Model Generalization on COVID19 Fake News Detection

Ye Jin Bang **Etsuko Ishii** **Samuel Cahyawijaya** **Ziwei Ji**
20319093 20625335 20667220 20618205
{yjbang, eishii, scahyawijaya, zjiad}@connect.ust.hk
Department of Electronic and Computer Engineering, HKUST

1 Introduction

As the whole world has been going through a tough time due to the pandemic COVID-19, the amount of information about COVID-19 online grew exponentially and also spread rapidly across all social media platforms. COVID-19 is the first global pandemic in the time of the world with 4th industrial revolution, which led increased access to the technology and enormous use of social media. It came along with *Infodemic*, defined as overabundance of information [1], including both misinformation and disinformation. The infodemic results in serious problems that even affects people's lives, for 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 mental health of public with restless disputes over online communities and anxiety or fear induced by the misinformation [38].

With the urgent calls to combat the infodemic regarding COVID-19, scientific community has produced intensive research and applications for analyzing contents, source, propagators, and propagation of the misinformation [26, 23, 12, 3] and providing accurate information through various user-friendly platforms [15, 30]. The early published fact sheet studied the COVID-19 misinformation in details suggested 59% of the sampled pandemic-related Twitter posts was evaluated as fake-news [3]. To address it with the primary focus on social media, a huge amount of tweets has been collected to disseminate misinformation [22, 24, 27, 14, 2]. Understanding the problematic consequences for the fake-news and misinformation spread through the social media platform, the platform providers have started flag all COVID-19 related information with an "alert" so viewers could be aware of content. However, human fact-checkers have the limited resources so they cannot check every single problematic content even until now. The automatic way to aid the human fact-checker is still in need, not just for COVID-19 but also for any infodemic that could happen continuously in the future.

In this work, we aim to achieve robust model for the COVID-19 fake-news detection task proposed by Patwa. et al. [25] by tackling it in two approaches 1) training with robust loss functions and 2) removing harmful training instances through influence calculation. In addition, we also further evaluate adaptability of our method by performing zero-shot evaluation on different COVID-19 misinformation test set [2]. Finally, we discuss current achievement and bottlenecks in automatic COVID-19 fake-news detection in social media platform, mainly Twitter.

2 Related Works

COVID-19 Infodemic Research in Natural Language Processing In recent years, researchers took various approaches to tackle the problem of Infodemic. Wang et. al [32] released centralized data that covers 59,000 scholarly articles about COVID-19, SARS-CoV-2, and other related corona viruses (CORD19) to encourage other studies and, using the data, [30] has built an end-to-end Question Answering system to provide accurate information. [28] first analyzed the global trend of tweets at

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

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Table 1: Dataset Statistics.

	FakeNews-19			Tweets-19	
Label	Train	Valid	Test	Valid	Test
Real	3360	1120	1120	45	178
Fake	3360	1020	1020	15	59
Total	6420	2140	2140	60	237

the first emergence of COVID19. We mainly focus on COVID-19 misinformation detection on social media platforms. To understand diffusion of information, [4, 27] analyze the patterns of spreading covid19 related information for each platform and also quantify the rumor amplification across different social media platforms including Twitter, Instagram, YouTube, Reddit and Gab. [2] focuses on fine-grained disinformation analysis on both English and Arabic tweets for the interests of multiple stakeholders such as journalists, fact-checkers, and policy makers. [10] proposes a multilingual approach to detect fake news about COVID-19 from Twitter posts for multiple Indic-Languages.

Generalization ability of models Along with the introductions of numerous tasks or datasets in various domains, the importance of model generalization ability with tiny amount or even without additional training datasets on the target domains has been intensely discussed, for example zero-shot learning or few-shot learning [29, 9, 19]. Recent works on model generalizability can be divided into two different directions: 1) model-based adaptation and 2) adaptive pre-training. In model-based adaptation, a pre-trained model is extended with a small set of parameters called adapter layer. [21, 16] employs adapter layer to help the generalization of the model over different domains. In adaptive pre-training approach, different adaptative pre-training methods have been developed and show promising result for improving the generalization of the model over different domains [6, 35, 18]. Another approach improve the generalization ability for low-resource domains by leveraging high-resource domain dataset [34]. In addition to these approaches, data cleansing technique with identifying influential instances in the training dataset is proposed to improve the accuracy and generalization ability of the models [7, 11].

