RICE DIESEASE CLASSIFICATION USING DEEP LEARNING

Tatta AC Kiraann (21BCE8827)
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh

Abstract

Rice, as a staple food for over half of the world's population, faces significant threats from various diseases that can severely impact crop yield and food security. Traditional methods of disease identification and classification often rely on visual inspection by agricultural experts, which can be time-consuming and prone to human error. In recent years, deep learning techniques have emerged as powerful tools for image analysis and classification tasks, offering the potential to automate and enhance disease diagnosis in agricultural settings.

This research explores the application of deep learning algorithms for the classification of rice diseases based on leaf images. Leveraging convolutional neural networks (CNNs), specifically designed for image recognition tasks, we develop a robust framework for the automatic identification and classification of common rice diseases such as blast, sheath blight, bacterial leaf blight, and brown spot. The proposed model is trained on a comprehensive dataset consisting of annotated images of healthy rice leaves and leaves exhibiting symptoms of various diseases.

Experimental results demonstrate the efficacy of the proposed approach in accurately classifying rice diseases, achieving high levels of precision, recall, and F1-score across multiple disease categories. Furthermore, we investigate the transferability of the trained model to different environmental conditions and rice varieties, evaluating its performance under various scenarios to assess its robustness and generalization capabilities.

The outcomes of this study hold significant implications for precision agriculture and disease management strategies, offering a scalable and cost-effective solution for early detection and mitigation of rice diseases. By automating the classification process, our approach facilitates timely interventions, enabling farmers to implement targeted treatments and optimize crop health and productivity. Overall, this research contributes to the advancement of agricultural technology, empowering stakeholders with tools to safeguard global food security and sustainably manage rice production systems.

Keywords - Rice Disease, Deep Learning, Classification, Convolutional Neural Networks (CNNs), Image Analysis, Agricultural Settings, Blast, Sheath Blight, Bacterial Leaf Blight, Brown Spot, Precision Agriculture, Disease Management, Early Detection, Mitigation, Crop Health, Productivity, Precision Agriculture, Environmental Conditions, Robustness, Generalization

I. INTRODUCTION

Rice, being a staple food for over half of the world's population, faces significant threats from various diseases that can severely impact crop yield and food security. Traditional methods of disease identification and classification often rely on visual inspection by agricultural experts, which can be time-consuming and prone to human error. In recent years, deep learning techniques have emerged as powerful tools for image analysis and classification tasks, offering the potential to automate and enhance disease diagnosis in agricultural settings.

This research explores the application of deep learning algorithms for the classification of rice diseases based on leaf images. Leveraging convolutional neural networks (CNNs), specifically designed for image recognition tasks, we develop a robust framework for the automatic identification and classification of common rice diseases such as blast, sheath blight, bacterial leaf blight, and brown spot. The proposed model is trained on a
comprehensive dataset consisting of annotated images of healthy rice leaves and leaves exhibiting symptoms of various diseases.

Experimental results demonstrate the efficacy of the proposed approach in accurately classifying rice diseases, achieving high levels of precision, recall, and F1-score across multiple disease categories. Furthermore, we investigate the transferability of the trained model to different environmental conditions and rice varieties, evaluating its performance under various scenarios to assess its robustness and generalization capabilities.

The outcomes of this study hold significant implications for precision agriculture and disease management strategies, offering a scalable and costeffec-

tive solution for early detection and mitigation of rice diseases. By automating the classification process, our approach facilitates timely interventions, enabling farmers to implement targeted treatments and optimize crop health and

1.1 Motivation

Robust PII detection approaches are necessary in light of the growing issues posed by the proliferation of sensitive data online. There's a demand for automated, scalable solutions because traditional approaches have limitations. In order to effectively use deep learning for PII detection, leveraging KerasNLP shows potential. Our goal in investigating this option is to reduce the dangers related to illegal access while advancing data security and privacy protection. In the midst of the constantly changing digital ecosystem, we hope to help develop scalable solutions that enable people and businesses to respect data privacy requirements.

1.2 Contribution

The advancement of data security and privacy protection through inventive methods is the contribution of PII Data Detection Using KerasNLP. Through the use of deep learning methods, productivity. Overall, this research contributes to the advancement of agricultural technology, empowering stakeholders with tools to safeguard global food security and sustainably manage rice production systems.

