Fine-tuning technique and data augmentation on transformer-based models for conversational texts and noisy user-generational content

Mike Tian-Jian Jiang  
Zeals Co, Ltd.  
Tokyo, Japan  
tmjiang@gmail.com

Shih-Hung Wu  
Department of CSIE  
Chaoyang University of Technology  
Taichung, Taiwan  
shwu@cyut.edu.tw

Yi-Kun Chen  
Department of CSIE  
Chaoyang University of Technology  
Taichung, Taiwan  
kun26712930@gmail.com

Zhao-Xian Gu  
Information Management  
Tamkang University  
New Taipei City, Taiwan  
406630136@s06.tku.edu.tw

Cheng-Jhe Chiang  
Information Management  
Tamkang University  
New Taipei City, Taiwan  
406630649@s06.tku.edu.tw

Yueh-Chia Wu  
Information Management  
Tamkang University  
New Taipei City, Taiwan  
406630102@s06.tku.edu.tw

Yu-Chen Huang  
Information Management  
Tamkang University  
New Taipei City, Taiwan  
406630193@s06.tku.edu.tw

Cheng-Han Chiu  
Information Management  
Tamkang University  
New Taipei City, Taiwan  
406630284@s06.tku.edu.tw

Sheng-Ru Shaw  
Information Management  
Tamkang University  
New Taipei City, Taiwan  
poppydumb1220@gmail.com

Min-Yuh Day†  
Information Management  
National Taipei University  
New Taipei City, Taiwan  
myday@gm.ntpu.edu.tw

Abstract—Transfer learning and transformer language models play important role in modern natural language processing research community. In this paper, we proposed fine-tuning technique and data augmentation (TMFTDA) on transformer-based models for conversational texts and noisy user-generational content. We used two NTCIR-15 sub-tasks, namely, dialogue evaluation (DialEval-1) and numeral attachment in financial Tweets (FinNum-2), to evaluate the efficacy of the proposed fine-tuning technique and data augmentation with transfer learning on Transformer-based models. Experimental results show that the proposed TMFTDA approach substantially outperform the baselines model of Bidirectional Long Short-Term Memory (Bi-LSTM) in multi-turn dialogue system evaluation at the NTCIR-15 DialEval-1 dialogue quality (DQ) and nugget detection (ND). Moreover, the proposed approaches with tokenization and fine-tuning techniques on the Bidirectional Encoder Representations from Transformers (BERT) and XLM-RoBERTa models perform satisfactory at NTCIR-15 FinNum-2. The research contribution of this paper is that we shed light on the transfer learning with fine-tuning technique and data augmentation on transformer-based models for conversational texts and noisy user-generational content in social media text analytics.

Keywords—conversational texts, data augmentation, fine-tuning technique, noisy user-generational content, transfer learning, Transformer-based models

I. INTRODUCTION

Transfer learning and Transformer-based language models play important role in modern natural language processing research community. Because of recent advances in natural language processing, more and more researchers and engineers are developing task-oriented dialogue systems. Customer services may benefit from such a chat-bot that responses to inquiries 24/7. However, assessing such systems often involves a costly and labor-intensive annotation process that defeats the purpose. The dilemma motivates the task organizers of NTCIR-14 STC-3 [38] and NTCIR-15 DialEval-1 [39] to come up with Dialogue Quality (DQ) and Nugget Detection (ND) subtasks that examine automatic evaluation systems for helpdesk conversations in Chinese or English.