3 Dataset

Fake-News COVID19 (`FakeNews-19`) We used a dataset released by [25], which aims to combat the infodemic 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 on COVID19, all in English textual data. The details of statistic is listed in Table 1. Since it is one of the subtasks on the hostile post detection proposed for Shared Task @CONSTRAINT 2021[2], most of the results and analysis in this paper are based on the evaluation on validation set. Each of social media posts is manually annotated either as "Fake" or "Real", depending on its veracity.

Tweets COVID19 (`Tweets-19`) For zero-shot test setting, we take the dataset from [2], 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 incorporate with our binary setting. The question is "*To what extent does the tweet appear to contain false information?*" Originally, it is answered in five labels based on the degree of falseness of the tweet, which are {"NO, definitely contains no false information", "NO, probably contains no false information", "Not sure", "YES, probably contains false information", "YES, definitely contains false information"}. We map the first two labels as "Real" while the last two are mapped to "Fake". For our cleansing experiment, we split our the dataset into validation and test set with equal label distribution. The details of statistic of the dataset is listed in Table 1.

[2] `https://constraint-shared-task-2021.github.io/`

4 Methodology

4.1 Task

In this work, the main task is a binary classification to determine a veracity for the given piece of text from social media platforms and assign label either "Fake" or "Real". This task is based on the assumption that the given text contains verifiable information. The primary aim of this task is to achieve the best performing models to predict labels on `FakeNews-19` task. Thus, the approach is to explore different large pre-trained language model classifiers and loss functions on full train set, as explained in Section 4.2. Considering the urgent need of the automated fake-news detection solution in real-world, it is also important to improve the generalization of a model. The second approach, described in Section 4.3, is taken to train model with a cleansed training set by less influential or noise-adding training instances.

To explore the adaptability of the best performing models in COVID-19 fake-news detection, we investigate the models from the two different approaches in a zero-shot setting through evaluation on another COVID-19 fake-news test set `Tweets-19`. This provides us understanding the models performance in real-life application situation. More details are discussed in later sections.

4.2 Approach 1: Fine-tuning Pre-trained Transformer based Language Models with Robust Loss Functions

When handling text datasets, Transformers [31] based language models are commonly used as feature extractors [5, 13, 17] thanks to publicly released large-scale pre-trained models. We adopt different Transformer based large-scale pre-trained language models with a feed-forward classifier trained on top of each model. The list and details of models are described in Section 5.2.

Robust Loss Function As reported in [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) [20]. 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 [8].

4.3 Approach 2: Data Noise Cleansing based on Training Instance Influence

We are inspired by the work of [11], which proposes an efficient method to estimate the influence of training instances given a target instance by introducing *turn-over dropout* mechanism. We define $D^{\mathrm{trn}} = \{d_1^{\mathrm{trn}}, d_2^{\mathrm{trn}}, \ldots, d_k^{\mathrm{trn}}\}$ as a 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 $h(d_i^{\mathrm{trn}})$ is a function to map a training data d_i^{trn} into the specific mask m_i. The influence score $I(d^{\mathrm{tgt}}, d_i^{\mathrm{trn}}, f)$ is calculated for a target sample d^{tgt} with the following equation:

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

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 $D^{\mathrm{tgt}} = \{d_1^{\mathrm{tgt}}, d_2^{\mathrm{tgt}}, \ldots, d_k^{\mathrm{tgt}}\}$, we accumulate each individual influence score by:

Table 2: Results on FakeNews-19 validation set using large language models. Underline indicates the best performance on each model. Note that the reported results are on validation set since the labels for test set have not been released. 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	CE		SCE		GCE		CL	
Models	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
SVM [25]	93.46	93.48	-	-	-	-	-	-
ALBERT-base	96.87	96.86	97.19	97.19	96.68	96.67	96.54	96.53
BERT-base	97.24	97.23	97.38	97.38	97.29	97.28	97.76	97.75
BERT-large	97.76	97.75	97.62	97.61	97.80	97.80	97.10	97.09
RoBERTa-base	97.76	97.75	97.52	97.51	97.34	97.33	97.38	97.38
RoBERTa-large	**98.13**	**98.13**	97.85	97.84	98.04	98.03	98.04	98.03

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

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 our data cleansing method as influence-based cleansing and with our cleansing approach we can remove noisy data and build a more robust model that can better adapt to the target domain.

5 Experiments

Our experiments are divided into two sections. We first explain evaluation metric and describe models and experimental set-ups for each approaches in following subsections.