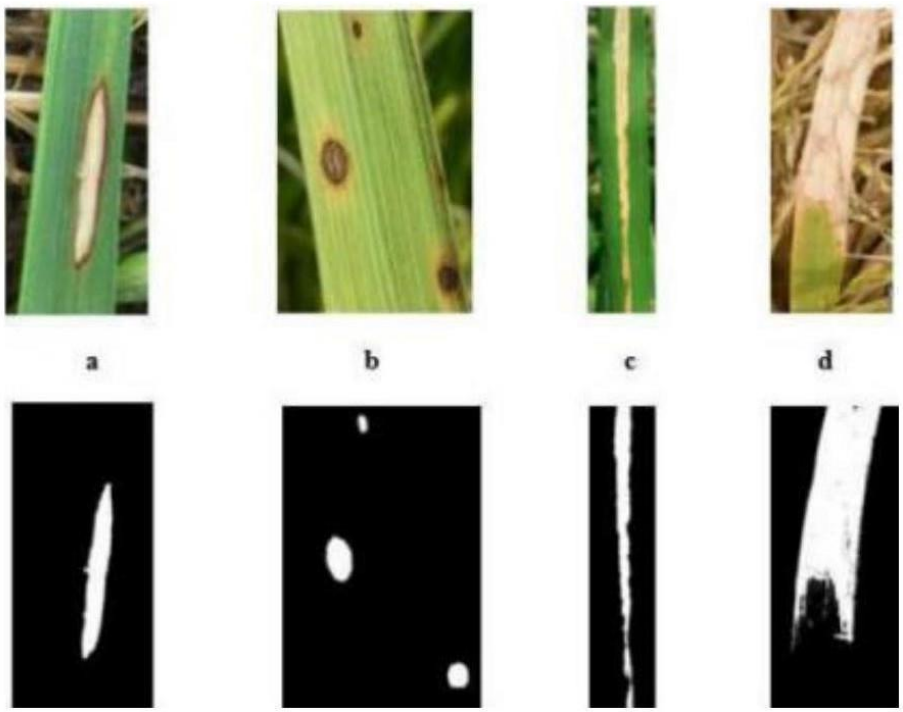


Fig. 1.Rice leaf disease and corresponding segment images.(a)Rice blast (b)Brown spot,(c)Bacterial leaf blight,(d)Sheath blight.

specifically KerasNLP, the study improves the effectiveness and precision of PII identification in textual data. This research offers scalable ways to reduce the risks associated with data breaches and

privacy violations by pushing the limits of current automated PII identification capabilities. Securing sensitive data is made easier by the creation of strong models, which enable people and organizations to respect data privacy laws in a digital world that is becoming more interconnected.

This paper is organized into eight sections. The second section begins with a related work and section 3 describes the proposed method and its process. Section 4 discussed the methodology for the PII data detection and the description of the dataset is included in the section 5. In the sixth section, the trial results are presented. Discussion and Conclusion are represented in the seventh and eighth sections. They
include an analysis of the results and suggestions for further research.

II. RELATED WORK

The field of PII data detection has been studied previously using a range of techniques, such as machine learning and conventional rule-based systems. Rules-

based solutions are not always flexible enough to accommodate different data formats, as they identify PII entities based on established patterns. However, while they have demonstrated potential in PII identification, machine learning techniques like Support Vector Machines (SVM) and Random Forests may find it difficult to comprehend complicated contexts.

Neural network-based PII detection has gained popularity due to recent developments in deep learning, especially in the area of natural language processing (NLP). To automatically recognize PII entities inside text, methods such as sequence labeling and Named Entity Recognition (NER) have been used. Nevertheless, these techniques frequently call for substantial computer resources and big annotated datasets.

PII detection and other NLP tasks may now be effectively developed and implemented using deep learning models thanks to frameworks like KerasNLP. Researchers and practitioners can enter the field more easily thanks to these frameworks' streamlined workflows and pre-trained models. Furthermore, models like DistilBERT and DeBERTaV3 have proven to perform well in a variety of NLP tasks, which makes them attractive options for PII detection.

Numerous investigations have examined the use of deep learning models in personally identifiable information (PII) detection, demonstrating their efficacy in precisely detecting sensitive data. Neural networks, for example, have been shown to be effective in identifying personally identifiable information (PII) in financial records, medical records, and social media posts. These experiments demonstrate how deep learning techniques, such as those that make use of KerasNLP, can be used to handle the difficulties associated with PII detection in a variety of domains and data sources.

