

Lecture Notes on Data Engineering
and Communications Technologies 191

Yap Bee Wah
Dhiya Al-Jumeily OBE
Michael W. Berry *Editors*



Data Science and Emerging Technologies

Proceedings of DaSET 2023

 Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 191

Series Editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

All books published in the series are submitted for consideration in Web of Science.

Yap Bee Wah · Dhiya Al-Jumeily OBE ·
Michael W. Berry
Editors


Data Science and Emerging Technologies

Proceedings of DaSET 2023

 Springer

Editors

Yap Bee Wah
School of Information Technology
UNITAR International University
Petaling Jaya, Malaysia

Dhiya Al-Jumeily OBE 
Faculty of Engineering and Technology
Liverpool John Moores University
Liverpool, UK

Michael W. Berry
University of Tennessee
Knoxville, TN, USA

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-97-0292-3 ISBN 978-981-97-0293-0 (eBook)
<https://doi.org/10.1007/978-981-97-0293-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Paper in this product is recyclable.

Conference Organization

Patron

Sahol Hamid Bin Abu Bakar, Vice-Chancellor, UNITAR International University, Malaysia

Honorary Chairs

Witold Pedrycz, University of Alberta, Canada
Mohd Uzir Mahidin, Department of Statistics, Malaysia

General and Founding Chairs

Yap Bee Wah, UNITAR International University, Malaysia
Dhiya Al-Jumeily OBE, Liverpool John Moores University, UK

International Advisory Committee

Mario Koeppen, Kyushu Institute of Technology, Japan
Heri Kuswanto, Institute Teknologi Sepuluh Nopember
Mohammed Bennamoun, University of Western Australia, Australia
Chidchanok Lursinsap, Chulalongkorn University, Thailand

Proceeding Editors

Yap Bee Wah, UNITAR International University, Malaysia
Dhiya Al-Jumeily OBE, Liverpool John Moores University, UK
Michael W. Berry, University of Tennessee, USA

Finance and Registration Committee

Chong Kim Loy, UNITAR International University, Malaysia
Normaiza Binti Mohamad, UNITAR International University, Malaysia
Simranpreet Kaur Hansaram, UNITAR International University, Malaysia

Secretary

Sharifah Nurul Aina Binti Sayed Burhanudin, UNITAR International University, Malaysia

Technical Program Committee

Abdulaziz Al-Nahari, UNITAR International University, Malaysia (Chair)
Azlin Ahmad, Universiti Teknologi MARA, Malaysia (Co-chair)
Aitizaz Ali, UNITAR International University, Malaysia
Hadi Naghavipour, UNITAR International University, Malaysia
Jamila Mustafina, Kazan Federal University, Russia
Sulaf Assi, Liverpool John Moores University, UK
Umar Farooq Khattak, UNITAR International University, Malaysia
Tahir Mehmood, UNITAR International University, Malaysia
Wasiq Khan, Liverpool John Moores University, UK
Abir Jaafar Hussain, Liverpool John Moores University, UK
Anazida Binti Zainal, Universiti Teknologi Malaysia, Malaysia
Wan Fairos Wan Yaacob, Universiti Teknologi MARA, Malaysia
Syerina Azlin Md Nasir, Universiti Teknologi MARA, Malaysia
Mohd Fadzil Hassan, Universiti Teknologi PETRONAS, Malaysia
Sapto Wahyu Indratno, Institut Teknologi Bandung, Indonesia
Farhad Nadi, UNITAR International University, Malaysia
Syarifah Bahiyah Rahayu Binti Syed Mansoor, Universiti Pertahanan Nasional Malaysia, Malaysia

Program Book Committee

Noor Lees Ismail, UNITAR International University, Malaysia
Iznora Aini Binti Zolkify, UNITAR International University, Malaysia
Rohaizah Abd Latif, UNITAR International University, Malaysia
Jan Lunn, Liverpool John Moores University, UK

IT Committee

Danny Ngo Lung Yao, UNITAR International University, Malaysia
Noor Azma Binti Ismail, UNITAR International University, Malaysia
Paridah Binti Daud, UNITAR International University, Malaysia
Intan Izzatul Fariza Binti Roslli, UNITAR International University, Malaysia
Ker Boon Chin, UNITAR International University, Malaysia
Muhamad Hadri Bin Mohd Hassan, UNITAR International University, Malaysia
Mohd Aiman Mohd Sani, UNITAR International University, Malaysia
Ahmad Ruzaini Bin Rahim, UNITAR International University, Malaysia

