Model Generalization on COVID-19 Fake News Detection

Yejin Bang(✉), Etsuko Ishii, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung

Center for Artificial Intelligence Research (CAiRE),
Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{yjbang,eishii,scahyawijaya,zjiad,}@connect.ust.hk

Abstract. Amid the pandemic COVID-19, the world is facing unprecedented *infodemic* with the proliferation of both fake and real information. Considering the problematic consequences that the COVID-19 fake-news have brought, the scientific community has put effort to tackle it. To contribute to this fight against the infodemic, we aim to achieve a robust model for the COVID-19 fake-news detection task proposed at CONSTRAINT 2021 (`FakeNews-19`) by taking two separate approaches: 1) fine-tuning transformers based language models with robust loss functions and 2) removing harmful training instances through influence calculation. We further evaluate the robustness of our models by evaluating on different COVID-19 misinformation test set (`Tweets-19`) to understand model generalization ability. With the first approach, we achieve 98.13% for weighted F1 score (W-F1) for the shared task, whereas 38.18% W-F1 on the `Tweets-19` highest. On the contrary, by performing influence data cleansing, our model with 99% cleansing percentage can achieve 54.33% W-F1 score on `Tweets-19` with a trade-off. By evaluating our models on two COVID-19 fake-news test sets, we suggest the importance of model generalization ability in this task to step forward to tackle the COVID-19 fake-news problem in online social media platforms.

Keywords: COVID-19 · Infodemic · Fake news · Robust loss · Influence-based cleansing · Generalizability

1 Introduction

As the whole world is going through a tough time due to the pandemic COVID-19, the information about COVID-19 online grew exponentially. It is the first global pandemic with the 4th industrial revolution, which led to the rapid spread of information through various online platforms. It came along with *Infodemic*. The infodemic results in serious problems that even affects people's lives, for

Y. Bang, E. Ishii, S. Cahyawijaya and Z. Ji—These authors contributed equally.

T. Chakraborty et al. (Eds.): CONSTRAINT 2021, CCIS 1402, pp. 128–140, 2021.
https://doi.org/10.1007/978-3-030-73696-5_13

instance, a fake news "Drinking bleach can cure coronavirus disease" led people to death[1]. Not only the physical health is threatened due to the fake-news, but the easily spread fake-news even affects the mental health of the public with restless anxiety or fear induced by the misinformation [38].

Table 1. Dataset statistics.

Label	FakeNews-19			Tweets-19	
	Train	Valid	Test	Valid	Test
Real	3360	1120	1120	51	172
Fake	3060	1020	1020	9	28
Total	6420	2140	2140	60	200

With the urgent calls to combat the infodemic, the scientific community has produced intensive research and applications for analyzing contents, source, propagators, and propagation of the misinformation [2,11,14,22,26] and providing accurate information through various user-friendly platforms [16,30]. The early published fact sheet about the COVID-19 misinformation suggested 59% of the sampled pandemic-related Twitter posts are evaluated as fake-news [2]. To address this, a huge amount of tweets is collected to disseminate the misinformation [1,21,23,27]. Understanding the problematic consequences of the fake-news, the online platform providers have started flag COVID-19 related information with an "alert" so the audience could be aware of the content. However, the massive amount of information flooding the internet on daily basis makes it challenging for human fact-checkers to keep up with the speed of information proliferation [28]. The automatic way to aid the human fact-checker is in need, not just for COVID-19 but also for any infodemic that could happen unexpectedly in the future.

In this work, we aim to achieve a robust model for the COVID-19 fake-news detection shared task proposed by Patwa. et al. [25] with two approaches 1) fine-tuning classifiers with robust loss functions and 2) removing harmful training instances through influence calculation. We also further evaluate the adaptability of our method out of the shared task domain through evaluations on different COVID-19 misinformation tweet test set [1]. We show a robust model with high performance over two different test sets to step forward to tackle the COVID-19 fake-news problem in social media platforms.