5.1 Evaluation Metric

We mainly evaluate the performances on FakeNews-19 Weighted-F1 (W-F1) score for the comparison on the leaderboard for the shared task[3]. For the performance on Tweets-19 test set, we takes additional consideration on binary-Recall (B-Rec.), binary-Precision (B-Prec.) and binary-F1 (B-F1) scores for both "Fake" and "Real" labels. Accuracy scores are reported for both test sets.

5.2 Experiment 1: Fine-tuning robust language models and loss functions

We setup the baseline of our experiment from [2], a SVM model trained with features extracted from tf-idf. For the Approach 1, we try with five different BERT-based models, including ALBERT-base [13], BERT-base, BERT-large [5], RoBERTa-base, and RoBERTa-large [17]. We fine-tune the models 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. In total, we have 20 settings to find the best performing combination of model and loss function on the task of FakeNews-19 . The hyper parameters are searched with learning rate of $1\text{e}{-6}, 3\text{e}{-6}, 5\text{e}{-6}$ and epoch of $1, 3, 5, 10$.

After getting the best fitting loss functions for each language model classifiers, we evaluate each of trained models on Tweets-19 test set.

5.3 Experiment 2: Zero-shot with Data Cleansing

We test the adaptability of our data cleansing method by performing zero-shot evaluation on Tweets-19. We first fine-tune a pre-trained RoBERTa-large model with FakeNews-19while

[3] https://constraint-shared-task-2021.github.io/

Table 3: Results on Tweets-19 test set of large language model classifiers with corresponding loss functions.

Model	ALBERT-base	BERT-base	BERT-large	RoBERTa-base	RoBERTa-large
Loss Function	SCE	CL	GCE	CE	CE
Acc.	29.62	32.69	32.69	28.08	**33.85**
W-F1	29.61	32.57	32.57	28.08	**33.65**

Table 4: Results on FakeNews-19 validation set and Tweets-19 test set using Data cleansing approach. Model performance is explored when n% of harmful instances are dropped from training. Note that Tweets-19 is zero-shot setting. The first row denotes pre-trained model without fine-tuning and the second row denotes model pre-trained without data cleansing.

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
-%	-	-	44.30	34.60	-	-	53.08	34.42	-	-
0%	0	6420	98.13	98.13	-	-	33.85	33.65	-	-
1%	64	6356	97.99	97.98	98.04	98.03	48.10	48.05	54.09	51.23
5%	321	6099	97.48	97.46	97.94	97.94	51.98	51.81	50.80	50.20
10%	642	5778	**98.08**	**98.08**	97.48	97.46	49.70	49.62	49.37	49.16
25%	1605	4815	97.80	97.80	97.90	97.89	56.71	56.38	52.40	48.92
50%	3210	3210	97.66	97.65	97.24	97.24	54.18	53.48	54.18	47.52
75%	4815	1605	96.17	96.15	96.12	96.10	61.60	60.19	59.41	58.22
90%	5778	642	94.95	94.93	93.92	93.89	66.50	63.57	64.14	62.19
99%	6356	64	90.98	90.89	86.44	86.43	**72.24**	**65.88**	69.20	62.28

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 set of Tweets-19. We investigate the effectiveness of our data cleansing approach by removing $n\%$ of training instances with smallest total influence score with $n = \{1, 5, 10, 50, 75, 90, 95, 99\}$ and retrain the models the remaining training data. All the models are trained with Cross Entropy loss function and learning rate of $3\mathrm{e}{-6}$.

As the baseline, we compare our method with two different approaches: 1) model trained without performing data cleansing and 2) model trained with random cleansing using the same cleansing percentage. We run each experiment five times with different random seed to measure the evaluation performance statistics from each experiment.

6 Results

6.1 Classification with language models

Table 2 reports result of Experiment 1 on FakeNews-19 task, described in Section 5.2. Across all settings, RoBERTa-large trained with cross entropy loss function achieved the highest accuracy and weighted F1 scores, 98.13 and 98.13 respectively, with a gain of 4.69 in W-F1 compared to the SVM baseline reported in [25]. In addition, regardless of loss functions, RoBERTa-large showed the best performance among all other models. This could be explained with its largest size of parameters 355M. On the contrary, ALBERT-base has 11M parameters, which is only 3% of RoBERTa-large, and it could still achieve 97.17 of F1-score, with a minimal difference of 0.96 from the best performance of RoBERTa-large. Interestingly, CE achieves the best performance with RoBERTa models, while the robust loss functions slightly improve the performance of ALBERT or BERT models. This implies that RoBERTa extracts feature well enough thus the robust loss functions contribute less than the other models.

