BERT based Transformers lead the way in Extraction of Health Information from Social Media

Sidharth Ramesh[1] **Abhiraj Tiwari**[1] **Parthivi Choubey**[1] **Saisha Kashyap**[2]

Sahil Khose[2] **Kumud Lakara**[1] **Nishesh Singh**[3] **Ujjwal Verma**[4]

{sidram2000, abhirajtiwari, parthivichoubey, saishakashyap8,
sahilkhose18, lakara.kumud, singhnisheshr4}@gmail.com
ujjwal.verma@manipal.edu

Abstract

This paper describes our submissions for the Social Media Mining for Health (SMM4H) 2021 shared tasks. We participated in 2 tasks: (1) Classification, extraction and normalization of adverse drug effect (ADE) mentions in English tweets (Task-1) and (2) Classification of COVID-19 tweets containing symptoms (Task-6). Our approach for the first task uses the language representation model RoBERTa with a binary classification head. For the second task, we use BERTweet, based on RoBERTa. Fine-tuning is performed on the pre-trained models for both tasks. The models are placed on top of a custom domain-specific pre-processing pipeline. Our system ranked first among all the submissions for subtask-1(a) with an F1-score of 61%. For subtask-1(b), our system obtained an F1-score of 50% with improvements up to +8% F1 over the score averaged across all submissions. The BERTweet model achieved an F1 score of 94% on SMM4H 2021 Task-6.

1 Introduction

Social media platforms are a feature of everyday life for a large proportion of the population with an estimated 4.2 billion people using some form of social media (Hootsuite and Social, 2021). Twitter is one of the largest social media platforms with 192 million daily active users (Conger, 2021). The 6th Social Media Mining for Health Applications Workshop focuses on the use of Natural Language Processing (NLP) for a wide number of tasks related to Health Informatics using data extracted from Twitter .

Our team, TensorFlu, participated in 2 tasks, (1) Task-1: Classification, extraction and normalization of adverse effect (AE) mentions in English tweets and (2) Task-6: Classification of COVID-19 tweets containing symptoms. A detailed overview of the shared tasks in the 6th edition of the workshop can be found in (Magge et al., 2021).

The classification and extraction of Adverse Drug Effects (ADE) on social media can be a useful indicator to judge the efficacy of medications and drugs while ensuring that any side effects that previously remained unknown can be found. Thus social media can be a useful medium to judge gauge patient satisfaction and well being.

According to the report in (Shearer and Mitchell, 2021), 15% of American adults get their news on Twitter while 59% of Twitter users get their news on Twitter itself. Thus during the spread of a pandemic like COVID-19, tracking reports by users as well as news mentions from local organizations can perform the function of tracking the spread of the disease in new regions and keep people informed.

Similar to the last edition of the workshop, the top performing model (Klein et al., 2020) for Task-1 with the highest score this year was RoBERTa (Liu et al., 2019). The biggest challenge while dealing with the dataset provided for this years competition was the huge class imbalance. The proposed approach handles this by the use of Random Sampling (Abd Elrahman and Abraham, 2013) of the dataset during finetuning. Named Entity Recognition (NER) for the extraction of text spans was performed using the RoBERTa based model provided in the spaCy (Honnibal et al., 2020) `en_core_web_trf` pipeline. For the classification of tweets with COVID-19 symptoms, we used a model called BERTweet (Nguyen et al., 2020) trained on 845 million English tweets and 23 million COVID-19 related English tweets as of the latest publically available version of the model. Fine-tuning was performed on the pretrained models for

[1] Dept. of Computer Science and Engineering
[2] Dept. of Information and Communication Technology
[3] Dept. of Mechanical and Manufacturing Engineering
[4] Dept. of Electronics and Communication Engineering
Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 33–38
June 10, 2021. ©2021 Association for Computational Linguistics

both tasks. Section 2 summarizes the methodology and results obtained for Task-1, while Section 3 summarizes the methodology and results for Task-6.