The DQ subtask uses subjective scales to quantify the quality of a whole dialogue. With 5-degree of rank each sorting from -2 to 2, the organizers define 3 score types:

1. A-score: Accomplishment  
   —to what extent has an inquiry resolved;
2. S-score: Satisfaction  
   —how assured a customer is with the conversation;
3. E-score: Effectiveness  
   —how helpful and economical a dialogue is.
The ND subtask first defines a nugget as a dialogue turn, determines whether it belongs to Customer side or Helpdesk side, and finally categorizes it into seven types of four groups:

1. CNaN / HNaN: Customer or Helpdesk’s non-nuggets that are irrelevant to the problem-solving situation;
2. CNUG / HNUG: Customer or Helpdesk’s regular nuggets that are relevant to the problem-solving situation;
3. CNUG* / HNUG*: Customer or Helpdesk’s goal nuggets that confirm and provide solutions, respectively;
4. CNUG0: Customer’s trigger nuggets that initiate a dialogue with certain problem descriptions.

Based on the above specifications, we formulate the DQ and the ND subtasks as a multilabel classification problem and a multiclass classification problem, respectively. Since STC-3 participants didn’t outperform the baselines model of Bidirectional Long Short-Term Memory (Bi-LSTM) [3,4,14,34], we take on the challenge to discover another strong baseline. To alleviate the high cost of architecture engineering and model training, our study pays more attention to tokenization and optimization for transfer learning. We apply well-established techniques of tokenization and fine-tuning to pretrained Transformer models. We find that some specific combinations of techniques work well with XLM-RoBERTa [5] and certain variations of the Bidirectional Encoder Representations from Transformers (BERT) [7], for English and Chinese, respectively.

II. RELATED WORKS

A. Dialogue system evaluation (DialEval-1)

Following the third Short-Text Conversation (STC-3) task at NTCIR-14, the first Dialogue Evaluation (DialEval-1) task continue examining, for Chinese and English, how well each participant’s system can tackle the two subtasks of Dialogue Quality (DQ) and Nugget Detection (ND). The former estimates the three quality scores of a dialogue, namely Accomplishment (A-score), Satisfaction (S-score), and Effectiveness (E-score), using integer ranks ranging from -2 to 2 each. The latter categorizes dialogue turns by seven nugget types. For DQ subtask, the task organizers measure performance by Normalised Match Distance (NMD) and Root Symmetric Normalised, Order-aware Divergence (RSNOD). For ND subtask, the metrics are Root Normalised Sum of Squares (RNSS) and Jensen-Shannon Divergence (JSD). We consider both subtasks classification problems and tackle them with several models of Transformer, to create a reliable and efficient process using the most recent advances of transfer learning. Our approaches involve various techniques of tokenization and fine-tuning for those Transformers. This paper describes their usages and usefulness of our official runs. In terms of NMD, our run2 for Chinese DQ subtask substantially outperforms the baselines. According to RSNOD, our run0 for English DQ subtask also achieve a significant difference of S-score statistically. Almost all of our runs for ND tasks reach the first places. NTCIR-15 DialEval-1 task. Those results suggest that one can easily optimize Transformers for DQ and ND subtasks.

In the past, researchers have relied on human to judge the quality of a dialogue system [1]. To overcome the inefficiency and the inconsistency of man-made assessments for spoken dialogue agents, one of the earliest works on learning an automatic evaluation function called PARADISE isolates task requirements from an agent’s conversational behavior, at the cost of measurable completeness and complexity of the task [31]. Since the measurement are not always available, instead a recent model called ADEM seeks to learn and predict the appropriateness of utterances [25]. ADEM and its successors keep evolving to adopt one new model by another, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) [9], and now BERT. It is then conceivable that many STC-3 participants have used LSTM or BERT. As one may argue that Bi-LSTM usually outperforms other architectures [8], STC-3 outcomes also suggest the bar set by a model of Bi-LSTM and GloVe [26] is uneasy to meet.

