Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 2, April 2021, pp. 988~998
ISSN: 2302-9285, DOI: 10.11591/eei.v10i2.2745

The COVID-19 fake news detection in Thai social texts

Pakpoom Mookdarsanit[1], Lawankorn Mookdarsanit[2]
[1]Faculty of Science, Chandrakasem Rajabhat University, Bangkok, Thailand
[2]Faculty of Management Science, Chandrakasem Rajabhat University, Bangkok, Thailand

Article Info

Article history:

Received Sep 22, 2020
Revised Dec 16, 2020
Accepted Jan 4, 2021

Keywords:

COVID-19 misinformation
Fake news detection
Feature shifting
Thai fake news filtering
Transfer learning

ABSTRACT

One important obstruction against Thai COVID-19 recovery is fake news shared on social media that is one of the “Artificial Intelligence Open Issues against COVID-19” reported by Montreal.AI. Misinformation spread is one of the main cyber-security threats that should be filtered out as the IDS for maintaining COVID-19 information quality. To detect fake news in Thai texts, Thai-NLP techniques are necessary. This paper proposes a state-of-the-art Thai COVID-19 fake news detection among word relations using transfer learning models. For pre-training from the global open COVID-19 datasets, the source dataset is constructed by English to Thai translating. The novel feature shifting is formulated to enlarge Thai text examples in target dataset. Machine translation can be used for constructing Thai source dataset to cope with the lack of local dataset for future Thai-NLP applications. To lead the knowledge in Thai text understanding forward, feature shifting is a promising accuracy improvement in fine-tuning stage.

Corresponding Author:

Pakpoom Mookdarsanit
Faculty of Science
Chandrakasem Rajabhat University
39/1 Rachadapisek Road, Chan Kasem District, Chatuchak, Bangkok, 10900, Thailand
Email: pakpoom.m@chandra.ac.th

NOMENCLATURE

Ada-SGD	Adaptive stochastic gradient descent
arXiv.CS	arXiv computer science
AWD-LSTM	Average-stochastic gradient descent weight-dropped long-short term memory
BERT	Bidirectional encoder representations from transformers
BEST	Benchmark for enhancing the standard of Thai language processing
bi-LSTM	Bidirectional long-short term memory
BLEU	Bilingual evaluation understudy
CoAID	COVID-19 healthcare misinformation dataset
COVID-19	Coronavirus disease 2019
English-NLP	English natural language processing
GPT	Generative pre-training
GRU	Gated recurrent unit
IDS	Intrusion detection system
iSAI-NLP	International Joint Symposium On Artificial Intelligence And Natural Language Processing
MLP	Multi-layer perceptron
Montreal.AI	Montreal artificial intelligence
NECTEC	Thailand's National Electronics And Computer Technology Center
NLP	Natural language processing
OOV	Out of vocabulary
PRICAI	Pacific Rim International Conference On Artificial Intelligence
PyThaiNLP	Python packages for Thai natural language processing
ReCOVery	Multimodal repository for COVID-19 news credibility research

RNN	Recurrent neural network
SCB-MT-EN-TH	Siam commercial bank's machine translation in English and Thai
State of AI	State of artificial intelligence report
SVM	Support vector machine
TCI	Thai citation index
Thai-NLP	Thai natural language processing
ULMFiT	Universal language model fine-tuning for text classification
VISTEC.AI	Vidyasirimedhi Institute of Science and Technology Artificial Intelligence

1. INTRODUCTION

Longer than one-hundred years ago, both Cholera and Spanish flu had been largely spreaded in Thailand [1] (as well as the COVID-19 in 2019-2020). Siriraj Hospital-known as the oldest hospital in Mekong society (including Myanmar, Laos, Vietnam, Cambodia, and Thailand), was served as the main medical hub [2, 3] provided to cue many infected people, during the reign of King Chulalongkorn and King Phramongkutklao. During 2019-2020, many world regions had been facing with [3, 4] the spread of new pandemic, known as COVID-19. As the main spreading zones in Thailand were also detected (e.g., *สนามมวยลุมพินี*: Lumpinee Boxing Stadium [5]), many supporting policies were proposed by Thai government sectors (e.g., *ตู้ปันสุข*: Thai Pantry of Sharing [6]). As dedicated to King Chulalongkorn's foundation [2] and Prince Mahidol's contribution [2] toward the high flourish of Thai public health and medical proficiency, Thailand ranked the world's top best ongoing COVID-19 recovery index [7] by Johns Hopkins Coronavirus Resource Center that was ready to be one of the best health and medical tourisms in the world after COVID-19 (known as the New Normal).

Thanks to all Thai health and medical personnel who had been continuously working hard in day and night time since December, 2019. Moreover, it could be obviously seen that many world-class medical centers (at the same level as hospitals in developed countries) to be treated COVID-19 could be found in Thailand, e.g., Siriraj Hospital, Bamrasnaradura Infectious Diseases Institute, King Chulalongkorn Memorial Hospital, Ramathibodi Hospital, Panyanunthaphikkhu Chonprathan Medical Center, Songklanagarind Hospital, Phramongkutklao Hospital and Chulabhorn Hospital. that could handle almost all Thai and other ASEAN patients. From the literature, there were more than 100 original COVID-19 papers from Thai health personnel [8] that exposed the discovery of new medical knowledge. To focus on another impact problem against cyber-security threats, many COVID-19 fake news and/or spambots (or other types of inaccurate information) on social media [9, 10]-composed by fraud social accounts; could be quickly shared/posted among millions social users, such as Facebook, Youtube, and Twitter that easily brought about the serious health misleading and harmfulness-known as one of the big problems against Thai COVID-19 recovery.