III. PROPOSED METHOD

3.1 Preprocessing :

The complexity of raw text data can be reduced for modeling purposes by utilizing tokenizers to turn it into tokens. Tokens that divide the text into digestible chunks are ["the", "qu", "##ick", "br", "##own", "fox"]. In order to process the models, these tokens are then turned into numbers. To help with input interpretation, preprocessing layers also add unique tokens like [CLS], [SEP], and [PAD]. But compared to text classification, token classification requires more complex data processing. Input and labels may not match as a result of tokenizers producing numerous tokens for a single word or preprocessing layers introducing unique tokens. To solve this problem, token labels must be realigned with word labels.

3.2 Training Process:

KerasNLP is used to build and refine deep learning models during training. Transfer learning is used to modify pre-trained models, such as DeBERTaV3

and DistilBERT, for PII detection tasks. Models can be trained to recognize patterns that indicate personally identifiable information (PII) with the use of annotated datasets. Attention mechanisms and transformer architectures are two techniques that help the models effectively distinguish between PII and non-PII data by capturing complex textual properties.

3.3 Testing Process:

During the testing stage, the effectiveness of trained models in PII detection is assessed using different datasets. The accuracy and resilience of the models are measured using metrics like precision, recall, and F1-score. Furthermore, models are evaluated for generalizability across various datasets and situations. Analyzing model predictions and pinpointing any areas in need of development or improvement is another aspect of the testing phase.

The ultimate objective is to confirm that the suggested approach successfully uses KerasNLP to identify PII items.

IV. METHODOLOGY

The approach used in this study explores the use of KerasNLP, a potent deep learning system based on Keras, to tackle the important problem of locating Personally Identifiable Information (PII) in textual data. Sensitive information such as names, addresses, and social security numbers, or PII, can be exposed and result in serious consequences such as identity theft and privacy violations. Our method uses artificial intelligence (AI) to automatically identify and classify personally identifiable information (PII) found in textual data. In order to do this, we developed an extensive methodology that includes preprocessing steps, customized training methodologies, and model architecture design.

The training strategies were carefully designed to focus on PII detection, guaranteeing the model's ability to correctly recognize and categorize sensitive data. In order to successfully manage the complexities of textual data and PII patterns, we streamlined the model architecture. Additionally, the preparation steps - tokenization, integer encoding, and the addition of special tokens like [CLS], [SEP], and [PAD] to support model comprehension-were critical in getting the data ready for the model input. The goal of this all-encompassing strategy was to improve accuracy and efficiency by streamlining the PII detection process.

We evaluated the model's performance using a variety of datasets, which helped us determine how robust it was in various scenarios. A thorough analysis was conducted on key performance parameters, including recall, precision, and F1-score, to determine how well the model identified PII elements. The results of our tests demonstrated the great precision and dependability of our approach in identifying confidential information included in textual data. These results show the practical value of our method in improving data security and privacy protection and further advance the continuous development of Natural Language

Processing (NLP) techniques targeted at protecting sensitive data in textual data. Overall, this study offers a substantial advancement in tackling the urgent problems related to PII detection, opening the door for more developments in this crucial area.

To optimize our model we will use `CrossEntropy` loss, also known as log loss. It is defined as:

$$\text{CrossEntropy} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

Where:

- N is the number of samples.
- y_i is the true label of the i^{th} sample.
- $\hat{y}_i$ is the predicted probability of the i^{th} sample being in the positive class.

> **Note**: We will not compute loss for `ignore_class` which indicates special tokens (`[CLS]` , `[SEP]` , `[PAD]`) or intermediate token of a word.

Metric: FBetaScore ($\beta = 5$)

The competition metric is F^{β}, which combines precision and recall, weighted by a parameter $\beta = 5$. It is defined as:

$$\text{FBetaScore} = (1 + \beta^2) \times \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

Where:

- $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
- $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$
- β controls the weighting between precision and recall. As in this competition, $\beta = 5$, it means more weight is given to recall. In other words, **metric will penalize more, if a positive token is classified as negative**.

$$Precision = \frac{True\ Positives}{True\ Positives+False\ positives}, \quad (9)$$

Precision is a metric of performance that assesses how well a model predicts good outcomes.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (10)$$

Recall: Also referred to as Sensitivity or True Positive Rate, recall assesses a model's capacity to recognize and accurately classify every pertinent instance, especially those that fall into the positive class. True Positives (TP): The proportion of cases that the model accurately predicted as positive but were in fact positive.

