



UNIVERSITY OF NAIROBI
SCHOOL OF COMPUTING AND
INFORMATICS

**Enhancing Named Entity Recognition in
Low Resource Domains using Deep
Transfer Learning: A Case of RT&B Crop
Diseases in Scientific and Online Text.**

Leroy Mwanzia
P58/9073/2006

Supervisor: Dr. Lawrence Muchemi

July, 2023

In partial fulfillment of the requirements for the degree of *Master of
Science in Computer Science*

Declaration

Student Declaration

I certify that the thesis report submitted is my own work. It was developed by me to meet the requirements for the Master of Science in Computer Science at the University of Nairobi. I have not submitted it for academic evaluation to any other higher-learning institution before.

Signature:  Date: 18th Aug 2023

Leroy Mwanzia,
P58/9073/2006

Supervisor Declaration

I, as the University Supervisor, have approved this thesis as a partial fulfilment of the requirements for the Master of Science degree in Computer Science at the University of Nairobi.

Signature:  Date: 18th Aug 2023

Dr Eng. Lawrence Muchemi,
School of Computing and Informatics,
University of Nairobi

I dedicate this work to my Heavenly Father, whose unconditional love makes everything possible.

My heartfelt thanks go to my family for their constant encouragement during this journey and the steadfast backing from my friends and loved ones during this endeavour. Especially, my wife for her patience and understanding and my children who have always been my motivation, reminding me that it is never too late to pursue my dreams.

Finally, I honour my mother and late father, whose passion for knowledge has been a beacon of inspiration, encouraging me to reach greater heights.

Acknowledgements

I want to convey my appreciation to my supervisor, Dr. Lawrence Muchemi, for providing invaluable guidance, assistance, and mentorship during my research. His wisdom and perspectives were indispensable to the advancement and achievement of this project.

I owe a debt of gratitude to Dr. Wilmer Cuellar for his expertise and commitment to combating pests and diseases in Roots, Tubers and Bananas crops. Working with him has been an enriching and fulfilling experience. I also thank my colleague, Dr. Steve Mutuvi, for his unwavering commitment to exploring artificial intelligence methods to combat plant diseases.

I also would like to acknowledge my old and new classmates, whose companionship, intellectual discussions, and shared experiences have enriched this journey. Their camaraderie, mutual support, and moments of levity during stressful times made the difficult process of completing this thesis much more bearable and enjoyable.

Finally, I express my gratitude for the constant support and encouragement my family has provided me. I want to thank you so much for your understanding extended to my busy schedule of managing work and school. It can be quite a challenge, but the support and prayers provided are greatly appreciated. Thank you for always believing in my dreams.

Abstract

Named Entity Recognition (NER) is important in fields where researchers have to review large amounts of scientific text, such as plant pathology. However, NER is especially difficult in low-resource domains, for example, domains with little annotated textual data. Roots, Tubers and Bananas (RT&B) crop disease monitoring is one such domain. This paper investigates the promise of transfer learning to enhance the effectiveness of NER in the identification of RT&B crop disease entities.

There is an increasing number of Pretrained Large Language Models (PLLMs) that have demonstrated better performance in Natural Language Processing (NLP) tasks. This study uses transfer learning to train new models for RT&B crop disease NER. It proposes a method for transferring knowledge from large language models in resource-rich domains to smaller, low-resource domains.

By creating scientific workflows to quickly train the growing number of PLLMs and evaluate them using key metrics including non-O accuracy and the F1 score. This research demonstrates the effectiveness of transfer learning in creating effective models for RT&B crop diseases. The final model, based on SciDeBERTa, outperforms the baseline model on all metrics, especially on non-O accuracy. The results underscore the huge potential of this approach in the surveillance of crop diseases.

This research makes a contribution towards more effective Named Entity Recognition in low-resource domains. It explores current advancements in NER and the use of transfer learning in these domains. The author acknowledges the limitations of the study, such as the lack of extensive hyperparameter tuning and the unknown nature of the generalisability of the models. Finally, the study proposes continuous benchmarking of new PLLMs, comprehensive hyperparameter tuning, and exploration of data augmentation techniques to improve data availability and impact of this innovative approach as further research opportunities.

Contents

Abbreviations	ix
1 Introduction	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Aim and Objectives	4
1.3.1 General Objectives	4
1.3.2 Specific Objectives	4
1.4 Research Questions	5
1.5 Justification	5
1.6 Assumptions	5
1.6.1 Scope and Limitations	5
2 Literature Review	7
2.1 Introduction	7
2.2 Named Entity Recognition and Extraction	7
2.2.1 Evaluation metrics for NER	8
2.2.2 Approaches to NER	8
2.2.2.1 Rule-based methods	8
2.2.2.2 Machine Learning-Based Methods	9
2.2.2.3 Hybrid Methods	9
2.3 Deep Learning in NER	9
2.3.1 Neural Network Architectures for NER	9
2.3.1.1 Recurrent Neural Networks	10
2.3.1.2 Convolutional Neural Network	10
2.3.1.3 Long Short-Term Memory	10
2.3.1.4 Bidirectional Long Short-Term Memory Networks . .	10
2.3.1.5 Conditional Random Fields	10
2.3.1.6 Bidirectional Long Short-Term Memory With a Con- ditional Random Field Layer	11
2.3.1.7 Transformer-based Models	11

2.3.2	Benefits and Challenges of Deep Learning for NER	11
2.4	Transfer Learning in NER	11
2.4.1	Transfer Learning Types	12
2.4.1.1	Domain Adaptation	12
2.4.1.2	Task Adaptation	12
2.4.2	Pre-trained Language Models	12
2.4.3	Fine-tuning Techniques for NER	12
2.5	NER in Low-Resource Domains	13
2.5.1	Challenges in Low-Resource NER	13
2.5.2	Techniques for NER in Low-Resource Settings	13
2.5.2.1	Data Augmentation	13
2.5.2.2	Multi-Task Learning	14
2.5.2.3	Cross-Lingual Learning	14
2.5.2.4	Applications in Low-Resource Domains	14
2.6	Deep Transfer Learning for NER	14
2.7	NER Applications in Agriculture and Plant Pathology	15
2.7.1	Importance of NER in agriculture and plant pathology	15
2.7.2	Existing NER systems for Agricultural Text	15
2.7.3	Challenges and Opportunities in Plant Pathology	15
2.8	Conceptual Model	16
2.9	Conclusion	16
2.9.1	Summary	16
2.9.2	Gaps and research directions for the current study	16
3	Methodology	19
3.1	Introduction	19
3.2	Research Design	19
3.3	Business Understanding	21
3.4	Data Understanding	21
3.5	Data Preparation	22
3.5.1	Data Collection	22
3.5.1.1	Data Sources	22
3.5.1.2	Data Collection Process	23
3.5.1.3	Data Preprocessing	23
3.5.2	Data Annotation	23
3.5.2.1	Data Preparation for Training	24
3.5.2.2	Data Export for Training	25
3.5.2.3	Annotated Data Summary	26
3.6	Model Design and Rationale for Selection of Models	28

3.6.1	Model Training Configuration and Automation	29
3.6.2	Experiment Tracking	30
3.6.3	HuggingFace Platform	31
3.6.4	Experimental Environment	32
3.7	Model Evaluation	32
3.8	Model Deployment	34
4	Results and Discussion	35
4.1	Summary of Results	35
4.2	Introduction to Results	36
4.3	Discussion of Results	36
4.3.1	Low Resource Domains can Make NER Task Difficult	39
4.3.2	Importance of Non-O Accuracy in Disease Recognition	39
5	Conclusions	43
5.1	Introduction	43
5.2	Conclusion and Limitations	43
5.3	Final Recommendations	44
	References	55
A	Code Snippets	56
A.1	Data Acquisition	57
A.1.1	Download News Text Using GoogleNews	57
A.1.2	Download Scientific Text Using SemanticScholar	57
A.2	Data Preprocessing	58
A.2.1	Pre-annotation Text Processing	58
A.2.2	Pre-annotation Creation of Gazzettters	58
A.2.3	Annotated Data Splitting and Conversion	59
A.3	Model Training and Evaluation	59
A.3.1	Training a baseline BiLSTM-CRF model	59
A.3.2	Fine Tuning, Evaluating and Testing LLMs using Hugging Face	60

List of Figures

2.1	Conceptual Model	17
3.1	Project workflow diagram illustrating the system architecture and process flow for the RT&B crop diseases NER model.	20
3.2	Diagram demonstrating the interconnections among the different stages of CRISP-DM (Jensen, 2012)	21
3.3	Prodigy Annotation Tool	24
3.4	Prodigy Train curve diagnostics	25
4.1	F1 Score for the 10 best models compared to the baseline	38
4.2	Non-O accuracy compared to baseline	40
4.3	Confusion Matrix for the Baseline Model	41
4.4	Confusion Matrix for SciDeBerta MOdel	42

List of Tables

3.1	Summary of annotated documents, tokens, and labels	26
3.2	Counts of each label in the annotated data	27
3.3	Summary of the overall count of documents and tokens use in experi- ments.	27
3.4	Summary of the label counts for each dataset.	27
3.5	Example of Normalized Multi-class Confusion Matrix	33
4.1	Experimental Results	37

List of Abbreviations

API Application Programming Interface. 22, 29, 31, 39

BERT Bidirectional Encoder Representations from Transformers. 11, 18, 19, 28–32, 34–36

BiLSTM Bidirectional Long Short-Term Memory. 9–11, 28

BiLSTM-CRF Bidirectional Long Short-Term Memory with Conditional Random Field Layer. 9, 11, 28, 30, 36, 37, 59

BioBERT Bidirectional Encoder Representations from Transformers for Biomedical Text Mining. 29

CNN Convolutional Neural Network. 9, 10, 31

CRF Conditional Random Fields. 9–11, 28

CRISP-DM CRoss Industry Standard Process for Data Mining. 19

DeBERTa Decoding-enhanced BERT with disentangled attention. 28, 35, 38

ELECTRA Efficiently Learning an Encoder that Classifies Token Replacements Accurately. 28

FN False negatives. 32–34

FP False positives. 32–34

GPT Generative Pre-trained Transformer. 11

HMMs Hidden Markov Models. 9

IE Information Extraction. 7

IOB Inside, Outside, Beginning. 25, 26, 39

- LLM** Large Language Model. 16
- Longformer** The Long-Document Transformer. 29, 30
- LSTM** Long Short-Term Memory. 9, 10
- MTL** Multi-Task Learning. 14
- NER** Named Entity Recognition. iii, vii, 4, 5, 7–16, 18–21, 24, 25, 28, 29, 31, 32, 35, 36, 38, 39, 43–45
- NLP** Natural Language Processing. iii, 3, 7, 9, 11, 19, 28, 31, 32, 34
- NPPO** National Plant Protection Agencies. 2, 21
- PLLM** Pretrained Large Language Model. iii, 4, 5, 7, 12, 15, 18, 19, 22, 23, 28, 29, 35, 36, 38, 39, 43, 44
- PubMedBERT** BERT model pre-trained over PubMed abstracts. 29, 38
- RNN** Recurrent Neural Network. 9, 10, 31
- RoBERTa** Robustly Optimized BERT. 28
- RT&B** Roots, Tubers and Bananas. ii, iii, vii, 1–5, 16, 18–23, 28, 35, 36, 38, 39, 43, 44
- SciBERT** BERT model trained on scientific text. 29
- SciDeBERTa** SciDeBERTa model trained on scientific text. iii, 29, 35, 38, 39
- SVMs** Support Vector Machines. 9
- TN** True negatives. 32, 33
- TP** True positives. 32–34
- WandB** Weights & Biases. 29–31

Chapter 1

Introduction

1.1 Background

Roots, Tubers and Bananas (RT&B) are important food crops that are propagated vegetatively (Thiele et al., 2017) that include cassava (*Manihot esculenta*), potatoes (*Solanum spp.*), sweet potatoes (*Ipomea batata*), yams (*Dioscorea spp.*), and bananas (*Musa spp.*). In developing countries, these crops play a vital role in ensuring that people have enough food, promote good nutrition, and create income opportunities. According to (Thiele et al., 2022, 2017), about 300 million people worldwide depend on the value chains of RT&B crops. RT&B crops play a crucial role in providing the necessary nutritional and dietary energy. This is due to their significant yield and high levels of carbohydrates. They provide more energy per hectare grown than cereals (RTB, 2016), and in sub-Saharan Africa, they contribute up to 50% of the total daily calorie intake (Petsakos et al., 2019).

To understand the significance of RT&B crops in maintaining food security in sub-Saharan Africa, it is essential to take into account the impact of climate change on agriculture in this region. Where agriculture is more vulnerable due to the effects of climate change (Girvetz et al., 2019); however, RT&B crops have characteristics that increase their ability to withstand the consequences of climate change. (Prain and Naziri, 2020). Although farmers in Africa use primarily RT&B crops for subsistence, these crops are also of global importance, as they are used as animal feed or for industrial production, such as ethanol production (Petsakos et al., 2019).

RT&B crops are all vegetatively propagated, which means that the planting materials, for example, stem cuttings or vines, are genetically identical to the parent plant (Andrade-Piedra et al., 2016). Therefore, these crops share similar breeding, seed systems, and post-harvest challenges (RTB, 2016). However, a significant challenge is the widespread occurrence of pests and diseases. These outbreaks of crop diseases and pests cost farmers and consumers significant losses yearly due to yield loss and poor

harvest quality. The yield loss is estimated to be between 20% and 40% (Kreuze et al., 2022; Savary et al., 2019). Vegetatively propagated crops are particularly vulnerable to pest infestation and pathogen infections because pests and pathogens tend to build up over time with each planting cycle (Thomas-Sharma et al., 2016). Researchers predict that this problem will deteriorate as climate change increases the geographic scope of pests and diseases or increases the severity of some diseases that affect these crops (Thiele et al., 2017). The global food trade network and the evolution of new pathogens will also contribute to the spread of plant diseases (Ristaino et al., 2021).

The relevant authorities can generally manage endemic plant diseases to avoid adverse agricultural effects. However, emerging plant diseases cause large-scale plant epidemics that devastate food security (Kreuze et al., 2022; Savary et al., 2019). These plant epidemics are transboundary and can affect yield in multiple countries at the same time. RT&B crops have several ongoing large-scale outbreaks in Africa and Asia, for example, Fusarium Wilt in bananas and Cassava Mosaic Disease (CMD) in cassava (Kreuze et al., 2022).

The surveillance responsibility for monitoring crops to protect them from pandemics is in the hands of National Plant Protection Agencies (NPPO)'s in different countries. For example, Kenya's NPPO, Kenya Plant Health Inspectorate Services, is acknowledged as a centre of excellence in eastern and southern Africa (Miller et al., 2009). In practice, plant disease diagnosis networks carry out disease monitoring and surveillance. NPPOs, research universities, international research organisations, development agencies, the private sector, and farmers make up these networks. (Miller et al., 2009; Ristaino et al., 2021). For example, such networks are in place to monitor Cassava viruses in Africa and Asia. Networks traditionally perform diagnostics using field surveys that use classical techniques such as diagnosis, grafting, and mechanical inoculation. However, with DNA sequencing becoming cheaper, whole genome sequencing has been added to these traditional testing programmes (Legg et al., 2015). The ability to detect outbreaks accurately and quickly is critical to implementing effective intervention measures. Early detection can minimise the impact and threat posed by disease, and a delayed response can have significant economic, social, and ecological impacts. The year 2020, was designated as the International Year of Plant Health by the General Assembly of the United Nations. This declaration was to encourage the creation of global surveillance networks and to increase public and policy makers' awareness (Seed World, 2018).