- Artificial intelligence open issues against COVID-19

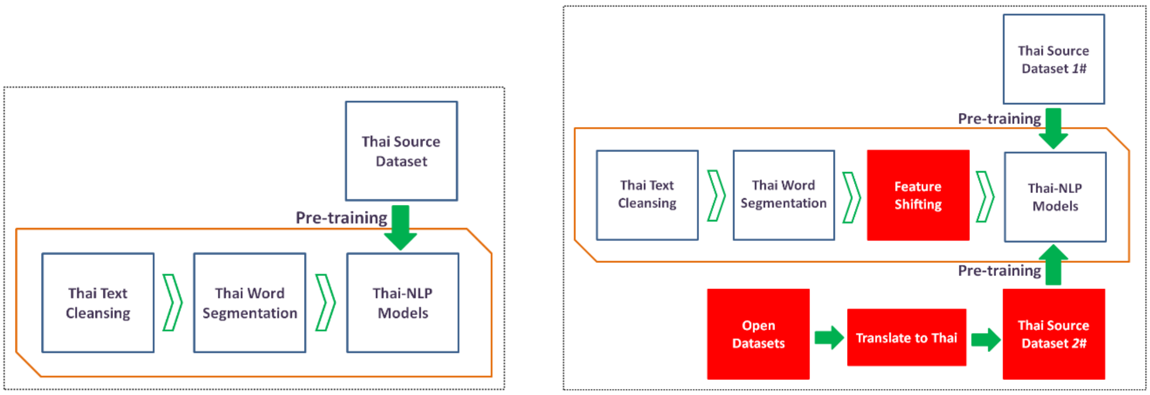
One of the "Artificial intelligence open issues against COVID-19" reported by Montreal.AI [11]-known as the world leading group in AI research and innovation, the COVID-19 fake news detection was still an open-world issue but it was possible to use NLP [12] to detect the misinformation [13] from the content. Many fake news detection tutorials and codes were also available on PapersWithCode [14]. Unquestionably, fake news referred to the misinformation mostly in informal posts [15] such as misinterpretation, personal bias, influence people. It is important to have the quality evalution of social information for maintaining information security by NLP [16], especially in COVID-19 duration. During the 2m-social-distancing, many state-of-the-art COVID-19 fake news detections were investigated [17]. Since both real and fake information has the same spreading pattern. The public hashtags, e.g., "#SocialDistancing" and "#WorkFromHome" [18] with user accounts on Twitter were crawled and analyzed to detect fake news. The multisource-based information e.g., contents, user accounts, spread patterns or URL was investigated on Twitter [19, 20]. Not only Twitter, Facebook page [21-23] was also found to be one of the largest misinformation sources [13] but it was found to be expensive to manually collect them all, instead of detection by the textual content. The COVID-19 ontology was designed as the medical knowledge retrieval by keywords [24]. Some COVID-19 fake news datasets are available e.g., CoAID [25, 26], ReCOVery [27, 28] and FakeCovid [29, 30]. It was obviously seen that arXiv.CS [31, 32] was the largest publication source for COVID-19 fake news detection. However, the prior papers were investigated in English texts. As well as Arabic and English [33], Thai and English have totally different syntactic writing (e.g., no space between any 2 Thai words [34-36]), the investigation in Thai texts should be totally categorized as another problem.

- The COVID-19 fake news detection as one of Thai-NLP applications

For the studies in misinformation detection in Thai social, a large number of Tweet texts with other multi-source-based information [37], e.g., URLs, FriendsCount, FavouritesCount, StatusesCount and RetweetCount. Were cleaned, collected and classified by traditional machine learning [38] like SVM [39], MLP [40] and Naïve Bayes method [41]. However, it could not directly detect the real-time COVID-19 fake news using only Thai textual content; it needed some prior multi-source-based information that was expensive to collect all attributes, as a brute-force compatibility for large-scale stream information. Practically, the COVID-19 fake news detection using only Thai textual content was totally one of Thai-NLP problems that needed Thai specific language's datasets and technical references (unlike English-NLP techniques or others). Thai-NLP and computer have been researching longer than 30 years [42, 43]. Many state-of-the-art papers on Thai-NLP had been being published by NECTEC [44-46] research group-known as an official research and development organization in Thailand that found the AI for Thai project [47], and also the important part of Thai-NLP foundation (e.g., the annual BEST competition [48-50] by NECTEC). Recently, one of the well-known local projects was to match similar Thai keywords between the list of experts and manuscript keywords for automatic assigning reviewers to read the paper [51] that was the collaboration between NECTEC and TCI-known as Thailand's largest publication metric [52] and index [53, 54]. The concept behind this project was Thai text classification [55, 56], also as a type of Thai-NLP problems. A large number of novel Thai-NLP papers can easily be searched on 2 main conferences (known as the largest Thai-NLP archives) as iSAI-NLP [57] and PRICAI [59]. There were so many raw Thai textual sources with the language usage revolution to build textual datasets, e.g., Pantip, Facebook and Youtube. Moreover, PyThaiNLP [58] tutorials with libraries and codes were also available for the beginners.

Since Thai-written language was a type of unsegment words, many state-of-the art papers were proposed to solve Thai word segmentation [34, 35] by bi-LSTM [60], adversarial example [35], pre-training model [61] or unsupervised method with optimization [62]. The word segmentation had been still the main shortfall in Thai-NLP society [63] that totally affected the correctness of other Thai-NLP tasks [46, 63], e.g., part of speech tagging, parsing, text classification, information extraction, semantic role labeling, machine translation, sentiment analysis, event extraction and question answering. Ontology with semantic web [64] was developed by NECTEC that was mostly used as the knowledge structure [65, 66] formulation of unsegmented words and sentences for Thai text knowledge and retrieval [67]. For other original Thai-NLP applications, a bullying detection in social Thai opinions could be detected by modified GRU [68]. Thai plagiarism detection was done [49, 69, 70] that could be used by TCI soon [51]. The state-of-the-art transfer learning with self-attention [71], like BERT [72], ULMFiT [73] and GPT [74] as semantic word embeddings were recently implemented in Thai opinion analysis [75] that could be outstretched by Thai synonyms.