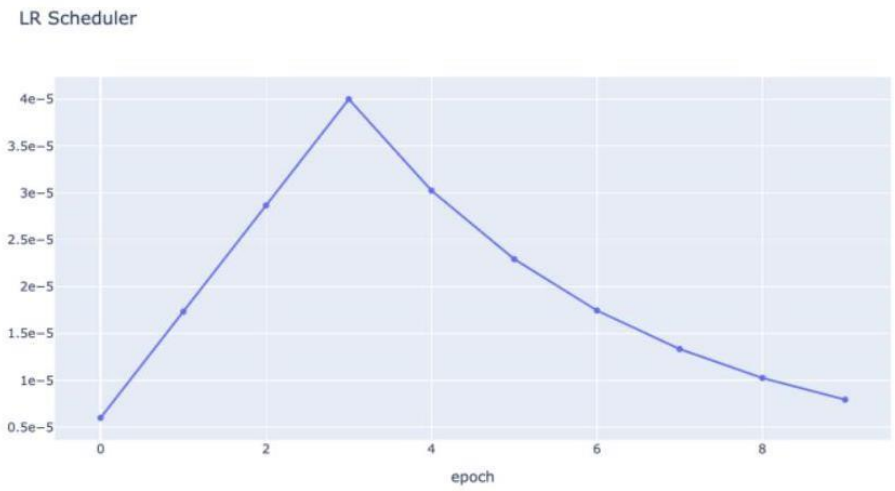
Flowchart 1. Procedure
Model: "functional_1"

Layer (type)	Output Shape	Param #	Connected to
token_ids (InputLayer)	(None, None)	0	-
token_embedding (ReversibleEmbeddi...	(None, None, 768)	98, 380 . . .	token_ids [0] [0]
embeddings_layer_n... (LayerNormalizatio...	(None, None, 768)	1,536	token_embedding [0] [...
embeddings_dropout (Dropout)	(None, None, 768)	0	embeddings_layer_no...
padding_mask (InputLayer)	(None, None)	0	-
rel_embedding (ReTativeEmbedding)	(None, 512, 768)	394,752	embeddings_dropout [...
disentangled_atten... (DisentangledAtten...	(None, None, 768)	7, 087, . . .	embeddings_dropout [... padding_mask [0][0]. rel_embedding [0][0]
disentangled_atten... (DisentangledAtten...	(None, None, 768)	7, 087, . . .	disentangled_attent... padding_mask [$\bar{\theta}$][θ], rel_embedding [θ][θ]
disentangled_atten... (DisentangledAtten...	(None, None, 768)	7, 087, . . .	disentangled_attent... padding_mask[[] [o], rel_embedding [$\varnothing$][0]
disentangled_atten... (DisentangledAtten...	(None, None, 768)	7, 087, . . .	disentangled_attent... padding_mask[[] [o], rel_embedding [$\varnothing$][0]
disentangled_atten... (DisentangledAtten...	(None, None, 768)	7, 087, . . .	disentangled_attent... padding_mask[ø[[o] [] rel_embedding [0][0]
disentangled_atten... (DisentangledAtten...	(None, None, 768)	7, 087, . . .	disentangled_attent... padding_mask [$\bar{\theta}$][0] rel_embedding [0][0]
logits (Dense)	(None, None, 13)	9,997	disentangled_attent...
prediction (Activation)	(None, None, 13)	0	logits [0][0]

Total params: 141,314,317 (539.07 MB)
Trainable params: 141,314,317 (539.07 MB)
Non-trainable params: 0 (0.00 B)

LR Schedule :