Logistics Committee

Mohd Farizudin Mohammad Fauzi, UNITAR International University, Malaysia
Muhammad Shahrir Aizat Muhammad Shuhaili, UNITAR International University, Malaysia
Roziyana Binti Bahrudin, UNITAR International University, Malaysia
Nurlisnawati Binti Mohd Hijazi, UNITAR International University, Malaysia
Syed Munir Barakbah Bin Syed Faozi Barakbah, UNITAR International University, Malaysia
Mohd Amar Bin Mohd Mokhtar, UNITAR International University, Malaysia
Farah Nazurah Binti Zainal, UNITAR International University, Malaysia

Publicity and Strategy Committee

Hadi Naghavipour, UNITAR International University, Malaysia (Chair)
Wan Zakiyatussariroh Wan Husin, Universiti Teknologi MARA, Malaysia
Bander Ali Al-Rimy, Universiti Teknologi Malaysia, Malaysia
Izzatdin Abdul Aziz, Universiti Teknologi PETRONAS, Malaysia
Norshahriah Binti Abdul Wahab, Universiti Pertahanan Nasional Malaysia, Malaysia

Corporate Committee

Stella Chua Ching Yee, UNITAR International University, Malaysia
Marc Kevin Natusch, UNITAR International University, Malaysia
Mohamad Shah Andrew Ibrahim, UNITAR International University, Malaysia
Nikki Poh Li Yi, UNITAR International University, Malaysia

Industry Committee

Badrie Abdullah, UNITAR International University, Malaysia
Visnuvarthen A/L. Sakayam, UNITAR International University, Malaysia

International Scientific Committee

Adel Al-Jumaily, Charles Sturt University, Australia (Chair)
Naomi Bt Salim, Universiti Teknologi Malaysia, Malaysia
Ku Ruhana Ku Mahmud, Universiti Utara Malaysia, Malaysia
Siddhivinayak A. Kulkarni, MIT-World Peace University, Pune, India
Simon Fong, University of Macau, China
Richard Millham, Durban University of Technology, South Africa
Layth Sliman, Efrei, Paris, France

Preface

This volume constitutes the proceedings of the Second International Conference on Data Science and Emerging Technologies (DaSET 2023) held from December 4 to 5, 2023, on a virtual platform. DaSET 2023 aims to provide a platform bringing together experts from academia, industries, government, and professional bodies to share recent trends in Artificial Intelligence and Emerging Technologies for Data-Driven Decisions. The theme of the conference is “Towards Green Artificial Intelligence and Sustainable Solutions.”

DaSET is committed to creating a forum that brings academic and industry practitioners to share and establish collaborations toward impactful innovative research for community development, business success, and economic prosperity. This conference is an international conference in collaboration with UK Malaysia University Consortium, Universiti Teknologi MARA, Universiti Teknologi Malaysia, Centre for Data Science (CerDaS), Universiti Teknologi PETRONAS, Institut Teknologi Sepuluh Nopember, Indonesia; Chulalongkorn University, Thailand; Charles Sturt University, Australia; Institut Teknologi Bandung, Indonesia; National Defence University of Malaysia and Data Analytics and Collaborative Computing Group, University of Macau, China. We also appreciate the strong support from Microsoft Malaysia, Malaysia Digital Economy Corporation (MDEC), Cybersecurity Malaysia, and Statworks (M) Sdn Bhd.

From a total of 80 submitted papers, 40 were selected after a rigorous review process for oral presentation, and the Best Paper Awards were given for each track. The authors and presenters for these 40 papers represented 10 different countries. We thank all the reviewers and Springer Editors for their time spent reviewing the papers.

