Social Media Mining for Health Monitoring

Anuja Raghunath Katkar*, Sagar Jitendra Thacker*
Department of Computer Science and Engineering
University at Buffalo, Amherst, NY 14260
{anujarag, sagarjit}@buffalo.edu

Abstract

This paper describes neural network based models designed for Social Media Mining for Health Monitoring 2020 and 2021 shared task. We worked on Task 2 and Task 3 from shared task 2020. Also, performed Task 5 and Task 6 from shared task 2021. We found that for three tasks i.e. task 2 from shared task 2020 and from shared task 2021 task 5, task 6, RoBERTa model gave the best results with F1 score of 63%, 98% and 75% respectively. Task 3 from shared task 2020 we got the best results for NER using BioBERT of 59% F1-score and for NER+Normalization with got 55% F1-score.

1 Introduction

This paper aims to tackle various problems in the health care domain. The rise of COVID-19 brought a new norm in ways in which humans communicate. Platforms such as Twitter have proven to be effective and generate loads of user generated data. As people have more freely shared their life experiences on the platform, it gives opportunity are the pharmaceutical firms to leverage insights from the data using natural language processing.

The Social Media Mining for Health Applications (SMM4H) shared tasks aims to tackle various healthcare application in social media texts. We worked upon total of 4 tasks, Task 2: Automatic classification of multilingual tweets that report adverse effects (only English language) and Task 3: Automatic extraction and normalization of adverse effects in English tweets extraction and normalization of adverse effects in English tweets from SMM4H 2020 maps to Task 1 & 2 in our problem statement. Also, Task 6: Perform three-way classification of COVID tweets containing symptoms and Task 5: Classification of tweets self-reporting potential COVID19 cases from SMM4H 2021 maps to Task 3 & 4 in our problem statement.

We worked upon different set of neural network based transformer models. As transformer models are based attention mechanism, empirically it has been observed that they perform better than recurrent and convolutional models. We experimented with different BERT (Devlin et al., 2018) based models and found fine-tuned RoBERTa performed the best of Task 1, 3, & 4. Task 2 was divided into two sub-tasks: We modelled the first sub-task as a Named Entity Recognition (NER) task to extract the span text and second sub-task as mapping the extracted text to standard MedDRA code.

The paper is organized as follows: In section 2 we describe the problem statement for each task. Section 3 describes the related work done to tackle the respective problem statement. Section 4 describes the Method & Model Architecture for each task which is followed by results in section 5. Section 6 provides details on the error analysis performed. Finally we conclude in section 7.

2 Task and Data Description

2.1 Task 1: Automatic classification of tweets that report adverse effects

This task involves distinguishing tweets that report adverse effect (AE) to a medication from those that do not. Table 1 shows the data distribution for the train and validation set, along with positive and negative split of data for both training and validation set.

*Equal contribution

2.2 Task 2: Automatic extraction and normalization of adverse effects in English tweets

This task is an end-to-end task involving extraction of span text that contain an AE of medication from the tweets that report an AE. After extraction we map the extracted AE to a standard concept ID in the MedDRA vocabulary. Table 1 shows the data distribution for the train and validation set, along with positive and negative split of data for both training and validation set.

2.3 Task 3: Classification of COVID19 tweets containing symptoms

This task is a multi-class classification problem. The target classes are self-reports, non-personal reports, and literature/news mentions. Self-reports are personal mentions of COVID19 symptoms, non-personal reports are mentions of symptoms experience by other person, and literature/news mentions are tweets containing symptoms mentioned in some news or scientific articles. Table 2 shows the data distribution for the train and validation set, along with per class distribution of data for both training and validation set.