Monitoring and controlling crop disease outbreaks is still an ongoing challenge worldwide. Plant disease surveillance is severely underfunded (Carvajal-Yepes et al., 2019). For example, global-scale surveillance is only conducted for wheat rust and late blight in potatoes (Ristaino et al., 2021). Despite the lack of funding, many modern and digital technologies are being applied to disease surveillance. This includes

geospatial and remote sensing systems, field sensors, data mining, and big data analytics, including NLP (Ristaino et al., 2021). Disease detection has also been used based on images from smartphones and drones (Kreuze et al., 2022). Disease surveillance networks are using all these methods to continuously monitor the spatial spread and incidence of pests and pathogens.

1.2 Problem Statement

Transboundary crop disease outbreaks are an ongoing challenge to food security and farmer income. Detecting outbreaks accurately and on time is critical for farmers and crop protection stakeholders to deploy efficient intervention measures. A delayed response can have significant economic, social, and ecological impacts. Researchers and crop protection around the world are advocating the use of modern and digital tools to build new disease surveillance networks (Carvajal-Yepes et al., 2019; Kreuze et al., 2022; Ristaino et al., 2021). For example, the National Academy of Sciences recently published an agricultural research agenda that underscored the need for breakthrough technologies to rapidly detect and prevent plant diseases (National Academies of Sciences, Engineering, and Medicine, 2019).

Through these efforts, digital data on crop and pest diseases will continue to grow exponentially. Although a lot of textual data is generated on crop diseases from structured sources, for example, published articles on databases such as PubMed (National Library of Medicine, 2023) or from free-form data sources such as social media or news media. According to recent research by (Ristaino et al., 2021), there is a lack of extensive attempts to utilise Natural Language Processing for tracking and charting the spread of plant diseases. In a recent survey of big data and digital tools for RT&B crops, all the tools that used machine learning for crop protection focused on image recognition, either through mobile phones or drones (Kreuze et al., 2022).

Transboundary diseases require coordinated effort because they cross political borders; however, disease surveillance methods are usually limited to specific geographical locations, and data is not always shared between political borders (Schermer et al., 2014). The availability of digital data sources provides an avenue for data-driven surveillance over borders. Advancements in Natural Language Processing (NLP) techniques have enabled the analysis of data from web sources, such as social networks, search queries, blogs, scientific literature, and online news articles for outbreak-related incidents related to diseases (O’Shea, 2017; Thomas et al., 2011). This form of surveillance involves collecting, analysing, and disseminating key information related to disease outbreaks to detect outbreaks and provide early warning to plant disease control experts. Such data-driven surveillance systems can boost the capacity of traditional surveillance approaches that rely largely on the visual and genomic identification of

crop disease outbreaks by domain experts in a local context. This can adversely affect the deployment of effective containment measures for global crop disease pandemics. For scientists and RT&B crop protection experts, the challenge becomes how to extract disease information from these growing digital text sources in an automated and efficient manner.

Relevant information that exists in these texts and that can be used for creating a disease surveillance system includes:

- Crop name
- Plant part name
- Pathogen name
- Disease name
- Symptom
- Geographic location
- Event Date
- Organisation

This information is similar to fine-grained named entities used in Agriculture NER studies, like (Malarkodi et al., 2016) and (Liu et al., 2020).

1.3 Aim and Objectives

1.3.1 General Objectives

The main objective of this research study was to develop and evaluate a deep learning model using transfer learning techniques for Named Entity Recognition to properly categorise and label specific entities associated with crop diseases in the context of Roots, Tubers and Bananas from diverse scientific literature and news media.

1.3.2 Specific Objectives

1. To design a scientific workflow that can use transfer-learning to train and evaluate multiple large-language models for the RT&B diseases NER task.
2. Find the most appropriate Pretrained Large Language Model (PLLM) that uses transfer learning to correctly recognise the named entities of RT&B crop diseases.

3. Assess how well the fine-tuned model performs in Named Entity Recognition of RT&B crop diseases in the scientific literature and online text.

1.4 Research Questions

1. What method allows for the efficient creation and execution of a scientific workflow to swiftly train and evaluate diverse large language models in deep transfer learning, especially for novel NER tasks in low-resource data domains?
2. What Pretrained Large Language Model emerges as the most suitable choice for transfer learning to produce a NER model aimed at RT&B crop diseases detection? How does the choice of this PLLM influence the effectiveness of the resultant model?
3. How does the fine-tuned model perform in terms of Named Entity Recognition for RT&B crop diseases when applied to scientific literature and online texts? What are the key factors that significantly impact this performance?

1.5 Justification

RT&B crops are crucial contributors to food security and income generation, with a particularly profound impact in Africa (RTB, 2016). Despite their importance, managing and controlling crop disease outbreaks pose formidable global challenges (Carvajal-Yepes et al., 2019). As technological advances in crop disease monitoring networks lead to a proliferation of relevant textual data for crop protection (Miller et al., 2009; Ristaino et al., 2021), the insights derived from this study are intended to significantly boost large-scale crop disease monitoring efforts. The focus is mainly on combating transboundary crop epidemics. Through the study, our objective was to facilitate the robust and continuous extraction of information from extensive online textual data sources. This approach is anticipated to enhance targeted crop protection initiatives and provide valuable support for resource-constrained crop disease networks, thus contributing to a more sustainable and secure agricultural future.

1.6 Assumptions

1.6.1 Scope and Limitations

1. The research study limited the evaluation of NER to diseases affecting only five RT&B crops: Cassava, Banana, Plantain, Potato, and Sweet Potato.

2. The study collected scientific data from abstracts available in the Semantic-Scholar open access literature database. However, the methodology can be applied to other databases such as PubMed or Google Scholar. The news items were collected from free news media indexed on Google News and refined using the crop name as the keyword search.
3. The project extracted data from online text and documents created digitally by the authors. This study did not use scanned PDF documents that required optical character recognition for text extraction.

Chapter 2

Literature Review

2.1 Introduction

The ability to rapidly detect the spread of crop epidemics is an integral part of crop disease surveillance networks. The growing availability of digital data in online sources provides an avenue for data-driven cross-border surveillance. Advancements in Natural Language Processing (NLP) techniques have enabled it to analyse data from Web sources, such as social networks, search queries, blogs, scientific literature, and online news articles for outbreak-related incidents related to diseases (O’Shea, 2017; Thomas et al., 2011). This review of the literature revolves around how deep transfer learning can be used to extract entities that are relevant to the monitoring and surveillance of crop diseases. The study focusses on understanding the developments in the natural language process, the emergence of Pretrained Large Language Model, and how these have been used in transfer learning to enhance deep learning models in areas with little training data. This review also focusses on NER in the agricultural sector.

2.2 Named Entity Recognition and Extraction

Natural Language Processing (NLP) explores the methods by which machines interpret and interact with human language in text or spoken form to achieve practical objectives (Chowdhary, 2020). A subset of NLP is Information Extraction (IE). Information Extraction works to automatically extract data from natural text to populate a structured database (Gaizauskas and Wilks, 1998). Named Entity Recognition (NER) is an essential component of Information Extraction (IE). At the Message Understanding Conference 6 (MUC-6) The term ”named entity”, a word that recognises elements with similar characteristics within a superset of elements, was introduced as part of Information Extraction (IE) as(Grishman and Sundheim, 1996). A named entity is termed a rigid designator. Elements such as person names, dates, and prod-

uct names can be considered entities. The goal of Named Entity Recognition is to recognise the mentions of these identifiers that belong to a predetermined text class (Nadeau and Sekine, 2007). The performance of NER tasks can be affected by factors such as language, type of entity and domain, for example biomedical or agriculture. Nested entities, ambiguity in the text, and the amount of annotated training data are challenges for NER (Goyal et al., 2018).

The goal of NER is to identify specific terms that are part of a predetermined category in a given text.

2.2.1 Evaluation metrics for NER

Evaluating the performance of NER models typically involves measuring precision, recall, and the F1 score, which provides a balanced assessment of the model’s accuracy in terms of false positives and false negatives. (Sang and De Meulder, 2003). Precision quantifies the fraction of named entities accurately pinpointed by the model of all entities predicted. In contrast, recall quantifies the proportion of named entities accurately detected relative to all the actual named entities present in the dataset. The F1 score is a measurement that combines recall and precision through the harmonic average. It helps balance the trade-off between these two factors into a single value. (Chinchor and Robinson, 1997). These evaluation metrics are often calculated for each named entity type and then averaged to assess the model’s performance across all entity types.

2.2.2 Approaches to NER

There are three main groups into which NER techniques can be broadly divided: hybrid, machine learning-based and rule-based methods.

2.2.2.1 Rule-based methods

Rule-based methods for NER depend on manually created rules, patterns, and dictionaries to classify entities in the text (Grishman, 1995). These methods often involve regular expressions or pattern matching techniques to capture specific syntactic or morphological structures associated with named entities (Chinchor and Robinson, 1997). Although rule-based methods can achieve high precision, they may require an improvement in recall due to the difficulty in creating comprehensive rules and dictionaries that encompass all potential variations of named entities (Nadeau and Sekine, 2007).

2.2.2.2 Machine Learning-Based Methods

Machine learning approaches for Named Entity Recognition use supervised learning strategies to autonomously discern patterns and characteristics from annotated datasets (Lafferty et al., 2001). Commonly used machine learning algorithms for NER include Support Vector Machines (SVMs), Conditional Random Fields (CRF) and Hidden Markov Models (HMMs) (McCallum and Li, 2003). More recently, deep learning-based methods, such as Recurrent Neural Network (RNNs), Long Short-Term Memory (LSTM) networks and Transformer-based models, have attained top-tier results on different NER benchmarks (Devlin et al., 2019; Lample et al., 2016).

2.2.2.3 Hybrid Methods

Hybrid methods utilise a combination of rule and machine learning based approaches to take advantage of the strengths of both techniques and mitigate their weaknesses (Finkel et al., 2005). These methods typically involve the use of rules to generate features or initial annotations, which are refined or combined with machine learning-based methods to produce the final output NER (Nadeau and Sekine, 2007). Hybrid methods can achieve improved performance by combining the high precision of rule and machine learning based approaches possessing the ability to adapt and generalise effectively.

2.3 Deep Learning in NER

Deep learning falls under the machine learning umbrella and emphasises layered data interpretations using multi-layered artificial neural networks (LeCun et al., 2015). These deep architectures allow models to extract intricate and high-level characteristics directly from raw data, making them particularly effective for tasks that involve large volumes of unstructured data, such as NLP and image and speech recognition (Goodfellow et al., 2016). Deep learning has significantly advanced NER in recent years, outperforming traditional rule and machine based techniques in various benchmarks and application areas (Lample et al., 2016; Ma and Hovy, 2016).

2.3.1 Neural Network Architectures for NER

Several deep learning architectures have been employed for NER tasks, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Long Short-Term Memorys (LSTMs), Bidirectional Long Short-Term Memorys (BiLSTMs), Conditional Random Fieldss (CRFs), Bidirectional Long Short-Term Memory with Conditional Random Field Layers (BiLSTM-CRFs), and transformer-based models.

2.3.1.1 Recurrent Neural Networks

RNNs fall under a class of neural networks designed especially for handling sequential data, preserving hidden states that retain details from prior stages (Elman, 1990). RNNs have been applied to NER tasks to capture contextual information and long-range model dependencies in input text (Chiu and Nichols, 2016).

2.3.1.2 Convolutional Neural Network

CNNs are a type of neural network which employs convolutional layers to capture local features in input data by applying filters (Lecun et al., 1998). Although originally designed for image recognition, CNNs underwent modifications for NER tasks by interpreting the text in a series of characters or words, using convolutions to grasp the local context and characteristics (Collobert et al., 2011).

2.3.1.3 Long Short-Term Memory

LSTMs are a type of RNNs that addresses the problem of vanishing gradients, which occurs when the network cannot learn long-term dependencies because of the decrease in gradients during training (Hochreiter and Schmidhuber, 1997). LSTMs utilised in NER tasks aim to more effectively grasp distant relationships and context within input sequences (Lample et al., 2016).

2.3.1.4 Bidirectional Long Short-Term Memory Networks

BiLSTMs, which extend LSTMs, process input series from both the preceding and the succeeding directions. This facilitates the model's ability to comprehend context from both earlier and upcoming tokens (Graves et al., 2005). BiLSTMs have been successfully applied to NER tasks, demonstrating improved performance over unidirectional LSTMs (Huang et al., 2015).

2.3.1.5 Conditional Random Fields

CRFs constitute a discriminative probabilistic graphical model, adept at articulating the interdependencies between input variables and their corresponding outputs in a structured prediction task, such as NER (Lafferty et al., 2001). CRFs have been combined with deep learning architectures, such as LSTMs and CNNs, to improve NER performance by capturing the dependencies between adjacent named entity labels (Ma and Hovy, 2016).

2.3.1.6 Bidirectional Long Short-Term Memory With a Conditional Random Field Layer

The BiLSTM-CRF model combines the strengths of BiLSTMs and CRFs. It employs BiLSTM to assimilate contextual specifics, while the CRF layer is used to articulate the interrelationships among the labels assigned to the named entities (Huang et al., 2015). This combination has outperformed individual BiLSTM or CRF models in NER tasks by effectively modelling both the input sequence context and the relationships between output labels (Ma and Hovy, 2016).

2.3.1.7 Transformer-based Models

Transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), are a family of deep learning architectures that rely on self-attention mechanisms to harness details of the context in input sequences (Devlin et al., 2019; Radford et al., 2018; Vaswani et al., 2017). These models achieved top-tier results in numerous NLP undertakings, such as NER, by refining pre-established models using data tailored for distinct applications (Devlin et al., 2019; Matthew et al., 2018).

2.3.2 Benefits and Challenges of Deep Learning for NER

Deep learning offers several benefits for NER tasks, including the ability to learn complex representations from raw data, capture long-range dependencies, and incorporate contextual information in input sequences (Goodfellow et al., 2016). Moreover, one can pre-train deep learning architectures utilising vast datasets, allowing them to leverage prior knowledge and achieve better performance on downstream tasks, such as NER, with limited labelled data (Devlin et al., 2019).

However, deep learning for NER also presents challenges, such as increased computational complexity, model interpretability, and the need for large amounts of labelled data for training (Zhang et al., 2015). In spite of these hurdles, progress in deep learning methodologies, especially transfer learning and unsupervised pre-training, have contributed to mitigating certain data requirements and computational limitations, making deep learning-based NER models an increasingly popular choice for researchers and practitioners (Matthew et al., 2018).

2.4 Transfer Learning in NER

Transfer learning involves using knowledge gained from one domain or machine learning task and applying it to another related function or domain. This process often

results in improved performance with fewer training data (Pan and Yang, 2010). In the context of NER, transfer learning allows models to leverage representations or structures pre-trained on expansive text databases. This decreases the volume of annotated data necessary for the intended NER activity (Ruder et al., 2019a).

2.4.1 Transfer Learning Types

Transfer learning can be broadly categorised into domain adaptation and task adaptation.

2.4.1.1 Domain Adaptation

Domain adaptation entails conveying knowledge acquired from an original domain to a desired domain that may have different data distributions to enhance the efficacy of the model within the desired domain (Pan and Yang, 2010). In NER, domain adaptation can benefit low-resource settings or when the target domain has unique linguistic features or terminology that differs from the source domain (Ruder et al., 2019a).

2.4.1.2 Task Adaptation

Task adaptation focusses on conveying knowledge acquired from an original task to enhance the results of a related target activity (Pan and Yang, 2010). In the context of NER, task adaptation can involve the use of pre-trained language models or embeddings developed on extensive unsupervised datasets or other NLP tasks and fine-tuning them on the NER task of interest (Devlin et al., 2019; Matthew et al., 2018).

2.4.2 Pre-trained Language Models

The introduction PLLM, such as BERT, GPT, and ELMo, has considerably advanced the field of NER by allowing models to exploit rich contextualised representations learnt from large-scale unsupervised data (Devlin et al., 2019; Matthew et al., 2018; Radford et al., 2018). These pre-trained models have achieved gold standard performance on various NER benchmarks by capturing long-range dependencies and semantic relationships in text, reducing the demand for manually designed features and extensive labelled data (Akbik et al., 2019).

