

KUL@SMM4H'23: Text Augmentations with R-drop for Classification of Tweets Self Reporting Covid-19

Sumam Francis and Marie-Francine Moens
KU Leuven, Belgium

Abstract

This paper presents models created for the Social Media Mining for Health 2023 shared task. Our team addressed the first task, classifying tweets that self-report Covid-19 diagnosis. Our approach involves a classification model that incorporates diverse textual augmentations and utilizes R-drop to augment data and mitigate overfitting, boosting model efficacy. Our leading model, enhanced with R-drop and augmentations like synonym substitution, reserved words, and back translations, outperforms the task mean and median scores. Our system achieves an impressive F1 score of 0.877 on the test set.

1 Introduction

The goal of the shared task as part of the Social Media Mining for Health (SMM4H-23) (AZ et al., 2023)¹ is to facilitate the use of Twitter data for monitoring personal experiences of Covid-19 in real-time and on a large scale. The task involves automatically distinguishing tweets that self-report a Covid-19 diagnosis, for example, a positive test, clinical diagnosis, or hospitalization from tweets that merely state that the user has experienced Covid-19 without presenting any evidence and thus would not be considered a diagnosis.

SMM4H Task 1 comprises datasets of English tweets with annotations indicating the presence or absence of Covid-19 diagnosis in the tweet. Recent research has highlighted the benefits of data augmentation, as it enhances the diversity of training data and improves the robustness of the model on downstream tasks. The majority of data augmentation techniques generate new examples through modifications of existing ones, rooted in previous task-specific knowledge. Augmented data functions as a form of regularization, for instance, R-drop (Wu et al., 2021), effectively mitigating over-

fitting during the training of machine learning models.

2 Data

The dataset (Klein et al., 2023) comprises a training set (7600 tweets), validation set (400 tweets), and test set (10,000 tweets). The dataset is imbalanced with only around 20% of the tweets consisting of Covid-19 diagnosis mentions. We use oversampling to deal with this class imbalance. The evaluation metric used is the F1-score for the “positive” class (i.e., tweets that self-report a Covid-19 diagnosis). The training data include the Tweet ID, the tweet text, and the annotated binary label.

As part of the pre-processing on the dataset, we removed URLs, retweets, mentions, extra space, non-ascii words, and characters. Further, we lower-cased and striped off white spaces at both ends. We inserted space between punctuation marks.

3 Model

For task 1, we combine two strategies: 1) textual augmentations to enhance training data diversity, and 2) fine-tuning with pretrained transformer models. Our approach uses a Covid-Twitter-BERT (CT-BERT (Müller et al., 2023)), which is trained on 97M unique tweets (1.2B examples) related to Covid-19 exhibiting superior performance in tweet-related NLP tasks. Further, to mitigate the effects of an imbalanced label distribution and reduce overfitting, we employed the following strategies to enhance the model performance.

R-drop (Wu et al., 2021): This technique is a simple yet highly effective regularization method built upon dropout. It minimizes the bidirectional Kullback-Leibler (KL) divergence between output distributions of dropout-generated sub-model pairs during training. Input data undergoes dual forward R-drop enforces regularization by minimizing the

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice. This paper has been peer-reviewed and accepted for presentation at SMM4H'23 at AMIA 2023 Annual Symposium

bidirectional KL divergence between these distributions for the same input, aiding data augmentation and mitigating overfitting.

The different textual augmentations used include the following. **Synonym substitution:** This augmentation involves the random replacement of words in the tweet with their synonymous WordNet counterparts. **Tense transformation:** This conversion alters the tense of English sentences, e.g., from present to past. **Reserved tokens replacement:** Tokens within tweets were substituted using predefined reserved tokens. For instance, the Covid-19 tokens were subject to a random replacement from a reserved list of alternatives like "Coronavirus," "Covid," "corona," and "SARS-CoV-2." etc. **Back translation:** This transformation translates an input sentence from the source language to the target language and subsequently translates it back to the source language. We use the English-German translation model utilizing it as a method for nuanced paraphrasing.

Table 1: Micro-average Precision (P), Recall (R) and F1 scores (F1) on the validation set of the SMM4H 2023 Task 1 with CT-BERT model.

Augmentation	P	R	F1
-	0.8438	0.9373	0.8925
+ Text-aug	0.8593	0.9649	0.9090
+ Text-aug + R-drop	0.9003	0.9473	0.9230

Table 2: Micro-average Precision (P), Recall (R) and F1 scores (F1) on the test set of the SMM4H 2023 Task 1 with CT-BERT model.

Augmentation	P	R	F1
+ Text-aug	0.859	0.869	0.864
+ Text-aug + R-drop	0.888	0.866	0.877
Task mean results	0.824	0.792	0.804
Task median results	0.853	0.861	0.865

4 Experiments and Results

For classification, each model is fine-tuned over 10 epochs with a learning rate of $5e - 5$ using Adam optimizer. Batch size is set to 32, and max sequence length to 128. We use PyTorch and HuggingFace² for CT-BERT training, applying cross-entropy loss. Model checkpoints are saved every 200 steps based on the validation set's F1-score.

²<https://huggingface.co/models>

We adapted CT-BERT's loss function, adding KL divergence for (R-drop) regularization. Training data was enriched with textual augmentations from section 3 (Text-aug).

Incorporating R-drop and textual augmentations into CT-BERT produces the highest accuracy for detecting Covid-19 diagnosis in tweets (see Table 1 and Table 2). Augmentations diversify data, boosting model robustness and generalization. Noise from augmented data reduces overfitting risk and the model's reliance on specific patterns. Augmented examples aid learning and capture linguistic nuances for better language understanding. Adding R-drop further enhances performance, akin to an ensemble of sub-models within a single model. These sub-models encourage the model to capture various aspects of the data and promote learning more generalized features. Results surpass the shared task's mean and median scores, showcasing the effectiveness of this approach.

5 Conclusion

In this work, we developed a classification model enhanced with R-drop and textual augmentations to mitigate label imbalance and avoid the risk of overfitting thereby improving model performance. We demonstrate the significance of the textual augmentations and R-drop in enhancing model generalizability. Our system achieves an impressive F1 score of 0.877 on the test set.

References

- Klein AZ, Banda JM, Guo Y, Flores Amaro JI, Rodriguez-Esteban R, Sarker A, Schmidt AL, Xu D, and Gonzalez-Hernandez G. 2023. Overview of the eighth social media mining for health applications (smm4h) shared tasks at the amia 2023 annual symposium. In *In Proceedings of the Eighth SMM4H Workshop and Shared Task; 2023*.
- Ari Z Klein, Shriya Kunatharaju, Karen O'Connor, and Graciela Gonzalez-Hernandez. 2023. Automatically identifying self-reports of covid-19 diagnosis on twitter: An annotated data set, deep neural network classifiers, and a large-scale cohort. *Journal of Medical Internet Research, 25*.
- Martin Müller, Marcel Salathé, and Per E Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*.
- Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems, 34*.