To extend beyond those prior works, in Figure 1, this paper proposes a state-of-the-art Thai-NLP formulation on word-level transfer learning (the ongoing NLP advancement [76], reported by State of AI) models, as a solution for the serious health misinformation problem against Thai COVID-19 recovery. The main contribution of this paper is concluded as: (i) The COVID-19 fake news detection based on transfer learning as a Thai-NLP problem that is introduced to be the IDS for filtering the COVID-19 information, (ii) The source dataset for pre-training stage can be constructed by English to Thai machine translation over those global open COVID-19 datasets, (iii) The feature shifting is formulated to generated more Thai text examples for target dataset. This can be done by synonyms and different Thai written styles that have the same semantical content, (iv) The future Thai-NLP based fake new detection and other applications can be exploited by using feature shifting to enlarge Thai text examples in target dataset; and machine translation from global datasets into Thai as the source dataset. The organization of this paper is categorized into 4 parts. Section 2 talks about research method. Measurement and results are in section 3. Finally, section 4 is conclusion.



(a) (b)

Figure 1. The proposed COVID-19 fake news detection based on Thai-NLP, (a) Prior Thai-NLP works, (b) The extension from those prior Thai-NLP works

2. RESEARCH METHOD

For the extension of prior works, Figure 1(a), this paper proposes a state-of-the-art COVID-19 fake news detection on Thai texts, based on transfer learning as shown in Figure 1(b) that can be seen as IDS for COVID-19 news quality. This section consists of Thai text preparation, English to Thai translating, feature shifting and Thai-NLP models based on transfer learning, respectively.

2.1. Thai text preparation

Unofficial Thai words/phrases on social media (tweets, posts or comments) can be written in variant styles. Thai raw texts can easily produce the noise/outlier for model learning. Thai text preparation is still one of the most important Thai-NLP tasks in order to the detection correctness.

2.1.1. Thai text cleansing

As to replace the grammatical error and delete the noise/outlier, a raw Thai text is necessary to be cleaned such as spelling corrections (e.g., โควิท, โควิต, โควีด, โคหวีด), redundant characters (e.g., โควิดดดด, กรุงเทพพพพ) and other irrelevant characters (e.g., #, !, *). Thai text cleansing obviously reduces the irrelevant and redundant data from the raw social text, prior to fine-tuning stage. In contrast, some posts or comments are such a long Thai text that can be formed to be a multi-semantic text (e.g., คลิปเสียงแพทย์หญิงระบุว่าห้างดังย่านพระราม 9 ใครผ่านไปแถวบริเวณนั้นควรกักตัวโดยด่วน). The multi-semantic text is manually divided into be many single-sentences (e.g., คลิปเสียงแพทย์หญิงระบุว่ามีผู้ป่วยโควิดที่ห้างดังย่านพระราม 9, ใครผ่านไปแถวห้างดังย่านพระราม 9 ควรกักตัวเองโดยด่วน) with the "fake" labeling; and trained to the model. Moreover, some words (e.g., ระบุ) can be combined with feature shifting (e.g., ระบุ, กล่าว, เปิดเผย) to enlarge the size of target dataset.

2.1.2. Thai word segmentation

Unlike English, Thai is naturally an unsegmented language that has no space between 2 words (e.g., กรุงเทพมีการระบาดของโควิดอย่างหนัก). Thai word segmentation (or tokenization) is still acknowledged that it is one of the main Thai-NLP issues. This problem has been waiting for the appropriate solution. The direct dictionary-based word segmentation [34] might not be a good solution for COVID-19 fake news detection. As shown in Figure 2, these 4 possibilities of Thai word segmentation are found to be correct in Thai grammars but need some Ada-SGD algorithms [34, 36] for the segmented optimization. This paper calls the API from PyThaiNLP [77] to segment all words in a sentence or phrase from the primary collection that consists of many Thai tokenization engines, e.g., DeepCut and Lextro.

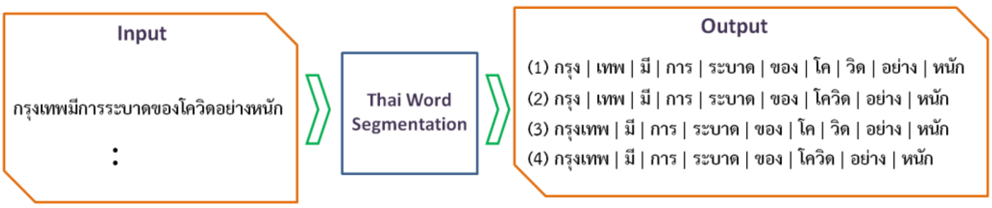


Figure 2. Thai word segmentation (or tokenization) as one of the main problems in Thai-NLP

2.2. English to Thai translating

Since there are so many open COVID-19 fake news datasets in Kaggle, those datasets are based on English. A large number of global information in Thai language is translated from the world-wide news publications that are described in English. The global COVID-19 fake news also has a chance to be translated and published in Thai social. The open COVID-19 fake news datasets are also translated to Thai as source dataset using SCB-MT-EN-TH translation by VISTEC.AI [78-80] as an external knowledge. The source dataset is used to pre-train those transfer learning models (BERT [72], ULMFiT [73] and GPT [74]).

2.3. Feature shifting

According to the language nature, Thai is so variant behavioral usage. The variance can be categorized into synonyms, and written styles, as shown in Figure 3, called feature shifting. This paper proposes feature shifting in Thai words/contents concerning COVID-19 to enlarge the target data volume based on the usage variances; and increase the model correctness.

There are so many Thai synonyms in Thai usage on social media. In this paper, Thai and COVID-19 usages words and their synonyms during December 2019 to June 2020 are manually collected in the datasets. The 1,362 Thai synonym groups are also stored in a database. Since the fake news can be arbitrarily composed using any whatever Thai synonyms (e.g., โควิด, โควิด-19, โคโรน่า). It is essential to train more generated texts with the same semantical content by replacing those synonyms for producing more data variety to make the model perception.