2.4.3 Fine-tuning Techniques for NER

Fine-tuning involves tailoring a PLLM to a designated NER activity by modifying the model parameters based on data specific to that task (Howard and Ruder, 2018).

Techniques for fine-tuning in the context of NER usually require running the model for several iterations using a reduced learning rate to prevent the erasure of the pre-learned information (Ruder et al., 2019a). Various strategies, such as layer-wise learning rate schedules, differential learning rates, and freezing specific layers during training, have been proposed as methods to enhance the fine-tuning procedure, ensuring the efficient transition of pre-trained insights to the specific NER objective (Howard and Ruder, 2018; Ruder et al., 2019a).

2.5 NER in Low-Resource Domains

2.5.1 Challenges in Low-Resource NER

NER in low-resource domains face several challenges that stem from the lack of available data and the unique characteristics of these domains. These challenges include data scarcity, limited annotated data, domain-specific language, and the complexity of handling rare entities (Pan et al., 2017). Some studies have shown that there are few corpora of agricultural-related documents tagged (Patil et al., 2013). Furthermore, creating annotated data in low-resource domains can be time consuming and expensive, requiring expert knowledge and manual annotation efforts (Ruder et al., 2019b). These challenges have motivated researchers to explore alternative techniques to improve NER performance in low-resource settings.

2.5.2 Techniques for NER in Low-Resource Settings

Numerous strategies have been put forth to tackle the difficulties of NER within low-resource domains. These techniques include data augmentation, multitask learning, and cross-lingual learning.

2.5.2.1 Data Augmentation

Techniques for data augmentation create new training instances by making varied modifications to the original data, including substituting with synonyms, randomly adding or deleting words, or interchanging their positions (Wei and Zou, 2019). In the context of NER, model performance can be improved by using data augmentation to enhance the quantity and diversity of training data. For instance, (Li et al., 2020) suggested an iterative data augmentation method that merges a rule-driven system with a neural network architecture to automatically generate labelled data for low-resource NER tasks.

2.5.2.2 Multi-Task Learning

Multi-Task Learning (MTL) is an approach where several tasks are trained concurrently to enhance the capability of generalising effectively (Caruana, 1997). In low-resource NER, MTL can take advantage of the commonalities between related tasks, for example, chunking, part-of-speech tagging and NER, to enhance the performance of the model (Plank et al., 2016). For instance, (Bingel and Sgaard, 2017) demonstrated that MTL could improve NER performance in low-resource languages by jointly learning-related tasks.

2.5.2.3 Cross-Lingual Learning

Cross-lingual learning involves the transfer of knowledge learnt from one language to another. In the context of low-resource NER, cross-lingual learning can help leverage the knowledge obtained from high-resource languages to enhance NER accuracy in lowly-resourced languages (Ruder et al., 2019b). As an example, (Conneau et al., 2018) proposed XNLI, which uses a pre-trained sentence encoder to transfer the knowledge from high to low resource languages for various NLP activities, such as NER.

2.5.2.4 Applications in Low-Resource Domains

Several applications have demonstrated the effectiveness of techniques such as data augmentation, cross-lingual learning, and multitask learning when addressing NER in poorly resourced domains. In the biomedical field, a study by (Yang et al., 2016), a data augmentation method that combines distant supervision and active learning, was proposed to enhance the efficacy of NER within the biomedical field. By adopting this approach, the model was able to learn from a limited amount of annotated data, which ultimately improved its overall generalisation capabilities. Another example is from low-resource languages, where (Lin et al., 2016) developed a multitask learning approach for NER that jointly learns NER and dependency parsing. The research indicated that the model could yield outstanding outcomes across multiple lowly-resourced languages, demonstrating the benefits multitasking learning brings to low-resource settings.

2.6 Deep Transfer Learning for NER

Deep transfer learning has shown its effectiveness in enhancing NER performance in various domains, including biomedical, legal, and environmental domains. Within the biomedical field, (Zhou et al., 2018) applied a multitask learning framework to enhance NER efficacy for biomedical items like genes, diseases, and chemicals. By concurrently training the model on various NER tasks, it attained top-tier results on benchmark

datasets. This approach highlights the potential of multitask learning to capture domain-specific language patterns and improve NER performance in specialised areas such as scientific domains. Regarding the Legal field, (Chalkidis et al., 2020) proposed Legal-BERT, a model fine-tuned using legal text corpora, which outperformed the original BERT model and conventional NER techniques in recognising legal concepts named entities. This work emphasises the importance of transfer learning of PLLMs for NER tasks in specialised domains.

2.7 NER Applications in Agriculture and Plant Pathology

2.7.1 Importance of NER in agriculture and plant pathology

Named Entity Recognition (NER) has significant potential in agriculture and plant pathology, as it allows the extraction of vital information from a significant amount of unstructured textual data. By identifying and classifying entities such as crop names, diseases, pests, and treatments, NER can facilitate the development of decision support systems, early warning systems, and advanced research methodologies in agriculture.

2.7.2 Existing NER systems for Agricultural Text

Several studies have explored the application of NER in agricultural text, focussing mainly on recognising diseases and pests in different languages and contexts. For example, a study by Li et al. (2019) investigated NER methods for Chinese forest disease texts, conversely, research by (Guo et al., 2020) explored NER methods tailored for Chinese agricultural texts for pests and diseases. The authors used multi-scale local context attributes coupled with a self-attention approach. Similarly, (Guo et al., 2021) suggested a method rooted in adversarial contextual embeddings for NER named ACE-ADP for agricultural diseases and pests. In another study, (Zhang et al., 2021) employed character augmentation to improve Chinese NER performance for apple diseases and pests. These works demonstrate the potential of NER in processing agricultural text data to provide insights into sector-specific trends and issues.

2.7.3 Challenges and Opportunities in Plant Pathology

One of the primary challenges in applying NER to agriculture and plant pathology is the need for domain-specific labelled data for model training. Furthermore, agricultural text often contains specialised terminology and complex relationships between

entities, making it difficult for general-purpose NER systems to perform well in this domain.

Despite these challenges, there are promising opportunities for NER in agriculture and plant pathology. Developing domain-specific NER systems can lead to better decision-making, early warning, and research methodologies. For instance, in research carried out by (Jiang et al., 2021) on fine-tuning BERT-based frameworks to compile plant health reports, revealed the efficacy of employing pre-trained linguistic models in classifying agricultural texts. Furthermore, integrating NER with other data sources, such as remote sensing, geospatial data, and expert knowledge, can result in more comprehensive information systems for agriculture and plant pathology.

2.8 Conceptual Model

The pictorial conceptual model (Jrvelin and Wilson, 2003) is a diagrammatic representation of the framework that illustrates the various components of the suggested NER framework will interact. This study collected data for model training from open-literature databases and online media. The data are preprocessed, annotated and used to pre-train different LLMs. The target Large Language Model knowledge is fin-tuned with our dataset and used for the new NER task of RT&B crop diseases.

2.9 Conclusion

2.9.1 Summary

This review of the literature covered the essential aspects of Named Entity Recognition (NER), focussing on low-resource domains, deep learning, transfer learning, and applications in agriculture and plant pathology. It discussed the definitions and tasks related to NER, multiple strategies related to NER, including hybrid, rule and machine learning based methods, and their evaluation metrics. The review also examined the impact of deep learning on NER, presenting different neural network architectures and their benefits and challenges. Furthermore, it addressed the concept of transfer learning, its types, and its application to NER. The review then delved into the challenges and techniques in low-resource NER and examined the importance, existing systems, and challenges of NER in agriculture and plant pathology.

2.9.2 Gaps and research directions for the current study

While the review of the literature provides a comprehensive overview of NER and its applications in agriculture and plant pathology, it also highlights several gaps and

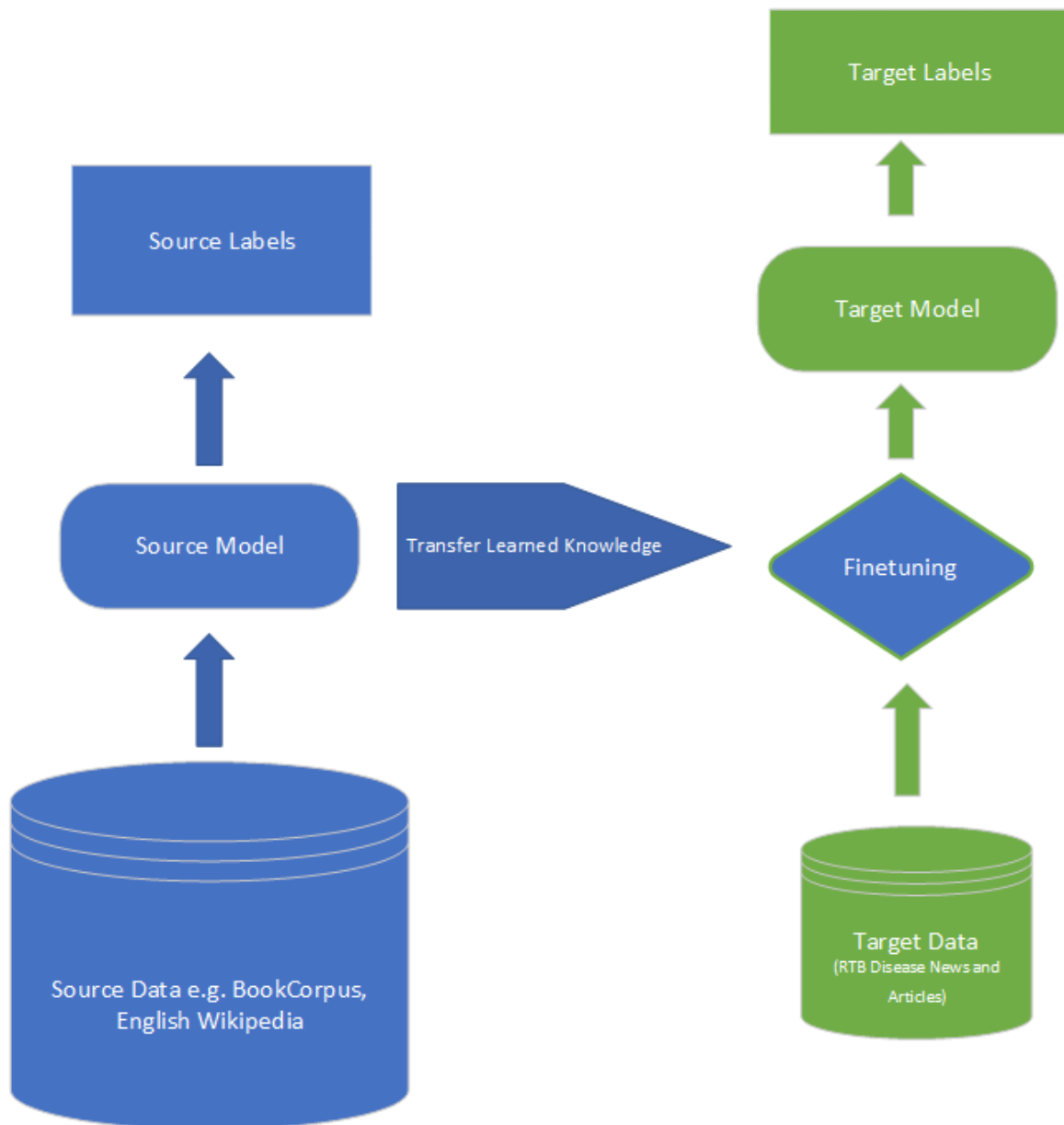


Figure 2.1: Conceptual Model

research directions that were addressed in the current study:

Domain-specific adaptation: Despite advances in deep learning and transfer learning for NER, there remains a need for domain-specific adaptation to improve the recognition and classification of entities in low-resource domains, Roots, Tubers and Bananas crop diseases.

Cross-lingual and low-resource language support: Most existing NER models and techniques have been developed for high-resource languages and domains. This fact provides an opportunity for further research and development of NER systems that can effectively handle low-resource languages and data domains. In this study we apply transfer learning to a low resource data domain of RT&B crop diseases.

Integration of deep learning and transfer learning: By combining deep learning architectures with transfer learning methods, the performance of NER models can be greatly improved in low-resource domains. The author investigated the employment of PLLMs, such as BERT and its successors, combined with transfer learning. Specifically, we use fine-tuning techniques and provide valuable insights into the adaptability and efficiency of these models within the framework of root, tuber, and banana crop diseases.

Addressing these gaps and research directions contributes to the advancement of NER models and techniques in the context of Roots, Tubers and Bananas crop diseases, ultimately benefiting agricultural research and practise in this low-resource data domain.

Chapter 3

Methodology

3.1 Introduction

This chapter outlines the concepts, principles, procedures, and techniques employed in this research. It outlines the process of collecting and analysing data and the model design. Primary data was obtained from open online databases and news sites and annotated by a researcher working in crop protection. Furthermore, it provides details of the experimental structure utilised to both train and evaluate different models. The study selected several BERT-based models, which have been shown to work well for NLP (Devlin et al., 2019), and used the transfer learning approach to create a deep learning architecture that detects named entities of Roots, Tubers and Bananas crop diseases from scientific and online texts. After the fine-tuning process, the chapter also details the method used to find the most appropriate Pretrained Large Language Model for NER and how the model was validated.

3.2 Research Design

The primary emphasis of the research centered on developing and implementing the RT&B crop diseases NER model. The study used the Cross Industry Standard Process for Data Mining (CRISP-DM) research process for mining data. This process was developed by a group of leading data mining suppliers and users, such as DaimlerChrysler, SPSS, NCR and OHRA (Wirth and Hipp, 2000) (see Figure 3.2). We use transfer learning to fine-tune selected models from the BERT family of transformer-based models to achieve this. The study was able to formulate a deep learning architecture which uses the power of these models for our specific task. (Vaswani et al., 2017). The model has been successfully deployed online for inference, allowing greater accessibility and more efficient use for the RT&B crop diseases NER task. The following is the NER project workflow diagram (refer to Figure 3.1).

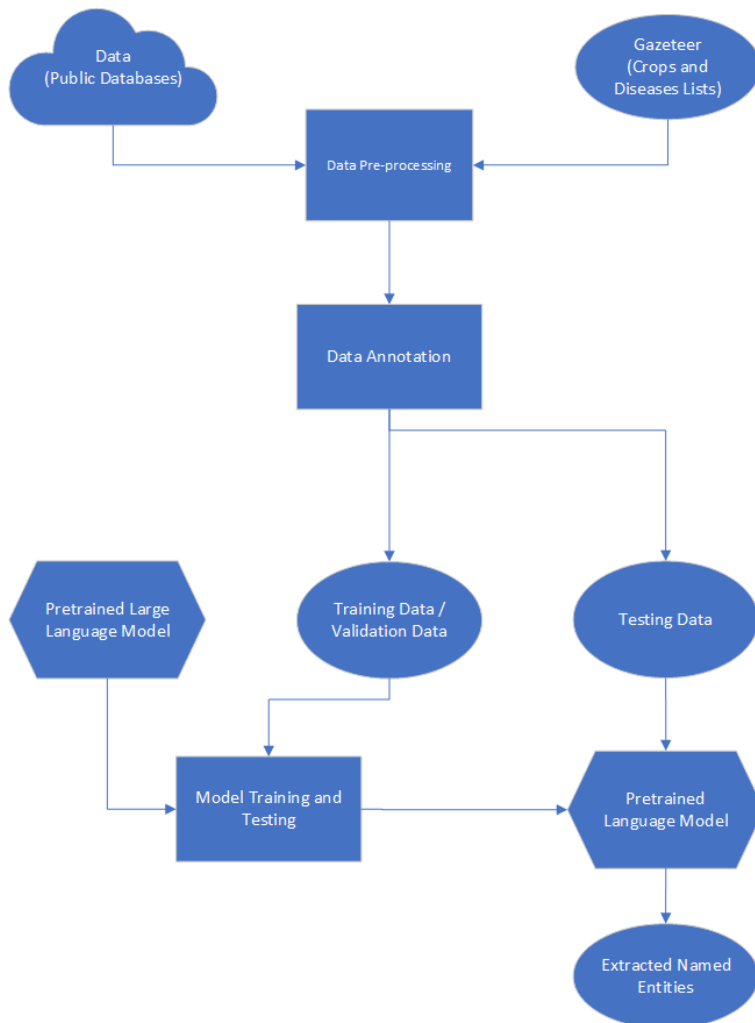


Figure 3.1: Project workflow diagram illustrating the system architecture and process flow for the RT&B crop diseases NER model.

