Approaching SMM4H with auto-regressive language models and back-translation

Joseph Cornelius[*†] **Tilia Ellendorff**[‡†] **Fabio Rinaldi**[*†]
[*]Dalle Molle Institute for Artificial Intelligence Research (IDSIA)
[†]Swiss Institute of Bioinformatics
[‡]University of Zurich, Department of Computational Linguistics
{joseph.cornelius,fabio.rinaldi}@idsia.ch
tilia.ellendorff@uzh.ch

Abstract

We describe our submissions to the 6th edition of the Social Media Mining for Health Applications (SMM4H) shared task. Our team (OGNLP) participated in the sub-task: Classification of tweets self-reporting potential cases of COVID-19 (Task 5). For our submissions, we employed systems based on auto-regressive transformer models (XLNet) and back-translation for balancing the dataset.

1 Introduction

The Social Media Mining for Health Applications (SMM4H) shared task 2021 (Magge et al., 2021) focuses on textual processing of noisy social media data in the health domain. Our team (OGNLP) participated in Task 5, a binary classification task to identify tweets self-reporting potential cases of COVID-19. Tweets are labeled as positive (marked "1") if they self-report potential cases of COVID-19, and negative (marked "0") otherwise.

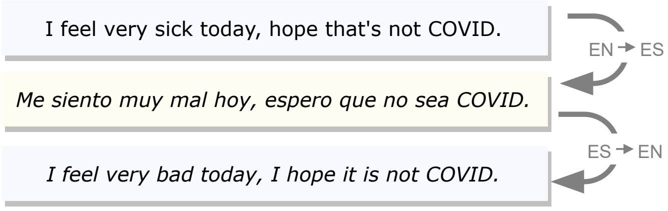
2 Dataset

The data provided by the organizers comprises tweets gathered from Twitter. As shown in Table 1, we have a total of 6465 training samples plus additional 716 validation samples. A test set with 10000 samples is provided for evaluation. On average, each tweet has 38 tokens and 155 characters. The dataset is unbalanced since the amount of negatively labeled tweets is five times higher than that of positively labeled tweets.

3 Methods

3.1 Preprocessing

Textual data from social media is often noisy since it contains many misspellings, abbreviations, emoticons, and non-standard wordings. Thus, preprocessing is a crucial part to de-noising the dataset and therefore increasing the performance. For this purpose, we modified the tweets as follows:


I feel very sick today, hope that's not COVID.

Me siento muy mal hoy, espero que no sea COVID.

I feel very bad today, I hope it is not COVID.

Figure 1: An example of back-translation.

- Hash symbols (#) were stripped from hash tags.
- All punctuation characters except ".,!?" were removed.
- URLs were eliminated from tweets.
- Emojies were stripped from tweets.
- The lowercase version of all tweets was used.

3.2 XLNet

As a baseline model, we used a pre-trained transformer language model, XLNet, which achieves state-of-the-art results on several sentiment analysis datasets (Yang et al., 2019). In contrast to the popular transformer-based BERT model (Devlin et al., 2019), XLNet is not pre-trained by predicting a masked token solely conditioned on its left (or right) tokens within the sentence, but instead the objective is to predict a masked token conditioned on all permutations of tokens within the sentence. Thus, XLNet's distinguishing feature is that it is able to learn context bidirectionally by permuting all the words in a sentence.

For the different trials, we used "XLNet-large-cased" from Huggingfaces python API[1] with a consistent setup of hyperparameters. We truncated each tweet to a maximum length of 256 characters, applied a batch size of four, and used a learning rate of 3e-6.

[1] https://huggingface.co/transformers/model_doc/xlnet.html

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 146–148
June 10, 2021. ©2021 Association for Computational Linguistics

Dataset	Neg	Pos	Total
Train	5,439	1,026	6,465
Valid	594	122	716
Test	-	-	10,000

Table 1: The number of tweets provided for Task 5, divided into training, validation, and test datasets.

