# Identification of Medication Tweets Using Domain-specific Pre-trained Language Models

**Yandrapati Prakash Babu**
Department of Computer Applications
NIT Trichy, India.
prakash.babu23@gmail.com

**Rajagopal Eswari**
Department of Computer Applications
NIT Trichy, India.
eswari@nitt.edu

## Abstract

In this paper, we present our approach for task1 of SMM4H 2020. This task involves automatic classification of tweets mentioning medication or dietary supplements. For this task, we experiment with pre-trained models like Biomedical RoBERTa, Clinical BERT and Biomedical BERT. Our approach achieves F1-score of 73.56%.

## 1 Introduction

In recent times, social media platforms like twitter, facebook, reddit attracted large number of internet users. The valuable information shared by internet users which also includes health related experiences is useful in many tasks including pharmacovigilance (Kalyan and Sangeetha, 2020b). User generated texts in social media are noisy with lots of slang words and misspelled words. We participate in task1 of SMM4H2020 which aims to develop a system that can identify tweets with medication or dietary supplement mentions. Example of tweet with medication or dietary supplement mention is 'It is good to take Vitamin C every day after lunch'. An example of a tweet without medication or dietary supplement mention is 'Vitamin C is good for health' (Wu et al., 2018). The main challenge in this task is that the system should be able to identify from the context of the tweet that the mention having drug or dietary supplement name is actually referring to the drug or dietary supplementary. This task is treated as binary classification and aims at training a model which can label the given tweet with 1 if it contains medication mention and 0 if there is no medication mention. The performance of models in this task is evaluated using F1-score of class 1.

Many research works (Kalyan and Sangeetha, 2020a; Subramanyam and S, 2020) show that models trained on medical text can better understand medical terms. So, the task is experimented with pre-trained models like Biomedical RoBERTa (Gururangan et al., 2020), Clinical BERT (Alsentzer et al., 2019) and Biomedical BERT (Lee et al., 2020). Biomedical BERT and Biomedical RoBERTa are obtained by further pre-training BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models on biomedical corpus while Clinical BERT is obtained by further pre-training BERT model on MIMIC-III corpus (Johnson et al., 2016). Among these, the model based on Biomedical RoBERTa achieves the highest F1-score of 73.56%.

## 2 Dataset and preprocessing

The organizers of this task released train, validation and test sets. The train set includes 55419 tweets ( 146 positive tweets and 55273 negative tweets), validation set includes 13853 tweets (35 positive tweets and 13818 negative tweets) and test set consists of 29687. As tweets are noisy in nature, the following basic steps are used to clean the tweets:

- User mentions and urls are replaced with $< user >$ and $< url >$ respectively.
- HTML characters and unnecessary punctuation symbols are removed.

*Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 128–130
Barcelona, Spain (Online), December 12, 2020.

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| Biomedical RoBERTa | 65.98 | 83.12 | **73.56** |
| Biomedical BERT | - | - | 71.00 |
| Clinical BERT | - | - | 67.00 |
| Average score of all Task-1 teams | 70.32 | 69.48 | 66.28 |

Table 1: Precision, Recall and F1-score of our models on test data.

- Emojis are replaced with their corresponding descriptions.
- Twitter slang words are replaced with corresponding standard words. For example, '*lol*' is replaced with '*laugh out loud*'.

## 3 Model Description

In recent times, researchers have focused on exploiting deep pre-trained language models like BERT, RoBERTa in most of the natural language processing tasks. These models are adapted to medical domain by means of additional training on large medical text. This paper investigates how well domain specific models like Biomedical RoBERTa, Biomedical BERT and Clinical BERT identify medication tweets. First, tweet representation $\boldsymbol{e_t} \in \mathbb{R}^n$ is generated using domain specific models, $n$ represents hidden state vector in pre-trained language model which is equal to 768. Then, sigmoid layer with parameters $W \in \mathbb{R}^{n \times 1}$ and $b \in \mathbb{R}$ is applied on $\boldsymbol{e_t}$ to transform it into single value which represents the predicted label $q$.

$$\boldsymbol{e_t} = PTLM(tweet) \tag{1}$$

$$q = Sigmoid(W^T e_t + b) \tag{2}$$

Here PTLM refers to pretrained language model and it can be Biomedical RoBERTa, Biomedical BERT or Clinical BERT.

## 4 Experiments and Results

To handle imbalance in the dataset, the dataset is augmented with 9622 tweets (4975 positive tweets and 4647 negative tweets) from SMM4H 2018 task1 dataset (Weissenbacher et al., 2018). Further, positive tweets are up sampled 15 times and 90% of the negative tweets are randomly chosen. We use validation set to find optimal values for various hyperparameters. We use batch size of 16, learning rate of 3e-5 and train the model for 3 epochs. All our models are implemented using transformers library in PyTorch (Wolf et al., 2019). The task organizers released precision and recall scores only for which model got the highest F1-score, the performance of our models and average score is listed in Table 1. Among the three models, the model based on Biomedical RoBERTa outperforms other models and achieves the highest F1-score of 73.56%. As a whole, our approach achieves good results which is much higher than the average scores.

## 5 Conclusion

The medication mentions in tweets are identified using domain specific deep pre-trained models. Experimental results show that the model based on Biomedical RoBERTa achieves the best F1-score of 73.56% which is significantly higher than the average F1-score of 66.28%.

## References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Katikapalli Subramanyam Kalyan and S Sangeetha. 2020a. Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network. Technical report, EasyChair.

Katikapalli Subramanyam Kalyan and S Sangeetha. 2020b. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101:103323.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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*.

Kalyan Katikapalli Subramanyam and Sangeetha S. 2020. Deep contextualized medical concept normalization in social media text. *Procedia Computer Science*, 171:1353 – 1362. Third International Conference on Computing and Network Communications (CoCoNet'19).

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Sixing Wu, Yongfeng Huang, and Xing Xie. 2018. Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 34–37.