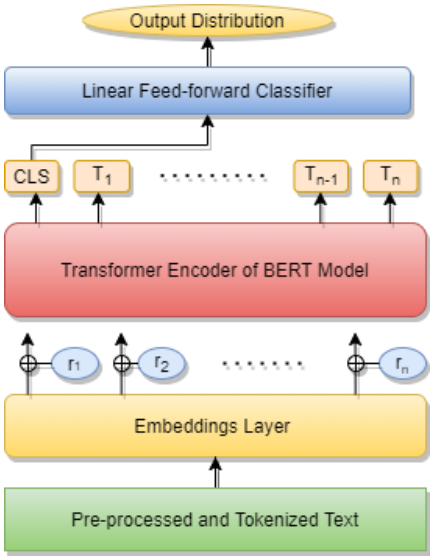

Figure 1: System Architecture

3 Result and Discussion

Table 1 shows the performance of different pre-trained models and approaches on Development Dataset, used in our experiment. As it is clear from Table 1, Blue BERT (Peng et al., 2019) fine-tuned using the Gradient-Based Adversarial Fine-tuning approach, outperforms other approaches and models on the development set. The results also suggest that adversarial fine-tuning, instead of normal fine-tuning, improves the performance of models, except for the BERT Large model. Two other important aspects to note are: improvement in the performance on using large models against the base models (which is expected given the increased number of parameters and model size) and the usefulness of Domain-specific Pre-trained model with Adversarial Fine-tuning. Table 2 shows the performance of our system entries in Shared Task - 8 on hold-on Test Dataset, which we selected after taking into consideration the above analysis.

Model	Dev F1 score
BERT base + FT	0.7826
BlueBERT base + FT	0.8025
BERT Large + FT**	0.8496
BlueBERT Large + FT	0.8205
BERT base + FT	0.8152
BlueBERT base + AFT	0.8289
BERT Large + AFT	0.8289
BlueBERT Large + AFT**	**0.86250**

Table 1: F1 score for Self Report Labelling on Development set (ϵ=1). **FT:** Fine-tuning and **AFT:** Adversarial Fine-tuning .Our final submission entries for Task 8 at SMM4H Shared Task is marked with **.

Model	F1	P	R
BERT+FT	0.8475	0.8754	**0.8214**
BlueBERT+AFT	**0.8508**	**0.8901**	0.8149

Table 2: Result on Hold-on test dataset on submission entries in SMM4H Shared Task-8. **F1:** F1 Score, **P**: Precision, **R**: Recall. Also, note both these submissions are with Large models, i.e. BERT-Large and BlueBERT Large models

4 Conclusion

In this paper, we described our approach of Adversarial fine-tuning on Domain-Specific Pre-Trained Model for classification of tweets as self-reports or not, which we used in our best submission at SMM4H-2021, Shared Task 8. Our ablation study demonstrates the usefulness of adversarial fine-tuning in improving the robustness of the model.

References

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Sahithi Lakamana, Jie Lin, Sabrina Li, Angel Xie, Whitney Hogg-Bremer, Mylin Torres, Imon Banerjee, and Abeed Sarker. 2020. Automatic breast cancer cohort detection from social media for studying factors affecting patient centered outcomes. *medRxiv*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Ben Chen, Bin Chen, Dehong Gao, Qijin Chen, Chengfu Huo, Xiaonan Meng, Weijun Ren, and Yang Zhou. 2021. Transformer-based language model fine-tuning methods for covid-19 fake news detection.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Giorgos Vernikos, Katerina Margatina, Alexandra Chronopoulou, and Ion Androutsopoulos. 2020. Domain adversarial fine-tuning as an effective regularizer.

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

A Supplemental Material

Links to the huggingface models used in the experiment:

- BERT base: `https://huggingface.co/bert-base-uncased`
- BERT Large: `https://huggingface.co/bert-large-uncased`
- BlueBERT Large: `https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-24_H-1024_A-16`
- blueBERT base: `https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-12_H-768_A-12`