System	Precision	Recall	F-Score
	Validation Set		
XLNet	0.83	0.81	0.82
XLNet+BT$_{k=1}$	**0.83**	**0.86**	0.**84**
XLNet+BT$_{k=1}$+PM	0.79	0.80	0.79
	Test Set		
XLNet	0.66	0.72	0.69
XLNet+BT$_{k=1}$	**0.70**	**0.72**	**0.72**
Mean	*0.74*	*0.74*	*0.75*

Table 2: Official and unofficial results of our systems, compared to the mean score of all competing systems.

3.3 Back-translation

Back-translation (BT) is a form of data argumentation and takes advantage of the advances in machine translation (Sennrich et al., 2015). BT allows us to balance the training set through the increase of the number of positive samples. Here our goal is to obtain a paraphrased tweet t' of a tweet t. To this end we automatically translate t into a different language (pivot) yielding $\tilde{t}$. Subsequently, we translate $\tilde{t}$ back to the source language and thus obtain the paraphrased tweet t'. BT leverages the fact that a translation often has several equivalent expressions in the target language. To obtain the BT dataset D_{BT}, we used the Google Translation API through TextBlob[2] and back-translated each English tweet from the minority class using the following ten languages as pivot: Bulgarian, Dutch, Gujarati, Hindi, Igbo, Japanese, Maltese, Pashto, Persian and Spanish. To increase the variance of the BT, we included pivots from low-resource languages and different language families.
Figure 1 shows that we can retrieve the paraphrased tweet t' *"I feel very bad today, I hope it is not COVID. "* from the original tweet t *"I feel very sick today, hope that's not COVID."* by using a BT from English → Spanish → English.

3.4 Parameter Merging

Parameter merging (PM) of equivalent models trained on different subsets of a dataset can be used to obtain a more robust and more generalized model (Utans, 1996; Ellendorff et al., 2019). For this purpose, we created a merged XLNet system from five XLNet models obtained by five-fold stratified cross-validation. For the merged XLNet system, we calculated the parameters' average across all five XLNet models.

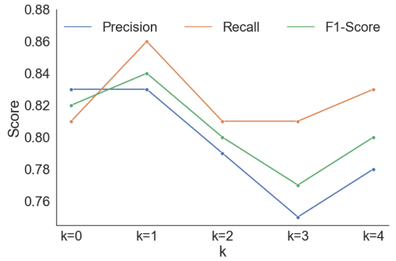


Figure 2: Influence of the number of BT samples used per tweet (k) on the performance of the XLNet system.

4 Results and Discussion

Table 2 shows the official results on the test set, as well as the unofficial results on the validation set. For models incorporating BT, the number of BT samples randomly drawn from D_{BT} used for each tweet is given by k, which means that for k=2, we triple the number of training samples. The XLNet model with BT k=1 has achieved the best results on both the test and validation dataset with an F-score of 0.72 and 0.84 respectively. In Figure 2 we can see that, contrary to expectation, the F-score of the models trained with BT does not constantly increase with an increase in k and has its optimum at k=1. The XLNet system generated from the PM of 5 XLNet models trained with cross-validation and BT k=1 achieved only the third-best F-score 0.80 on the validation dataset. Hence, we did not select the PM system for official submission as the submission was limited to the results of two runs.
For the second official submission, we used the XLNet system trained for four epochs without BT. However, this model achieved a significantly lower F-score with 0.69 on the official test set and 0.82 on the unofficial test set.
We assume that PM did not lead to an improvement as we had to set the number of folds for cross-validation very low (5) due to the limited GPU computing power available to us. Furthermore, we can conclude that back-translation leads to a significant improvement, but the number of additional generated samples plays a decisive role.

[2] https://textblob.readthedocs.io

References

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*, pages 4171–4186. Association for Computational Linguistics.

Tilia Ellendorff, Lenz Furrer, Nicola Colic, Noëmi Aepli, and Fabio Rinaldi. 2019. Approaching smm4h with merged models and multi-task learning. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 58–61. University of Zurich.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Joachim Utans. 1996. Weight averaging for neural networks and local resampling schemes. In *Proc. AAAI-96 Workshop on Integrating Multiple Learned Models. AAAI Press*, pages 133–138. Citeseer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.