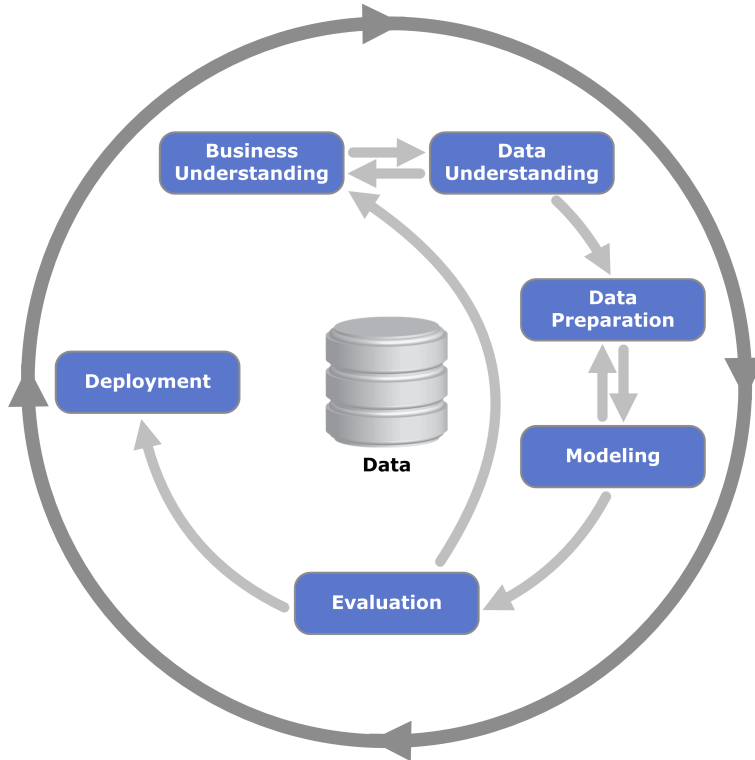


Figure 3.2: Diagram demonstrating the interconnections among the different stages of CRISP-DM (Jensen, 2012)

3.3 Business Understanding

RT&B crop disease monitoring networks collect much textual data relevant to crop protection to combat transboundary crop disease outbreaks (Miller et al., 2009; Ristaino et al., 2021). These data collected by NPPOs, CGIAR centres, stakeholder organisations and farmers are geotagged data on pests and diseases that affect many food crops, including the RT&B crops. These data are stored online in databases of scientific literature such as SemanticScholar (The Allen Institute for Artificial Intelligence, 2022) or in unstructured data sources such as social media or news media. From this information, a rapid and efficient understanding of the spread of crop pandemics allows crop protection stakeholders to act quickly on new outbreaks and reduce the chances of spreading or reintroducing pests and pathogens. Given the availability of these textual data and the lack of textual annotated data in agriculture (Patil et al., 2013). The study’s objective was to use transfer learning using an annotated dataset from this source to create a NER model for recognising disease entities in text.

3.4 Data Understanding

Open-access research abstracts related to Roots, Tubers and Bananas crops were retrieved from the SemanticScholar database using keyword searches. The data set was

extensively analysed to identify articles that contained specific entities targeted for extraction as positive examples and those that did not have disease-related information as negative examples. A similar process was used to obtain news articles, utilising Google News keyword searches.

Both scenarios involved the utilisation of Python to perform the search queries and download the scientific text. See Appendix A for more details. Only English texts were considered for this study.

3.5 Data Preparation

3.5.1 Data Collection

As noted in Patil et al. (2013), there is a noticeable lack of tagged corpora that focus specifically on agriculture-related documents. Therefore, this study collected a relatively small but significant dataset consisting of scientific and online texts that are relevant to crop diseases associated with Roots, Tubers and Bananas. The collected data were used as the basis for the creation of a corpus for experimentation. The data was first used to train a baseline model and later used for transfer learning, specifically, for fine-tuning the Pretrained Large Language Model (PLLM)s.

To accomplish this, data were collected from two primary sources: SemanticScholar and Google News. This subsection outlines the data collection process, the pre-processing techniques used, and the final format for storing the acquired data.

3.5.1.1 Data Sources

SemanticScholar (The Allen Institute for Artificial Intelligence, 2022) served as the main data source for this study. It is an AI-powered research tool that assists researchers in discovering pertinent publications and extracting information from an extensive corpus of scientific literature. Data was collected from SemanticScholar using their API, which grants access to a plethora of metadata and abstracts related to the research topic.

Google News, an online news aggregation service that compiles and presents news articles from various sources, was also used as a data source. Google News API was used to collect links to the data. Python was used to download the information contained in the news articles, their headlines and snippets relevant to the research topic from the source news website.

3.5.1.2 Data Collection Process

Python is used as the primary programming language for data collection. Requests are made to the SemanticScholar and Google News APIs, which return relevant publications and news articles as JSON objects. A set of search queries is designed to retrieve the maximum number of pertinent documents. Keywords such as "*cassava diseases*", "*potato diseases*", and "*banana crop diseases*" were used to ensure comprehensive coverage. The data collection scripts are available in Appendix A.

3.5.1.3 Data Preprocessing

After obtaining the raw data, it was necessary to preprocess the text to ensure that it was in a suitable format for annotation. The study developed a Python script to clean and preprocess the data. The script performed the following tasks:

- **Text Cleaning:** The script was designed to eliminate unnecessary whitespaces and non-standard symbols from the text. This ensured that the text was consistent with UTF-8 encoding, which is crucial for the subsequent processing and analysis stages. This step also improved the readability and interpretability of the data, facilitating more accurate annotation.
- **Storage Format and Annotation Tool:** After preprocessing, the data were stored in the JSON Lines (JSONL) format, which is convenient for handling large volumes of text data. Each line in the JSONL file represents a single document or article and includes the required fields for annotation. Every JSONL data entry was split into two segments: the text field, which held the subject matter and a unique identifier field. Scientific articles had a DOI identifier, whereas news articles had a URL identifier. The JSONL format is compatible with Prodigy (ExplosionAI, 2023), a popular annotation tool used in this study to annotate named entities related to RT&B diseases.

The preprocessing script was designed to keep the data as close to the original format as possible, ensuring that the context and meaning of the text were preserved. The data obtained served as the basis for the subsequent stages of the study. The details of the Python script are described in Appendix A.

3.5.2 Data Annotation

We annotated the data using the Prodigy annotation tool (ExplosionAI, 2023); see Figure 3.3 for an example of the annotator screen. Annotated data was used to train a base model and fine-tune PLLMs as a transfer learning technique. Prodigy provides a free research licence to bonafide research students.

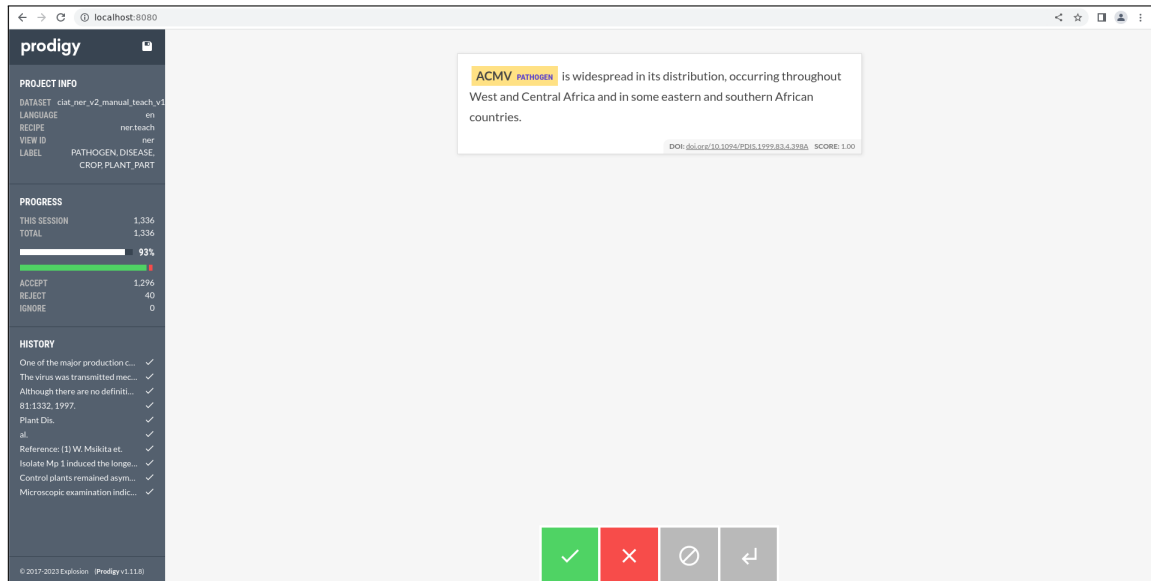


Figure 3.3: Prodigy Annotation Tool

3.5.2.1 Data Preparation for Training

Initially, gazetteers containing plant names, diseases, and other key information were used to partially preannotate a small set of 50 abstracts using Prodigy. These abstracts were then manually reviewed, further annotated, and corrected as necessary. These manually annotated data were exported from Prodigy and used to train a named entity recognition model using the spaCy library (Honnibal and Montani, 2021). The study used this intermediate model to improve the data annotation process through active learning. The goal was to suggest entities to the annotators who could verify the predictions’ accuracy and make changes or additions. This saved time compared to annotating all the text without any suggestions. To assess whether increasing the volume of data would increase the model’s efficacy, the initial spaCy model was trained incrementally with data amounts of 25%, 50%, 75% and 100%. This was accomplished using Prodigy’s train-curve recipe, which sends chunked data to spaCy. The aim was to test the viability of creating a model that could effectively learn the new entities.

Following this process, a total of 300 abstracts and 256 news items were annotated for the study. These annotated data served as the basis for the subsequent stages of the investigation, offering a firm foundation for developing and assessing the NER model. The study used a fine-grained set of tags adopted from others that have been used in NER research in agriculture, like (Liu et al., 2020; Malarkodi et al., 2016).

- Crop name
- Plant part name
- Pathogen name

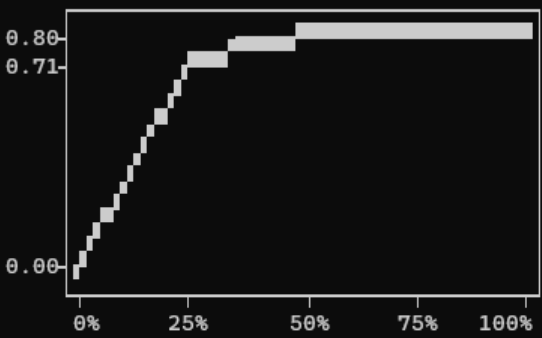
- Symptoms
- Disease name
- Geographic location
- Event Date
- Organization

```
(deep_gpu) leroy@CIATKEIT01820L:~/Dev/Code/prodigy_data_pdp$ prodigy train-curve
--ner ciat_ner_v1 --gpu-id 0 --show-plot

===== Generating Prodigy config =====
■ Auto-generating config with spaCy
✓ Generated training config

===== Train curve diagnostic =====
Training 4 times with 25%, 50%, 75%, 100% of the data

%      Score      ner
----  -
0%     0.00      0.00
25%    0.71 ▲    0.71 ▲
50%    0.78 ▲    0.78 ▲
75%    0.79 ▲    0.79 ▲
100%   0.80 ▲    0.80 ▲
```



```
✓ Accuracy improved in the last sample
As a rule of thumb, if accuracy increases in the last segment, this could
indicate that collecting more annotations of the same type will improve the
(deep_gpu) leroy@CIATKEIT01820L:~/Dev/Code/prodigy_data_pdp$
```

Figure 3.4: Prodigy Train curve diagnostics

3.5.2.2 Data Export for Training

A set of data transformation steps was added to the workflow to prepare the data for the subsequent stages of this study. An essential transformation involved converting the data into the Inside, Outside, Beginning (IOB) format. The structured delineation of each token's position in a named entity makes this format commonly used in Named

Entity Recognition tasks. The generation of the IOB format was facilitated by a Python script, the details of which are described in Appendix A.

This script was designed with flexibility and adaptability in mind and was able to export the IOB format with different separators to meet the requirements of an array of machine-learning models and libraries. In addition, the script has the ability to divide annotated data into training, validation, and testing batches. It is crucial to follow this step when creating machine learning models. This involves fitting the model to a specific dataset (training set), fine-tuning the model for transfer learning using another set (development set), and then assessing its performance on a completely distinct dataset (validation set) that the model has not encountered before. By following the above process, the model’s performance and capability to apply knowledge to data that have not been previously encountered can be reliably assessed.

3.5.2.3 Annotated Data Summary

After going through the pre-processing and transformation steps, the data used in this study were represented in terms of the various data labels and their corresponding counts. These details are summarised in Table 3.1. The data was also classified according to specific labels, providing a more granular view of the data distribution. This categorisation is presented in Table 3.4. These tables give a summary of the data and offer key insights into the characteristics and composition of the dataset employed in this research.

Data/Label	Count
Total number of annotation documents	556
Total number of tokens	289387
CROP	5996
LOC	1498
PLANT_PART	1576
GPE	3615
DATE	2496
DISEASE	1801
SYMPTOM	620
PATHOGEN	2409
ORG	2171

Table 3.1: Summary of annotated documents, tokens, and labels

Below is an overview of the data set partitioned into three distinct subsets, training, validation, and evaluation, with a proportion of 75%, 15%, and 15%, in that order. By using this partitioning strategy, it was possible to train, fine-tune, and assess how well the model performs on new and previously untested data. It also provided insight into the composition and balance of the data used during every phase of

Label	Count
B-CROP	5996
I-CROP	1240
B-LOC	1498
I-LOC	1622
B-PLANT_PART	1576
B-GPE	3615
I-GPE	909
B-DATE	2496
I-DATE	2511
B-DISEASE	1801
I-DISEASE	2177
B-SYMPTOM	620
I-SYMPTOM	2479
B-PATHOGEN	2409
I-PATHOGEN	2760
B-ORG	2171
I-ORG	3803
I-PLANT_PART	29
O	249675

Table 3.2: Counts of each label in the annotated data

the model’s evolution. It is important to highlight the significance of this partitioning approach in ensuring the generalisability and robustness of the model. More research can be conducted to understand and improve its efficiency.

Dataset	Number of Documents	Number of Tokens
Train	389	204579
Validation	84	37123
Test	83	47685

Table 3.3: Summary of the overall count of documents and tokens use in experiments.

Dataset	Train	Validation	Test
DISEASE	1345	235	221
CROP	4257	756	983
DATE	1644	345	507
ORG	1479	260	432
GPE	2498	475	642
PLANT_PART	1123	215	238
LOC	1027	185	286
PATHOGEN	1724	375	310
SYMPTOM	454	88	78

Table 3.4: Summary of the label counts for each dataset.

3.6 Model Design and Rationale for Selection of Models

To design the NER model, we started by training a Bidirectional Long Short-Term Memory with Conditional Random Field Layer (BiLSTM-CRF) model, that merges the best features of BiLSTMs and CRFs, making it a popular choice for NER tasks (Huang et al., 2015). This model has been shown to exceed the performance of individual BiLSTM or CRF models in NER tasks by efficiently modelling the context of the input sequence and the relationships between output labels (Ma and Hovy, 2016). That’s why it was once considered the gold standard in NER. Using this as the baseline model, we could compare and improve it with our transfer learning approach. The study trained the baseline model using the corpus previously tagged with the RT&B diseases fine-grained entities.