2 Dataset

Fake-News COVID-19 (`FakeNews-19`). A dataset released for the shared task of CONSTRAINT 2021 workshop [24], which aims to combat the infodemic

[1] https://www.bbc.com/news/world-53755067.

regarding COVID-19 across social media platforms such as Twitter, Facebook, Instagram, and any other popular press releases. The dataset consists of 10,700 social media posts and articles of real and fake news, all in English. The details of the statistic are listed in Table 1. Each social media post is manually annotated either as “Fake” or “Real”, depending on its veracity.

Table 2. Top-10 most frequent words on FakeNews-19 and Tweets-19

Dataset	Label	Most frequent words
Real	FakeNews-19	cases, #covid19, new, covid, tests, people, states, deaths, total, testing
	Tweets-19	#coronavirus, covid, cases, #covid19, people, virus, corona, health, spread, us
Fake	FakeNews-19	covid, coronavirus, people, virus, vaccine, #coronavirus, trump, says, new, #covid19
	Tweets-19	virus, corona, coronavirus, covid, #coronavirus, fake, news, get, really, media

Tweets COVID-19 (`Tweets-19`). To evaluate the generalizability of trained models test setting, we take the test set from [1], which is also released for fighting for the COVID-19 Infodemic tweets. The tweets are annotated with fine-grained labels related to disinformation about COVID-19, depending on the interest of different parties involved in the Infodemic. We took the second question, “*To what extent does the tweet appear to contain false information?*”, to incorporate with our binary setting. Originally, it is answered in five labels based on the degree of the falseness of the tweet. Instead of using the multi-labels, we follow the binary setting as the data releaser did to map to “Real” and “Fake” labels for our experiments. For our cleansing experiment, we split the dataset into validation and test set with equal label distribution. The detail is listed in Table 1. The most frequent words after removing stopwords on each dataset is listed in Table 2.

3 Methodology

3.1 Task and Objective

The main task is a binary classification to determine the veracity for the given piece of text from social media platforms and assign the label either “Fake” or “Real”. We aim to achieve a robust model in this task with a consideration on both high performance on predicting labels on `FakeNews-19` shared task and generalization ability through performance on `Tweets-19` with two separate approaches described in the following Sects. 3.2 and 3.3. Note that models are trained only with `FakeNews-19` train set.

3.2 Approach 1: Fine-Tuning Pre-trained Transformer Based Language Models with Robust Loss Functions

When handling text data, Transformers [31] based language models (LM) are commonly used as feature extractors [4,13,17] thanks to publicly released large-scale pre-trained language models (LMs). We adopt different Transformer LMs with a feed-forward classifier trained on top of each model. The list and details of models are described in Sect. 4.1. As reported in [9,12,37], robust loss functions help to improve the deep neural network performance especially with noisy datasets constructed from social medium. In addition to the standard cross-entropy loss (CE), we explore the following robust loss functions: symmetric cross-entropy (SCE) [33], the generalized cross-entropy (GCE) [39], and curriculum loss (CL) [19]. Inspired by the symmetric Kullback-Leibler divergence, SCE takes an additional term called reverse cross-entropy to enhance CE symmetricity. GCE takes the advantages of both mean absolute error being noise-robust and CE performing well with challenging datasets. CL is a recently proposed 0–1 loss function which is a tighter upper bound compared with conventional summation based surrogate losses, which follows the investigation of 0–1 loss being robust [7].

3.3 Approach 2: Data Noise Cleansing Based on Training Instance Influence

This approach is inspired by the work of Kobayashi et al. [10], which proposes an efficient method to estimate the influence of training instances given a target instance by introducing *turn-over dropout* mechanism. We define $D^{\text{trn}} = \{d_1^{\text{trn}}, d_2^{\text{trn}}, \ldots, d_k^{\text{trn}}\}$ as a training dataset with k training sample and $\mathcal{L}(f, d)$ as a loss function calculated from a model f and a labelled sample d. In turn-over dropout, a specific dropout mask $m_i \in \{0, \frac{1}{p}\}$ with dropout probability p is applied during training to zeroed-out a set of parameters $\theta \in \mathbb{R}^n$ from the model f for each training instance d_i^{trn}. With this approach, every single sample in the training set is trained on a unique sub-network of the model.

