Fake-EmoReact 2021 Challenge

Luo He Zhou
sam8811288@gmail.com

Hsu Huan Yu
max870121@gmail.com

Kuo Jing Chen
ku0231@gmail.com

Abstract

This paper provide a method for Fake-EmoReact 2021 challenge. The task of this challenge is to predict the true and false of the given tweet data.We use the powerful language powerful pre-trained model **RoBERTa** as our backbone networks,then use the preprocessed data to fine-tune the model.To reduce domain bias,we leverage different classifier with ensemble methods.From the experiment,our RoBERTa-based model show positive effect on labeling tweet data.Besides,this approach achieve 0.997 F1-score in evaluate data and 0.701 F1-score in competition phase

1 Introduction

It is a common issue on Natural Language field to distinguish the true from the false due to natural language is often a indicative of one's emotion. In fact,the epidemic spread of fake news is a side effect of the expansion of social networks to circulate news, in contrast to traditional mass media such as newspapers, magazines, radio, and television. Human inefficiency to distinguish between true and false facts exposes fake news as a threat to logical truth, democracy, journalism, and credibility in government institutions.There are various tasks of identifying Fake News on social networks based on Natural Language Processing. Nicollas R. de Oliveira briefly introduce the models and tools in NLP which is applied in detecting fake news.With the evolution of computing ability,the deep neural networks network framework come to be realized,and a lot of powerful models have been released such as BERT,ELMO,GPT[Ming-Wei Chang,2018, Matthew Peters.2018, Alec Radford.2020].All of them use numerous data and resources to build the model.Thanks to these pretained weight,these models are fine-tuned for many tasks,which accelerates the development of the NLP field.

In Fake-EmoReact 2021,the challenge is to use text and reply of the tweet to distinguish if the tweet is true or fake.Tweet is a form of short message,containing any kinds of topics, latest information and culture around the world.According the feature of tweet,we propose an architecture to deal with this issue.Below section,we will introduce our methods to improve the classification result

- use random sample to deal with the data imbalance
- design the good preprocess methods on tweet data
- Fine-tune pre-trained model on the classification tasks

2 Related Work

Our work can mainly categorized into three parts,pre-trained models,classification,and ensemble methods

2.1 Pre-trained Models

Pre-trained models have been widely applied in a variety of NLP systems and achieve dramatically performance for downstream tasks. There are three major advantages for pre-trained models. First of all, since they are unsupervised learning, there will be unlimited corpus can be trained. Secondly, a strength pre-trained language model can generate deep contextual word representation which means a word token can have several representation in different sentences. Hence, through fine-tuning we improve downstream tasks more efficiently.Thirdly, using pre-trained models can reduce huge architecture engineering. This allows us don't need to design a deep learning network by our- selves and train with massive cost.

2.1.1 BERT

BERT,**B**idirectional **Encoder** **R**epresentations from **T**ransformers,is one of state-of-the-art pre-trained model released by Google researchers.There are two main tasks in pre-training stages.At the first task, called Masked LM (MLM), is to replace 15% of the words in each sequence to a [MASK] token and model need to predict these masked tokens. Encoder learns contextual representations during this stage. Second task, Next Sentence Prediction (NSP), the model takes pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original documents. In details, 50% of the inputs will be a pair in original documents in training, while the other 50% a random sentence from the corpus is chosen as the second sentence.

2.1.2 RoBERTa

RoBERTa,**R**obustly **o**ptimized **BERT** **a**pproach is a new training recipe that improves on BERT.The authors of the model found that BERT is undertrained,so they improved the training process of the original BERT models,modify key hyperparameters to make it better.After ameliorated,the main differences between BERT and RoBERTa are showed below

- Use dynamic masking instead of static masking,which avoid using the same mask for every epoch.Therefore,the model sees different versions of the same sentence with masks on different positions
- Remove NSP loss.This change can slightly improve downstream tasks
- Use bytes instead of unicode as the base subword units,which makes the model learn larger subword vocabulary

These modifications help RoBERTa outperform BERT in most of the downstream tasks

2.2 Ensemble Methods

Traditional learning algorithms that only output one hypothesis may lead three major problems

- **Statistical**:The Statistical Problem arises when the hypothesis space is too large for the amount of available data. Hence, there are many hypotheses with the same accuracy on the data and the learning algorithm chooses only one of them. There is a risk that the accuracy of the chosen hypothesis is low on unseen data.
- **Representational**:Representational Problem arises when the hypothesis space does not contain any good approximation of the target class(es).
- **Computational**:The Computational Problem arises when the learning algorithm cannot guarantees finding the best hypothesis.

Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. Basic idea is to construct a set of classifiers and then classify new data points by taking a vote of their predictions which could usually address the three problems just mentioned.There is some reason why ensemble help dealing those issues.First,By constructing an ensemble out of all of these accurate classiers the algorithm can average their votes and reduce the risk of choosing the wrong classier.Second,in most applications of machine learning, the true function cannot be represented by any of the hypotheses.By forming weighted sums of hypotheses,it maybe possible to expand the space of representable functions.Third,many learning algorithms implement local search which may stop even if the best solution found by the algorithm is not optimal. An ensemble constructed by running the local search from many different starting points may provide a better approximation to the true unknown function than any of the individual classifiers.

3 Dataset

The task contain 168,521 training Twitter tweet.Each tweet contain 6 columns which are **idx**, **text**,**reply**,**categories**,**mp4**,**label** individually.
The following items are the explanation of 5 columns

- idx:identifiable label of the tweet
- text:the text content of the tweet
- reply: the text content of the tweet response
- categories: the categories of the response GIF,which totally have 43 categories

- mp4: response GIF's hash file name
- label: True or False of the tweet

In addition to the training data,there are also other two JSON files **dev.json** and **eval.json**,which is validation data and test data,respectively.The difference between train data,validation data,and test data is the former consists of 6 columns while the latter two only consist 5 columns which miss the label attribute.
After doing a simple statistics,we can found there is a severe data imbalance problem as shown in Table 1,Figure 1.

label	count
True	136722
False	31799

Table 1: Label Distribution

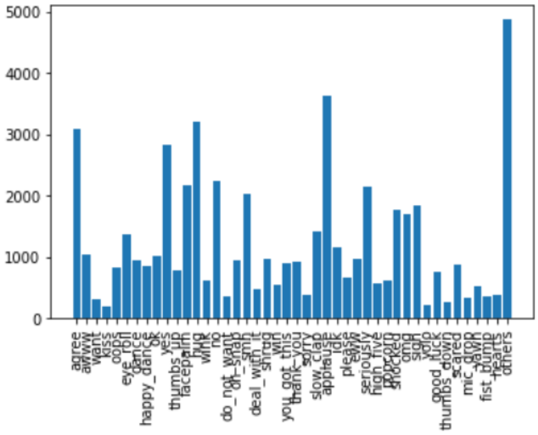

Figure 1: Categories Distribution

4 Method

The method of this research composed of 3 part.The first part is preprocessing the data. The second part is to pretrain a language model. The third part is using the pretrained weight to train the classifier. The overview of the method is shown in figure 1.

4.1 Preprocessing the tweet

Tweet data have it's special structure. It is a social media and people would use emoji on it and may used tweet style abbreviations. In order to transform the original tweet data to model readable data we do the following steps:

- Convert to the lower case.

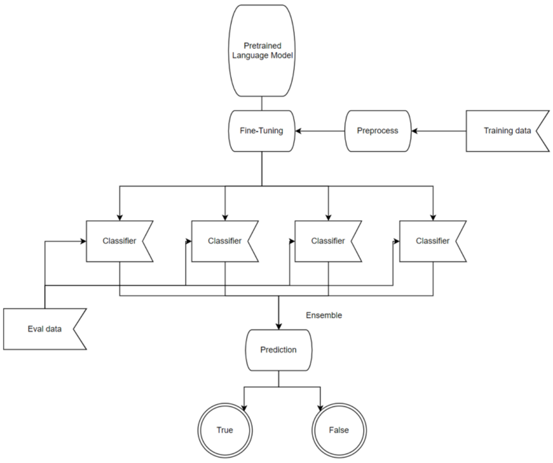

Figure 2: Overview of method

- Transform weird punctuation for example """ and "'".
- Convert abbreviations to normal word, such as "ain't" to "are not", "you've" to you have.
- Demojize: convert the emoji to its meaning like happy, sad and other word describing the emoji.
- Convert company names trending word to its category such as "Covid-19" to virus, and "Netflix" to streaming service

Then, the text and reply of the tweet is concatenate to form the final data as following:

$$['<s>'] + text + ['</s></s>'] + reply + ['<s>']$$

The ['<s>'] and [' </s></s>'] is the token used by the RoBERTa model which indicate the start and end of a sentence.

4.2 Pre-training the language model

The tweet data may has it special corpus other than the formal corpus like Wikipedia. Therefore, if we used the pre-trained model trained by formal corpus may present domain bias. Therefore, we used the 168,521 training Twitter tweet to fine-tune the pre-train the language model. The pre-trained language model understood more about the tweet data. We fine-tune the pretrained RoBERTa model provided by the Simple Transformers, a python Natural Language Processing library established by Rajapakse et al,

with the training data-set.