2 Task-1: Classification, extraction and normalization of adverse effect (AE) mentions in English tweets

2.1 Sub-Task 1a: Classification

The goal of this sub-task is to classify tweets that contain an adverse effect (AE) or also known as adverse drug effect (ADE) with the label ADE or NoADE.

2.1.1 Data and Pre-processing

The organizers of SMM4H provided us with a training set consisting of 18,256 tweets with 1,297 positive examples and 16,959 negative examples. Thus, the dataset has a huge class imbalance. The validation dataset has 913 tweets with 65 positive examples and 848 negative examples.

To overcome the class imbalance we performed random oversampling and undersampling (Abd Elrahman and Abraham, 2013) on the provided dataset. The dataset was first oversampled using a sampling strategy of 0.1 i.e. the minority class was oversampled so that it was 0.1 times the size of majority class, then the resultant dataset was undersampled using a sampling strategy of 0.5 i.e. the majority class was undersampled so that the majority class was 2 times the size of minority class

Removal of twitter mentions, hashtags and URLs was performed, but it negatively affected the performance of the model. Hence, this pre-processing step was not performed in the final model. The tweets were then preprocessed using fairseq (Ott et al., 2019) preprocessor which tokenizes the sentences using GPT-2 byte pair encoding(Radford et al., 2019) and finally converts them into binary samples.

2.1.2 System Description

Fairseq's (Ott et al., 2019) pretrained RoBERTa (Liu et al., 2019) large model was used for the task with a binary classification head. The RoBERTa model was pretrained over 160GB of data from BookCorpus (Zhu et al., 2015), CC-News (Nagel, 2016), OpenWebText (Gokaslan* et al., 2019) and Stories.

2.1.3 Experiments

RoBERTa and BioBERT (Lee et al., 2019) were trained for ADE classification and extensive hyperparameter tuning was carried out. The hyperparameters tested on the validation split included the learning rate, batch size, and sampling strategy of the dataset. The RoBERTa model was trained for 6 epochs with a batch size of 8. The learning rate was warmed up for 217 steps with a weight decay of 0.1 and a peak learning rate of 10^{-5} for the polynomial learning rate scheduler. A dropout rate of 0.1 is used along with the Adam optimizer having (β_1, β_2)=(0.9, 0.98).

2.1.4 Results

Precision is defined as the ratio between true positives and the sum of true positives and false positives.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall is defined as the ratio between true positives and the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Our primary objective is to create a model that prevents incorrect classification of ADE tweets. A model with higher recall than precision is more desirable for us as the former tends to reduce the total number of false negatives. F1 Score is chosen to be the evaluation metric for all our models.

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Table 1 showcases the performance of the different models which performed well on the validation set. The RoBERTa model that was finally chosen after hyperparameter tuning achieved the highest score on the leaderboard among all teams participating in the subtask. The score obtained on the test set can be found in Table 2.

It can be seen in the results of the validation set and test for the ADE class that the recall is 0.92 for the validation set and 0.752 for the test set. The results show that the model has learnt features for classifying ADE samples from a small amount of data. Although it might classify some amount of NoADE tweets incorrectly as evidenced by the low precision, the greater number of correctly classified ADE tweets aligns with our objective of classifying the maximum number of ADE tweets correctly as

S.No.	Model	Arch	Label	Precision	Recall	F1
1.	**RoBERTa**	$BERT_{LARGE}$	NoADE	**0.99**	0.95	0.97
			ADE	0.59	**0.92**	**0.72**
2.	BioBERT	$BERT_{BASE}$	NoADE	0.97	**0.99**	**0.98**
			ADE	**0.78**	0.60	0.68

Table 1: Comparing different models used for task 1a on the **Validation Set**. **RoBERTa** is chosen owing to its higher F1- score while predicting the ADE label correctly.