We define $h(d_i^{\text{trn}})$ is a function to map a training data d_i^{trn} into the specific mask m_i. The influence score $I(d^{\text{tgt}}, d_i^{\text{trn}}, f)$ for each target sample d^{tgt} is defined as follow:

$$I(d^{\text{tgt}}, d_i^{\text{trn}}, f) = \mathcal{L}(f^{\widetilde{h(d_i^{\text{trn}})}}, d^{\text{tgt}}) - \mathcal{L}(f^{h(d_i^{\text{trn}})}, d^{\text{tgt}}),$$

where $\widetilde{m_i}$ is the flipped mask of the original mask m_i, i.e., $\widetilde{m_i} = \frac{1}{p} - m_i$, and f^{m_i} is the sub-network of the model with the mask m_i applied. Intuitively, the influence score indicates the contribution of a training instance d_i^{trn} to the target instance d^{tgt}. A positive influence score indicates d_i^{trn} reduces the loss of d^{tgt} and a negative influence score indicates d_i^{trn} increases the loss of d^{tgt}, and the magnitude of the score indicates how strong the influence is. To calculate the total influence score of a training data d_i^{trn} over multiple samples from a given target set $D^{\text{tgt}} = \{d_1^{\text{tgt}}, d_2^{\text{tgt}}, \ldots, d_k^{\text{tgt}}\}$, we accumulate each individual influence score by:

$$I_{\text{tot}}(D^{\text{tgt}}, d_i^{\text{trn}}, f) = \sum_{j=1}^{K} I(d_j^{\text{tgt}}, d_i^{\text{trn}}, f).$$

The total influence score I_{tot} can be used to remove harmful instances, which only add noise or hinder generalization of the model, from the training set by removing top-n% of training instances with the smallest total influence score from the training data. We refer to our data cleansing method as influence-based cleansing which can remove noisy data and further improve model robustness and adaptability.

Table 3. Results on `FakeNews-19` test set using large language models. Underline indicates the best performance on each model. Acc. and W-F1 stands for Accuracy and weighted F1 respectively. SVM is placed under the column of CE for ease of comparison.

Loss functions models	CE		SCE		GCE		CL	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
TF-IDF SVM [25]	93.32	93.32	–	–	–	–	–	–
ALBERT-base	97.34	97.33	96.82	96.82	96.45	96.44	96.73	96.72
BERT-base	97.99	97.99	97.15	97.14	97.66	97.66	97.71	97.7
BERT-large	97.15	97.14	96.92	96.91	97.29	97.28	97.24	97.23
RoBERTa-base	97.94	97.94	97.52	97.51	97.57	97.56	97.62	97.61
RoBERTa-large	**98.13**	**98.13**	97.90	97.89	97.48	97.47	97.48	97.47

4 Experiment 1: Fine-Tuning LMs with Robust Loss Functions

4.1 Experiment Set-Up

We set up the baseline of our experiment from [25], an SVM model trained with features extracted from extracted by using TF-IDF. We try five different pre-trained BERT-based models, including ALBERT-base [13], BERT-base, BERT-large [4], RoBERTa-base, and RoBERTa-large [17]. We fine-tune the models on `FakeNews-19` train set with the classification layers on the top exploiting the pre-trained models provided by [36]. We train each model with four different loss functions, which are CE, SCE, GCE, and CL. The hyperparameters are searched with learning rate of 1e−6, 3e−6, 5e−6 and epoch of 1, 3, 5, 10 and the best combination is chosen based on performance on `FakeNews-19` validation set. The robustness of fine-tune models is then evaluated on both `FakeNews-19` and `Tweets-19` test sets. In this experiment, we mainly focus our evaluation on the Weighted-F1 (W-F1) score.