Subsequently, we extended the training, using transfer learning, to BERT and several of its derivatives Pretrained Large Language Models. The BERT family models are renowned for their exceptional performance in NER tasks and have consistently been the best-performing models in multiple NLP challenges, such as NER. (Devlin et al., 2019).

The PLLMs utilise the Transformer framework. (Vaswani et al., 2017). They are typically pre-trained on extensive text collections, like Wikipedia and BookCorpus. This pre-training allows them to learn long-term dependencies in the text and perform exceptionally well on NLP tasks. Transfer learning has been shown to enhance the efficiency of downstream tasks, particularly when there are inadequate labelled data to train the model. (Ruder et al., 2019a). To tackle our NER task, we used fine-tuning as the transfer learning method with the Pretrained Large Language Models. Furthermore, the encoder-only transformer architecture of BERT lends itself well to the task of NER, as demonstrated in the original BERT paper where NER was used as an example task. The research showed that it is more computationally efficient to fine-tune BERT for a NER task rather than to build a model from the ground up (Devlin et al., 2019). Taking into account the proven efficacy of BERT and its derivatives in NER, the study examines the use and performance of the following models:

1. BERT (Devlin et al., 2019)
2. RoBERTa (Liu et al., 2019)
3. ELECTRA (Clark et al., 2020)
4. DeBERTa (He et al., 2021)

5. Longformer (Beltagy et al., 2020)

In addition, we explore the following models trained with scientific information to understand whether domain-specific versions of the PLLMs improve performance in novel low-resource tasks.

1. SciBERT (Beltagy et al., 2019)
2. BioBERT (Lee et al., 2020)
3. PubMedBERT (Gu et al., 2022)
4. SciDeBERTa (Jeong and Kim, 2022)

The implementation of transfer learning in our study allowed for a rapid and efficient methodology for handling NER tasks in a low-resource data domain. The methodology involved rapidly evaluating multiple Pretrained Large Language Model (PLLM) to determine the most effective option. As Pretrained Large Language Models continue to evolve, our scientific workflow for training models can be readily modified to incorporate future PLLM that will surpass current models and achieve superior performance. Using this approach, the development of more accurate and efficient NER models will be greatly improved, especially in low-resource data domains.

3.6.1 Model Training Configuration and Automation

This project used the Python HuggingFace API and the HuggingFace Transformers toolkit (Wolf et al., 2020) to train, using transfer learning, several BERT based models. The process was designed to be easily automated and highly configurable, allowing multiple experiments with different models and parameters. A separate JSON file defined each model training and evaluation configuration. The configuration file was then passed as an argument to the training script as listed in Appendix A. For example, the command listed below would commence the training procedure for the DeBERTa v3 model with the specified configuration:

```
python run_ner.py ./data_30/deberta_v3_large/train_config_deberta_v3_large_128.json
```

The configuration file contained key parameters for model training, including model name or path, labels, data directory, output directory, maximum sequence length, number of training epochs, batch size, save steps, logging steps, and seed. It also specified whether to report to the Weights & Biases (WandB) Machine Learning tracking platform (Biewald, 2020). Finally, it also configured whether to perform training only or include evaluation and prediction. The overwriting of the output directory and cache was also configurable.

An example configuration for the SciBERTa model is as follows:

```
{
"model_name_or_path": "KISTI-AI/Scideberta-full",
"labels": "./data_30/labels.txt",
"data_dir": "./data_30/sciberta_full/128",
"output_dir": "./output/sciberta_full/128",
"max_seq_length": 256,
"num_train_epochs": 14,
"per_device_train_batch_size": 32,
"save_steps": 500,
"logging_steps": 500,
"seed": 3,
"report_to": "wandb",
"do_train": true,
"do_eval": true,
"do_predict": true,
"overwrite_output_dir": true,
"overwrite_cache": true
}
```

Each model was pre-processed with its own tokeniser using the Hugging Face library, ensuring that the input data were appropriately formatted for each specific model architecture. This approach provided a flexible and efficient framework for conducting various experiments with different BERT models and configurations.

There are limitations on the quantity of tokens that can be handled by BERT and its related models. The original BERT model can only handle up to 512 tokens (Devlin et al., 2019). To ensure consistency in our preprocessing, we have limited the number of tokens to 128 and 256 for all models except the Longformer model. Longformer is specifically designed to handle larger documents and can process up to 4096 tokens (Beltagy et al., 2020). Documents larger than the specified token size were chunked using a preprocessing script as detailed in the appendix A.

3.6.2 Experiment Tracking

In the study, we used Weights & Biases (WandB), a platform to track experiments in machine learning. WandB allows users to track the hyperparameters, metrics, and artifacts of their experiments and to visualise the results in various ways. WandB provided the study with a robust framework for experiment tracking, and other studies have used a similar approach to experiment tracking (Bir et al., 2023).

The experimental plan was straightforward, yet comprehensive. A baseline BiLSTM-CRF model was trained and evaluated for comparison purposes, serving as a point of

reference for the performance of the models developed using transfer learning. Subsequently, we trained the different BERT models, each with different configurations, and compared their performance metrics, including the F1 score, the accuracy, the non-O accuracy, and the precision with the baseline.

WandB played a pivotal role in this process. It allowed us to log various information about our experiments, including hyperparameters, metrics, and artifacts such as model weights and predictions. This comprehensive logging facilitated real-time tracking of our experiments, enabling us to monitor the progress and contrast the outcomes of various models efficiently.

In addition, the visualisation tools offered by WandB presented the experiments with different methods to display the results, such as graphs, tables, and interactive dashboards. By comparing these metrics, the study could determine the most efficient setup of the model and ultimately gauge the efficacy of the fine-tuned models when compared to the baseline. The knowledge gained from WandB simplified the process of training and evaluating the models.

3.6.3 HuggingFace Platform

The advent of the transformer architecture (Vaswani et al., 2017) revolutionised the NLP field, outperforming traditional models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (CNN). This architecture has facilitated the development of high-capacity models and, due to its amenability to pretraining, has allowed the effective use of this capacity across a broad spectrum of downstream tasks, including text classification and token classification tasks such as Named Entity Recognition (Wolf et al., 2020).

The HuggingFace Transformers library is an open source platform that supports Transformer-based models and promotes the dissemination of pre-trained models. This library includes cutting-edge Transformer models under a unified API, providing a streamlined and standardised interface for model interaction (Wolf et al., 2020).

To ensure consistency and reliable results, the study used the Hugging Face Transformers library to experiment with different pre-trained models. The library offered a unified API to load, train and save the NER models, making it easier to compare and evaluate them. By using HuggingFace, the training, testing and evaluation of various models was automated and conducted in a standardised manner. This improved the reliability and comparability of the results, because all experiments used the same approach once the configuration files were set.

3.6.4 Experimental Environment

The study was run in dedicated environments with the same GPU capabilities. The environment was hosted on RunPod, a GPU cloud platform (RunPod, 2023). This environment was equipped with a high performance NVIDIA A100 GPU with 80 GB of VRAM.

In addition to the GPU, the environment was also provisioned with 125 GB of RAM and 12 virtual CPUs. This configuration provided us with a powerful and flexible platform to run our experiments. The substantial amount of VRAM allowed us to train large BERT models, while the generous allocation of RAM and CPUs ensured smooth model training and evaluation.

The use of such a cloud-based environment offered the flexibility to scale resources according to the needs of the experiments. It also ensured that our experiments were not limited by the constraints of local hardware, enabling us to focus on the experimentation and model-tuning process. It also saved on costs as the servers were only provisioned for the experiment and thereafter shut down.

3.7 Model Evaluation

In NLP, we approach NER as a token classification task. Therefore, during experimentation, when the study evaluated the performance of the baseline and different BERT models, there were four possible outcomes.

- True positives (TP): If an entity was predicted to belong to a class and it indeed matched that class.
- False positives (FP): If an entity was predicted to belong to a class and it did not match that class.
- True negatives (TN): If an entity was predicted to belong to a class and it truly did not match that class.
- False negatives (FN): If an entity was predicted not to belong to a class, while it actually did so.

We used the confusion matrix to gauge the effectiveness and obtain a critical assessment of the model's correct and incorrect classifications. The matrix offers insight into the errors made by the classifier and the types of errors that occur. This is critical as certain entities, such as crop disease, may require correct prediction for effective monitoring, while others, such as the plant part, may not be as essential. Our study specifically employed a normalised multiclass confusion matrix to analyse how the models performed. The confusion matrix visually displayed the performance

of a classification model across multiple classes, the Y-axis displayed the actual labels of the entities, while the X-axis showed the predicted labels.

Assuming that a study has n entity classes, the confusion matrix would be an $n \times n$ table. Every cell within the matrix denotes the proportion of model predictions, categorised by the actual and predicted classes.

	Predicted: Entity 1	Predicted: Entity 2	...
Actual: Entity 1	TP_1	FP_{12}	...
Actual: Entity 2	FP_{21}	TP_2	...
...

Table 3.5: Example of Normalized Multi-class Confusion Matrix

In the normalised confusion matrix, each cell value is a number between 0 and 1, representing the proportion of predictions for each class. The study used the normalised confusion matrix instead of a standard confusion matrix. The standard confusion matrix, which presents absolute counts, could have led to misleading interpretations due to the high prevalence of nonentities (O entities) and the imbalance among other entities. The normalised confusion matrix, on the other hand, provided a more accurate and fair evaluation of our model’s performance across all entity classes.

The study also used the following metrics to validate the robustness of the model: **Accuracy**: The percentage of correct predictions in the test data set. We calculated it as follows:

$$\mathbf{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision: The percentage of positive instances among all total predicted positive instances.

$$\mathbf{Precision} = \frac{TP}{TP + FP}$$

Recall: The percentage of positive instances among all actually positive instances.

$$\mathbf{Recall} = \frac{TP}{TP + FN}$$

F1-Score: Average precision and recall, weighted by their inverses. Therefore, the higher the F1 score, the better; a perfect model would have an F1 score of 1.

$$\begin{aligned} \mathbf{F1} &= \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall} \\ &= \frac{TP}{TP + \frac{1}{2}(FP + FN)} \end{aligned}$$

The study evaluated all models in the experiment, including the baseline model and the BERT models that used transfer learning, using these metrics.

3.8 Model Deployment

The trained models have been deployed and made publicly accessible via the HuggingFace Model Hub. The Model Hub is a platform that facilitates the sharing and collaboration of machine learning models. These models are hosted as public repositories in Hugging Face Spaces, allowing for easy accessibility and usage by the broader research and development community.

For example, the DeBERTa v2 model, fine-tuned as part of this research, is readily available for download and inference (Mwanzia, Leroy, 2023). This model can be utilised directly from the Hugging Face Model Hub, enabling researchers and developers to leverage the model’s capabilities without retraining. The model repository provides comprehensive details about its configuration, training, and performance, ensuring transparency and reproducibility of the results.

The deployment of these models on the Hugging Face Model Hub not only promotes the open sharing of resources within the machine learning community, but also provides a platform for continuous improvement and collaboration. The DeBERTa v2 model, along with other models trained in this research, provides significant value as a resource for further research and application in the field of NLP.

Chapter 4

Results and Discussion

4.1 Summary of Results

In this chapter, the research results and significant findings made in developing an Named Entity Recognition (NER) model using transfer learning will be discussed. The NER model developed was for the task Roots, Tubers and Bananas (RT&B) crop diseases which is a low-resource domain.

The experimentation was set up as described in Chapter 3. Using Python and HuggingFace to build the scientific workflow, the study created specific configuration files for each variation of the BERT model trained using transfer learning. The study assessed 25 variations of the PLLM outlined in Section 3.6, demonstrating the ability to compare multiple models using the developed workflow efficiently.

The findings of the experiment, summarised in Table 4.1, demonstrate the performance of various transformer-based models trained, using transfer learning, on our NER task on RT&B crop diseases. The "SciDeBERTa-full-128" model, a fine-tuned DeBERTa model by the Korea Institute of Science and Technology Information AI, outperformed all models, including the baseline model. Based on the evaluation, the study determined that SciDeBERTa is the most appropriate option to perform NER in RT&B crop diseases. The model performed better in non-O accuracy, accuracy, precision, F1 score, and recall metrics. However, due to some problems that are areas of further research, some models did not return any accuracy score during model testing and thus did not exceed the baseline metrics. Specifically, the models labelled "electra-large-256", "roberta-large-256", and "deberta-v2-xlarge-128" in Table 4.1 failed to identify any entities correctly.

The findings show that the use of transfer learning to train Pretrained Large Language Models in low-resource NER domain tasks produces better results than machine learning methods used in low domain Named Entity Recognition.

4.2 Introduction to Results

This research examined the efficacy of various Pretrained Large Language Models trained using transfer learning to improve Named Entity Recognition in situations where annotated data resources were limited. The goal was to identify models that can make use of transfer learning to improve transformer-based PLLMs (such as BERT and its variants) to achieve better results in NER tasks, even with limited available annotated data.

Different performance metrics were used to evaluate the baseline model and those fine-tuned using transfer learning. The chosen metrics were non-O accuracy, accuracy, precision, F1 score, and recall. When it comes to NER tasks, these metrics are considered standard and provide a thorough evaluation of the effectiveness of each model. Accuracy is the percentage of predictions that a model makes correctly, while non-O accuracy focusses specifically on the prediction correctness of named entities, excluding non-entities. Precision assesses the model’s ability to minimise the number of false positives, while Recall assesses the model’s capability to identify all relevant instances. The F1 score is a measure of the overall performance of a model that takes into account both precision and recall.

A baseline Bidirectional Long Short-Term Memory with Conditional Random Field Layer (BiLSTM-CRF) model was used to establish a reference point of comparison. This model’s performance was previously considered the benchmark in NER tasks due to its ability to efficiently model the connections between the input sequence and the output labels (Ma and Hovy, 2016). The experiments aimed to determine which transformer-based PLLM trained for RT&B disease entities through transfer learning could surpass baseline performance. The goal was to push the boundaries of what is currently possible in a low-resource domainNER task.

In Table 4.1, you can see the findings of our experiments. These results offer significant information about the abilities and restrictions of the different fine-tuned PLLMs in low-resource data environments. The results demonstrate that these models trained using transfer learning improve NER efficiency even in low resource domains. However, selecting the most suitable model and accurately fine-tuning it is essential. The following sections provide further discussion of these results.

4.3 Discussion of Results

The results of our experiment are presented in Table 4.1. A comparative analysis was conducted on various transformer-based models trained by transfer learning, using the annotated data set of RT&B crop diseases. The aim was to perform a task of Named Entity Recognition. The study considered multiple performance metrics to evaluate

the models, including non-O accuracy, accuracy, precision, F1 score, and recall. It evaluated the models’ overall performance by analysing metrics such as their ability to accurately identify entities, classify non-entities, and strive to achieve a trade-off between precision and recall. Note that the model prefix listed by Hugging Face has been omitted from the table due to space constraints.