Despite the architecture differences, almost all of them have modeled the ND and DQ subtasks as classification problems. We adopt the same tactic for DialEval-1, such that our efforts may focus on developing a recipe of transfer learning that comprises the state-of-the-art ingredients. For that matter, we look into various works of transfer learning, especially on optimization algorithms and loss functions. Layer-wise Adaptive Rate Scaling (LARS) [36] aims to implicitly adapt various learning rates for different layers of convolutional networks with large batches, and soon spawns a version called LAMB [37] for BERT training. As the name suggests, however, they are designed for relatively big size of batches for the efficiency of pretraining, we fail to find significant improvements using them for fine-tuning. The fact that we’re already using discriminative fine-tuning may further complicate the behavior of convergence.

Another perspective on taming the behavior of convergence is about stabilizing gradient updates. Lookahead [40], Rectified Adam [20], and Gradient Centralization [35] fall into this category. Ranger further combines them together as one optimizer. Again, based on our pre-trials for the DQ and ND subtasks, they are neither faster nor stabler.

Last but not least, if we see the tokenization tricks as feature engineering for deep neural networks, whilst being seldom used for text classification and fine-tuning, it is a common approach for text generation and pretrained. CTRL [15] and GPT-3 [2] have many designated “prompts” that enable conditioned generations. Feature engineering done in such a preprocessing manner may be easier for adapting different tasks or pretrained models than specialized embeddings.

B. Numeral attachment in financial Tweets (FinNum-2)

FinNum-2 is a shared task to analyze financial tweets, these tweets are discussing stock prices and the companies [5]. There are many stock names and many numbers in these tweets. The goal of the shared task is to uncover whether names and numbers in a tweet associates or not. In the training data, we can see that there is at least one pair of target numeral and cashtag in a tweet, therefore the problem definition can be a binary classification to tell if the target numeral is relevant to the given cashtag.
For Chinese DQ and ND subtasks, we test the official one experimentation, all pretrained models are the base versions. BERT, RoBERTa, and XLM-A.

multiclass classifications of the procedures that are conceptually related to multilabel and learning.

fastai models available on HuggingFace’s content.

models for conversational texts and noisy user tuning technique and data augmentation on transformer models, we also apply several finetuning models pretrained

inspires us to further Research (BERT)

Bidirectional Encoder Representations from Transformer

word/character embeddings to represent token information of DQ and ND
texts and noisy user
tuning technique and data augmentation on transformer

models, we also introduce model specifications and training

different datasets and tokenization

tokenization techniques of learning

rate and optimizer.

III. PROPOSED METHODS

Figure 1 shows the proposed research framework of finetuning technique and data augmentation on transformer-based models for conversational texts and noisy user-generational content. Firstly, we establish our tool-chain. To go through the trial-and-error phase as quick as possible, we only try pretrained models available on HuggingFace’s Transformers [32], and use fastai [10,11] to control the quality and the speed of transfer learning. We introduce model specifications and training procedures that are conceptually related to multilabel and multiclass classifications of the dialogue evaluation (DialEval-1) DQ and ND as well as numeral attachment in financial Tweets (FinNum-2) subtasks.

A. Transformer-based Models Selection for Transfer Learning

We conduct transfer learning by fine-tuning pretrained BERT, RoBERTa, and XLM-RoBERTa models for text sequence classification. To meet our goal of rapid experiments, all pretrained models are the base versions. For Chinese DQ and ND subtasks, we test the official one (denoted as bert-chinese when necessary) and a whole-word masking version (bert-chinese-wwm) [6] of BERT. The official XLM-RoBERTa model (xlm-roberta) runs for both Chinese and English. Finally, the runs of the official RoBERTa model (roberta) [21] and the case-reserved BERT (bert-cased), are merely control groups for the English ND subtask. The principle behind the choices is simple: they cover representative differences of the pretraining scheme and the token specification.