4.2 Experimental Results

Table 3 reports the result of on `FakeNews-19` task. Across all settings, RoBERTa-large trained with CE loss function achieved the highest W-F1 scores, 98.13%, with a gain of 4.81% in W-F1 compared to the TF-IDF SVM baseline. Except for BERT-large, all other models achieved their best performance when fine-tuned with CE loss function. The robust loss functions did not contribute in terms of improving the performance of predicting the labels. In other words, the large-scale LMs could extract high-quality features that the noise with FakeNews-19 was barely available for the robust loss functions to contribute.

In Table 4, we show the inference results on Tweets-19; unlike the successful result on FakeNews-19 RoBERTa-large with CE scores only 33.65% of W-F1 on Tweets-19, showing that the generalization of the model is not successful. Instead, the highest performance could be achieved with BERT-large with SCE with 38.18%, which is 4.53% gain compared to RoBERTa-large with CE. Interestingly, across all models, the highest performance when fine-tuned with the robust loss functions, SCE, GCE, and CL. This shows the robust loss functions help to improve the generalization ability of models. For instance, the RoBERTa-large could gain 3.85% with CL loss function, compared to its performance with CE. Considering that RoBERTa-large with CL achieves 97.47%, which is only 0.66% loss from the highest performance, it can be considered as a fair trade-off for selecting RoBERTa-large with CL could as a robust model, which achieves high performance on FakeNews-19 as well as generalizes better on Tweets-19.

Overall, while LMs with robust loss functions could achieve the highest 98.13% and lowest 96.44% on FakeNews-19, performance on Tweets-19 is comparatively poor as lower than 40% and even results in 22.85% lowest for W-F1. It could be inferred that the test set distributions are distinct although they are both related to COVID-19 infodemic and share the same data source, Twitter. This could be explained that CL is more robust to noisy labels, where FakeNews-19 labels are considered to be noisy to Tweets-19 test set. Further analysis is in Sect. 6.1.

Table 4. Results on Tweets-19 test set of large language model classifiers. Underlined results indicate the highest performance within each model.

Loss functions models	CE		SCE		GCE		CL	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
ALBERT-base	35.38	35.07	36.15	35.69	37.69	37.16	33.85	33.59
BERT-base	23.08	22.85	33.08	32.93	31.15	31.10	24.62	24.50
BERT-large	32.69	32.57	**38.85**	**38.18**	32.69	32.57	31.54	31.47
RoBERTa-base	28.08	28.08	36.92	36.38	33.46	33.24	29.62	29.61
RoBERTa-large	33.85	33.65	31.54	31.47	31.92	31.84	38.08	37.50

5 Experiment 2: Data Cleansing with Influence Calculation

5.1 Experiment Set-Up

We first fine-tune a pre-trained RoBERTa-large model with `FakeNews-19` train set while applying *turn-over dropout* to the weight matrix on the last affine transformation layer of the model with dropout probability of $p = 0.5$. We calculate the total influence score from the resulting model to the validation sets of `FakeNews-19` and `Tweets-19`. We investigate the effectiveness of our data cleansing approach by removing n% of training instances with the smallest total influence score with $n = \{1, 25, 50, 75, 99\}$. Then, we retrain the models from the remaining training data and perform an evaluation of the retrained model. All the models are trained with Cross-Entropy loss function with a fixed learning rate of 3e−6. We run the model for 15 epochs with the early stopping of 3. As the baseline, we compare our method with three different approaches: 1) pre-trained RoBERTa-large model without additional fine-tuning, 2) RoBERTa-large model fine-tuned with all training data without performing any data cleansing, and 3) model trained with random cleansing using the same cleansing percentage. We run each experiment five times with different random seeds to measure the evaluation performance statistics from each experiment.