	Precision	Recall	F1
RoBERTa	**0.515**	**0.752**	**0.61**
Median	0.505	0.409	0.44

Table 2: Comparing our best-performing model to the median for task 1a.

possible so that we don't lose valuable information about adverse drug effects that might be found. Our model achieved a significantly higher recall than the median of all other teams (Table 2), indicating that a majority of ADE tweets are correctly classified.

2.2 Task-1b: ADE Span Detection

The goal of this subtask is to detect the text span of reported ADEs in tweets.

2.2.1 Data and Pre-processing

The given dataset consisted of 1,712 spans across 1,234 tweets. For the purpose of better training of the model, all tweets with duplicate or overlapping spans were manually removed. The decision to do this manually was to ensure that spans providing better context were kept instead of just individual words that would have been less helpful in discerning the meaning of the sentence.

2.2.2 System Description

The dataset was passed through a Named Entity Recognition (NER) pipeline made using the `en_core_web_trf` model. The pipeline makes use of the `roberta-base` model provided by Huggingface's Transformers library (Wolf et al., 2020). The algorithm for extracting Adverse Effects from tweets is provided in Algorithm 1.

Algorithm 1: Algorithm for Extraction of Adverse Drug Effects from Tweets

```
Input: Input raw tweet $T$;
Output: $Label$, Start char, End char, Span;
Given ($T$), Classify the tweet with fairseq
  RoBERTa into ADE or NoADE;
if $Label$ is ADE then
    Perform NER on $T$ using spaCy NER
      pipeline;
    Return Start char, End char, Span;
end
```

2.2.3 Experiments

Two Named Entity Recognition (NER) pipelines, `en_core_web_trf` (`https://spacy.io/models/en#en_core_web_trf`) and `en_core_web_sm` (`https://spacy.io/models/en#en_core_web_sm`) were tried. The first is a RoBERTa based model while the second is a fast statistical entity recognition system trained on written web text that includes vocabulary, vectors, syntax and entities. After hyperparameter tuning, the transformer model was chosen. The model was trained for 150 epochs with a dropout of 0.3, Adam optimizer (Kingma and Ba, 2014) and a learning rate of 0.001 with (β_1, β_2)=(0.9, 0.999).

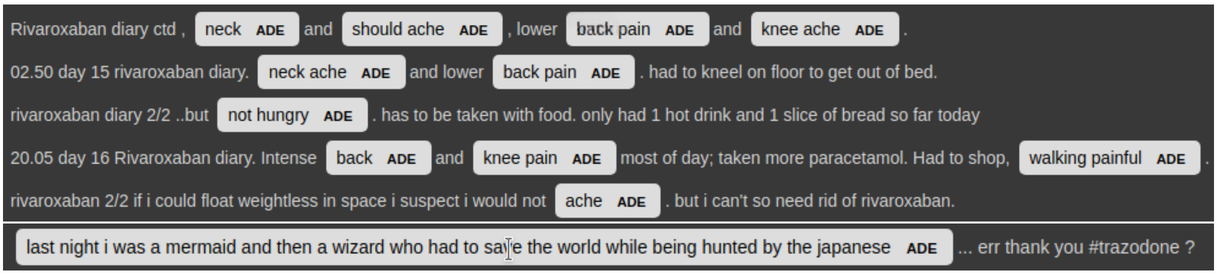
2.2.4 Results

The models have been evaluated with two metrics, the Relaxed F1 score, and the Strict F1 score. The Relaxed metrics evaluate the scores for spans that have a partial or full overlap with the labels. The Strict metrics only evaluate the cases where the spans produced by the model perfectly match the span in the label.