BERT by default tokenizes each input sequence using WordPiece [33]. Its pretraining typically relies on two objectives: masked language modeling (MLM) and next sentence prediction (NSP). The former requires the model to predict tokens that have been randomly masked in a 15% chance per input sentence, and the latter demands the model to predict whether two randomly concatenated sentences are actually adjacent to each other or not. XLM-RoBERTa, on the other hand, combines and revises techniques of cross-lingual language model (a.k.a. XLM) pretraining schemes [19] and a robustly optimized BERT pretraining approach (a.k.a. RoBERTa). In terms of optimization, RoBERTa builds on BERT and modifies key hyperparameters such as the MLM objectives, removing the NSP objective and training with much larger mini-batches and learning rates. As for tokenization, it differs from BERT by using a byte-level Byte Pair Encoding (BPE) [28] as a tokenizer, and dynamically changing the masking pattern applied to the training data. XLM-RoBERTa follows most of XLM approaches, except it removes language embeddings for a better code-switching ability. It also differs from RoBERTa by tokenizing with unigram-level sentencepiece [17,18] instead of BPE.

B. Tokenization Tricks for Transformer-based Models

To better represent the structure of a dialogue, using XLM-RoBERTa’s markups as example, we not only utilize special tokens for the beginning of a sentence (</s>), the end of a sentence (</s>), and the separator of sentences (</s> </s>), but also customize a couple of tokens in the fastai convention of “xx” prefix that provides context. For example, consider a tokenized turn below:

xxlen _3 </s> xxstrn _1 xxsdrr _customer _@ _China _Uni com _Customer _Service in _Gu ang dong … _Middle _Road . </s>

The special tokens xxlen and xxstrn stand for length of the dialogue in turns and the position of each turn of the dialogue, respectively. The numbers right next to them provide certain features of turns. The same trick goes with xxsdrr that differentiates whether the sender is Customer or Helpdesk. When a turn’s context says “xxstrn _1 xxsdrr _customer”, the nugget type is almost definitely CNUG0. As for DQ, a whole dialogue can be tokenized in a similar fashion, where xxlen could be useful for certain quality scores, should it be about the time.turns spent on resolving a problem:

xxlen _3 </s> xxstrn _1 xxsdrr _customer _@ _China _Uni com _Customer _Service in _Gu ang dong … _Middle _Road . </s>/s>

xxsdrr _help desk _Hello ! … _Thank you ! </s>/s> xxstrn _3 xxsdrr _customer _The _Uni com … _No _phone _call _is _answered ! </s>
Although we don’t apply the default tokenizer of fastai, it might be worthwhile to explain what it is and why we don’t use it. The fastai convention of “xx” prefix denotes special context tokens. By default, fastai tokenizes English texts using SpaCy and inserts special tokens before uncapitalized or originally repeated words/characters. For instance, consider the following utterance from the test set:

… Beijing Unicom Unicom still …

If we apply fastai’s default tokenization to it, the outcome will have “Unicom Unicom” converted into “xxwrep 2 xxmaj unicm” for title case and word duplication simultaneously. As lossless as the conversion may be, since pretrained Transformer models are unaware of those special context tokens, we must ask whether they can still help fine-tuning for a specific task or not. In our opinions, if the task were sentiment analysis of utterance, repetitions and capitalization could be important clues. However, it is hard to imagine that the recurring word/character can help semantically or syntactically, not to mention that XLM-RoBERTa already preserves letters of subword tokens. Based on the above observations, we don’t apply them for the DialEval-1 task.

C. Fine-tuning Techniques on Downstream Tasks

We adopt recently advanced fine-tuning techniques as much as possible. Some of them are originally designed for AWD-LSTM and QRNN [22,23] by ULMFiT, such that we must assess their usefulness for XLM-RoBERTa. Based on our preliminary tests, discriminative fine-tuning and fastai’s version of one-cycle policy work well, but graduate unfreezing produces little effect, which is consistent with the findings of similar studies [13,27]. Techniques other than the above mainly involve choosing the most promising combination of optimization algorithms and loss functions. For the FinNum-2 task in a binary classification setting, we find none of more recent optimizers and loss functions work better than Adam optimizer with class weights. We will list configuration values of finally used techniques in the next section of experiments. The section of related works will briefly describe what optimizers and loss functions we have evaluated.

