CHAAI@SMM4H'22: RoBERTa, GPT-2 and Sampling - An interesting concoction

Christopher Palmer, Sedigheh Khademi, Muhammad Javed,
Gerardo Luis Dimaguila, Jim Buttery

Murdoch Children's Research Institute

{chris.palmer, sedigh.khademi, muhammad.javed, gerardoluis.dimaguil, jim.buttery}@mcri.edu.au

Abstract

This paper describes the approaches to the SMM4H 2022 Shared Tasks that were taken by our team for tasks 1 and 6. Task 6 was the "Classification of tweets which indicate self-reported COVID-19 vaccination status (in English)". The best test F1 score was 0.82 using a CT-BERT model, which exceeded the median test F1 score of 0.77, and was close to the 0.83 F1 score of the SMM4H baseline model. Task 1 was described as the "Classification, detection and normalization of Adverse Events (AE) mentions in tweets (in English)". We undertook task 1a, and with a RoBERTa-base model achieved an F1 Score of 0.61 on test data, which exceeded the mean test F1 for the task of 0.56.

1 Introduction

The shared tasks of the Social Media Mining for Health (SMM4H) are focused on overcoming difficult challenges in utilizing natural language processing techniques for deriving health-related information from social media data (Weissenbacher et al., 2022).

Task 6 of the seventh SMM4H (in 2022) was the "Classification of tweets which indicate self-reported COVID-19 vaccination status (in English)". Our team, "Champions of Health and Artificial Intelligence" (CHAAI), found task 6 particularly interesting, as the challenge is similar to a key research area we are conducting in SAEFVIC (SAEFVIC, 2022) to leverage social media monitoring to improve vaccine safety surveillance. Detecting self-reporting in health-related online posts helps to identify genuine descriptions of health events. It is as important to know *who* is speaking as it is to know *what* is being spoken about; as in detecting adverse events following immunization (AEFI) (Habibabadi et al., 2022; Wang et al., 2019; Lian et al., 2022); experiences of COVID-19 disease (Valdes et al., 2021); and drug effects (Aji et al., 2021).

Task 1 was the "Classification, detection and normalization of Adverse Events (AE) mentions in tweets (in English)". The adverse events mentioned in the task description are Adverse Drug Events (ADE), which are unexpected side effects following consumption of medications. Task 1a was for the classification of ADE, task 1b was to identify the spans of the relevant text, and task 1c required matching the colloquially expressed reactions to MedDRA standard concept IDs. Whereas task 6 was akin to obtaining an online data stream of likely candidates for AEFI detection, task 1a's challenge of precisely identifying an ADE was akin to differentiating specific AEFI in that data stream.

In both tasks, we had to deal with massive imbalances between the under-represented positive class and the negative class, which matches real-world conditions. Dealing with class imbalances was a major focus of our response to the tasks.

2 Task 6: Classification of tweets which indicate self-reported COVID-19 vaccination status

Identifying tweets of users reporting on their COVID-19 vaccination status.

2.1 Task 6 Data and Pre-processing

There were 13,692 records in the supplied training data (SMM4H, 2022), with 1,495 records with the positive label of "Self_reports", and 12,197 with the negative "Vaccine_chatter" label. There were 2,783 unlabeled records in the validation dataset. However, 2,478 records in the validation dataset were also found in the training data.

Text preparation consisted of replacing user mentions and URLs with placeholder expressions, converting emoji into text equivalents, and splitting words on apostrophes.

The dataset specification described the positive class as "unambiguous tweets of users clearly stating that they have been vaccinated." However, the training data did not align with this description.

We found that almost half of the positive labels (705/1495) included general descriptions of non-personal vaccination, personal statements of *not* being vaccinated, and even declarations and sentiment opposing vaccination. There were also personal reports of recent vaccination in the negative labels (392/12197).

We responded this labelling inconsistency by added additional clearly positive (and some clarifying negative) labelled texts to help train a model more consistently on the characteristics of a text discussing personal vaccination status. These were obtained from our previous work on COVID-19 vaccination reactions data (Khademi Habibabadi et al., 2022), which had been based on prior research that combined topic modelling and classification for filtering tweets to obtain *vaccine adverse event mentions* (VAEM) (Khademi Habibabadi et al., 2019), (Khademi et al., 2022).

After evaluation of classification using the added data, we also bolstered the existing shared task positive labelled data (despite its lack of accuracy), by generating additional similar examples. We used a GPT-2 (Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, 2020) model to learn from the positively labelled texts, then generated new texts by using a seed phrase of the initial few words of each existing positive record. Various combinations of the separately obtained texts and generated texts were used and assessed with F1 scoring. We found that adding the generated texts improved the score. Eventually, an enlarged and balanced dataset of 29,470 records was obtained. This was split 90/10 to create training and validation datasets of 26,524 and 2,946 records respectively.