Table 3 showcases the performance of both NER pipelines on the validation set. It can be observed that the RoBERTa model provides a higher F1 score than the statistical model and is able to make much more accurate classifications of the ADE class. The statistical model however provides a higher recall which indicates it has fewer false negatives and is thus misclassifying the ADE samples as NoADE less often. The RoBERTa model is however far superior to the statistical model when considering the strict F1 scores. This implies that it is able to produce a perfect span more often and has learnt a

Model	Relaxed P	Relaxed R	Relaxed F1	Strict P	Strict R	Strict F1
`en_core_web_sm`	0.516	**0.551**	0.533	0.226	0.241	0.233
`en_core_web_trf`	**0.561**	0.529	**0.544**	**0.275**	**0.253**	**0.263**

Table 3: Scores on the Validation Set for the model for task 1b.



Rivaroxaban diary ctd , neck ADE and should ache ADE , lower back pain ADE and knee ache ADE .

02.50 day 15 rivaroxaban diary. neck ache ADE and lower back pain ADE . had to kneel on floor to get out of bed.

rivaroxaban diary 2/2 ..but not hungry ADE . has to be taken with food. only had 1 hot drink and 1 slice of bread so far today

20.05 day 16 Rivaroxaban diary. Intense back ADE and knee pain ADE most of day; taken more paracetamol. Had to shop, walking painful ADE .

rivaroxaban 2/2 if i could float weightless in space i suspect i would not ache ADE . but i can't so need rid of rivaroxaban.

last night i was a mermaid and then a wizard who had to save the world while being hunted by the japanese ADE ... err thank you #trazodone ?

Figure 1: Example span extraction from TensorFlu's model for task 1b

	Precision	Recall	F1
`en_core_web_trf`	0.493	**0.505**	**0.50**
`en_core_web_sm`	**0.521**	0.458	0.49
Median	0.493	0.458	0.42

Table 4: Comparing our best-performing model to the median for task 1b.

better representation of the data.

The final test set result achieved by the model placed on the leaderboard was achieved by the RoBERTa based NER model. The results obtained by both models are compared to the median in Table 4. The transformer pipeline provides a higher recall than the statistical pipeline thus showcasing the fact that a higher number of tweets were correctly classified as ADE while having overlapping spans. A few example images showing the performance of the entire adverse effect extraction pipeline are provided in Figure 1.

3 Task-6: Classification of COVID-19 tweets containing symptoms

The goal of this task is to classify tweets into 3 categories: (1) Self-reports (2) Non-personal reports (3) Literature/News mentions.

3.1 Data and Pre-processing

The SMM4H organizers released a training dataset consisting of 9,567 tweets and test data consisting of 6,500 tweets. The training dataset consisted of 4,523 tweets with Literature/News mentions, 3,622 tweets with non-personal reports and 1,421 tweets with self-reports. There is very little class imbalance in the given dataset. Tokenization of tweets was done using VinAI's `bertweet-base` tokenizer from the *Huggingface* API (Wolf et al., 2020). In order to use the BERTweet model, the tweets were normalized by converting user mentions into the @USER special token and URLs into the HTTPURL special token. The *emoji* package was used to convert the emoticons into text. (Nguyen et al., 2020)

3.2 System Description

BERTweet (Nguyen et al., 2020) uses the same architecture as BERT base and the pre-training procedure is based on RoBERTa, (Liu et al., 2019) for more robust performance, as it optimizes the BERT pre-training approach. BERTweet is optimized using Adam optimizer (Kingma and Ba, 2014), with a batch size of 7K and a peak learning rate of 0.0004, and is pre-trained for 40 epochs (using first 2 epochs for warming up the learning rate). The `bertweet-covid19-base-uncased` model was used for our application, which has 135M parameters, and is trained on 845M English tweets and 23M COVID-19 English tweets.

For training the BERTweet model on our train dataset, (`https://github.com/VinAIResearch/BERTweet`) was used with number of labels set to 3.