1) Discriminative Fine-tuning

As different layers may capture various types of information, we shall fine-tune them to different extents. Instead of using the same learning rate for all layers of the model, discriminative fine-tuning enables us to tune each layer with different learning rates. We use blurr to split the model layers into groups automatically corresponding to architectures. For both BERT and XLM-RoBERTa, it results in four groups: the top layer of classifier, the pooling layer, the Transformer layers, and the bottom layer of embeddings. Intuitively, the lower groups may contain more general information while the higher ones contain more specific information. Therefore, we set a base learning rate for the top group and then assign linearly decreased learning rates per lower groups.

2) One-cycle Policy

A cycle wraps an arbitrary number of epochs for sharing the same policy of hyperparameters, especially for learning rates and momentums. For training a deep neural network with stochastic gradient decent or similar algorithms, a policy of cyclical learning rates, meaning it periodically increases for a step size and then decreases the learning rates, may converge faster and better [29,30]. In addition, the fastai version of the One-cycle Policy comprises three complementary techniques that balance the trade-off between fast convergence and overshooting. The Slanted Triangular Learning Rates (STLR) [12] and the Cyclical Momentum [29,30] allow us to micro-manage iterations/updates within a cycle, whereas changing maximum learning rate (max_lr) per cycle let us control the quality of each. Empirically, STLR and cyclical momentum together work best when they simultaneously change in a reversed direction. SLTR uses a warm-up and annealing for the learning rate while doing the opposite with the momentum. On the other hand, it indicates that we apply a simply decay on max_lr per cycle.

3) Other Optimization Schemes

We test several optimizers and find none of them improve the convergence stability significantly than Adam [16]. For the choice of loss function, we realize that the label smoothing function [24] suits multilabel/multiclass classification better than typical cross-entropy one.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

We present our experimental results on NTCIR-15 DialEval-1 Dialogue Quality (DQ) and Nugget Detection (ND) subtasks that examine automatic evaluation systems for helpdesk conversations in both Chinese and English and Numeral attachment in financial Tweets (FinNum-2).

A. Dialogue Quality (DQ) and Nugget Detection (ND) at DialEval-1

Table 1 shows the mapping between our official runs, the designated models, the batch sizes (B), and the recipes of hyperparameters. Important hyperparameters include the cycle schemes and their max_lr’s of discriminative learning rates, while they share the same reduction rate: the lower bound is always max_lr/1000, and every cycle contains just one epoch.

<table>
<thead>
<tr>
<th>Task</th>
<th>Lang</th>
<th>Run</th>
<th>Model</th>
<th>B.</th>
<th>Recipe</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ</td>
<td>en</td>
<td>0</td>
<td>xlm-roberta</td>
<td>12</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>zh</td>
<td>0</td>
<td>xlm-roberta</td>
<td>1</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bert-chinese-wwm</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td>en</td>
<td>0</td>
<td>xlm-roberta</td>
<td>12</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bert-cased</td>
<td>8</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>roberta</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>zh</td>
<td>0</td>
<td>xlm-roberta</td>
<td>12</td>
<td>e</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bert-chinese-wwm</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bert-chinese</td>
<td>16</td>
<td>d</td>
</tr>
</tbody>
</table>

Table 1. Configurations of DialEval-1 Official Runs
The factor of 1000 hints that we hope the four layer-groups may roughly have the rates distributed evenly. However, it comes to our attention that, after the timing of the official runs, the version 3.3.0 and above of HuggingFace’s Transformers has removed the pooling layer, because in theory they are unrelated to classification. Should any reader want to reproduce the outcome, please be advised that it will definitely vary if using different versions.