We are very honored to have Dato’ Sri. Dr. Mohd Uzir Mahidin, Chief Statistician of Malaysia to officiate the opening of DaSET2023. We are privileged to have Prof. Witold Pedcryz, University of Alberta, for his special keynote address. We are proud to have eight distinguished international and local keynote speakers: Prof. Dr. Muhammad Khurram Khan, King Saud University, Saudi Arabia; Prof. Dr. Schahram Dustdar, Tu Wien, Vienna University of Technology, Austria; Prof. Dr. Seifedine Kadry, Noroff Education AS, Norway; Prof. Dr. Naomie Salim, Universiti

Teknologi Malaysia, Assoc. Prof. Dr. Chin Kim On, Universiti Malaysia Sabah, Dato' Dr. Amirudin Abdul Wahab, Cybersecurity Malaysia, Ms. Puteri Anis Aneesa binti Zakaria, Statworks Group, and Mr. Raja Segaran, MDEC, Malaysia. All the distinguished speakers shared various data science and emerging technologies perspectives and projects which are beneficial for academics and industry practitioners.

We would like to thank Professor Emeritus Tan Sri Dato' Sri. Ir. Dr. Sahol Hamid Bin Abu Bakar, Vice-Chancellor of UNITAR International University for his great leadership, advice, and support of local and international academic activities to foster collaborations that lead to the exchange of knowledge and skills for research with impactful outcomes for social and economic prosperity.

We also thank the Series Editor, Springer, Lecture Notes on Data Engineering and Communications Technologies, for the opportunity to organize this guest-edited volume. We are grateful to Mr. Aninda Bose (Senior Publishing Editor, Springer India Pvt. Ltd.) and Mr. Radhakrishnan Madhavamani for the excellent collaboration, patience, and help during the preparation of this volume. We are confident that the volume will provide insightful information to researchers, practitioners, and graduate students in the areas of data science, artificial intelligence, and emerging technologies which are important in this digital information era. Last but not least, we thank all the DaSET 2023 committees for working tirelessly to ensure a successful conference.

Petaling Jaya, Malaysia
Liverpool, UK
Knoxville, USA

Yap Bee Wah
Dhiya Al-Jumeily OBE
Michael W. Berry

About This Book

The book presents selected papers from the Second International Conference on Data Science and Emerging Technologies (DaSET 2023), held online at UNITAR International University, Malaysia, from December 4–5, 2023. This book aims to present current research and applications of data science and emerging technologies. The deployment of data science and emerging technology contributes to the achievement of the Sustainable Development Goals for social inclusion, environmental sustainability, and economic prosperity. Emerging technologies such as artificial intelligence and blockchain are useful for various domains such as marketing, health care, finance, banking, environmental, and agriculture. Innovations in the field of artificial intelligence continue to shape the future of work across nearly every industry. Data Science has a transformative effect on the economy, industry, and society. An important grand challenge in data science is to determine how developments in computational and social-behavioral sciences can be combined to improve well-being, emergency response, sustainability, and civic engagement in a well-informed, data-driven society. The topics of this book include, but are not limited to: artificial intelligence, machine and deep learning, statistical learning, and health and industrial applications.

Contents

Artificial Intelligence

A Comparative Study of Lemmatization Approaches for Rojak Language	3
Liu Jun Yoon, Xuan Yi Tan, Khai Yin Lim, Chi Wee Tan, Ling Ern Cheng, and Jenny Tan	
Multi-aspect Extraction in Indonesian Reviews Through Multi-label Classification Using Pre-trained BERT Models	17
Nur Hayatin, Suraya Alias, Lai Po Hung, and Yuliana Setiowati	
Artificial Intelligence (AI) Empowered Sign Language Recognition Using Hybrid Neural Networks	33
Ambar Saxena, Nailya Sultanova, Jamila Mustafina, and Noor Lees Ismail	
The Performance of GPT-3.5 in Summarizing Scientific and News Articles	49
Sabkat Arshad, Muhammad Yaqoob, and Tahir Mehmood	
Wound Stage Recognition Using YOLOv5	63
Clair Abela and Frankie Inguanez	
Harvest Palm Tree Based on Detection Through 2D LiDAR Sensor Using Power Equation	79
Luqman Hakim Bin Yusof, Abdulaziz Yahya Yahya Al-Nahari, Danny Ngo Lung Yao, and Normaiza Mohamad	
Enhancing Security Surveillance Through Business Intelligence with NVIDIA DeepStream	91
Vishal Pednekar, Nidhi Shettigar, and Sayli Tawhare	
Fuzzified Hybrid Metaheuristics for QoS-Aware Service Composition	105
Hadi Naghavipour, Farhad Nadi, and Ali Aitizaz	