5.2 Experiment Result

Based on our experiment results in Table 5, our influence-based cleansing method performs best for `Tweets-19` when the cleansing percentage is at 99% by only using 64 most influential training data. When cleansing percentage ≥25%, our influence-cleansed model outperforms the model without cleansing and the model with the random cleansing approach in terms of both accuracy and W-F1. The pre-trained model without fine-tuning (i.e. 0 training instance) results in 34.36% and 46.24% W-F1 on `FakeNews-19` and `Tweets-19` respectively. Our best model produces a significantly higher F1-score compared to the pre-trained model without fine-tuning by a large margin on both `FakeNews-19` and `Tweets-19`, which means that the small set of the most influential training data helps to significantly boost the generalization ability on both datasets. Furthermore, even with a high cleansing percentage, our model can maintain high evaluation performance on the `FakeNews-19`. Specifically, our model with a 99% cleansing percentage can produce an evaluation performance of 61.10% accuracy score and 54.33% W-F1 score on `Tweets-19` and 87.79% accuracy score and 87.69% W-F1 score on `FakeNews-19`. With this method, we could achieve an absolute gain of 20.69 W-F1 on `Tweets-19`, a much-improved generalization ability. Compared to the highest score achieved with using the full data for training, however, there is a trade-off with 10.44% loss for `FakeNews-19`. This trade-off in performances on two test sets suggests a potential for handling unseen data set during the training phase.

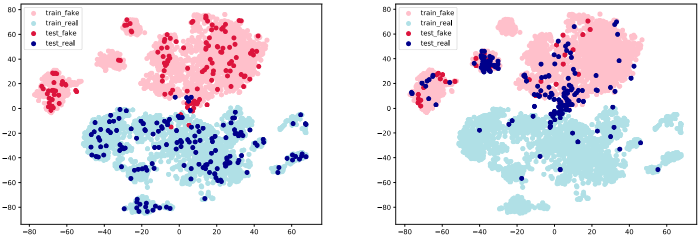
Table 5. Results on FakeNews-19 test set and Tweets-19 test set using Data cleansing approach. Model performance is explored when $n\%$ of harmful instances are dropped from the training. We run the experiments 5 times and report the mean. The underlined value indicates a higher value for comparing Influence vs. Random for each test set and each row.

Drop of instance		Training instance	**FakeNews-19**				**Tweets-19**			
			Influence		Random		Influence		Random	
%	#	#	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
0%	0	6420	**98.13**	**98.13**	98.13	98.13	33.85	33.65	33.85	33.65
1%	64	6356	97.96	97.96	97.40	97.40	32.00	31.76	30.60	30.39
25%	1605	4815	97.25	97.24	97.14	97.13	36.70	36.12	32.60	32.33
50%	3210	3210	97.01	97.00	88.29	86.38	37.70	37.09	30.80	30.19
75%	4815	1605	96.27	96.26	96.34	96.32	39.50	38.62	38.50	37.58
99%	6356	64	87.79	87.69	89.13	89.09	**61.10**	**54.33**	48.00	45.45

6 Discussion

6.1 Data Distribution Between Different FakeNews-19 and Tweets-19 Test Sets

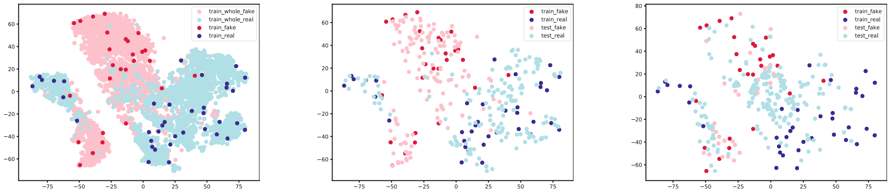
Although both data set built to address COVID-19 fake-news and share the same data collection source, tweets, the results show that the models trained on FakeNews-19 could achieve relatively lower performance on Tweets-19 test set. (Note that the Tweets-19 consists of the only test set with relatively smaller scale compared to FakeNews-19.) For further understanding, we visualize features extracted by the best performing model right before the classification layers with t-SNE. As shown in Fig. 1, even though the features of FakeNews-19 test set can distinguish the "Fake" and "Real" labels, the features of Tweets-19 cannot separate the two labels quite well.



(a) FakeNews-19 test set.

(b) Tweets-19 test set.

Fig. 1. Datasets distribution comparison with FakeNews-19 training set using t-SNE. While the distributions within FakeNews-19 kept to be similar, the distribution of Tweets-19 is significantly different.