To best our knowledge, each word from Thai word segmentation is further search ($search_{Thai\ syn}(word_i)$); and all synonyms are listed ($list_{\forall syn}$) in order to generate more Thai written styles, modeled by recursive function (1).

$$search_{Thai\ syn}(word_i) = \begin{cases} list_{\forall Thai\ syn} = \{(word_i, syn_1), (word_i, syn_2), \ldots, (word_i, syn_n)\}; & if\ (list_{\forall Thai\ syn} \notin \phi) \\ search_{Thai\ syn}(word_{i+1}); & otherwise \end{cases} \quad (1)$$

The same semantical content can be further composed by many Thai written styles as a word combination problem (e.g., การระบาดของโควิดอย่างหนักที่กรุงเทพ, ที่กรุงเทพโควิดระบาดอย่างหนัก) that depend on the language written behaviors. As to variant behaviors, it can be combined with Thai synonyms (e.g., การแพร่ระบาดของไวรัสสายพันธุ์ใหม่อย่างหนักในกรุงเทพ, การแพร่กระจายของเชื้อไวรัสโคโรน่าอย่างหนัก ณ เมืองกรุง). Thai COVID-19 fake news detection should handle the written variants in both sentence and phrase. To enlarge the target dataset, a number of possible writing styles are generated and trained to the model.

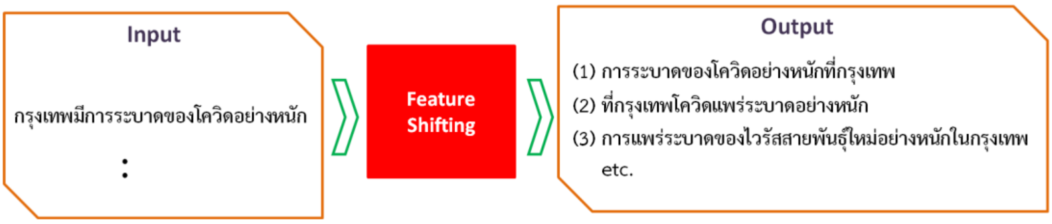


Figure 3. Feature shifting: Thai synonyms with Thai written styles

2.4. Thai-NLP models based on transfer learning

In this paper, transfer learning refers to two important steps of word2vec deep neural training: pre-training Thai COVID-19 model and fine-tuning Thai COVID-19 model as shown in Figure 4. Word2vec is to make neural network to learn the word relations within a Thai text. The first training is randomly initialized the network parameters and learnt by some translated Thai texts. The model's parameters are further reused and learn more other Thai texts crawled from the social media (with feature shifting to increase data volume and variety) in fine-tuning. Transfer learning in sequential data (e.g., word-level sequence) and language process is mostly applied in transformer-based model [71, 81] and later other sequential models: BERT [72], ULMFiT [73], and GPT [74].

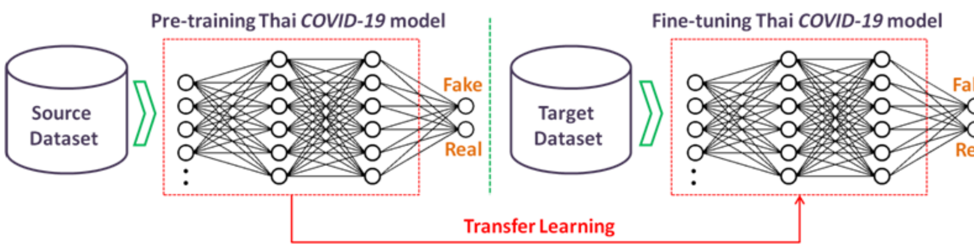


Figure 4. Transfer learning steps of deep neural training

2.4.1. BERT
Based on Transformer, BERT [72] has a well-known “self-attention” mechanism that enables neural networks highlight the essential weights for each token within a sequence. BERT is a bi-directional representation that use both left and right context in all network layers. Normally, transformer is designed for sequential pre-trained model (known as “masked language modeling”). The output of first word is still used as a parameter in the next steps for language modeling. For the configuration, the BERT is set as the 768-sized hidden nodes in each layer, 12 attention heads as amount of 110M parameters.

2.4.2. ULMFiT
ULMFiT [73] is proposed to be transfer learning for language processing as well as ImageNet pre-trained model in computer vision. ULMFiT inherits AWD-LSTM [82] that has 3 stages: pre-training, discriminative fine-tuning and classifier fine-tuning. ULMFiT shows the state-of-the-art text classification performance under the low processing resource. Moreover, many novel techniques are still proposed to solve the catastrophic interference problem. The model has 3-layer AWD-LSTMs with the embedding size as 400 and hidden node size as 1,150 in each hidden layer.

2.4.3. GPT
GPT [74] demonstrates state-of-the-art results in many NLP applications. Since Transformer can tackle with the long-range dependencies in sequential data (better than RNN), it can take in the entire sequence as once time. GPT is also a transformer-based model as well as BERT but it is such unidirectional feed-forward architecture. For this work, the model is configured as 12 transformer layers with 12 attention heads and 768 dimensional states.

3. MEASUREMENT AND RESULTS
Based on Thai-NLP measurement, this section demonstrated the promising results of the proposed Thai COVID-19 fake news detection according to the paper contributions (as one of the “Artificial Intelligence Open Issues against COVID-19” by Montreal.AI [11]). In summary, the pre-trained models were fine-tuned into Thai COVID-19 fake news classifiers based on both pre-training and fine-tuning Thai COVID-19 datasets. These transfer learning models (BERT [72], ULMFiT [73] and GPT [74]) on Thai texts were run on Tesla V100 GPU on GCP with Colab environment. This section could be categorized into 3 sub-sections: dataset construction, pre-training Thai COVID-19 models and fine-tuning Thai COVID-19 models.

3.1. Dataset construction
According to the data source from COVID-19 news open datasets (as source dataset) and the crawled Thai texts from social media (as target dataset), the labels could be fake and real. This section talked about the dataset construction. The datasets in transfer learning could be divided into 2 types.