Name	NonOAcc	F1	Accuracy	Precision	Recall
SciDeberta-full-128	91.39	85.62	97.80	82.13	89.42
PubMedBert-buncft-256	91.11	86.09	97.92	83.09	89.31
PubMedBert-buncft-128	90.39	86.34	97.91	84.17	88.62
DeBerta-v3-large-256	90.33	86.29	97.63	83.47	89.31
SciDeberta-full-256	89.83	84.66	97.73	82.08	87.41
bert-large-cased-128	89.56	84.11	97.51	79.73	87.35
electra-large-128	89.50	84.70	97.66	81.58	88.07
scibert-uncased-128	89.39	83.80	97.63	80.31	87.61
deberta-v3-large-128	89.33	86.18	97.58	83.26	89.32
scibert-uncased-256	89.28	83.33	97.59	79.62	87.41
bert-large-uncased-256	89.17	83.04	97.54	79.62	87.81
longformer-base-4096	88.87	84.94	97.38	82.88	87.10
bert-large-cased-256	88.61	83.04	97.57	79.58	86.81
bert-large-uncased-128	88.56	83.52	97.53	80.66	87.87
electra-base-128	88.28	83.46	97.41	80.82	86.28
electra-base-256	88.00	83.39	97.34	80.06	87.01
scibert-cased-256	88.00	81.94	97.38	78.91	85.21
scibert-cased-128	87.94	83.20	97.41	80.02	86.64
biobert-base-cased-256	87.22	82.52	97.39	79.65	85.61
biobert-base-cased-128	86.94	82.49	97.37	79.47	85.76
roberta-large-128	82.22	73.16	96.26	69.15	77.66
electra-large-256	0.00	0.00	87.46	0.00	0.00
roberta-large-256	0.00	0.00	87.46	0.00	0.00
deberta-v2-xlarge-128	0.00	0.00	87.46	0.00	0.00
Baseline-128	75.52	80.66	96.02	84.16	77.44

Table 4.1: Experimental Results

Metric Performance Legend

	Top Performance in Metric
	2 nd Performance in Metric
	3 rd Performance in Metric
	4 th Performance in Metric
	5 th Performance in Metric
	Lowest Performance in Metric
	Invalid Results

The "Baseline-128" BiLSTM-CRF model is used as a benchmark to evaluate the

performance of PLLMs that have been trained using transfer learning. The baseline model achieved a non-O accuracy of 75.52%, an F1 score of 80.6%, an accuracy of 96.02%, a precision of 84.16%, and a recall of 77.44%. Although these results are respectable, the objective of the experiments was to explore whether the transformer-based models could outperform this baseline.

As demonstrated in the results shown in Table 4.1, two families of models trained with transfer learning showed very promising results for Named Entity Recognition of RT&B crop diseases. First, are DeBERTa based models, especially SciDeBERTa (Jeong and Kim, 2022) and version 3 of DeBERTa (He et al., 2023). The second was PubMedBERT (Gu et al., 2022). These models performed well on all observed metrics, especially on the F1 score and the non-O accuracy.

The SciDeBERTa model 'Scideberta-full-128' trained on data with a maximum length of 128 tokens displayed superior performance across all metrics. It achieved the highest non-O accuracy of 91.39% and accuracy of 97.80%, indicating a significant improvement in correctly identifying both entities and non-entities. Furthermore, the F1 score of this model was one of the best, which means an effective balance between identifying true positives and minimising both false negatives and false positives.

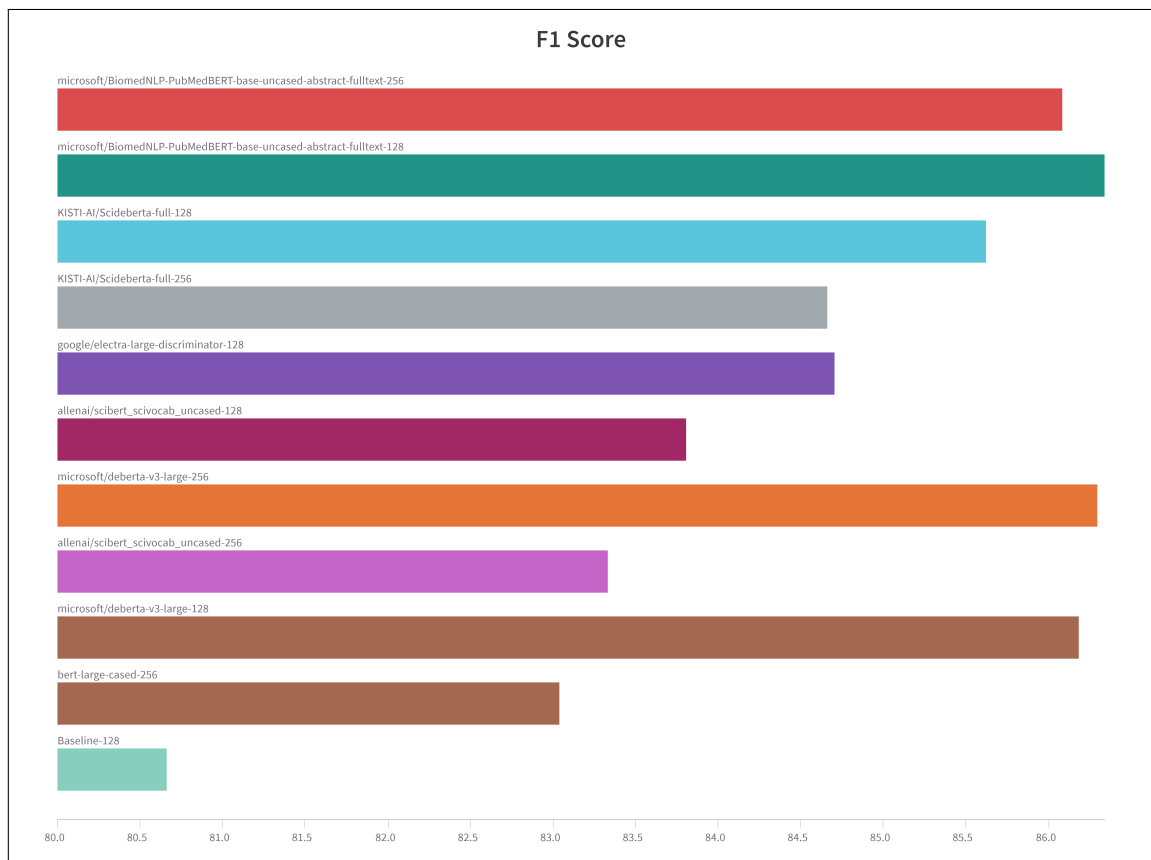


Figure 4.1: F1 Score for the 10 best models compared to the baseline

It is important to mention that certain models used in the study did not show

a significant advantage over the baseline. Specifically, the models 'electra-large-256', 'roberta-large-256' and 'deberta-v2-xlarge-128' had a non-O accuracy and precision of 0.00%, meaning they could not correctly identify any entities. There may be a few reasons for this underperformance, such as the models' inability to adapt to low-resource data characteristics, inadequate hyperparameters, or potential issues with the training process. More research is necessary to understand why certain models did not perform well.

These findings emphasise the importance of selecting appropriate models and using transfer learning techniques when dealing with low-resource domains in NER tasks. Additionally, it highlights the potential of transformer-based Pretrained Large Language Models, such as SciDeBERTa, to greatly enhance performance in such tasks. However, the findings also caution against taking a universal approach since some models were unable to surpass the baseline. Further research could investigate the factors that affect these performance disparities, which may result in more effective approaches for NER tasks in poorly resourced domains. To do this, the study created a workflow that can quickly assess large language models for NER, as long as they are accessible in the HuggingFace API.

4.3.1 Low Resource Domains can Make NER Task Difficult

Named Entity Recognition is often considered a solved problem because it has achieved high performance on established data. However, the WNUT2017 Shared Task on Recognition of Novel and Emerging Entities (Derczynski et al., 2017) shows that these scores can be misleading. Systems achieving these high scores often struggle with infrequent or novel entities, and their success is largely attributed to familiar and predictable entities (Augenstein et al., 2017; Derczynski et al., 2017). This is the same challenge with low-resource NER domains. New entities can be difficult to recognise with existing NER models. This was evident in our experiments, where many models struggled with novel entities such as symptoms.

4.3.2 Importance of Non-O Accuracy in Disease Recognition

When it comes to predicting crop disease entities in NER, the non-O accuracy metric plays a significant role. The metric measures the accuracy of identifying named entities, excluding the "O" class. The "O" is used to label tokens that are not part of any named entities and is commonly used in annotation schemes such as the IOB scheme, used in this study. In the NER task in this study for RT&B crop diseases, correct recognition of entities such as the name of the disease or pathogen is crucial. For example, during a transboundary crop epidemic, the cost of failing to detect a

disease (a false negative) can be much higher than the cost of incorrectly identifying a non-entity as a disease (a false positive).

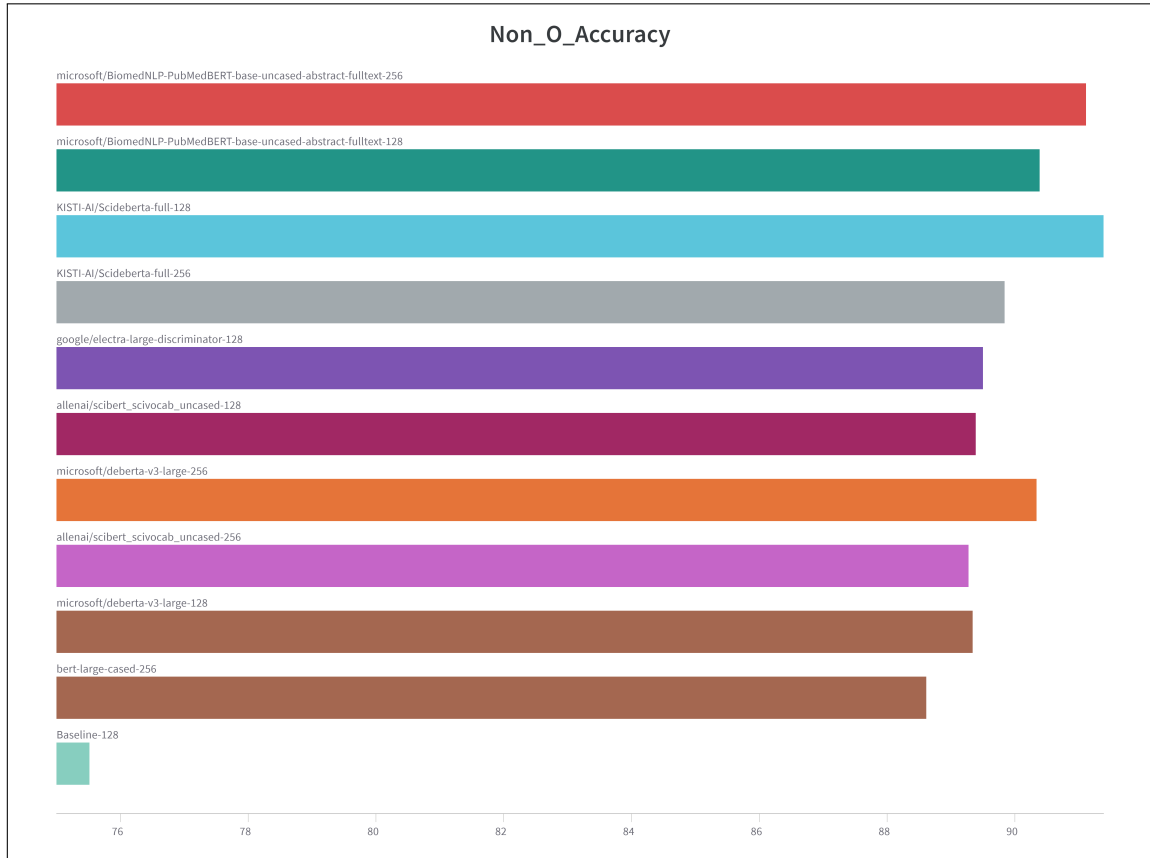


Figure 4.2: Non-O accuracy compared to baseline

It is essential to optimise for the recognition of entities in critical tasks such as crop disease detection, even if it increases the likelihood of false positives. This is why non-O accuracy, which gives more weight to the correct identification of entities, is a particularly relevant metric in this study. To continue the example, if the crop disease is not detected due to a false negative, it could spread unchecked, leading to widespread crop damage and significant economic loss. However, a false positive, although still undesirable, could lead to further testing and verification, thus mitigating potential damage.

However, it is important to note that while non-O accuracy is crucial for evaluating model performance, it should not be the only metric used. F1 score, precision, and recall are additional metrics that provide valuable information on a model's performance. Taking a balanced approach that takes into account all of these metrics results can lead to a more robust and reliable model for disease identification.