Machine/Deep Learning

Fraudulent E-Commerce Website Detection Using Convolutional Neural Network Based on Image Features 123

Nurfazrina Mohd Zamry, Anazida Zainal, Eric Khoo, Mohamad Nizam Kassim, and Zanariah Zainudin

A Generic Framework for Ransomware Prediction and Classification with Artificial Neural Networks 137

Saaman Nadeem, Tahir Mehmood, and Muhammad Yaqoob

Leveraging Gamification for Engaged Learning in Online Teaching and Learning Experiences 149

Norshahriah Abdul Wahab, A'tifah Hanim Rosli, Syarifah Bahiyah Rahayu Syed Mansoor, Norazliana Akmal Jamaludin, and Siti Hajar Adam

Sentiment Analysis Using Large Language Models: A Case Study of GPT-3.5 161

Farhad Nadi, Hadi Naghavipour, Tahir Mehmood, Alliesya Binti Azman, Jeetha A/P Nagantheran, Kezia Sim Kui Ting, Nor Muhammad Ilman Bin Nor Adnan, Roshene A/P Sivarajan, Suita A/P Veerah, and Romi Fadillah Rahmat

Telecom Customer Experience Analysis Using Sentiment Analysis and Natural Language Processing—Comparative Study 169

Ahmed Mohamed Abdou Ahmed, Abdulaziz Al-Nahari, Raghad Al-Shabandar, Chong Kim Loy, and A. H. Mohammed

Efficient Time Series of Smoothing and Auto-regressive Forecasting Models for Predicting Police Officer Fatalities in the USA 181

Danush Nagappan, Manoj Jayabalan, Ahmad Alanezi, Farhad Nadi, and Thomas Coombs

Multimodal Emotion Recognition Using Attention-Based Model with Language, Audio, and Video Modalities 193

Disha Sharma, Manoj Jayabalan, Nailya Sultanova, Jamila Mustafina, and Danny Ngo Lung Yao

Comparative Analysis of Emotion Recognition Using Large Language Models and Conventional Machine Learning 211

Mangu Soujanya Rao, Thomas Coombs, Normaiza Binti Mohamad, Vinay Kumar, and Manoj Jayabalan

The Impact of Clustering-Based Sequential Multivariate Outliers Detection in Handling Missing Values 221

Mety Agustini, Kartika Fithriasari, and Dedy Dwi Prastyo

Sarcasm Detection in Newspaper Headlines 237
 Vishnu Sai Reddy Chilpuri, Saaman Nadeem, Tahir Mehmood,
 and Muhammad Yaqoob

**Transformer-Based Named Entity Recognition Model—Tamil
 Language** 251
 Karthi Dhayalan, Nailya Sultanova, Jamila Mustafina, and Paridah Daud

**A Comparative Study of Methods for Topic Modelling in News
 Articles** 269
 Swapna D. Rajan, Thomas Coombs, Manoj Jayabalan,
 and Noor Azma Ismail

**Application of Deep Learning Algorithms to Terahertz Images
 for Detection of Concealed Objects** 279
 Soumen Sardar, Sulaf Assi, Iznora Aini Zolkify, Manoj Jayabalan,
 Manea Alsaleem, Ammar H. Mohammed, and Dhiya Al-Jumeily OBE

**Multivariate Comparative Analysis of Statistical and Deep
 Learning Models for Prediction Hardware Failure** 291
 Saurabh Gupta, Raghad Alshabandar, Chong Kim Loy,
 and Ammar H. Mohammed

Statistical Learning

**A Case Study via Bayesian Network: Investigating Factors
 Influencing Student Academic Performance in Online Teaching
 and Learning During COVID-19 Pandemic** 303
 Zheng Ning Looi, Poh Choo Song, Huai Tein Lim, and Sing Yan Looi

**Harnessing the XGBoost Ensemble for Intelligent Prediction
 and Identification of Factors with a High Impact on Air Quality:
 A Case Study of Urban Areas in Jakarta Province, Indonesia** 319
 Wahyu Wibowo, Harun Al Azies, Susi A. Wilujeng,
 and Shuzlina Abdul-Rahman