Task	Training Set			Validation Set		
	% Positive	% Negative	Total examples	% Positive	% Negative	Total examples
Task 1	9.26%	90.74%	20544	9.23%	90.74%	5134
Task 2	34.81%	65.19%	2246	34.82%	65.18%	560
Task 4	15.87%	84.13%	6465	17.04%	82.96%	716

Table 1: Distribution of data for task 1, 2, and 4

Task	Data Distribution			
	% Lit-News	% Nonpersonal	% Self	Total examples
Task 3 (Train)	47.17%	37.96%	14.87%	9067
Task 3 (Validation)	49.4%	36.0%	14.6%	500

Table 2: Distribution of data for Task 3

2.4 Task 4: Classification of tweets self-reporting potential COVID19 cases

This task involves distinguishing tweets that self-report potential cases of COVID19 from those that do not. “Potential case” tweets discuss topics such as testing, symptoms, traveling, or social distancing that indicates higher risk of exposure to COVID-19. “Other” tweets are related to COVID-19 and may discuss on the same topics but do not indicate that the user or a member of the user’s household may be infected. Table 1 shows the data distribution for the train and validation set, along with positive and negative split of data for both training and validation set.

3 Related Work

Over multiple tasks there were many techniques and approaches that are common that includes: Standard text pre-processing steps applied to clean the tweets. Main difference were either to remove the token or replace it with some arbitrary tag name. Baseline models such as Logistic Regression, and SVM were used, although their performance was not at par with state-of-the-models (SOTA) such as BERT and its variations. SOTA models like BERT, RoBERTa, EnDRBERT, BioBERT, SciBERT, Bi-directional LSTM, and many more were used. Various word representation techniques such as Tf-IDF, Fast text embedding, GloVe embeddings and unigram, bigram tokens were used.

3.1 Task 1: Automatic classification of tweets that report adverse effects

Data augmentation was performed by using 2018 and 2019 competition data (Kalyan and Sangeetha, 2020). Ensemble BERT models with voting schema was used to classify the labels (Miftahutdinov et al., 2020). The best result on the test set was using RoBERTA using replacement strategy for pre-processing and negex algorithm to identity negated findings and diseases (Wang et al., 202). The best F1-score achieved was 0.64 on the test data set where as the average score was 0.46.

3.2 Task 2: Automatic extraction and normalization of adverse effects in English tweets

CNN & Bi-LSTM CRF + MedDRA and it's variations with different hyperparameters and modified version of MedDRA was developed (Vydiswaran et al., 2020). RoBERTa + BIO tagging and multi-task learning based RoBERTA used for both tasks (Kalyan and Sangeetha, 2020). The best score was achieved using EnDRBERT with BIO tagging, gazetteer features and additional data from CSIRO adverse drug event corpus (CADEC) was used. Normalization task used classification over concepts, meta learning, and combined (Miftahutdinov et al., 2020). The best F1-score for NER was 0.755 with average score as 0.564; NER+Normalization was 0.463 with average score as 0.292.

3.3 Task 3: Classification of COVID19 tweets containing symptoms

Some of the paper used traditional ML approaches and compared the same SOTA Deep Learning methods which gave almost same F1 score as median score. Some of the preprocessing tool like preprocess-twitter and for oversampling python tool imbalance-term was used. With SOTA model BERT autoregressive model like XLNet was also implemented. Up till now the best results were found in CT-BERT model which gave 0.95 F1 score where median score was 0.93 (Valdes et al., 2021).

3.4 Task 4: Classification of tweets self-reporting potential COVID19 cases

BERT was modified in different ways such as Enhancing of BERT model with augmentation, data cleaning etc and Domain specific BERT model such as CT-BERT(COVID-Twitter-BERT) was introducted. Pretrained models like RoBERTa, CT-BERT (Müller et al., 2020), and Twitter-RoBERTa were ensemble to yield better performance (Luo et al., 2021). Up till now the best results were found in enhanced BERT model which gave 0.79 F1 score where median score was 0.74 (Aji et al., 2021)

4 Method & Model Architecture

This section describes the techniques and algorithms used to build the baseline architecture.