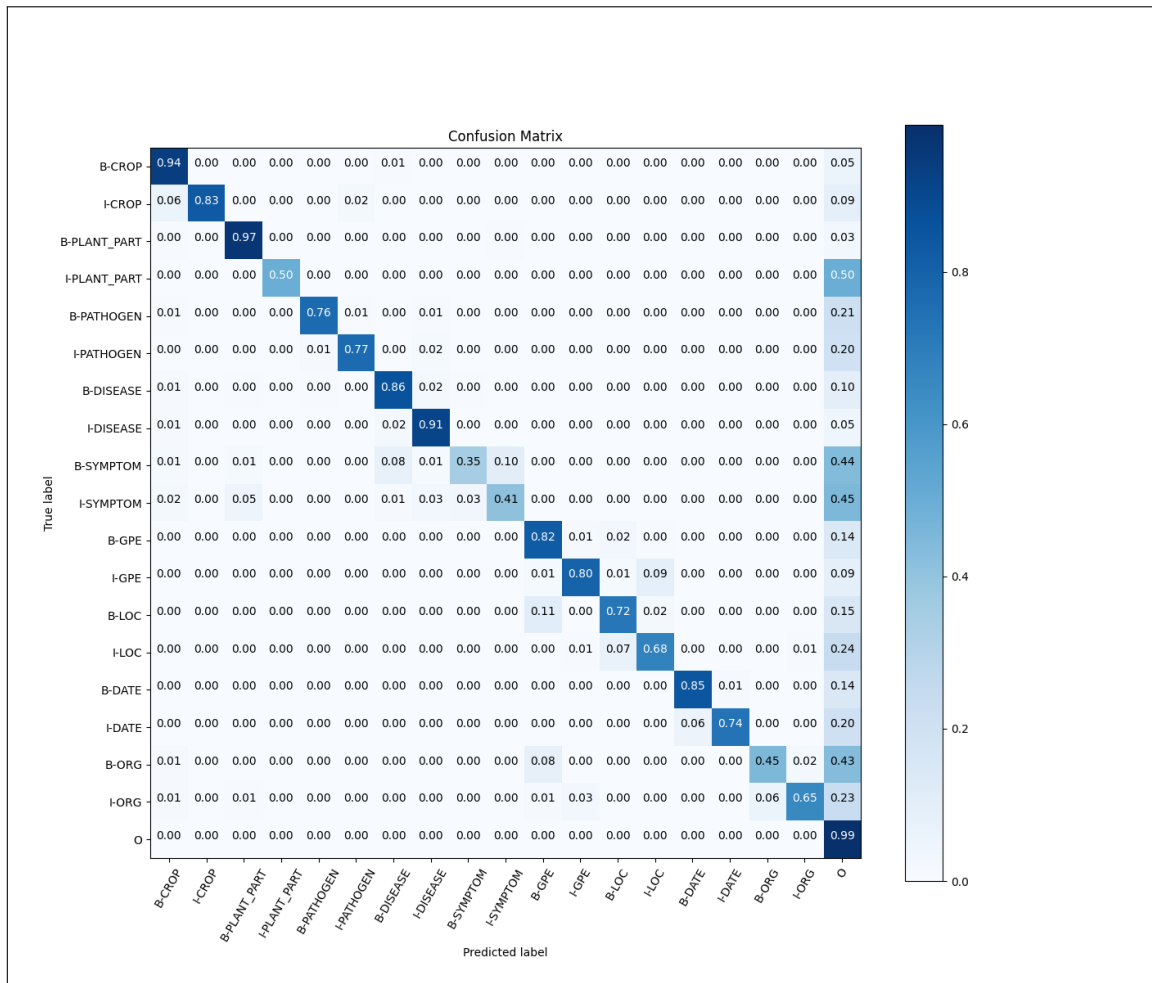


Figure 4.3: Confusion Matrix for the Baseline Model

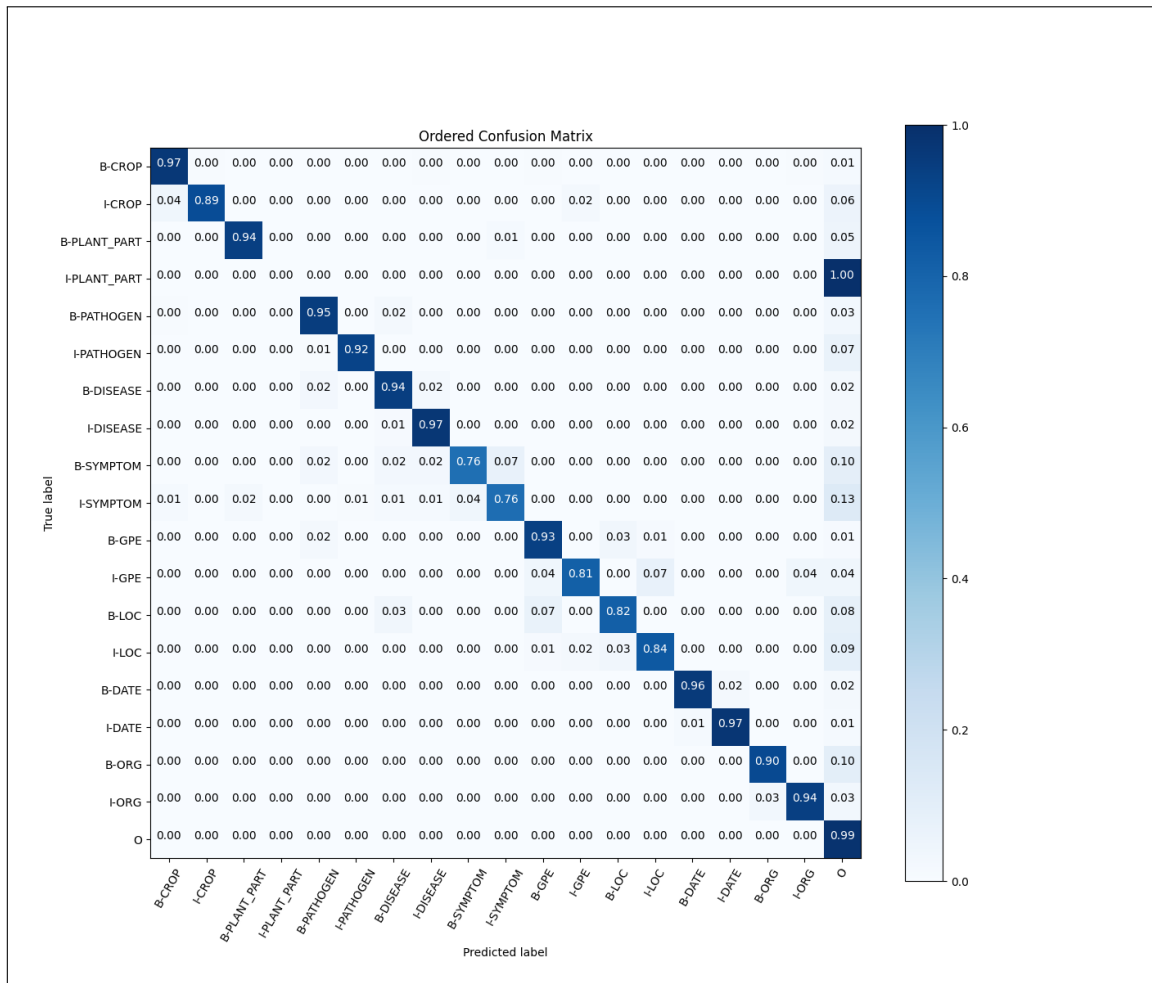


Figure 4.4: Confusion Matrix for SciDeBerta Model

Chapter 5

Conclusions

5.1 Introduction

This chapter presents an assessment of the research conducted in this study, focussing on the application of Named Entity Recognition (NER) in fields with low machine learning resources. The study specifically looked at the recognition of crop disease entities for Roots, Tubers and Bananas crops. The study's conclusions, drawn from empirical evidence gathered through experimentation, offer valuable insights into the effectiveness of transfer learning in creating new models from Pretrained Large Language Model for NER tasks. The chapter recognises the study's limitations but also emphasises the potential for further research. It also recognises that our knowledge of NER in low-resource domains is constantly developing. The chapter concludes with future work recommendations, with the aim of guiding subsequent research towards enhancing our capabilities in NER tasks, particularly in high-stakes, low-resource domains such as crop disease recognition.

5.2 Conclusion and Limitations

This research project was carried out with the aim of tackling the task of Named Entity Recognition (NER) in lowly-resourced domains, focussing specifically on Roots, Tubers and Bananas (RT&B) crop diseases. Due to the insufficiency of annotated data in this field, transfer learning techniques were utilised to transfer useful knowledge from models in better-resourced domains to the target domain. The study approach was shown to be successful in improving the correctness and effectiveness of the Named Entity Recognition (NER) task in this challenging context. The findings of this study have led to new insights and understandings for the field of machine learning, particularly in the area of transfer learning, and can potentially inform future research in this area.

This study's unique contribution lies in its creative use of transfer learning on pre-existing Pretrained Large Language Model (PLLM)s, generating new models specif-

ically for a NER task in the specialised and underexplored domain of RT&B crop diseases. The generated models were evaluated using several NER metrics, including non-O accuracy and F1 score, providing a quantitative assessment of their performance. These results demonstrate the effectiveness and great potential of leveraging transfer learning techniques to create novel models, significantly improving NER performance in low-resource domains.

It is worth mentioning that there are certain limitations to consider when interpreting the results of the study. A primary concern in the use of the models is their ability to generalise accurately to related or different data and domains. Given this challenge, it is important to exercise caution when applying models to broader applications. Furthermore, the study was unable to perform extensive hyperparameter tuning, a missed chance to improve the models' performance, which is unfortunate. Finally, the ever-evolving landscape of PLLM development and innovation presents an inherent challenge in maintaining the relevance of the study.

Future work could explore alternative transfer learning techniques or other PLLM architectures that may be more effective in low-resource settings. Research could also investigate methods for generating or augmenting data in specialised domains to alleviate data scarcity. This research has applications beyond RT&B crop diseases and can be extended to other crops and other low-resource domains, presenting more possibilities for future research. This study serves as a stepping stone toward more advanced and effective NER solutions for low-resource domains.

5.3 Final Recommendations

The results obtained from the research demonstrate the potential of transfer learning to improve NER tasks in low-resource domains, such as RT&B crop diseases. By using transfer learning techniques, it is possible to effectively overcome the challenges associated with the insufficient availability of extensively annotated data for model development. The models trained with our approach provide practical and innovative solutions to these domains.

Constant benchmarking against newly released PLLM must become an integral part of ongoing research in this area. This will not only ensure alignment with cutting-edge methods but will also sustain the relevance and impact of the findings in the ever-changing landscape of model development. This study presents a workflow to efficiently evaluate new models. The emphasis on benchmarking can uncover novel findings and reinforce the study's promising results.

Finally, to address the challenge of data scarcity in low-resource domains, more research should explore innovative methods for data augmentation or synthetic data generation. Unsupervised data annotation techniques should be investigated to gen-

erate additional annotated data for use with transfer learning and to expand the available training data. The study also suggests that further research should go deeper into hyperparameter tuning compared to the current study. This process could uncover more potential for transfer learning models to make use of the limited data resources that are present in these domains.

Following these recommendations, the research community can push the boundaries of Named Entity Recognition in low-resource domains, providing valuable tools and insights for various stakeholders in agriculture and beyond.

References

- Akbik, A., Bergmann, T., and Vollgraf, R. (2019). Pooled Contextualized Embeddings for Named Entity Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics. 226 citations (Semantic Scholar/DOI) [2023-03-27]. 12
- Andrade-Piedra, J. L., Bentley, J. W., Almekinders, C. J. M., Jacobsen, K., Walsh, S., and Thiele, G. (2016). Case studies of Roots, Tubers and Bananas seed systems. Working Paper, International Potato Center. Accepted: 2017-05-16T20:48:07Z ISSN: 2309-6586. 1
- Augenstein, I., Derczynski, L., and Bontcheva, K. (2017). Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83. 24 citations (Crossref) [2023-07-15]. 39
- Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics. 1502 citations (Semantic Scholar/DOI) [2023-03-27]. 29
- Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv:2004.05150 [cs]. 29, 30
- Biewald, L. (2020). Experiment Tracking with Weights and Biases. 29
- Bingel, J. and Sgaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. arXiv:1702.08303 [cs]. 14
- Bir, A., Cuesta-Vargas, A. I., Martn-Martn, J., Szilgyi, L., and Szilgyi, S. M. (2023). Synthetized Multilanguage OCR Using CRNN and SVTR Models for Realtime Collaborative Tools. *Applied Sciences*, 13(7):4419. 0 citations (Crossref) [2023-08-03] Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. 30

-
- Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1), 41-75. *Google Scholar Google Scholar Digital Library Digital Library*. 14
- Carvajal-Yepes, M., Cardwell, K., Nelson, A., Garrett, K. A., Giovani, B., Saunders, D. G. O., Kamoun, S., Legg, J. P., Verdier, V., Lessel, J., Neher, R. A., Day, R., Pardey, P., Gullino, M. L., Records, A. R., Bextine, B., Leach, J. E., Staiger, S., and Tohme, J. (2019). A global surveillance system for crop diseases. *Science*, 364(6447):1237–1239. Publisher: American Association for the Advancement of Science. 2, 3, 5
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. arXiv:2010.02559 [cs]. 15
- Chinchor, N. and Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21. 8
- Chiu, J. P. and Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370. 1519 citations (Semantic Scholar/DOI) [2023-03-27]. 10
- Chowdhary, K. R. (2020). Natural Language Processing. In Chowdhary, K., editor, *Fundamentals of Artificial Intelligence*, pages 603–649. Springer India, New Delhi. 7
- Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555 [cs]. 28
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *NATURAL LANGUAGE PROCESSING*. 10
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jgou, H. (2018). Word Translation Without Parallel Data. arXiv:1710.04087 [cs]. 14
- Derczynski, L., Nichols, E., van Erp, M., and Limsopatham, N. (2017). Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics. 55 citations (Crossref) [2023-07-15]. 39

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. 9994 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv: 1810.04805. 9, 11, 12, 19, 28, 30
- Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, 14(2):179–211. 9987 citations (Semantic Scholar/DOI) [2023-03-27] eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1. 10
- ExplosionAI (2023). Prodigy: A Python library for efficient annotation. 23
- Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics. 3357 citations (Semantic Scholar/DOI) [2023-03-27]. 9
- Gaizauskas, R. and Wilks, Y. (1998). Information extraction: beyond document retrieval. *Journal of Documentation*, 54(1):70–105. 258 citations (Semantic Scholar/DOI) [2023-03-27] Publisher: MCB UP Ltd. 7
- Girvetz, E., Ramirez-Villegas, J., Claessens, L., Lamanna, C., Navarro-Racines, C., Nowak, A., Thornton, P., and Rosenstock, T. S. (2019). Future Climate Projections in Africa: Where Are We Headed? In Rosenstock, T. S., Nowak, A., and Girvetz, E., editors, *The Climate-Smart Agriculture Papers: Investigating the Business of a Productive, Resilient and Low Emission Future*, pages 15–27. Springer International Publishing, Cham. 1
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press. 9, 11
- Goyal, A., Gupta, V., and Kumar, M. (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. *Computer Science Review*, 29:21–43. 159 citations (Semantic Scholar/DOI) [2023-03-27]. 8
- Graves, A., Fernández, S., and Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. In *Artificial Neural Networks: Formal Models and Their Applications ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part II 15*, pages 799–804. Springer. 10
- Grishman, R. (1995). The NYU System for MUC-6 or Where's the Syntax? Technical report, NEW YORK UNIV NY DEPT OF COMPUTER SCIENCE. 8

- Grishman, R. and Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. 7
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23. 173 citations (Crossref) [2023-07-13] arXiv:2007.15779 [cs]. 29, 38
- Guo, X., Hao, X., Tang, Z., Diao, L., Bai, Z., Lu, S., and Li, L. (2021). ACE-ADP: Adversarial Contextual Embeddings Based Named Entity Recognition for Agricultural Diseases and Pests. *Agriculture*, 11(10):912. 2 citations (Semantic Scholar/DOI) [2023-03-27] Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. 15
- Guo, X., Zhou, H., Su, J., Hao, X., Tang, Z., Diao, L., and Li, L. (2020). Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism. *Computers and Electronics in Agriculture*, 179:105830. 12 citations (Semantic Scholar/DOI) [2023-03-27]. 15
- He, P., Gao, J., and Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543 [cs]. 38
- He, P., Liu, X., Gao, J., and Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654 [cs]. 28
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780. Publisher: MIT press. 10
- Honnibal, M. and Montani, I. (2021). spaCy: Industrial-strength Natural Language Processing in Python. 24
- Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. 245 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv:1801.06146 [cs, stat]. 12, 13
- Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. 2997 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv:1508.01991 [cs]. 10, 11, 28
- Hugging Face Contributors (2023). huggingface/transformers: Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX. 60

- Jensen, K. (2012). Process diagram showing the relationship between the different phases of CRISP-DM. CC BY-SA 3.0. vii, 21
- Jeong, Y. and Kim, E. (2022). SciDeBERTa: Learning DeBERTa for Science Technology Documents and Fine-Tuning Information Extraction Tasks. *IEEE Access*, 10:60805–60813. 0 citations (Crossref) [2023-07-13] Conference Name: IEEE Access. 29, 38
- Jiang, S., Angarita, R., Cormier, S., and Rousseaux, F. (2021). Fine-tuning BERT-based models for Plant Health Bulletin Classification. *arXiv:2102.00838 [cs]*. 0 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv: 2102.00838. 16
- Jrvelin, K. and Wilson, T. D. (2003). On conceptual models for information seeking and retrieval research. *Information research*, 9(1):9–1. 16
- Kreuze, J., Adewopo, J., Selvaraj, M., Mwanzia, L., Kumar, P. L., Cuellar, W. J., Legg, J. P., Hughes, D. P., and Blomme, G. (2022). Innovative Digital Technologies to Monitor and Control Pest and Disease Threats in Root, Tuber, and Banana (RT&B) Cropping Systems: Progress and Prospects. In Thiele, G., Friedmann, M., Campos, H., Polar, V., and Bentley, J. W., editors, *Root, Tuber and Banana Food System Innovations: Value Creation for Inclusive Outcomes*, pages 261–288. Springer International Publishing, Cham. 2, 3
- Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Number of pages: 8. 9, 10
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural Architectures for Named Entity Recognition. 3338 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv:1603.01360 [cs]. 9, 10
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521 (7553), 436-444. *Google Scholar Google Scholar Cross Ref Cross Ref*. 9
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. 9994 citations (Semantic Scholar/DOI) [2023-03-27] Conference Name: Proceedings of the IEEE. 10
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical

- text mining. *Bioinformatics*, 36(4):1234–1240. 2800 citations (Semantic Scholar/DOI) [2023-03-27]. 29
- Legg, J. P., Lava Kumar, P., Makesh Kumar, T., Tripathi, L., Ferguson, M., Kanju, E., Ntawuruhunga, P., and Cuellar, W. (2015). Chapter Four - Cassava Virus Diseases: Biology, Epidemiology, and Management. In Loebenstein, G. and Katis, N. I., editors, *Advances in Virus Research*, volume 91 of *Control of Plant Virus Diseases*, pages 85–142. Academic Press. 2
- Li, J., Sun, A., Han, J., and Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*. 526 citations (Semantic Scholar/arXiv) [2023-03-27] Number: arXiv:1812.09449 arXiv:1812.09449 [cs]. 13
- Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016). Neural Relation Extraction with Selective Attention over Instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics. 14
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]. 28
- Liu, Z., Luo, M., Yang, H., and Liu, X. (2020). Named Entity Recognition for the Horticultural Domain. *Journal of Physics: Conference Series*, 1631(1):012016. 3 citations (Semantic Scholar/DOI) [2023-03-27] Publisher: IOP Publishing. 4, 24
- Ma, X. and Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. 2248 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv:1603.01354 [cs, stat]. 9, 10, 11, 28, 36
- Malarkodi, C., Lex, E., and Sobha, L. D. (2016). Named Entity Recognition for the Agricultural Domain. In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016); Research in Computing Science*. 4, 24
- Matthew, P., Mark, N., Mohit, I., Matt, G., Christopher, C., Kenton, L., and Luke, Z. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics. 11, 12