4.3 Training the Classification and ensemble model

In Fake-EmoReact 2021 Challenge, every tweet can be classified into 2 labels, fake or real. We defined the fake label as 0 and real label as 1. After fine-tune the pre-training RoBERTa model, we used the training data-set with the label to train the classification model. As mentioned in Section 3. Data, the the data-set is severely imbalance. In order to solve the data imbalanced problem, we sample data from fake news. The amount of sampled fake tweet matches the total number of real tweet. The sampled fake tweet and all real tweet formed a sub-training data-set. We sampled data from the fake tweet 4 times to form 4 different sub-training data-set. Then, we used 4 different sub-training data-set to train 4 different classification model. The classification model output the probability of each label class. The output of 4 classification model is summed. If the summed value of real is larger than the summed value of fake, the output is real and vice versa. Equation (1) describe the above method.

$$argmax(\sum_{n=1}^{4}[real\_p_n, fake\_p_n]) \qquad (1)$$

5 Experiment

5.1 Experimental Setup

For the pre-trained language model and the classification model we used Adam (**?**) with epsilon = 1e-8 and learning rate = 4e-5 to train the model. The batch size is set to 16 and max sequence length is set to 113. The block size is set to 512. All model is trained for 4 epochs. The model was trained using a graphics processing unit (RTX 2080 Ti, NVidia, Santa Clara, CA, USA), for about 4 hours.

5.2 Evaluation and Test Results

model	Precision	Recall	F1-score
RoBERTa	0.949	0.982	0.964
RoBERTa (pretrain)	0.986	0.994	0.990
RoBERTa (pretrain+ensemble)	0.994	0.996	0.997

Table 2: Validation result

The Validation result is shown in Table 2. If we do not pre-train the language model we can get F1-score=0.964. If we used the pre-train model the F1-score raised to 0.990. If we combine all the method we used including the pretrain and the ensemble. The F1-score can further improved to 0.997. It proves the method we proposed can really improve the performance of the model.

model	Precision	Recall	F1-score
proposed model	0.7480	0.7142	0.7016

Table 3: Test result

Table 3. shows the test result of our proposed model the F1-score is 0.7016 which is much lower than the validation result. The possible explanation of it may include different time domain, and different distribution of data. The time domain of the training and the evaluation data-set may be similar, but the testing data-set may be different. The time domain may affect the context of the data, for example in 2016 lots of fake tweets contain Donald Trump, but in 2010 almost none of the tweets contain Donald Trump. We got the 10^{th} place at the final test result. The following are some improvement we may make:

- Replace url in the tweet
- Combine Training, validation and testing dataset to pretrain the language model
- Ensemble with more model

6 Conclusion

In this work, we propose an an system architecture combining with preprocessing, model framework,and ensemble models for Fake-EmoReact task. We use the preprocessing for tweets,fine-tune model to handle the specific task in this challenge and increase the coverage of words recognized by tokenizer. Based on preprocessing data, We apply pre-trained model and ensemble models with power weighted sum.

Besides,We observe that the tweet with categories always be true. However, we are not sure whether it is a data leakage,we don't deal with it but only concatenate the reply next to text.Furthermore,due to the computing resources and time limit,we can't ensemble with various model used in NLP field.However,we are planning to ensemble with BERT-cased and BERT-uncased if there is a chance.According to Emran Al-Bashabsheh,we assume that is would better our model.

7 Question and Answer

Question 1. Do you compare the change of the model performance when transform tweet word to its' Category, for example "covid" to "virus" ? No, we don't compare it, because these word may not include in the token of RoBERTa model, but we believe it may affect the result.
Related work can be referenced in A Natural Language Processing Approach

References

[2] Jose Camacho-Collados, Mohammad Taher Pilehvar *On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis*. Addison-Wesley, Reading, Massachusetts, 1993.

[3] *Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019* [*a distilled version of bert: smaller, faster, cheaper and lighter.*]. arXiv preprint arXiv:1910.01108..

[4] Preprocessing Techniques for Text Mining
https://www.researchgate.net/publication/273127322$_{P}reprocessing_{T}echniques_{f}or_{T}ext_{M}ining$

[5] Predicting Fake News using NLP and Machine Learning — Scikit-Learn — GloVe — Keras — LSTM
https://towardsdatascience.com/predicting-fake-news-using-nlp-and-machine-learning-sci

[6] 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, pages 5753–5763.

[7] Zhongwu Zhai, Hua Xu, Bada Kang, and Peifa Jia.2011. *Exploiting effective features for chinese sentiment classification.*.Expert Systems with Applications, 38(8):9139–9146.

[8] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. *Distilbert, a distilled version of bert:smaller,faster,cheaper and lighter.*.Expert Systems with Applications,38(8):9139–9146.

[9] Mohammed Jabreel and Antonio Moreno. 2016. *Sentirich: Sentiment analysis of tweets based on a rich set of features.*.In CCIA, pages 137–146.

8 Task Allocation

Student ID	Name	Tasks
0616316	Kuo Jing Chen	Data preprocessing, Model training
0816166	Luo He Zhou	Data preprocessing, Model training, Architectural design
309264009	Hsu Huan Yu	Data preprocessing,Model training,Architectural design