3.1.1 Source dataset by machine translation
The pre-training Thai COVID-19 models was trained by the 123,762 Thai-translated single texts from source dataset (as described in section 3.2). For the data collection, the global COVID-19 fake (and real) news in English from well-known open datasets: CoAID [25, 26], ReCOVery [27, 28] and FakeCovid [29, 30] were collected and translated to Thai using SCB-MT-EN-TH translation by VISTEC.AI [77-79]. To evaluate the quality of English to Thai translation, the BLEU [83] was used and compared to other well-known Eng-to-Thai machines: AI for Thai by NECTEC [47] and Google Translate. The results demonstrated that the transformer-based SCB-MT-EN-TH translation outperformed other neural-based machine translations for Eng-to-Thai translation in any text length variation, based COVID-19 open datasets, as shown in Figure 5.

3.1.2. Target dataset with feature shifting improvement
The language network models were fine-tuned by sharing the pre-trained weights learnt by 123,762 Thai texts from the source dataset. For the target dataset construction, the local COVID-19 fake (and real) news in Thai from posts and/or links shared on Facebook, Twitter and fake URLs were collected during December 2019 to June 2020. All collected information was extracted into 45,372 single Thai texts as pre-training Thai COVID-19 dataset by Thai text cleansing and Thai word segmentation, respectively. According to Thai variant usage in social media, the target dataset is enlarged to 73,280 Thai text examples with the OOV as 3.26%. These texts were generated by feature shifting. The target dataset was used in Fine-tuning Thai COVID-19 models (section 3.3).

Feature shifting also helped to enlarge the size of Thai text examples into target dataset train on fine-tuning models. Figure 6 demonstrated the accuracy enhancement by feature shifting in all transfer learning models for Thai COVID-19 fake news classification. GPT provided the highest improvement by feature shifting as 36.43%.

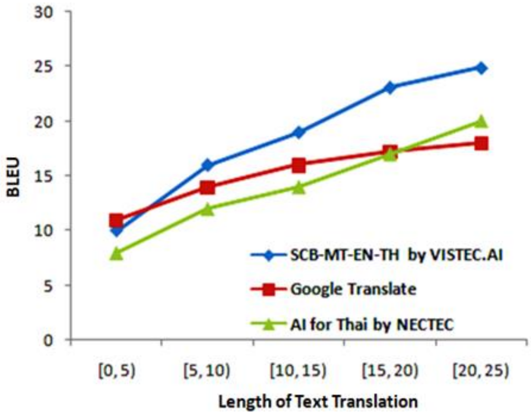


Figure 5. Comparison between English to Thai machine translations based on global COVID-19 open datasets

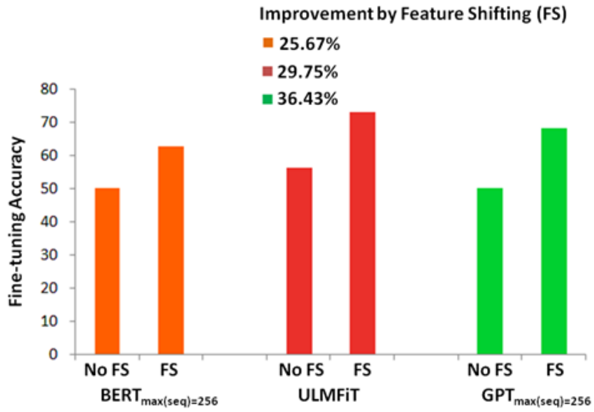


Figure 6. Transfer learning improved by feature shifting in fine-tuning stage

3.2. Pre-training Thai COVID-19 models

The source dataset was proposed to train the pre-training Thai COVID-19 models that could be splited into 3 partitions: training (70%), validation (15%), and testing (15%) as well as [80] configuration. From Table 1, ULMFiT ran the least pre-training time. Since the ULMFiT was based on AWD-LSTM [82] and well-designed for pre-trained model, especially in state-of-the-art Thai text classification under the low processing resource. And ULMFiT also had the lowest training loss based on the 123,762 Thai text examples from source dataset that was constructed by SCB-MT-EN-TH translation on those open COVID-19 datasets (CoAID [25, 26], ReCOVery [27, 28], and FakeCovid [29, 30]).

Table 1. Comparison between pre-training Thai COVID-19 models

Transfer learning model	Pre-training time (hours)	Loss
BERT $_{max(seq)=256}$	22	4.0719
ULMFiT	**12**	**3.8517**
GPT$_{max(seq)=256}$	19	4.5365

3.3. Fine-tuning Thai COVID-19 models

From Table 2, the fine-tuning Thai COVID-19 models were shared parameters from pre-trained models and learnt more 73,280 Thai text examples from target dataset. The partition between training and testing was set as 70:30 [80]. Although GPT had the highest improvement by feature shifting, ULMFiT was well-performed in different writing styles and synonyms in Thai texts that gave highest accuracy based on this target data. Although BERT was a bidirectional model that could leverage both left and right words (called tokens) for embeddings [72, 75], ULMFiT contrastly achieved higher performance, especially for Thai feature shifting.

Table 2. Comparison between fine-tuning Thai COVID-19 models

Transfer learning model	Fine-tuning time (minute per epoch)	Accuracy
BERT $_{max(seq)=256}$	4	0.6286
ULMFiT	**2**	**0.7293**
GPT$_{max(seq)=256}$	4	0.6819

4. CONCLUSION

As it relates to the problem stated fake news against Thai COVID-19 recovery, this paper addresses fake news detection in Thai texts based on transfer learning, using Thai-NLP techniques. Transfer learning models consist of BERT, ULMFiT and GPT; where ULMFiT is shown to be higher performance for Thai COVID-19 fake news detection. With the help of English to Thai machine translation, the COVID-19 open datasets in English can be leveraged to construct the source dataset for pre-training. Feature shifting is proposed to enlarge the target dataset for fine-tuning that absolutely improves the classification accuracy in all transfer learning models, where GPT has the highest improvement rate. For the limitation, Thai text noisy/outlier are still a main obstacle in the full-supervised learning. To label all million Thai texts without noisy is totally impossible. Mislabeling easily comes with the full-supervision in large volume that also makes the model learnt some wrong labels. The future Thai fake news detection should be semi-supervised learning that has 2 parts: (i) Firstly trains only the high quality of labeling in some Thai texts as partial-supervised model and (ii) Automatically label those unknown Thai texts by the partial-supervision. Not only the wrong labeled prevention but also long speed of full-training is problems in full-supervision. To extend this work, multi-task transfer learning with Thai synonyms can be used for Thai language understanding to make model learnt by many tasks from multi-datasets.