During the evaluation phase we obtained scores from the CodaLab system, which used the separate task 6 validation dataset. Labels were only supplied to competitors *after* evaluation completion.

The task 6 hold-out test dataset was used to obtain the final scores of the models, which are used by the task examiners to rank the entries.

2.2 Task 6 System Description

Models' results are presented in Table 1. We have previously used the BERTweet-Large (Nguyen et al., 2020) model for similar tasks, so we evaluated fine tuning of the original model for Model 1, abbreviated as BERTweet-Lg in the table. A checkpoint of a BERTweet-Large model, previously fine-tuned by our team (Khademi Habibabadi et al., 2022) was used for Model 2 (abbreviated as BTL-PrevFT), and likewise a checkpoint of a previously fine-tuned (Khademi et al., 2022) RoBERTa-Large model (Liu et al., 2019) was used for Model 3 (RL-PrevFT), resulting in two new viable checkpoints (Model 3 and Model 3a). Additionally, we evaluated fine-tuning of the COVID-Twitter-BERT (CT-BERT) model (Müller et al., 2020) for Model 4.

We used the HuggingFace Trainer with an AdamW optimizer, for 3 to 5 epochs, and hyper-parameters of learning rate of 2e-5 and weight decay of 0.01. Other settings were Trainer defaults, including batch size of 8. We retained the best checkpoints of each training run, some of which favored recall, and others precision, as prior experience indicated models for classifier ensembles should emphasize recall and precision differently. We created a high-scoring ensemble that included the two versions of Model 3.

2.3 Task 6 Results

Model #, Descrip	Precision	Recall	F1
1 – BERTweet-Lg	0.82	0.75	0.78
2 – BTL-PrevFT	0.75	0.74	0.75
3 – RL-PrevFT	0.79	0.84	0.82
3a – RL-PrevFT	0.88	0.74	0.81
4 – CT-BERT	0.80	0.64	0.71
Ensemble (1,3,3a)	0.86	0.79	0.83

Table 1: Task 6 validation scores

Table 1 shows precision, recall and F1 scores on the positive class in the supplied validation data, for the best (by F1) of each the four models we assessed, and for an ensemble of 3 models using max voting.

Upon receiving the validation labels, we analyzed our incorrect predictions and found many validation records that we considered as incorrectly labelled. Although we had 103/2783 incorrect predictions, by our judgement of the labels, we concluded that the model was incorrect for only 15 of the 103 records.

The false positives tended to have the language of self-reports, but the vaccination mentions were either for someone else such as a parent, or for a future appointment – e.g., "*I just qualified to be eligible for the #COVID19 vaccine*", and "*Today he received his first dose of a COVID-19 vaccine*". The false negatives often had an indirect reference that implied that a previous vaccination had occurred – e.g., "*I am going to get my booster*", and "*I will not get the second shot*".

3 Task 1a: Classification of Adverse Events mentions in tweets

Task 1a was to classify English tweets containing one or more Adverse Drug Events (ADE) or no ADE.

3.1 Task 1a Data and Pre-processing

The training dataset consisted of 17,385 tweets, with 1,235 positive labels (ADE) and 16,150 negative labels (noADE). To overcome the class imbalance, we performed oversampling and under-sampling on the training dataset.

Oversampling: Applied data augmentation (Lemaitre et al., 2017) to increase the size of positive samples in the training dataset. We used contextual word embedding techniques (Ma, 2019) to insert or substitute words randomly in copies of the positively labelled texts, using a pretrained Roberta base model. After insertion and substitution of words in the posts, the positive samples were increased to 2,905.

Under-sampling: 0.5 under-sampling was applied to the majority negative class, which reduced the negative samples to 8,075.

In the pre-processing step, we removed the hashtags, user mentions, and special characters but as these changes did not make a positive contribution to model scores, so we progressed to fine-tune the models without pre-processing.

3.2 Task 1a System Description

We evaluated the RoBERTa-base transformer model (Liu et al., 2019), the BERT-base-uncased and BERT-large-uncased models (Devlin et al., 2018), and the distilBERT-base-uncased (Sanh et al., 2019) model.

3.3 Task 1a Experiments

We fine-tuned and evaluated the various models, dividing the dataset into 80%, 10%, and 10% for training, testing, and validation respectively. After analyzing the results of these trainings, we decided to use the Roberta-base model and trained for 5 epochs with a batch size of 32, a learning rate of 1e-5, a dropout rate of 0.1, and the Adam optimizer.