Table 5: Binary evaluation results of influence-based 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 and Underline denotes the best performance on each cleansing setting.

Drop	Real			Fake		
%	B-F1	B-Rec.	B-Pre.	B-F1	B-Rec.	B-Pre.
0%	49.24 ± 3.01	98.65 ± 2.21	32.87 ± 2.77	48.37 ± 10.65	32.58 ± 9.29	98.90 ± 1.83
1%	48.83 ± 1.34	**99.32 ± 0.93**	32.38 ± 1.21	47.28 ± 4.55	31.12 ± 4.01	**99.32 ± 0.94**
5%	50.37 ± 1.65	97.63 ± 3.51	33.99 ± 1.73	53.25 ± 6.89	36.85 ± 6.73	98.26 ± 2.42
10%	49.47 ± 1.93	98.65 ± 1.42	33.04 ± 1.74	49.78 ± 6.39	33.48 ± 5.66	98.72 ± 1.27
25%	53.18 ± 2.60	98.31 ± 2.07	36.51 ± 2.62	59.58 ± 6.33	42.92 ± 6.82	98.79 ± 1.28
50%	50.83 ± 3.16	93.90 ± 8.18	35.17 ± 4.28	56.14 ± 12.92	41.01 ± 14.32	96.51 ± 4.18
75%	53.10 ± 2.80	86.78 ± 4.70	38.40 ± 3.33	67.28 ± 6.33	53.26 ± 8.06	92.55 ± 2.25
90%	53.91 ± 1.89	78.98 ± 13.80	41.64 ± 3.53	73.22 ± 6.11	62.36 ± 10.83	90.72 ± 4.79
95%	**56.34 ± 2.62**	67.12 ± 11.60	**49.77 ± 5.22**	**81.56 ± 3.59**	76.63 ± 8.00	87.82 ± 2.75
99%	51.20 ± 1.93	58.65 ± 5.57	45.66 ± 2.15	80.56 ± 1.49	**76.74 ± 3.38**	84.89 ± 1.34

In Table 3, we show the inference results of the best combinations of models and loss functions on `Tweets-19`; even RoBERTa-large score only 33.85% of accuracy and 33.65% of weighted F1. It could be inferred that the dataset distributions are distinct although they are both related to COVID-19 infodemic. Further analysis can be found in Section 7.1.

6.2 Zero-shot with Data Cleansing

Based on our zero-shot experiment results on Table 4, our influence-based cleansing method performs best for zero-shot setting when the cleansing percentage is at 99% by only using 64 most influential training data. Our influence-cleansed model outperforms the model without cleansing and the model with random cleansing approach in terms of both accuracy and W-F1. Our model also produce significantly higher score compared to pre-trained model without fine-tuning by a large margin on both `FakeNews-19`and `Tweets-19`, which means that the 64 most influential training data helps to significantly boost the generalization ability on both datasets.

Based on our binary evaluation results on Table 5 the best performance is achieved by the model with 95% data cleansing with only 321 most influential samples according to the validation set of `Tweets-19`. In general, the "Fake" B-Rec and "Real" B-Pre scores increase as the cleansing percentage increase, while "Real" B-Rec and "Fake" B-Pre behave the other way around. Overall, the B-F1 for each labels increases as the cleansing percentage increase. Specifically, our influence-based cleansing method significantly outperforms to the model without data cleansing by around 7% for the "Real" B-F1 and around 33% for the "Real" B-F1.

7 Discussion

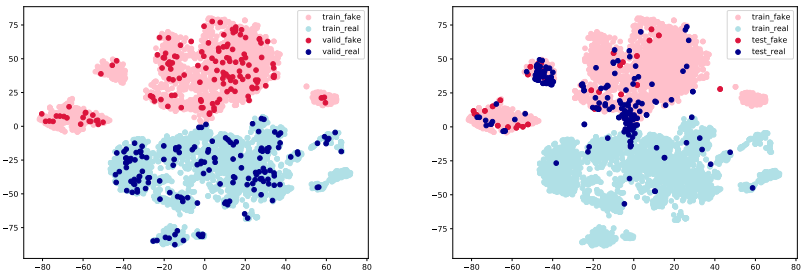
7.1 Data Distribution between different Test Sets

The dataset `FakeNews-19` is the only large-scale COVID-19 fake news dataset that consists of complete train, valid, and test splits. It covers a broad range of media platforms, however, it does not take a fine-grained approach. This has been shown in test on another twitter fake-news data set `Tweets-19` in Table 3.