ACKNOWLEDGEMENTS

Thanks to all Thai health and medical personnel for their hard working on COVID-19. This paper addresses the filtering Thai fake news related to COVID-19 using only Thai textual content as the IDS for filtering the COVID-19 misinformation. Do not hesitate to contact the corresponding author for the local data and code (and please referred to this paper). Under the dynamic information, the higher accuracy can be further improved by more preparation process and higher parameters in transfer learning models. Thanks to the expert from Semantic Technology Laboratory, NECTEC for advising all technical knowledge and data. With the quote “Rajabhat means the king’s men”, all computational resources are gratefully supported by Chandrakasem Rajabhat University.

REFERENCES

[1] SCB, “300 Years of History of Deadly Pandemics,” 2020. [Online]. Available at: https://www.scb.co.th/en/personal-banking/stories/tips-for-you/300-years-pandemics.html. [Accessed: 20-Sept-2020].
[2] S. Charuluxananan and V. Chentanez, “History and Evolution of Western Medicine in Thailand,” *Asian Biomedicine*, vol. 1, no. 1, pp. 97-101, 2007.
[3] S. Sukpanichnan, “Emergence of Influenza Pandemic in Bangkok in 1918: Historical Review,” *Siriraj Medical Journal*, vol. 72, no. 4, pp. 368-370, 2020.
[4] S. Sukpanichnan, “Influenza Pandemic in Siam during World War I: Perspective from COVID-19 Situation,” *Disease Control Journal*, vol. 46, no. 2, pp. 106-114, 2020.
[5] Thai PBS World, “Lumpini Boxing Stadium’s Board Fired over Cluster of COVID-19 Infections,” 2020. [Online]. Available at: https://www.thaipbsworld.com/lumpini-boxing-stadiums-board-fired-over-cluster-of-covid-19-infections. [Accessed: 20-Sept-2020].
[6] The Nation Thailand, “‘Pantry of sharing’ Brings out Spirit of Helping in Thais,” 2020. [Online]. Available at: https://www.nationthailand.com/news/30387585. [Accessed: 20-Sept-2020].
[7] Panorama Destination, “Thailand Tops Global COVID-19 Recovery Index,” 2020. [Online]. Available at: http://www.panorama-destination.com/thailand-news/thailand-tops-global-covid-19-recovery-index. [Accessed: 20-Sept-2020].
[8] I. M. Shettar and G. S. Hadagali, “Coronavirus: A Scientometrics Study of World Research Publications,” *International Journal of Information Dissemination and Technology*, vol. 10, no. 1, pp. 8-16, 2020.
[9] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, “The COVID-19 Social Media Infodemic,” *arXiv preprint arXiv:2003.05004*, 2020.
[10] J. Chen, K. Li, Z. Zhang, K. Li, and P. S. Yu, “A Survey on Applications of Artificial Intelligence in Fighting against COVID-19,” *arXiv preprint arXiv:2007.02202*, 2020.
[11] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz, “Mapping the Landscape of Artificial Intelligence Applications against COVID-19,” *arXiv preprint arXiv:2003.11336*, 2020.
[12] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, “Natural Language Processing Advancements by Deep Learning: A Survey,” *arXiv preprint arXiv:2003.01200*, 2020.
[13] T. Quandt, S. Boberg, T. Schatto-Eckrodt, and L. Frischlich, “Pandemic News: Facebook Pages of Mainstream News Media and the Coronavirus Crisis-A Computational Content Analysis,” *arXiv preprint arXiv:2005.13290*, 2020.
[14] Papers with Code, “Fake News Detection: Natural Language Processing,” [Online]. Available at: https://paperswithcode.com/task/fake-news-detection. [Accessed: 20-Sept-2020].
[15] O. J. Ying, M. M. A. Zabidi, N. Ramli, and U. U. Sheikh, “Sentiment Analysis of Informal Malay Tweets with Deep Learning,” *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 212-220, 2020.