(a) FakeNews-19 train set. **(b)** FakeNews-19 test set. **(c)** Tweets-19 test set.

Fig. 2. Datasets distribution comparison with top 1% influential training samples using t-SNE. Top 1% influential samples are distributed fairly evenly over the whole training set (a), thus the extracted test features remain separable (b), and the Tweets-19 distribution is captured better than trained with the full training set (c).

6.2 How Did Smaller Data Help for Generalization Ability of the Model?

As mentioned in Subsect. 5.2, higher cleansing percentage tends to lead to higher evaluation F1 score. By using the model trained with top 1% influential instances, we extract sentence representation as depicted in Fig. 2. Similar to in Fig. 1, the same number of instances from the test set are randomly selected for better understanding. Top 1% influential instances are fairly evenly sampled from the whole training set, and this small subset of the training set is enough to produce the distribution to separate the test features, which supports the effectiveness of the influential score. Moreover, since the top 1% samples are more sparse, the trained model can flexibly deal with samples from unseen distributions, resulting in extracted features of higher quality.

Table 6. Binary evaluation results of influence-based data cleansing model on Tweets-19 test set. B-F1, B-Rec., and B-Pre. denotes binary F1, binary recall, and binary precision scores respectively. Bold denotes the best performance over all experiments.

Drop %	Fake			Real		
	B-F1	B-Rec.	B-Pre.	B-F1	B-Rec.	B-Pre.
0%	28.80 ± 1.06	99.29 ± 1.60	16.85 ± 0.71	33.33 ± 5.25	20.12 ± 3.80	99.44 ± 1.24
1%	29.06 ± 1.17	99.29 ± 1.60	17.03 ± 0.82	34.46 ± 7.49	21.05 ± 5.43	99.58 ± 0.93
25%	30.56 ± 1.23	99.29 ± 1.60	18.07 ± 0.88	41.67 ± 6.11	26.51 ± 5.01	99.65 ± 0.78
50%	31.02 ± 0.75	**100.0 ± 0.00**	18.36 ± 0.52	43.16 ± 3.02	27.56 ± 2.49	**100.0 ± 0.00**
75%	31.51 ± 0.85	99.29 ± 1.60	18.73 ± 0.66	45.72 ± 4.47	29.77 ± 3.97	99.69 ± 0.70
99%	**37.17 ± 2.20**	81.43 ± 9.24	**24.28 ± 2.53**	**71.50 ± 6.92**	**57.79 ± 9.59**	95.23 ± 1.65

For the performance on Tweets-19 test set, we take additional consideration on binary-Recall (B-Rec.), binary-Precision (B-Prec.), and binary-F1 (B-F1) scores to further analyze the generalization ability of the model. As shown in

Table 6, the model with around 99% data cleansing achieves the best per class F1-score with 37.17% B-F1 score on the fake label and 71.50% on the real label. In general, the “Fake” B-Pre and “Real” B-Rec scores increase as the cleansing percentage increase, while “Real” B-Pre and “Fake” B-Rec behave the other way around, which means the model with higher cleansing percentage capture more real news and reduce the number of false “Fake” label with the trade-off of capturing less true ‘Fake” label. Overall, the B-F1 for each labels increases as the cleansing percentage increase. Our influence-based cleansing method outperforms the model without data cleansing by a large margin with 8.37% for the “Fake” B-F1 and 38.17% for the “Real” B-F1.

7 Related Works

COVID-19 Infodemic Research in Natural Language Processing. In recent months, researchers took various approaches to tackle the problem of COVID-19 Infodemic. Wang et al. [32] released centralized data CORD-19 that covers 59,000 scholarly articles about COVID-19 and other related coronaviruses to encourage other studies. Singh et al. [29] analyzed the global trend of tweets at the first emergence of COVID-19. To understand the diffusion of information, [3, 27] analyze the patterns of spreading COVID-19 related information and also quantify the rumor amplification across different social media platforms. Alam et al. [1] focuses on fine-grained disinformation analysis on both English and Arabic tweets for the interests of multiple stakeholders such as journalists, fact-checkers, and policymakers. Kar et al. [8] proposes a multilingual approach to detect fake news about COVID-19 from Twitter posts.