4.1 Pre-processing

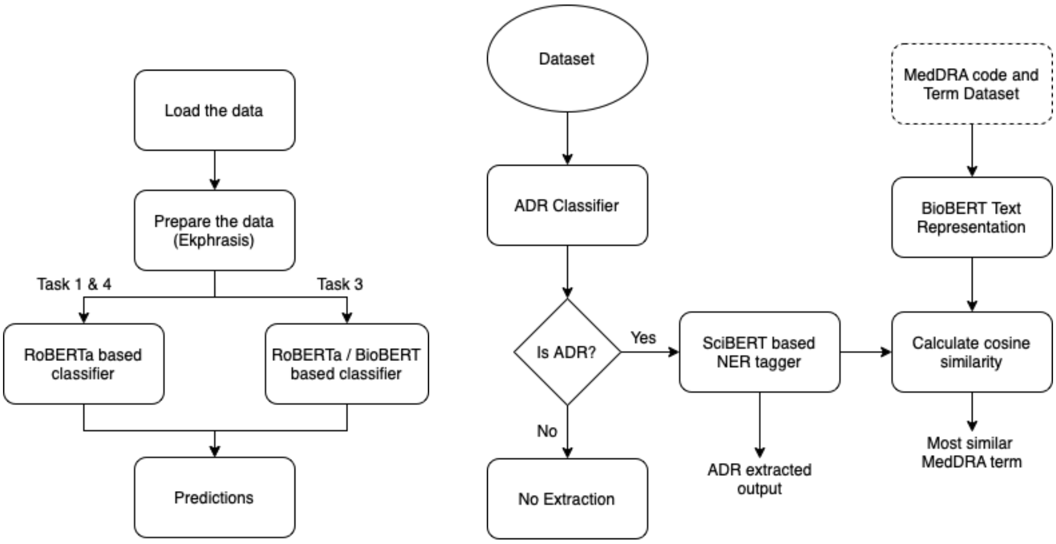
Twitter data contains a lot of noise, to clean the data we applied Ekphrasis (Baziotis et al., 2017) text-preprocessing techniques that is geared towards text from social media. Using this technique includes lower case the tweets, normalize & annotate URL's, mentions, hashtags, redundant characters, emojis, emoticons, punctuation's, repeated characters, extra white-spaces, and expand contractions. This technique is common across task 1, 3, and 4.

4.2 Task 1: Automatic classification of tweets that report adverse effects

We experimented with different transformer models like BERT base uncased, RoBERTa base, SciBERT with scivocab, and BioBERT base v1.1. We fine-tuned each model on the pre-processed data and performed classification through a linear layer. The model were trained for 5 epochs with batch size 32 and learning rate as 2e-5. We found the best result for RoBERTa base on the validation set. Table 4 shows the performance of the model on validation set.

4.3 Task 2: Automatic extraction and normalization of adverse effects in English tweets

The given task can be divided into two subtasks:
1. Extraction of Adverse Effects from tweets
2. Normalization of extracted Adverse Effect to a standard concept ID



(a) Task 1, 3 & 4

(b) Task 2

Figure 1: Model Architecture

1. **Extraction**

- **Classification of tweets containing ADR mentions**
  In the first stage of the pipeline, before extracting any span of text we pre-processed the tweets with the same preprocessing techniques mentioned in section 4.1. These tweets would be passed to train RoBERTa to classify whether tweets contain ADR mentions or not.

- **NER tagger**
  We posed the problem as a Custom Named Entity Recognition problem using the training dataset. We found that the start and end index of the extracted text did not align with the indexes in the tweets. To rectify these we found the correct start and end indexes and used those indexes for building our model. We fine tuned SciBERT and BioBERT for extracting the ADR mention extract. Before training the model, we tagged each tweet using the 'BIO' tagging scheme, where 'B' stands for begin of the token, 'I' stands for inside the token, and 'O' stands for outside the token. We only extracted text from tweets that were classified containing ADR mentions from the previous step. We trained our model for 5 epochs with a learning rate of 3e-5 and found the best results using BioBERT as our model.