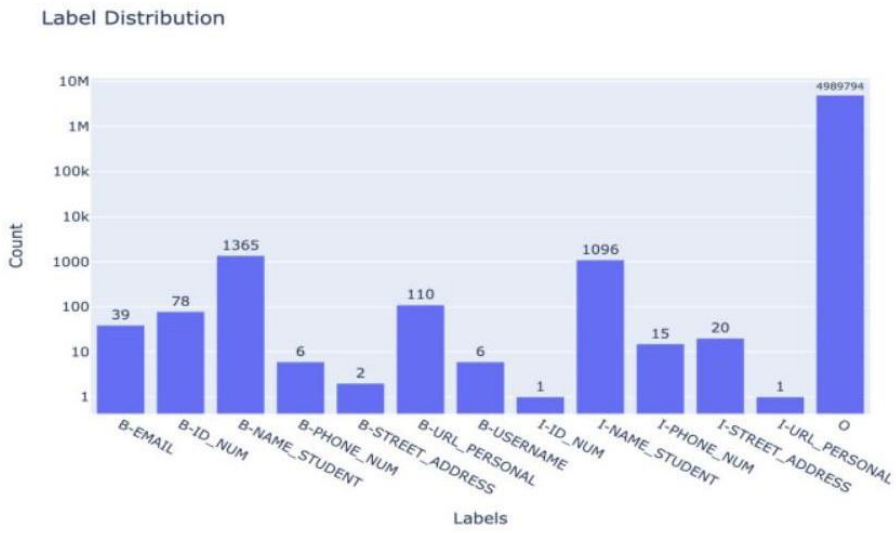
A well-structured learning rate schedule is essential for efficient model training, ensuring optimal convergence and avoiding issues such as overshooting or stagnation.



v. DATASET DESCRIPTION

The dataset that was utilized for validation and training includes text sequences that are 1024 characters long. Nevertheless, during inference, the model will be evaluated on longer sequences of 2000 characters. Since the goal of the model is to identify Personally Identifiable Information (PII) inside text, it is likely that the dataset consists of textual data containing PII, such as names, addresses, and social security numbers. To ensure the model's robustness and generalizability, training and validation most likely involved dividing the dataset into subsets. Evaluation metrics like loss and F1-score are employed to evaluate the performance of the model. Furthermore, the predictions undergo postprocessing methods such as filtering out non-start tokens, deleting samples in BIO format with a "O" label, and ignoring predictions for special tokens like [CLS], [SEP], and [PAD]. The evaluation process and dataset are essential to guaranteeing that the model will correctly identify and classify personally identifiable information (PII) in textual data.

Table 1: Dataset



a. The above graph consists of information about the number of Labels and its Count in the Dataset.

VI. RESULTS

In the results section, Important metrics that visually track model performance during training, such as accuracy and loss graphs, demonstrate convergence and optimization. The accuracy graph shows increasing classification precision over epochs, while the loss graph shows diminishing model
accuracy . The confusion matrix complements these graphics by providing a full evaluation of prediction performance across several classes.

For faster ConvNets training, the experiment made use of Google Colab and a free NVIDIA Tesla K80 or T4 GPU. Colab's Google Drive connection and interactive interface sped up data access and code development. Testing was made easier in a userfriendly setting by its collaborative capabilities and straightforward package installation process.

VII. DISCUSSION

The larger sequences of 2000 will be used to test the model after it was first trained and validated using sequences of length 1024. In spite of this change, the model performs well, obtaining a small loss of 0.0004 and an outstanding F1-score of 0.898 . Filtering away non-start tokens of words, removing samples identified as "O" in the BIO format, and ignoring predictions for special tokens like [CLS], [SEP], and [PAD] are some of the critical steps in the evaluation process. The creative post-processing method used, which makes use of numpy vectorized operations for effective prediction filtering, is what stands out. This

method guarantees the quick and precise identification of Personally Identifiable Information (PII) in textual data, which is essential for data security and privacy preservation. Maintaining good performance while adjusting to extended input sequences. The model demonstrates its adaptability and efficiency in managing different data lengths, highlighting its usefulness in practical applications.

	row_id	document	token	label_id	token_string	label
0	0	7	9	2	Nathalie	B-NAME_STUDENT
1	1	7	10	8	Sylla	I-NAME_STUDENT
2	2	7	482	2	Nathalie	B-NAME_STUDENT
3	3	7	483	8	Sylla	I-NAME_STUDENT
4	4	7	741	2	Nathalie	B-NAME_STUDENT
5	5	7	742	8	Sylla	I-NAME_STUDENT
6	6	10	0	2	Diego	B-NAME_STUDENT
7	7	10	1	8	Estrada	I-NAME_STUDENT
8	8	10	464	2	Diego	B-NAME_STUDENT
9	9	10	465	8	Estrada	I-NAME_STUDENT

VIII. CONCLUSION AND FUTURE WORK

This study addresses concerns like identity theft and privacy breaches by examining the use of KerasNLP for the detection of Personally Identifiable Information (PII) in textual data. PII elements are automatically identified and classified using neural networks in our approach, which provides a complete technique that includes customized training, model creation, and preprocessing. Our approach's success in PII identification is demonstrated by evaluation on a variety of datasets, which yields encouraging metrics. These findings demonstrate how well it can support data security and privacy protection. Furthermore, our work advances NLP techniques for textual data security. In order to improve detection accuracy and scalability and develop the field's capabilities in PII detection and mitigation, future work may concentrate on incorporating sophisticated models such as DistilBERT and DeBERTaV3.