[16] P. Tanwar and P. Rai, "A Proposed System for Opinion Mining using Machine Learning, NLP and Classifiers," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 4, pp. 726-733, 2020.
[17] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A First Look at COVID-19 Information and Misinformation Sharing on Twitter," *arXiv preprint arXiv:2003.13907*, 2020.
[18] K. Sharma, S. Seo, C. Meng, S. Rambhatla, and Y. Liu, "COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations," *arXiv preprint arXiv:2003.12309*, 2020.
[19] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, "An Exploratory Study of COVID-19 Misinformation on Twitter," *arXiv preprint arXiv:2005.05710*, 2020.
[20] B. Huang and K. M. Karley, "Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak," *arXiv preprint arXiv:2006.04278*, 2020.
[21] K. Ding, K. Shu, Y. Li, A. Bhattacharjee, and H. Liu, "Challenges in Combating COVID-19 Infodemic-Data, Tools, and Ethics," *arXiv preprint arXiv:2005.13691*, 2020.
[22] S. Laato, A. K. M. Najmul Islam, M. N. Islam, and E. Whelan "Why do People Share Misinformation during the COVID-19 Pandemic?," *arXiv preprint arXiv:2004.09600*, 2020.
[23] A. Mourad, A. Srour, H. Harmanani, C. Jenainatiy, and M. 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.
[24] A. Groza, "Detecting Fake News for the New Coronavirus by Reasoning on the COVID-19 Ontology," *arXiv preprint arXiv:2004.12330*, 2020.
[25] L. Cui and D. Lee, "CoAID: COVID-19 Healthcare Misinformation Dataset," *arXiv preprint arXiv:2006.00885*, 2020.
[26] C. Meng, "CoAID," [Online]. Available at: https://github.com/cuilimeng/CoAID. [Accessed: September 20, 2020].
[27] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research," *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3205-3212, 2020.
[28] A. Mulay, "ReCOVery," [Online]. Available at: https://github.com/apurvamulay/ReCOVery. [Accessed: September 20, 2020].
[29] G. K. Shahi and D. Nandini, "FakeCovid-A Multilingual Cross-domain Fact Check News Dataset for COVID-19," *arXiv preprint arXiv:2006.11343*, 2020.
[30] hkefka385, "FakeCovid-A Multilingual Cross-domain Fact Check News Dataset for COVID-19," [Online]. Available at: https://github.com/hkefka385/paper_reading/issues/74. [Accessed: September 20, 2020].
[31] Cornell University, "arXiv: Computer Science," [Online]. Available at: https://arxiv.org/archive/cs. [Accessed: 20-Sept-2020].
[32] C. Sutton and L. Gong, "Popularity of arXiv.org within Computer Science," [Online]. Available at: http://groups.inf.ed.ac.uk/cup/csarxiv.
[33] G. S. Cheema, S. Hakimov, and R. Ewerth "Check square at CheckThat! 2020: Claim Detection in Social Media via Fusion of Transformer and Syntactic Features," *arXiv preprint arXiv:2007.10534*, 2020.
[34] C. Haruechaiyasak, S. Kongyoung, and M. N. Dailey, "A Comparative Study on Thai Word Segmentation Approaches," *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pp. 125-128, 2008.
[35] P. Netisopakul, G. Wohlgenannt, and A. Pulich, "Word Similarity Datasets for Thai: Construction and Evaluation," *IEEE Access*, vol. 7, pp. 142907-142915, 2019.
[36] C. Udomcharoenchaikit, P. Boonkwan, and P. Vateekul, "Adversarial Evaluation of Robust Neural Sequential Tagging Methods for Thai Language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 4, pp. 1-25, 2020.
[37] S. Aphiwongsophon and P. Chongstitvatana, "Identifying Misinformation on Twitter with a Support Vector Machine," *Engineering and Applied Science Research*, vol. 47, no. 3, pp. 306-312, 2020.
[38] S. Aphiwongsophon and P. Chongstitvatana, "Detecting Fake News with Machine Learning Method," *2018 International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pp. 528-531, 2018.
[39] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
[40] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533-536, 1986.
[41] D. Heckerman, "A tutorial on learning with Bayesian networks," *Innovations in Bayesian networks*, pp. 33-82, 2008.
[42] H. Thaweesak Koanantakool, T. Karoonboonyanan, and C. Wutiwiwatchai, "Computers and the Thai Language," *IEEE Annals of the History of Computing*, vol. 31, no. 1, pp. 46-61, Jan.-March 2009.
[43] V. Sornlertlamvanich, "A 29-year Journey of Thai NLP," [Online]. Available at: https://www.slideshare.net/virach/a-29year-journey-of-thai-nlp. [Accessed: 20-Sept-2020].
[44] C. Wutiwiwatchai, V. Chunwijitra, S. Chunwijitra, P. Sertsi, S. Kasuriya, P. Chootrakool, and K. Thangthai, "The NECTEC 2015 Thai Open-Domain Automatic Speech Recognition System," *International Symposium on Natural Language Processing*, pp. 124-136, 2016.
[45] NECTEC, "Publication," [Online]. Available at: https://www.nectec.or.th/hccru/publication/lst. [Accessed: 20-Sept 2020].