Generalization Ability of Models. As described in the previous section, several NLP studies involve emerging COVID-19 infodemic yet the generalization aspect is neglected although it is essential to accelerate industrial application development. In recent years, along with the introduction of numerous tasks in various domains, the importance of model generalization ability with a tiny amount or even without additional training datasets has been intensely discussed. In general, recent works on model generalizability can be divided into two different directions: 1) adaptive training and 2) robust loss function. In adaptive training, different meta-learning [5] and fast adaptation [18, 20, 35] approaches have been developed and show promising result for improving the generalization of the model over different domains. Another meta-learning approach, called meta transfer learning [34], improves the generalization ability for a low-resource domain by leveraging a high-resource domain dataset. In robust loss function, different kind of robust loss functions such as symmetric cross-entropy [33], generalized cross-entropy [39], and curriculum loss [19] have been shown to produce a more generalized model compared to cross-entropy loss due to its robustness towards noisy-labeled instances or so-called outliers from the training data. In addition to these approaches, data de-noising could actually improve model performance [15], thus, a data cleansing technique with identifying influential instances in the training dataset is proposed to further improve the evaluation performance and generalization ability of the models [6,10].

8 Conclusion

We investigated the COVID-19 fake-news detection task with an aim of achieving a robust model that could perform high for the CONSTRAINT shared task and also have high generalization ability with two separate approaches. The robust loss functions, compared to the traditional cross-entropy loss function, do not help much in improving F1-score on `FakeNews-19` but showed better generalization ability on `Tweets-19` with a fair trade-off as shown with the result comparison between RoBERTa-large with CE and CL. By performing influence data cleansing with high cleansing percentage (≥25%), we can achieve a better F1-score over multiple test sets. Our best model with 99% cleansing percentage can achieve the best evaluation performance on `Tweets-19` with 61.10% accuracy score and 54.33% W-F1 score while still maintaining high enough test performance on `FakeNews-19`. This suggests how we could use the labeled data to solve the problem of fake-news detection while model generalization ability should also be taken into account. For future work, we would like to combine the adaptive training, robust loss function with the influence score data cleansing method such that the resulting influence score can be made more robust for handling unseen or noisy data.

References

1. Alam, F., et al.: Fighting the COVID-19 infodemic in social media: a holistic perspective and a call to arms (2020)
2. Brennen, J.S., Simon, F., Howard, P.N., Nielsen, R.K.: Types, sources, and claims of COVID-19 misinformation. Reuters Institute **7**, 3–1 (2020)
3. Cinelli, M., et al.: The COVID-19 social media infodemic. arXiv preprint arXiv:2003.05004 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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: Human Language Technologies, vol. 1, pp. 4171–4186. ACL, Minneapolis, June 2019. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
5. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135. PMLR, International Convention Centre, Sydney, 06–11 August 2017 (2017)
6. Hara, S., Nitanda, A., Maehara, T.: Data cleansing for models trained with SGD. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 4213–4222. Curran Associates, Inc. (2019)