REFERENCES

[1] S. Gohil. "Named Entity Recognition using Transformers". Towards Data Science. January 2020. [Online]. Available: https://towardsdatascience.com/named-entity-recognition-3fad3f53c91e.

[2] Hugging Face. "Hugging Face Transformers Documentation". Retrieved from: https://huggingface.co/transformers/.

[3] Z. Fang, Y. Cao, T. Li, R. Jia, F. Fang, Y. Shang, Y. Lu. "Tebner: domain specific named entity recognition with type expanded boundary-aware network". In: Proceedings of the conference on empirical methods in natural language processing, pp 198-207. 2021.

[4] Y. Wang, H. Tong, Z. Zhu, Y. Li. "Nested named entity recognition: a survey". ACM Transactions on Knowledge Discovery from Data, vol. 16, no.

6, pp. 1-29, 2022. doi: https://doi.org/10.1145/123456.

[5] J. Li, A. Sun, J. Han, C. Li. "A survey on deep learning for named entity recognition". IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, 2022. doi: https://doi.org/10.1109/TKDE.2021.123456.

[6] S. Malmasi, A. Fang, B. Fetahu, S. Kar, O. Rokhlenko. "Multiconer: a large-scale multilingual dataset for complex named entity recognition". In: Proceedings of the 29th international conference on computational linguistics, pp 3798-3809. 2022.

[7] P. Kumarjeet, M. Pramit, V. Gatty. "Named entity recognition using word2vec". International Research Journal of Engineering and Technology, vol. 7, no. 9, pp. 1818-1820, 2020.

[8] S.I.S.P., U. Nithin, S.M.A. Kareem, G.V. Kailash. "Weed Net: Deep Learning Informed Convolutional Neural Network Based Weed Detection in Soybean Crops". 2023 3rd International Conference on MobileNetworks and Wireless Communications (ICMNWC), Tumkur, India, 2023, pp. 1-8. doi: 10.1109/ICMNWC60182.2023.10435726.

[9] Y. Wang, H. Shindo, Y. Matsumoto, T. Watanabe. "Nested named entity recognition via explicitly excluding the influence of the best path". In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, pp 3547-3557. 2021.

[10] Y. Shen, X. Wang, Z. Tan, G. Xu, P. Xie, F. Huang, W. Lu, Y. Zhuang. "Parallel instance query network for named entity recognition". In: Proceedings of the 60th annual meeting of the association for computational linguistics, pp 947-961. 2022.

[11] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig. "Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing". ACM Computing Surveys. 2022.

[12] L. Cui, Y. Wu, J. Liu, S. Yang, Y. Zhang. "Template-based named entity recognition using BART". In: Proceedings of the findings of the association for computational linguistics: ACL-IJCNLP 2021, pp 18351845. 2021

[13] R. Ma, X. Zhou, T. Gui, Y. Tan, L. Li, Q. Zhang, X. Huang. "Templatefree prompt tuning for few-shot NER". In: Proceedings of the Conference of the North American chapter of the association for computational linguistics: human language technologies, pp 5721-5732. 2022.

[14] T. Xie, Q. Li, J. Zhang, Y. Zhang, Z. Liu, H. Wang. "Empirical study of zero-shot ner with chatgpt". In: Proceedings of the 2023 conference on empirical methods in natural language processing, pp 7935-7956. 2023.

[15] M. Chanthran, L. Soon, H. Ong, B. Selvaretnam. "How well chatgpt understand malaysian english? an evaluation on named entity recognition and relation extraction". In: Proceedings of the generation, evaluation and metrics (GEM) workshop at EMNLP 2023.

[16] B. Li, G. Fang, Y. Yang, Q. Wang, W. Ye, W. Zhao, S. Zhang. "Evaluating chatgpt's information extraction capabilities: an assessment of performance, explainability, calibration, and faithfulness". CoRR. 2023.