[46] NECTEC, “NLP, Information Retrieval and Text Mining,” [Online]. Available at: http://www2.it.kmutnb.ac.th/teacher/FileDL/maleerat178255420125.pdf. [Accessed: 20-Sept 2020]. [in Thai].
[47] NECTEC, “"AI for Thai" Transforming Digital Transformation with Artificial Intelligence,” [Online]. Available at: https://www.nectec.or.th/research/research-project/aiforthai-digitaltransformation.html. [Accessed: 20-Sept-2020]. [in Thai].
[48] K. Kosawat, M. Boriboon, P. Chootrakool, A. Chotimongkol, S. Klaithin, S. Kongyoung, K. Kriengket, S. Phaholphinyo, S. Purodakananda, T. Thanakulwarapas, and C. Wutiwiwatchai, “BEST 2009: Thai Word Segmentation Software Contest,” *2009 Eighth International Symposium on Natural Language Processing*, pp. 83-88, 2009.
[49] NECTEC, “BEST 2018: Automatic Copy Verification System Competition,” [Online]. Available at: https://thailang.nectec.or.th/best/2017/03/20/best2018/. [Accessed: 20-Sept-2020]. [in Thai].
[50] NECTEC, “BEST 2020: Handwritten Recognition,” [Online]. Available at: https://thailang.nectec.or.th/best/2019/06/20/best-2020-handwrittenrecognition-objective-2/. [Accessed: 20-Sept-2020]. [in Thai].
[51] S. Kasuriya, “TCI-TSRI-SCOPUS Collaboration Project Phases 1-2: Online Submission, Reviewer Pool and Reviewer Finder by NECTEC,” [Online]. Available at: https://tci-thailand.org/?p=5199. [Accessed: 20-Sept-2020]. [in Thai].
[52] TCI, “TCI History,” Available at: https://tci-thailand.org/?page_id=21. [Accessed: 20-Sept-2020]. [in Thai].
[53] N. Sombatsompop, P. Ratchatahirun, V. Surathanasakul, N. Premkamolnetr, and T. Markpin, “A Citation Report for Thai Academic Journals Published during 1996-2000,” *Scientometrics*, vol. 55, no. 3, 445-462, 2002.
[54] N. Sombatsompop, S. Chancheewa, T. Markpin, N. Premkamolnetr, S. Ittiritmeechai, C. Wongkaew, W. Yochai, and P.Ratchatahirun, “Thai-Journal Citation Index (TCI) Centre: 10 Years of Experiences, Lessons Learned and Ongoing Development,” *Malaysian Journal of Library & Information*, vol. 17, no. 3, pp. 17-33, 2012.
[55] A. Kongthon, C. Haruechaiyasak, M. Buranarach, S. Thaiprayoon, and N. Angkawattanawit, “A Framework for Managing R&D for Thai Research Community using Text Information Exploitation,” *PICMET '08 - 2008 Portland International Conference on Management of Engineering & Technology*, pp. 1470-1476, 2008.
[56] A. Kongthon, C. Haruechaiyasak, and K. Trakultaweekoon, "Using Bibliometric Analysis and Text Mining to Improve the Thai Talent Database,” *2016 Portland International Conference on Management of Engineering and Technology (PICMET)*, pp. 1928-1934, 2016.
[57] Artifical Intelligence Association of Thailand, “The International Joint Symposium on Artificial Intelligence and Natural Language Processing: Timeline,” [Online]. Avialable at: https://aiat.or.th/isai-nlp/. [Accessed: 20-Sept-2020].
[58] Thai NLP, “Py Thai NLP,” [Online]. Available at: https://www.thainlp.org/pythainlp/tutorials/notebooks/pythainlp_get_started.html. [Accessed: 20-Sept-2020].
[59] PRICAI, “PRICAI History,” [Online]. Available at: https://www.pricai.org/. [Accessed: 20-Sept-2020].
[60] T. Lapjaturapit, K. Viriyayudhakom, and T. Theeramunkong, “Multi-Candidate Word Segmentation using Bi-directional LSTM Neural Networks,” *2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES)*, Khon Kaen, pp. 1-6, 2018.
[61] S. Seeha, I. Bilan, L. M. Sanchez, J. Huber, M. Matuschek, and H. Schütze, “ThaiLMCut: Unsupervised Pretraining for Thai Word Segmentation,” *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6947-6957, 2020.
[62] J. Sunkpho and M. Hofmann, “Thai Words Segmentation Using an Unsupervised Learning Technique,” *International Conference on Computing and Information Technology*, pp. 79-92, 2020.
[63] C. Tapsai, P. Meesad, and H. Unger, “An Overview on the Development of Thai Natural Language Processing,” *Information Technology Journal*, vol. 15, no. 2, 45-52, 2019.
[64] M. Buranarach, T. Supnithi, Y. M. Thein, T. Ruangrajitpakorn, T. Rattanasawad, K. Wongpatikaseree, A. O. Lim, Y. Tan, and A. Assawamakin, “OAM: An Ontology Application Management Framework for Simplifying Ontology-Based Semantic Web Application Development,” *International Journal of Software Engineering and Knowledge Engineering*, vol. 26, no. 1, pp. 115-145, 2016.
[65] T. Ruangrajitpakorn, C. Prombut, and T. Supnithi, “A Development of an Ontology-based Personalised Web from Rice Knowledge Website,” *2018 Thirteenth International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, pp. 1-6, 2018.
[66] M. Buranarach, C. Anutariya, N. Kalayanapan, T. Ruangrajitpakorn, V. Wuwongse, and T. Supnithi, “An Ontology-Based Approach to Supporting Knowledge Management in Government Agencies: A Case Study of the Thai Excise Department,” *IEICE Transactions on Information and Systems*, vol. 101, no. 4, pp. 884-891, 2018.
[67] J. Panawong, T. Ruangrajitpakorn, and M. Buranarach, “An Automatic Database Generation and Ontology Mapping from OWL File,” *Conference: The Third International Workshop on Practical Application of Ontology for Semantic Data Engineering*, pp. 20-27, 2016.
[68] P. Mookdarsanit, “TGF-GRU: A Cyber-bullying Autonomous Detector of Lexical Thai across Social Media,” *NKRAFA Journal of Science and Technology*, vol. 15, pp. 50-58, 2019.
[69] S. Thaiprayoon, P. Palingoon, K. Trakultaweekoon, S. Klaithin, C. Haruechaiyasak, A. Kongthon, S. Thatpitakkul, and S. Kasuriya, “Developing a Framework for a Thai Plagiarism Corpus,” *International Conference of the Pacific Association for Computational Linguistics*, pp. 513-522, 2019.
[70] W. Massagram, S. Prapanitisatian, and K. Kesorn, “A Novel Technique for Thai Document Plagiarism Detection using Syntactic Parse Trees,” *Engineering and Applied Science Research*, vol. 45, no. 4, pp. 290-300, 2018.

[71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000-6010, 2017.

[72] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171-4186, 2019.

[73] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 328-339, 2018.

[74] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," *Pre-print*, 2018.

[75] T. Horsuwan, K. Kanwatchara, P. Vateekul, B. Kijsirikul, "A Comparative Study of Pretrained Language Models on Thai Social Text Categorization," *Asian Conference on Intelligent Information and Database Systems*, pp. 63-75, 2020.

[76] N. Benaich and I. Hogarth, "State of AI Report 2020," 2020. [Online]. Available at: https://www.stateof.ai. [Accessed: 17-Oct-2020].

[77] ThaiNLP, "PyThaiNLP: pythainlp.tokenize," [Online]. Available at: https://thainlp.org/pythainlp/docs/2.0/api/tokenize.html. [Accessed: 20-Sept-2020].

[78] L. Lowphansirikul, C. Polpanumas, A. T. Rutherford, and S. Nutanong, "SCB-MT-EN-TH-2020: A Large English-Thai Parallel Corpus," *arXiv preprint arXiv:2007.03541*, 2020.

[79] Vistec-AI, "mt-opus," 2020. [Online]. Available at: https://github.com/vistec-AI/mt-opus. [Accessed: 20-Sept-2020].

[80] Vistec-AI, "mt-datasets," 2020. [Online]. Available at: https://github.com/vistec-AI/mt-datasets. [Accessed: 20-Sept-2020].

[81] J. C. B. Cruz, J. A. Tan, and C. Cheng, "Localization of Fake News Detection via Multitask Transfer Learning," *Proceedings of the 12th Conference on Language Resources and Evaluation*, pp. 2596-2604, 2020.

[82] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and Optimizing LSTM Language Models," *arXiv preprint arXiv:1708.02182*, 2017.

[83] K. Papineni, S. Roukos, T. Ward, and W-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318, 2002.