7. Hu, W., Niu, G., Sato, I., Sugiyama, M.: Does distributionally robust supervised learning give robust classifiers? In: International Conference on Machine Learning, pp. 2029–2037. PMLR (2018)
8. Kar, D., Bhardwaj, M., Samanta, S., Azad, A.P.: No rumours please! A multi-indic-lingual approach for COVID fake-tweet detection (2020)
9. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. **65**, 101759 (2020)
10. Kobayashi, S., Yokoi, S., Suzuki, J., Inui, K.: Efficient estimation of influence of a training instance. In: Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pp. 41–47 (2020)
11. Kouzy, R., et al.: Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. Cureus **12**(3), e7255 (2020)
12. Kumar, H., Sastry, P.: Robust loss functions for learning multi-class classifiers. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 687–692. IEEE (2018)
13. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
14. Lee, N., Bang, Y., Madotto, A., Fung, P.: Misinformation has high perplexity (2020)
15. Lee, N., Liu, Z., Fung, P.: Team yeon-zi at SemEval-2019 task 4: hyperpartisan news detection by de-noising weakly-labeled data. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 1052–1056 (2019)
16. Li, Y., et al.: Jennifer for COVID-19: an NLP-powered chatbot built for the people and by the people to combat misinformation. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. ACL, July 2020. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.9
17. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692
18. Liu, Z., et al.: CrossNER: evaluating cross-domain named entity recognition (2020)
19. Lyu, Y., Tsang, I.W.: Curriculum loss: robust learning and generalization against label corruption. arXiv preprint arXiv:1905.10045 (2019)
20. Madotto, A., Lin, Z., Bang, Y., Fung, P.: The adapter-bot: all-in-one controllable conversational model (2020)
21. Medford, R.J., Saleh, S.N., Sumarsono, A., Perl, T.M., Lehmann, C.U.: An "infodemic": leveraging high-volume Twitter data to understand public sentiment for the COVID-19 outbreak. medRxiv (2020)
22. Mian, A., Khan, S.: Coronavirus: the spread of misinformation. BMC Med. **18**(1), 1–2 (2020)
23. Mourad, A., Srour, A., Harmanani, H., Jenainatiy, C., Arafeh, M.: Critical impact of social networks infodemic on defeating coronavirus COVID-19 pandemic: Twitter-based study and research directions. arXiv preprint arXiv:2005.08820 (2020)
24. Patwa, P., et al.: Overview of CONSTRAINT 2021 shared tasks: detecting English COVID-19 fake news and Hindi hostile posts. In: Chakraborty, T., et al. (eds.) CONSTRAINT 2021. CCIS, vol. 1402, pp. 42–53. Springer, Cham (2021)
25. Patwa, P., et al.: Fighting an infodemic: COVID-19 fake news dataset (2020)
26. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., Rand, D.G.: Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. Psychol. Sci. **31**(7), 770–780 (2020)

27. Shahi, G.K., Dirkson, A., Majchrzak, T.A.: An exploratory study of COVID-19 misinformation on Twitter (2020)
28. Shao, C., et al.: Anatomy of an online misinformation network. PLoS ONE **13**(4), e0196087 (2018)
29. Singh, L., et al.: A first look at COVID-19 information and misinformation sharing on Twitter. arXiv preprint arXiv:2003.13907 (2020)
30. Su, D., Xu, Y., Yu, T., Siddique, F.B., Barezi, E.J., Fung, P.: CAiRE-COVID: a question answering and multi-document summarization system for COVID-19 research. arXiv preprint arXiv:2005.03975 (2020)
31. Vaswani, A., et al.: Attention is all you need. CoRR abs/1706.03762 (2017). http://arxiv.org/abs/1706.03762
32. Wang, L.L., et al.: CORD-19: the COVID-19 open research dataset. arXiv (2020)
33. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 322–330 (2019)
34. Winata, G.I., Cahyawijaya, S., Lin, Z., Liu, Z., Xu, P., Fung, P.: Meta-transfer learning for code-switched speech recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3770–3776. ACL, July 2020. https://doi.org/10.18653/v1/2020.acl-main.348. https://www.aclweb.org/anthology/2020.acl-main.348
35. Winata, G.I., et al.: Learning fast adaptation on cross-accented speech recognition. In: Meng, H., Xu, B., Zheng, T.F. (eds.) InterSpeech 2020, 21st Annual Conference of the International Speech Communication Association, pp. 1276–1280. ISCA (2020). https://doi.org/10.21437/Interspeech.2020-0045
36. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, October 2020
37. Xia, X., et al.: Part-dependent label noise: towards instance-dependent label noise (2020)
38. Xiong, J., et al.: Impact of COVID-19 pandemic on mental health in the general population: a systematic review. J. Affect. Disord. **277**, 55–64 (2020)
39. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in Neural Information Processing Systems, pp. 8778–8788 (2018)