For further understanding, we visualize features extracted by the best performing model right before the classification layers with t-SNE. As shown in Figure 1, even though the features of `FakeNews-19` validation set can be separated, the features of `Tweets-19` are not captured well.

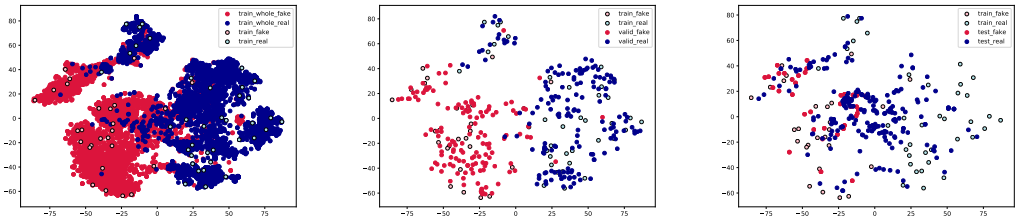
7.2 Why smaller data helped for generalization?

As mentioned in Subsection 6.2, 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



(a) FakeNews-19 validation set. (b) Tweets-19 test set.

Figure 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) Whole FakeNews-19 train set. (b) FakeNews-19 validation set. (c) Tweets-19 test set.

Figure 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 validation features remain separable (b), and the Tweets-19 distribution is captured better than trained with the full training set (c).

as depicted in Figure 2. Similar to in Figure 1, the same number of instances from the validation 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 validation 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.

7.3 Preprocessing ablation

In the whole experiment, we did not have any additional pre-processing step. However, the characteristic twitter data is hashtags (#), mentions (starting with @) and link shares. We deliberately kept the original text as much as possible because the certain hashtags and mentions tend to be related to veracity of the content, according to our training data analysis as well as previous studies. For instance, mention of WHO To understand if the pre-processing could be helpful for the generalization or increasing the performance on Tweets-19 , we added pre-processing step and ran an experiment with same RoBERTa-large setting. The preprocessing includes, replacing all links (e.g. http://caire.ust.hk) to be <LINK>, removing non-ascii characters including emoji, and replacing escaped text symbol into its original form (e.g. “&” into “&”) . We could achieve 34.62 and 34.37 for accuracy and W-F1 score respectively. The preprocessing step showed improved results in both scores, although it is marginal, for instance, 0.77% gain in accuracy and 0.72% gain in W-F1. This tells us that meticulous preprocessing steps can be useful to further improve the performance of the model.

Table 6: Results using different classifiers on top of fine-tuned RoBERTa-large. RoBERTa-large + NN refers to our best model listed on Table 2 and Table 3.

Models	FakeNews-19		Tweets-19	
	Acc.	W-F1	Acc.	W-F1
RoBERTa-large + NN	**98.13**	98.13	33.85	33.65
+ Logistic Regression	98.08	98.18	42.08	35.42
+ SVM	**98.13**	**98.22**	41.08	35.42
+ Gradient Boosting	98.04	98.14	**43.43**	**39.57**

7.4 Traditional classifiers instead of Neural Network

As reported in [25], the baseline SVM with the tf-idf feature extraction already achieves the somewhat satisfactory evaluation performance with accuracy of 93.46% and F1 score of 93.48%. Moreover, the features extracted by the fine-tuned RoBERTa are linearly separable as shown in the Figure 1, which suggests the potential of high performance with classical classifies. Therefore, we try three traditional classifiers with features extracted by the fine-tuned RoBERTa: logistic regression, SVM, and gradient boosting, which results are listed in Table 6. In general, these classifiers perform as good as the RoBERTa model on `FakeNews-19`, and even outperform on `Tweets-19`, especially for the gradient boosting. The results indicate that our NN classifier has too many parameters thus causing overfitting to `FakeNews-19`. In addition, adopting the traditional algorithms is preferable in terms of computational efficiency and model transparency while it does not harm the performance.

8 Conclusion

In this project, we explored COVID-19 fake news detection problem. We fine-tuning different pre-trained large language models methods with different robust loss functions and test on in-domain and out-domain dataset to see the generalization of the model. We figure out that training with robust loss function doesn't really help the model to generalize better and only achieve evaluation performance of 33.85% accuracy and 33.65% W-F1.