- McCallum, A. and Li, W. (2003). Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191. 9
- Miller, S. A., Beed, F. D., and Harmon, C. L. (2009). Plant Disease Diagnostic Capabilities and Networks. *Annual Review of Phytopathology*, 47(1):15–38. 2, 5, 21
- Mutuvi, S. (2023). BLSTM-CRF-NER. original-date: 2020-10-03T12:27:39Z. 60
- Mwanzia, Leroy (2023). lmwanzia/deberta.v2_xlarge_ner_rtb_diseases Hugging Face. 34
- Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguistic Investigations*, 30(1):3–26. ISBN: 0378-4169 Publisher: John Benjamins Type: <https://doi.org/10.1075/li.30.1.03nad>. 8, 9
- National Academies of Sciences, Engineering, and Medicine (2019). *Science breakthroughs to advance food and agricultural research by 2030*. The National Academies Press, Washington, DC. 3
- National Library of Medicine (2023). PubMed. 3
- O’Shea, J. (2017). Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International Journal of Medical Informatics*, 101:15–22. 3, 7
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359. Publisher: IEEE. 12
- Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics. 13
- Patil, S., Pawar, S., and Palshikar, G. (2013). Named Entity Extraction using Information Distance. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1264–1270, Nagoya, Japan. Asian Federation of Natural Language Processing. 13, 21, 22
- Petsakos, A., Prager, S. D., Gonzalez, C. E., Gama, A. C., Sulser, T. B., Gbegbelegbe, S., Kikulwe, E. M., and Hareau, G. (2019). Understanding the consequences of changes in the production frontiers for roots, tubers and bananas. *Global Food Security*, 20:180–188. 1

- Plank, B., Sgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. arXiv:1604.05529 [cs]. 14
- Prain, G. and Naziri, D. (2020). The role of root and tuber crops in strengthening agrifood system resilience in Asia. A literature review and selective stakeholder assessment. Report, International Potato Center. Accepted: 2020-01-21T20:16:13Z ISBN: 9789290605393. 1
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI. 11, 12
- Ristaino, J. B., Anderson, P. K., Bebbler, D. P., Brauman, K. A., Cunniffe, N. J., Fedoroff, N. V., Finegold, C., Garrett, K. A., Gilligan, C. A., Jones, C. M., Martin, M. D., MacDonald, G. K., Neenan, P., Records, A., Schmale, D. G., Tateosian, L., and Wei, Q. (2021). The persistent threat of emerging plant disease pandemics to global food security. *Proceedings of the National Academy of Sciences*, 118(23):e2022239118. 2, 3, 5, 21
- RTB (2016). Roots, Tubers and Bananas (RTB) Full Proposal 2017-2022. Technical report, CGIAR Research Program on Roots, Tubers and Bananas (RTB). Accepted: 2016-04-10T16:41:29Z. 1, 5
- Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019a). Transfer Learning in Natural Language Processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics. 289 citations (Semantic Scholar/DOI) [2023-03-27]. 12, 13, 28
- Ruder, S., Vuli, I., and Sgaard, A. (2019b). A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65:569–631. 13, 14
- RunPod (2023). RunPod. 32
- Sang, E. F. T. K. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. 3353 citations (Semantic Scholar/arXiv) [2023-03-27] arXiv:cs/0306050. 8
- Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3(3):430–439. Number: 3 Publisher: Nature Publishing Group. 2

- Scherm, H., Thomas, C., Garrett, K., and Olsen, J. (2014). Meta-Analysis and Other Approaches for Synthesizing Structured and Unstructured Data in Plant Pathology. *Annual Review of Phytopathology*, 52(1):453–476. eprint: <https://doi.org/10.1146/annurev-phyto-102313-050214>. 3
- Seed World (2018). Global Initiative Announced to Protect Worlds Plants from Pests. 2
- The Allen Institute for Artificial Intelligence (2022). Semantic Scholar. 21, 22
- Thiele, G., Friedmann, M., Campos, H., Polar, V., and Bentley, J. W., editors (2022). *Root, Tuber and Banana Food System Innovations: Value Creation for Inclusive Outcomes*. Springer International Publishing, Cham. 1
- Thiele, G., Khan, A., Heider, B., Kroschel, J., Harahagazwe, D., Andrade, M., Bonierbale, M., Friedmann, M., Gemenet, D., Cherinet, M., Quiroz, R., Faye, E., and Dangles, O. (2017). Roots, Tubers and Bananas: Planning and research for climate resilience. *Open Agriculture*, 2(1):350–361. Publisher: De Gruyter Open Access. 1, 2
- Thomas, C. S., Nelson, N. P., Jahn, G. C., Niu, T., and Hartley, D. M. (2011). Use of media and public-domain Internet sources for detection and assessment of plant health threats. *Emerging Health Threats Journal*, 4(1):7157. Publisher: Taylor & Francis eprint: <https://doi.org/10.3402/ehth.v4i0.7157>. 3, 7
- Thomas-Sharma, S., Abdurahman, A., Ali, S., Andrade-Piedra, J. L., Bao, S., Charkowski, A. O., Crook, D., Kadian, M., Kromann, P., Struik, P. C., Torrance, L., Garrett, K. A., and Forbes, G. A. (2016). Seed degeneration in potato: the need for an integrated seed health strategy to mitigate the problem in developing countries. *Plant Pathology*, 65(1):3–16. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ppa.12439>. 2
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. 11, 19, 28, 31
- Wei, J. and Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv:1901.11196 [cs]. 13
- Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.*, volume 1, page 11. 19

-
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). HuggingFace’s Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs]. 29, 31
- Yang, Z., Salakhutdinov, R., and Cohen, W. (2016). Multi-Task Cross-Lingual Sequence Tagging from Scratch. arXiv:1603.06270 [cs]. 14
- Zhang, J., Guo, M., Geng, Y., Li, M., Zhang, Y., and Geng, N. (2021). Chinese named entity recognition for apple diseases and pests based on character augmentation. *Computers and Electronics in Agriculture*, 190:106464. 6 citations (Semantic Scholar/DOI) [2023-03-27]. 15
- Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc. 11
- Zhou, H., Ning, S., Yang, Y., Liu, Z., Lang, C., and Lin, Y. (2018). Chemical-induced disease relation extraction with dependency information and prior knowledge. *Journal of Biomedical Informatics*, 84:171–178. 18 citations (Semantic Scholar/DOI) [2023-03-27]. 14

Appendix A

Code Snippets

In this paper, we used the Python programming language for data collection and pre-processing, as well as training and evaluating the model. In this section, we provide snippets that highlight key aspects of these processes. These code snippets were instrumental in implementing the proposed methodology and conducting the experiments. In Appendix A, the reader will find links to a comprehensive collection of Python code used in the project as open source on Github. These snippets offer valuable insight into the technical implementation and serve as a valuable resource for readers interested in replicating or further exploring the methodology presented in this thesis. All the Python files with main functionality can be called and used from the console; this allows the files to be used in an unattended manner or in a notebook.

A.1 Data Acquisition

All Python code used for data acquisition is available in the following public repository: <https://github.com/Kaboi/PDPDataUtils/>.

A.1.1 Download News Text Using GoogleNews

These is the main code components for donwloading news using the NewsPaper3k package after searching using GoogleNews. The full code to download the news can be accessed online at: https://github.com/Kaboi/PDPDataUtils/blob/master/data_acquisition/download_news.py.

The following is the usage of the download.py Python script available in the github repository.

```
usage: download_news.py [-h] -c CROP -s SEARCH [-sd STARTDATE] [-ed ENDDATE] [-p PAGESIZE]
```

Download news articles from Google News

optional arguments:

```
-h, --help            show this help message and exit
-c CROP, --crop CROP  Crop name
-s SEARCH, --search SEARCH
                        Search string
-sd STARTDATE, --startdate STARTDATE
                        Start date
-ed ENDDATE, --enddate ENDDATE
                        End date
-p PAGESIZE, --pagesize PAGESIZE
                        Page size
```

A.1.2 Download Scientific Text Using SemanticScholar

This section highlights code blocks for downloading abstracts of scientific papers from the Semantic Scholar academic database. The full Python script for the download of articles is available online at: https://github.com/Kaboi/PDPDataUtils/blob/master/data_acquisition/download_papers.py.

```
# %% add search parameters
searchCrop = "potatosp_1"
searchString = 'first report potato'
searchFields = ['url', 'externalIds', 'year', 'title', 'abstract']
# searchLimit ideally should be multiple of pagesize and > than pagesize
# max pagesize is 100
pageSize = 10
searchLimit = 50

# %% search for the papers
print("searching for the papers...")
articles = search_semantic_scholar(searchString, searchFields, pageSize)

# %% load papers into Data Frame
print("populating data to a limit of ", searchLimit)
articles_dataframe = populate_article_df(articles, searchLimit, pageSize, searchCrop)
# print(articles_dataframe)

# %% save data frame as CSV
filename = "data/" + searchCrop + "_Output.xlsx"
print("saving the file to ", filename)
articles_dataframe.to_excel(filename, index=False, engine='xlsxwriter')
```

A.2 Data Preprocessing

A.2.1 Pre-annotation Text Processing

This section provides utility functions that clean up acquired text before being processed by the Annotation software.

```
# define a normalization function
def normalize_text(text):
    # original_text_remove = text
    # join words split by a hyphen or line break
    text = preprocessing.normalize.hyphenated_words(text)
    # remove any unnecessary white spaces
    text = preprocessing.normalize.whitespace(text)
    # substitute fancy quotation marks with an ASCII equivalent
    text = preprocessing.normalize.quotation_marks(text)
    # normalize unicode characters in text into canonical forms
    text = preprocessing.normalize.unicode(text)
    # remove any accents character in text by replacing them with ASCII equivalents or
    #   ↪ removing them entirely
    text = preprocessing.remove.accents(text)

    return text

def normalize_scitext(scitext):
    # Replace three or more consecutive line breaks (accounting for spaces) with two
    scitext = re.sub(r'((\r\n|\r|\n)\s*){3,}', '\n\n', scitext)

    normalize_text(scitext)
```

A.2.2 Pre-annotation Creation of Gazetteers

The data downloaded from online news and scientific databases are stored in CSV files by scripts. This code processes the CSV files into JSONL annotation files required by the Prodigy annotation software. The full Python script is available on Github: https://github.com/Kaboi/PDPDataUtils/blob/master/data_acquisition/create_pattern_files.py

The code snippet below shows the actual processing of the csv data to create the gazetteers:

```
# %% process data

musa_disease_code = musa_df['Disease'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'disease'))
potato_disease_code = potato_df['Disease'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'disease'))
sweetpotato_disease_code = sweetpotato_df['Disease'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'disease'))
musa_pathogen_code = musa_df['Pathogen'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'pathogen'))
potato_pathogen_code = potato_df['Pathogen'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'pathogen'))
sweetpotato_pathogen_code = sweetpotato_df['Pathogen'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'pathogen'))
cassava_disease_code = cassava_df['Disease'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'disease'))
cassava_pathogen_code = cassava_df['Pathogen'].dropna().apply(
    lambda cropdf: create_code_text_line(cropdf, 'pathogen'))
```

A.2.3 Annotated Data Splitting and Conversion

This Python code provides the functionality to convert a JSONL file created by our annotation tool and gives the option to create training, validation, and evaluation sets. The full Python script is located on Github: https://github.com/Kaboi/PDPDataUtils/blob/master/data_acquisition/convert_offset_iob.py

The following is how to use the data splitting and conversion script:

```
usage: convert_offset_iob.py [-h] -f FILE_PATHS [FILE_PATHS ...] [-min MIN_LENGTH] [-max
  ↪ MAX_LENGTH] [-s] [-sm SPACY_MODEL] [-sep SEPARATOR]
```

Convert jsonl annotations to IOB format.

optional arguments:

```
-h, --help            show this help message and exit
-f FILE_PATHS [FILE_PATHS ...], --file_paths FILE_PATHS [FILE_PATHS ...]
                        Input file names, can be multiple space separated files.
-min MIN_LENGTH, --min_length MIN_LENGTH
                        Minimum sentence length, default is the document length
-max MAX_LENGTH, --max_length MAX_LENGTH
                        Maximum ballpark sentence length, default is the document length
-s, --split           Split the data into train, test and validation sets.
-sm SPACY_MODEL, --spacy_model SPACY_MODEL
                        Spacy language model.
-sep SEPARATOR, --separator SEPARATOR
                        Separator for the output file. Allowed values are ",", (comma), \t
                        ↪ (tab) and \s (space) Default is comma.
```

A.3 Model Training and Evaluation

A.3.1 Training a baseline BiLSTM-CRF model

The following is a BiLSTM-CRF model for Named Entity Recognition (NER) tasks. This code, written in Python, is designed to train a model to recognise and classify entities in text.

The model will default to using a GPU if available. Moreover, this code integrates with the Weights and Biases (wandb) platform for experiment logging and tracking, helping in performance evaluation, comparison between different runs, and keeping track of the experiments' hyperparameters and outputs.

The parameters of the model can be configured via command-line options, including:

- Data locations (training, development, test set, and score file).
- Tagging scheme (IOB or IOBES).
- Text processing parameters, such as lowercase words and replacing digits with zeros.
- Dimensions of character and token embedding.

- LSTM parameters for characters and tokens, including hidden layer size and whether to use a bidirectional LSTM.
- Pre-trained embeddings and whether to load all embeddings or not.
- Use of capitalisation feature.
- Use of CRF.
- Dropout rate for input.
- Option to reload the last saved model.
- Whether to use a GPU.
- Loss file location.
- Model name.
- Character mode: either 'CNN' or 'LSTM'

The source code was forked from (Mutuvi, 2023) and improved to work for this task.

The complete repository is available on Github: <https://github.com/Kaboi/RTB-Diseases-NER-Baseline>

A.3.1.1 Example Usage

Following is how we use the baseline training Python script:

```
python train.py --train dataset/ner_diseases-output-iob-tags-train.txt --dev dataset/  
↪ ciat_ner_diseases-output-iob-tags-validate.txt --test dataset/ciat_ner_diseases-  
↪ output-iob-tags-30-test.txt --char_mode 'LSTM' --name 'Baseline'  
  
python eval.py --test dataset/ner_diseases-output-iob-tags-test.txt --char_mode 'LSTM' --  
↪ model_path models/Baseline
```

A.3.2 Fine Tuning, Evaluating and Testing LLMs using Hugging Face

The code below trains any large language model available on Hugging Face. The code is forked from the Transformers example Github repository (Hugging Face Contributors, 2023) and modified and enhanced for our task. The complete repository can be found in the following Github link Github: https://github.com/Kaboi/RTB_Disease_NER_Transfer

A.3.2.1 Configuration Files

Below is an example of a configuration file needed to fine-tune an LLM using this code.

```
{
  "model_name_or_path": "KISTI-AI/Scideberta-full",
  "labels": "./data_30/labels.txt",
  "data_dir": "./data_30/sciberta_full/128",
  "output_dir": "./output/sciberta_full/128",
  "max_seq_length": 256,
  "num_train_epochs": 14,
  "per_device_train_batch_size": 32,
  "save_steps": 500,
  "logging_steps": 500,
  "seed": 3,
  "report_to": "wandb",
  "do_train": true,
  "do_eval": true,
  "do_predict": true,
  "overwrite_output_dir": true,
  "overwrite_cache": true
}
```

A.3.2.2 Example Usage

Following is how we use the Python script to do the actual transfer learning using the configured JSON file: As is evident all the code in this project can be automated using Bash scripts.

```
python run_ner.py ./data_30/sciberta_full/train_config_sciberta_full_128.json
```