2. **Normalization**

Considering that there are more MedDRA terms present beyond the ones from the train set, to incorporate the wide range of terms we added the whole list MedDRA code and term mapping. We found the word embedding (average of the last two hidden layers) for each MedDRA term using SciBERT and created aa dictionary mapping for each code and its representation. To normalize we find the cosine similar of the word embedding of the extracted text with all the MedDRA terms and tag the term that is the most similar.

4.4 Task 3: Classification of COVID19 tweets containing symptoms

We experimented with different transformer models like BERT base uncased, RoBERTa base, SciBERT with scivocab, and BioBERT base v1.1. Also, tried domain specific BERT model such as CT-BERT. We fine-tuned each model on the pre-processed data and performed classification through a linear layer. The model were trained for 5 epochs with batch size 32 and learning rate as 2e-5. We found the best result for RoBERTa base and BioBERT on the validation set. Table 4 shows the performance of the model on validation set.

4.5 Task 4: Classification of tweets self-reporting potential COVID19 cases

We experimented with different transformer models like BERT base uncased, RoBERTa base, SciBERT with scivocab, and BioBERT base v1.1. Also, tried domain specific BERT model such as CT-BERT. We fine-tuned each model on the pre-processed data and performed classification through a linear layer. The model were trained for 5 epochs with batch size 32 and learning rate as 2e-5. We found the best result for RoBERTa base on the validation set. Table 4 shows the performance of the model on validation set.

5 Results

Below we summarize the performance different models on all tasks. The best score among the different models is highlighted in bold for each task. Table 3 and 4 shows the performance of the models on validation set. For Task 1, the best F1-score we achieved was 0.64 using RoBERTa model which was an 16% improvement over the baseline model. Task 2, the best F1-score for NER was achieved using BioBERT which was an 11.5% and 16.9% improvement over the baseline model for strict and relaxed measure. For NER+Normalization we got the best F1-score as 59% and 55% which was an 49% and 40.3% improvement over the baseline for strict and relaxed scores respectively. Task 3, RoBERTa and BioBERT both gave us the best F1-micro score which was 0.984 which was an 2% improvement over the baseline model. Finally, task 4, RoBERTa model gave us the best F1-score with 0.75 which was an 23% improvement over the baseline model

Task	**Strict**			**Relaxed**		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Baseline						
Task 2 (NER)	0.175	0.24	0.137	0.221	0.291	0.178
Task 2 (NER + Normalization)	0.100	0.146	0.076	0.147	0.204	0.114
Final Model						
Task 2 (NER)	0.29	0.28	0.30	0.29	0.27	0.32
Task 2 (NER + Normalization)	0.59	1.00	0.42	0.55	1.00	0.38

Table 3: Results for Task 2 on validation set

Model	**Task 1**	**Task 3**	**Task 4**
	F1-score	**Micro F1-score**	**F1-score**
Baseline	0.47	0.964	0.52
Transformer models			
BERT	0.58	0.98	0.74
RoBERTa	**0.63**	**0.984**	**0.75**
SciBERT	0.55	0.982	0.66
BioBERT	0.56	**0.984**	0.57
CT-BERT	-	0.964	0.74

Table 4: Results for Task 1, 3, & 4 on validation set

6 Discussion and Error Analysis

6.1 Task 1: Automatic classification of tweets that report adverse effects

Using SOTA transformer models results in better performance of the model over baseline models. Upon analysis of the tweets that were misclassified we find various pattern that emerge. Example, 'did you know twitter is as addicting as nicotine ?' and 'i swear food has nicotine in it, cause im addicted to that shit.', the word 'nicotine' has been used in various context such as nicotine patches, nicotine has an addition to food, drugs, etc,.