We further explored influence data cleansing technique and with 99% cleansing percentage, our model can produce the best evaluation performance on `Tweets-19`with 72.24% accuracy score and 65.88% W-F1 score while still maintaining high enough validation performance on `FakeNews-19`. We also analyze the the effect of preprocessing and different model classifiers to further boost the generalization ability of the model. For future work, we would like to combine the robust loss function with the influence score data cleansing method such that the resulting influence score can be made more robust for removing outlier data in multiple domains setting.

Contribution

All members contributed to equally on this project and the project was done through active discussions among team members. The task distribution details for each member is as following:

- Ye Jin Bang: Analyze `FakeNews-19`and `Tweets-19`dataset, conduct preprocessing ablation, conduct traditional classifier experiment, write report, prepare presentation content, and record presentation
- Etsuko Ishii: Prepare visualization for discussion part, analyze result on zero-shot experiment, conduct preprocessing ablation, conduct traditional classifier experiment, and write report
- Samuel Cahyawijaya: Analyze `FakeNews-19`and `Tweets-19`, implement code for influence-based cleansing, conduct zero-shot experiment, analyze result on zero-shot experiment, and write report
- Ziwei Ji: Implement code for robust loss functions, conduct robust loss function experiment, analyze result on robust loss experiment, write report, and prepare presentation content

References

[1] Managing the covid-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation, Sep 2020.

[2] Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms, 2020.

[3] J Scott Brennen, Felix Simon, Philip N Howard, and Rasmus Kleis Nielsen. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7:3–1, 2020.

[4] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[7] Satoshi Hara, Atsushi Nitanda, and Takanori Maehara. Data cleansing for models trained with sgd. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 4213–4222. Curran Associates, Inc., 2019.

[8] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2029–2037, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[9] Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection, 2020.

[11] Sosuke Kobayashi, Sho Yokoi, Jun Suzuki, and Kentaro Inui. Efficient estimation of influence of a training instance. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 41–47, Online, November 2020. Association for Computational Linguistics.

[12] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3), 2020.

[13] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[14] Yan Leng, Yujia Zhai, Shaojing Sun, Yifei Wu, Jordan Selzer, Sharon Strover, Julia Fensel, Alex Pentland, and Ying Ding. Analysis of misinformation during the covid-19 outbreak in china: cultural, social and political entanglements. *arXiv preprint arXiv:2005.10414*, 2020.

[15] Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. Jennifer for covid-19: An nlp-powered chatbot built for the people and by the people to combat misinformation. 2020.

[16] Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online, November 2020. Association for Computational Linguistics.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[18] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition, 2020.

[19] Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi. ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8105–8117, Online, July 2020. Association for Computational Linguistics.

[20] Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. *arXiv preprint arXiv:1905.10045*, 2019.

[21] Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. The adapter-bot: All-in-one controllable conversational model, 2020.

[22] Richard J Medford, Sameh N Saleh, Andrew Sumarsono, Trish M Perl, and Christoph U Lehmann. An" infodemic": Leveraging high-volume twitter data to understand public sentiment for the covid-19 outbreak. *medRxiv*, 2020.

[23] Areeb Mian and Shujhat Khan. Coronavirus: the spread of misinformation. *BMC medicine*, 18(1):1–2, 2020.

[24] Azzam Mourad, Ali Srour, Haidar Harmanani, Cathia Jenainatiy, and Mohamad Arafeh. Critical impact of social networks infodemic on defeating coronavirus covid-19 pandemic: Twitter-based study and research directions. *arXiv preprint arXiv:2005.08820*, 2020.

[25] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset, 2020.

[26] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.

[27] Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. An exploratory study of covid-19 misinformation on twitter, 2020.

[28] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*, 2020.

[29] Shashank Srivastava, Igor Labutov, and Tom Mitchell. Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[30] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. Caire-covid: A question answering and multi-document summarization system for covid-19 research. *arXiv preprint arXiv:2005.03975*, 2020.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[32] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, Alex D Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.

[33] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 322–330, 2019.

[34] Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung. Meta-transfer learning for code-switched speech recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3770–3776, Online, July 2020. Association for Computational Linguistics.

[35] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung. Learning fast adaptation on cross-accented speech recognition. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1276–1280. ISCA, 2020.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[37] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise, 2020.

[38] Jiaqi Xiong, Orly Lipsitz, Flora Nasri, Leanna MW Lui, Hartej Gill, Lee Phan, David Chen-Li, Michelle Iacobucci, Roger Ho, Amna Majeed, et al. Impact of covid-19 pandemic on mental health in the general population: A systematic review. *Journal of affective disorders*, 2020.

[39] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pages 8778–8788, 2018.