Table 2 to 9 compare our runs with the best baseline per task-language-metrics. Table 2 shows the Chinese Nugget Detection Results of IMTKU official runs. Table 3 shows the English Nugget Detection Results of IMTKU official runs. In the ND subtask for both Chinese and English, corresponding run0 results of XLM-RoBERTa are only slightly better than the LSTM baselines. For that matter, we closely examine the outcomes and then notice intriguing phenomenon, such as

“Are you from a security software manufacturer?”

and

“Do you think if it would be better for me to complain to the Ministry of Industry and Information Technology?”

of IDs 4245108926487325 and 4392549047578258, respectively. The types of turns like the above examples are mostly C NaN, but the models predict them as CNUG. We anticipate that the word "you" has caused confusions. The models might have taken it literally for Customer replying to Helpdesk, but the turns and similar are likely sarcasm hence unrelated to the problem-solving situation.

For the DQ subtask, we manually compare the differences among models for different runs. Table 4, 5, and 6 present the A-score, E-score, and S-score of Chinese Dialogue Quality results of IMTKU official runs. Table 7, 8, and 9 present the A-score, E-score, and S-score of Chinese Dialogue Quality results of IMTKU official runs. Although the Chinese versions of BERT outperform XLM-RoBERTa, they all share the same recipe of cycle schemes. In addition, since we know that the English datasets are translations of the Chinese ones, it is as expected that XLM-RoBERTa seems equally competitive for both languages.

**B. Numerical attachment in financial Tweets at FinNum-2**

For financial tweets analysis, we proposed BERT-FN-PS, which is based on the BERT model with proposed preprocessing strategy. We also proposed XLM-RoBERTa-FN-FTT, which is fine-tuned system based on the XLM-RoBERTa pretraining model with more tokenization and fine-tuning techniques.

The macro-F1 of the proposed XLM-RoBERTa-FN-FTT model is 95.99% on development set, and 71.90% on formal test which ranked second best in NTCIR-15 FinNum-2.

We fine-tune two Transformer-based models, namely BERT and XLM-RoBERTa. We proposed XLM-RoBERTa-FN-FTT model, which further apply techniques developed by fastai for ULMFiT, such as discriminative fine-tuning and a variation of one-cycle policy.

The tokenization trick we used in XLM-RoBERTa-FN-FTT model play important role in text data preprocessing. To better represent the structure of a financial tweet, we not only utilize the XLM-RoBERTa’s special tokens, namely the beginning of a sentence (<s>), the end of a sentence </s>, and the separator of sentences </s> </s>, but also customize a couple of tokens in the fastai convention of “<<” prefix that provides context. For example, consider a tokenized tweet below:

<s> $ xxtag_ _RAD_ _about xxnum_ _9 _million _more _share s _than _the _90 _day _average _... </s>

The special tokens xxnum and xxtag annotate the numeral (_9 but not _90) and the cashtag (_RAD) in question, respectively. Combining with the actual subwords of