Tweets ranging from sarcastic comments to casual mentions of ADR or drug such as 'geez this vyvanse makes me talk a mile a minute haha' and 'vyvanse make me so hyper and creative and i think of so many tweets' cause models to perform poorly. 'vyvanse' is a medication used to treat ADHD and binge-eating disorder. It has been used in context ranging from not able to sleep to its benefits. Also, effects related to depression, anxiety disorders, obsessive-compulsive disorder (OCD), bipolar disorder and drug such as 'paxil' used in context such as 'screw you paxil! you do wonders for anxiety but you make me a ultra lightweight' proves to be difficult for the model to evaluate. Upon further analysis, tweets that contains common mentions of medications and symptoms in both the classes; or various words used in both the classes in different context proved to be difficult. Unable to capture the contextual meaning/representation cause the model to perform poorly.

Attempts: We tried training RoBERTa-large on our data but failed to do so. The reason behind that was we could not use a big enough batch size as we would get Cuda-Out-of-Memory error in Google Colab. We tried batch-size ranging from 32, 16, 8, 4, and finally tried training using batch-size as 2. With batch size so small each epoch took longer time which lead to random disconnection from Google Colab and at times ran out of GPU allocation limit.

6.2 Task 2: Automatic extraction and normalization of adverse effects in English tweets

Task 2 proved to be the most difficult task so far. Implementing basic custom NER model performed poorly on tweets contains multiple extractions. Single drug can also have multiple adverse effects leading to difficulty for the model to learn. Also, we found that based on our preprocessing and BERT sub-work tokenization methodology the model could map the tag for the first term in the subword but subsequent subwords were not marked with the same tag. Hence, the model found the correct span but only gave us the first token. Also, we found inconsistencies in the dataset where extracted phrase is 'withdrawal' but it is mapped to 10023222 which is code for 'joint pain'. Extractions that contain numeric figures were points where the model struggled.

For Normalization, we found MedDRA codes alone are not a good indicator for normalization. Rather than the correct meddra term, model would tag it to terms with similar meaning. Example: MedDRA code is 'bizarre dreams' was tagged to 'bad dreams'.

Attempts: We tried ensembling SciBERT and BioBERT but we struggled on how to do so. The pain points we faced was both model gave different tokenization output and it was difficult for us to map the labels as they won't align if we wanted to perform voting mechanism. We then tried detaching the head of the model and concatenate the outputs but we weren't successful. We also tried using the CADEC dataset to improve the model training process where we found that CADEC dataset contain multiple span of texts and reuses overlapping indexes separating them by ';' symbol. We tried curating the dataset and tag it to BIO tagging scheme. After training the model perform poorly and we suspect errors in curating the dataset we made some mistakes.

6.3 Task 3: Classification of COVID19 tweets containing symptoms

Using SOTA transformer models results in better performance of the model over baseline models. Although the model performed well it struggled to distinguish between Lit/News vs Non-personal and Non-personal vs Self-reports in certain cases.

Example, 'I had crippling body aches, fatigue and couldn't concentrate' was mention as part of a Lit/News tweets with mention of The Guardian (Newspaper) but predicted as Non-personal. Phrases such as 'Loss of taste and smell' are part of both Lit/News and Non-personal in many instances causes model to interchange class labels between them. Also, Lit/News and Self-reports tend to mention various coronavirus symptoms in their tweet which makes it different to distinguish between them. Also, it is difficult to identify tweets that talk in third person or self view with long distance dependencies.

Result by (Mondal et al., 2021) also show performance of Machine Learning models vs Deep Leaning models such as Bi-directional LSTM's where they produce similar results. In our case, we also saw marginal increase in F1-score compared to the baseline. It can be argued in many cases to use baseline because of it's simplicity and faster performance and overlook the marginal increase in results.

6.4 Task 4: Classification of tweets self-reporting potential COVID19 cases

Using SOTA transformer models results in better performance of the model over baseline models. Although the performance gains was significant, the model still struggles due to class imbalance as shown in Table 1. Also, we identified the major reasons behind the poor performance as word overlap between classes. Words such as 'flight', 'quarantine', 'coronavirus', 'hospital', 'sick', symptoms like 'fever', 'cough', 'cold', and many more are used heavily in both the classes. This causes it difficult to distinguish between the two classes.