3.3 Experiments

A number of experiments were carried out to reach the optimal results for the task. Other models besides BERTweet were trained for the task such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and Covid-Twitter-BERT (Müller et al., 2020). A majority voting ensemble with

S.No.	Model	Arch	Label	Precision	Recall	F1
1.	RoBERTa	$BERT_{LARGE}$	Lit-News mentions	0.98	0.97	0.98
			Nonpersonal reports	0.95	0.97	0.96
			Self reports	0.97	0.96	0.97
2.	**BERTweet**	$BERT_{BASE}$	Lit-News mentions	**0.99**	0.99	**0.99**
			Nonpersonal reports	**0.99**	**0.98**	**0.98**
			Self reports	0.97	**1.00**	**0.99**
3.	DeBERTa	$BERT_{BASE}$	Lit-News mentions	0.95	**1.00**	0.98
			Nonpersonal reports	**0.99**	0.95	0.97
			Self reports	**1.00**	0.95	0.97
4.	Covid-Twitter BERT	$BERT_{LARGE}$	Lit-News mentions	0.98	0.98	0.98
			Nonpersonal reports	0.97	0.97	0.97
			Self reports	0.97	0.99	0.98
5.	Majority Voting	NA	Lit-News mentions	0.98	0.99	**0.99**
			Nonpersonal reports	0.98	0.97	0.97
			Self reports	0.99	0.99	**0.99**

Table 5: Comparing different models used for task 6 on the **Validation Set**

all 4 models was also evaluated. After a lot of tuning, BERTweet was found to be the best performing model on the dataset.

The ideal hyperparameters for the model were found empirically following many experiments with the validation set. The best results were obtained with the following hyperparameters: the model was finetuned for 12 epochs with a batch size of 16; the learning rate was warmed up for 500 steps with a weight decay of 0.01.

Due to little class imbalance in the given dataset and pretrained BERT based models performing very well on classification tasks, almost all models achieved a relatively high F1-score.

3.4 Results

The results on the validation set for all the trained models are reported in Table 5. As mentioned in section 2.1.4 the models have been compared on the basis of Precision, Recall and F1-score. The best performing model as seen in Table 5 is BERTweet. The same model was also able to achieve an F1 score above the median on the test set as seen in Table 6.

	Precision	Recall	F1
BERTweet	**0.9411**	**0.9411**	**0.94**
Median	0.93235	0.93235	0.93

Table 6: Comparing our best-performing model to the median for task 6

4 Conclusion

In this work we have explored an application of RoBERTa to the task of classification, extraction and normalization of Adverse Drug Effect (ADE) mentions in English tweets and the application of BERTweet to the task of classification of tweets containing COVID-19 symptoms. We have based our selection of these models on a number of experiments we conducted to evaluate different models. Our experiments have shown that RoBERTa outperforms BioBERT, achieving state of the art results in ADE classification. For the second task, we found that BERTweet outperformed all the other models including an ensembling approach (majority voting).

We foresee multiple directions for future research. One possible improvement could be to use joint learning to deal with Task-1(a) and Task-1(b) simultaneously.

5 Acknowledgements

For Task 1a, *Sidharth R, Abhiraj Tiwari, Parthivi Choubey and Saisha Kashyap* contributed equally to the work. For Task 1b, *Sidharth R, Abhiraj Tiwari, Sahil Khose, and Kumud Lakara* contributed equally to the work. All authors contributed equally to the work done for Task 6. We would also like to thank *Mars Rover Manipal and Project Manas* for providing the necessary resources to train our models.

References

Shaza M Abd Elrahman and Ajith Abraham. 2013. A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1(2013):332–340.

Kate Conger. 2021. Twitter shakes off the cobwebs with new product plans. *The New York Times*.

Aaron Gokaslan*, Vanya Cohen*, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hootsuite and We Are Social. 2021. Digital 2021: Global overview report.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36, Barcelona, Spain (Online). Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. 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.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Sebastian Nagel. 2016. Cc-news.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Elisha Shearer and Amy Mitchell. 2021. News use across social media platforms in 2020. *Pew Research Center*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.