<table>
<thead>
<tr>
<th>Run</th>
<th>JSD</th>
<th>Run</th>
<th>RNSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run0</td>
<td>0.0674</td>
<td>IMTKU-run0</td>
<td>0.1636</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.0709</td>
<td>BL-lstm</td>
<td>0.1673</td>
</tr>
<tr>
<td>IMTKU-run1</td>
<td>0.0726</td>
<td>IMTKU-run1</td>
<td>0.1700</td>
</tr>
<tr>
<td>IMTKU-run2</td>
<td>0.0752</td>
<td>IMTKU-run2</td>
<td>0.1754</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>JSD</th>
<th>Run</th>
<th>RNSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run0</td>
<td>0.0707</td>
<td>IMTKU-run0</td>
<td>0.1699</td>
</tr>
<tr>
<td>IMTKU-run2</td>
<td>0.0757</td>
<td>IMTKU-run2</td>
<td>0.1753</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.0762</td>
<td>BL-lstm</td>
<td>0.1781</td>
</tr>
<tr>
<td>IMTKU-run1</td>
<td>0.0789</td>
<td>IMTKU-run1</td>
<td>0.1804</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>RSNOD</th>
<th>Run</th>
<th>NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run0</td>
<td>0.2197</td>
<td>IMTKU-run0</td>
<td>0.1437</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.2271</td>
<td>BL-lstm</td>
<td>0.1591</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>RSNOD</th>
<th>Run</th>
<th>NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run0</td>
<td>0.1657</td>
<td>IMTKU-run0</td>
<td>0.1221</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.1687</td>
<td>BL-lstm</td>
<td>0.1248</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>RSNOD</th>
<th>Run</th>
<th>NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run0</td>
<td>0.1892</td>
<td>IMTKU-run0</td>
<td>0.1250</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.2111</td>
<td>BL-lstm</td>
<td>0.1413</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>RSNOD</th>
<th>Run</th>
<th>NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run2</td>
<td>0.2130</td>
<td>IMTKU-run2</td>
<td>0.1392</td>
</tr>
<tr>
<td>IMTKU-run0</td>
<td>0.2163</td>
<td>IMTKU-run0</td>
<td>0.1406</td>
</tr>
<tr>
<td>IMTKU-run1</td>
<td>0.2204</td>
<td>IMTKU-run1</td>
<td>0.1442</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.2305</td>
<td>BL-lstm</td>
<td>0.1598</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>RSNOD</th>
<th>Run</th>
<th>NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run1</td>
<td>0.1631</td>
<td>IMTKU-run1</td>
<td>0.1165</td>
</tr>
<tr>
<td>IMTKU-run0</td>
<td>0.1648</td>
<td>IMTKU-run0</td>
<td>0.1181</td>
</tr>
<tr>
<td>IMTKU-run2</td>
<td>0.1655</td>
<td>IMTKU-run2</td>
<td>0.1194</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.1782</td>
<td>BL-lstm</td>
<td>0.1386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>RSNOD</th>
<th>Run</th>
<th>NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMTKU-run2</td>
<td>0.1918</td>
<td>IMTKU-run2</td>
<td>0.1254</td>
</tr>
<tr>
<td>IMTKU-run1</td>
<td>0.1964</td>
<td>IMTKU-run1</td>
<td>0.1284</td>
</tr>
<tr>
<td>IMTKU-run0</td>
<td>0.1977</td>
<td>IMTKU-run0</td>
<td>0.1290</td>
</tr>
<tr>
<td>BL-lstm</td>
<td>0.2088</td>
<td>BL-popularity</td>
<td>0.1442</td>
</tr>
</tbody>
</table>
In this paper, we have proposed fine-tuning technique and data augmentation (TMFTDA) on transformer-based models for conversational texts and noisy user-generational content. We used two NTCIR-15 sub-tasks, namely, dialogue evaluation (DialEval-1) and numeral attachment in financial Tweets (FinNum-2), to evaluate the efficacy of the proposed fine-tuning technique and data augmentation with transfer learning on Transformer-based models.

Experimental results show that the proposed TMFTDA approaches substantially outperform the baselines model of Bidirectional Long Short-Term Memory (Bi-LSTM) in multi-turn dialogue system evaluation at the NTCIR-15 DialEval-1 dialogue quality (DQ) and nugget detection (ND). Moreover, the proposed approaches with tokenization and fine-tuning techniques on the Bidirectional Encoder Representations from Transformers (BERT) and XLM-RoBERTa models perform satisfactorily at NTCIR-15 FinNum-2.

The research contribution of this paper is that we shed light on the transfer learning with fine-tuning technique and data augmentation on transformer-based models for conversational texts and noisy user-generational content in social media text analytics.

ACKNOWLEDGMENT

This research was supported in part of Tamkang University (TKU) research grant, National Taipei University (NTPU), Zeals Co., Ltd., and Ministry of Science and Technology (MOST) under the grant number MOST 109-2221-E-324-024.

REFERENCE