Examples, 'Day 3 into my #covid_19 quarantine and I made an experimental short film about this creepy chandelier that I found at an estate sale...', 'me knowing I just dusted and am highly allergic to dust and get coughing fits when i clean: me immediately thinking it's coronavirus'. Sarcastic or mocking tweets such as 'Jeremiah asking what's wrong I'm like i have a headache .. he's like "you got the coronavirus "? Ayooooo.' are also difficult to predict. One could say that words such as 'coronavirus', 'quarantine', 'cough', etc. act like stopwords in our problem statement as they are used heavily in both classes and doesn't help much in identifying distinguishing factors.

7 Conclusion

In this project, we experimented with different transformer models for each task and found that they perform better than the models used in our baseline. We also found that pre-trained models such as BioBERT, CT-BERT and SciBERT for particular domain perform better than other models.

For future improvements in task 1, incorporating different drug and it's effects; along with some indicator for the tone of the tweets might help improve the performance of the models. For task 2, correcting the tagging scheme, mapping correct MedDRA term with MedDRA code, adding more data for training can improve NER performance. NER+Normalization can be improved be incorporating some relation between the drug and the extracted text. Also, MedDRA code follows a hierarchical structure, finding where our tweet or drug and extracted text map to in the structure can give us some better insights. For task 4, improving data pre-processing techniques could help as we identified various words used heavily in both classes could be considered as stop words in this case.

8 Work Distribution

Table 5 describes the work distribution within our team.

Team Member Name	Contribution	Work Done
Anuja Raghunath Katkar	50%	Task 2, 3, 4
Sagar Jitendra Thacker	50%	Task 1, 2, 3

Table 5: Work Distribution

References

Anupam Mondal, Sainik Kumar Mahata, Monalisa Dey, Dipankar Das 2021. Classification of COVID19 tweets using Machine Learning Approaches. In *Proceedings of the Sixth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 135-137

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, August. Association for Computational Linguistics.

Martin Müller, Marcel Salathé, Per E Kummervold 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. arXiv preprint arXiv:2005.07503

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha 2020. Want to Identify, Extract and Normalize Adverse Drug Reactions in Tweets? Use RoBERTa In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 121–124, Barcelona, Spain (Online). Association for Computational Linguistics.

Chen-Kai Wang, Hong-Jie Dai, You-Chen Zhang, Bo-Chun Xu, Bo-Hong Wang, You-Ning Xu, Po-Hao Chen, and Chung-Hong Lee 2020. ISLab System for SMM4H Shared Task 2020 In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 42–45, Barcelona, Spain (Online). Association for Computational Linguistics.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina 2020. KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56, Barcelona, Spain (Online). Association for Computational Linguistics.

Alberto Valdes, Jesus Lopez, and Manuel Montes 2021. UACH-INAOE at SMM4H: a BERT based approach for classification of COVID-19 Twitter posts In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 65–68, Mexico City, Mexico. Association for Computational Linguistics.

V.G.Vinod Vydiswaran, Deahan Yu, Xinyan Zhao, Ermioni Carr, Jonathan Martindale, Jingcheng Xiao, Noha Ghannam, Matteo Althoen, Alexis Castellanos, Neel Patel, and Daniel Vasquez 2020. Identifying Medication Abuse and Adverse Effects from Tweets: University of Michigan at #SMM4H 2020 In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 90–94, Barcelona, Spain (Online). Association for Computational Linguistics.

Alham Fikri Aji, Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Radityo Eko Prasojo, and Tirana Fatyanosa 2021. BERT Goes Brrr: A Venture Towards the Lesser Error in Classifying Medical Self-Reporters on Twitter In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 58–64, Mexico City, Mexico. Association for Computational Linguistics.

Ying Luo, Lis Pereira, and Kobayashi Ichiro 2021. OCHADAI at SMM4H-2021 Task 5: Classifying self-reporting tweets on potential cases of COVID-19 by ensembling pre-trained language models In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 123–125, Mexico City, Mexico. Association for Computational Linguistics.