3.4 Task 1a Results

Model	Precision	Recall	F1
RoBERTa-base	0.72	0.83	0.77
RoBERTa-large	0.75	0.73	0.74
BERT-base-uncased	0.63	0.62	0.62
Distilbert-uncased	0.52	0.82	0.63

Table 2: Task 1a validation scores

The validation scores on the positive label of the best model are in Table 2. The F1 score of the selected model on the validation dataset was 0.77.

One of the major reasons for false negatives in the validation dataset are subtleties in the expressions to explain ADEs – e.g., "*debating on taking a trazodone and literally passing out for the day*". In some tweets people wanted to know the causes of adverse events with an unclear description of an experienced ADE – e.g., "*21y.o. w/ sickle-cell anemia and taking trazodone presents w/ priapism. what's the cause?*".

4 Conclusion

For task 6, although our best model on validation data was an ensemble of a BERTweet-Large and two RoBERTa-Large models, surprisingly, the CT-BERT model (Model 4) was our best model on test data. Its F1 score of 0.82 was close to the 0.83 F1 score of the baseline SMM4H model, and significantly better than the median F1 score of 0.77 for the task. However, the ensemble was only marginally behind – to 3 decimals its F1 score was 0.814, compared to 0.819 for the CT-BERT.

For task 1a, our best result on the test dataset was from the RoBERTa-base model. Notably, thanks to its balance of precision and recall, its F1 score matched last year's winning F1 score of 0.61 (Ramesh et al., 2021). Moreover, it exceeded this year's mean F1 score of 0.56. Scores are presented in Table 3.

Task	Model	Precision	Recall	F1
6	CT-BERT	0.86	0.78	0.82
	Ensemble	0.87	0.77	0.81
	Baseline	0.90	0.77	0.83
	Median	0.90	0.68	0.77
1a	RoBERTa-base	0.61	0.61	0.61
	Mean	0.65	0.50	0.56

Table 3: Test scores for both tasks

References

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. :58–64.

Ilya Sutskever Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei. 2020. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(May):1–7.

Jacob Devlin, Ming-Wei Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Sedigheh Khademi Habibabadi, Pari Delir Haghighi, Frada Burstein, and Jim Buttery. 2022. Vaccine Adverse Event Mining of Twitter Conversations: 2-Phase Classification Study. *JMIR Med Inform 2022;10(6):e34305 https://medinform.jmir.org/2022/6/e34305*, 10(6):e34305.

Sedigheh Khademi Habibabadi, Pari Delir Haghighi, Sedigheh Khademi Habibabadi, Pari Delir Haghighi, Sedigheh Khademi, and Pari Delir Haghighi. 2019. Topic Modelling for Identification of Vaccine Reactions in Twitter. *ACM International Conference Proceeding Series*:31.

Sedigheh Khademi Habibabadi, Christopher Palmer, Gerardo Luis Dimaguila, Muhammad Javed, Hazel Clothier, and Jim Buttery. 2022. Automated social media surveillance for detection of vaccine safety signals: a validation study. *Applied Clinical Informatics*, in press.

Sedigheh Khademi, Pari Delir Haghighi, Frada Burstein, and Jim Buttery. 2022. Vaccine adverse event mentions in social media: Mining the language of Twitter conversations. *JMIR Medical Informatics preprint*.

Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17):1–5.

Andrew T Lian, Jingcheng Du, and Lu Tang. 2022. Using a Machine Learning Approach to Monitor COVID-19 Vaccine Adverse Events (VAE) from Twitter Data. *Vaccines*, 10(1):103.

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. (1).

Edward Ma. 2019. nlpaug: NLP Augmentation.

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

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Sidharth Ramesh, Abhiraj Tiwari, Parthivi Choubey, Saisha Kashyap, Sahil Khose, Kumud Lakara, Nishesh Singh, and Ujjwal Verma. 2021. BERT based Transformers lead the way in Extraction of Health Information from Social Media. :33–38.

SAEFVIC. 2022. Surveillance of Adverse Events Following Vaccination in the Community.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. :2–6.

SMM4H. 2022. Social Media Mining for Health 2022 (#SMM4H) | Health Language Processing Lab @ Penn IBI.

Alberto Valdes, Jesus Lopez, and Manuel Montes. 2021. UACH-INAOE at SMM4H: a BERT based approach for classification of COVID-19 Twitter posts. :65–68.

Junxiang Wang, Liang Zhao, and Yanfang Ye. 2019. Semi-supervised Multi-instance Interpretable Models for Flu Shot Adverse Event Detection. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*(October):851–860.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the Seventh Social Media Mining for Health Applications #SMM4H Shared Tasks at COLING 2022. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.