**Modeling Earthquake Catalog in Sumatra by Space–Time
 Epidemic-Type Aftershock Sequences Model: Combining
 Davidon–Fletcher–Powell and Stochastic Declustering Algorithms** 335
 Christopher Andreas, Achmad Choiruddin, and Dedy Dwi Prastyo

**Small Area Estimation of Mean Years of Schooling Under Time
 Series and Cross-sectional Models** 353
 Reny Ari Noviyanti, Setiawan, and Agnes Tuti Rumiati

**Probabilistic Seismic Hazard Analysis for Sulawesi-Maluku
 Region of Indonesia Using the Space–Time Epidemic-Type
 Aftershock Sequence Model** 369
 Sonia Faradilla, Achmad Choiruddin, and Bambang Widjanarko Otok

Application of Time Series Regression, Double Seasonal ARIMA, and Long Short-Term Memory for Short-Term Electricity Load Forecasting 385
Hafez Afghan and Hidayatul Khusna

A Bayesian Network for Classifying and Predicting Ship Collision 403
Iis Dewi Ratih, Ketut Buda Artana, Heri Kuswanto, Emmy Pratiwi, and Muhammad Farhan Nuari

Outlier Detection in Simultaneous Equations with Panel Data 415
Suci Ismadyaliana, Setiawan, and Jerry Dwi Trijoyo Purnomo

Assessing Departmental Efficiency at Sepuluh Nopember Institute of Technology: A Comparative Study Using Classical and Advanced Data Envelopment Analysis 429
Zakiatul Wildani, M. Naufal Mubarik, Sri Pingit Wulandari, Lucia Aridinanti, and Muhammad Alifian Nuriman

Multivariate Adaptive Fuzzy Clustering Means Regression Splines Model Using Generalized Cross-Validation (GCV) on Stunting Cases in Southeast Sulawesi 447
Mira Meilisa, Bambang Widjanarko Otok, and Jerry Dwi Trijoyo Purnomo

Statistical Inferences for Multivariate Generalized Gamma Regression Model 463
Hasbi Yasin, Purhadi, and Achmad Choiruddin

Health and Industrial Applications

W@rk: Attendance Application Framework Using Blockchain Technology 479
Putra Roskhairul Fitri Kaha, Syarifah Bahiyah Rahayu, Afiqah M. Azahari, Mohd Hazali Mohamed Halip, and K. Venkatesan

Exploring the Impact of COVID-19 on Individuals' Mental Health Through Cluster Analysis 493
Azlin Ahmad, Siti Nabilah Mohd Abdul Hakim Amir, Ezzatul Akmal Kamaru Zaman, and Abdulaziz Al-Nahari

Tracking High Potential Transmission Risk Spots of Infectious Disease Using Spatial Social Network Analysis and Visualisation (SSNAV) Techniques 505
Ilham Abdul Jalil and Abdul Rauf Abdul Rasam

Evaluation of Machine Learning Algorithms for Early Prediction of Liver Disease 521
Sushmitha Geddam, Sulaf Assi, Hadi Naghavipour, Manoj Jayabalan, Abdullah Al-Hamid, and Dhiya Al-Jumeily OBE

An Agricultural Information Recommendation Method Based on Matrix Decomposition Knowledge Graph Algorithm 531
Ruipeng Tang, Narendra Kumar Aridas, and Mohamad Sofian Abu Talip

Performance of the Auxiliary Information Based Hybrid EWMA Chart with Fast Initial Response 545
Peh Sang Ng, Huai Tein Lim, Wai Chung Yeong, and Sajal Saha

Evaluation of Machine Learning Models for Breast Cancer Detection in Microarray Gene Expression Profiles 563
Mohammad Nasir Abdullah and Yap Bee Wah

Author Index 577

About the Editors

Professor Yap Bee Wah is the director of the Research and Consultancy Center at UNITAR International University Malaysia. She is the founding and general chair for DaSET2022: International Conference on Data Science and Emerging Technologies and the editor of the proceedings published in Lecture Notes on Data Engineering and Communications Technologies published by Springer. She was the conference chair of the International Conference on Soft Computing in Data Science (2015–2019 and 2021) and an editor of the SCDS conference proceedings published in the Springer CCIS series. She is also one of the editors of the book titled *Supervised and Unsupervised Learning for Data Science* published by Springer Nature Switzerland AG 2020. She actively published papers in ISI and Scopus journals such as *Expert Systems with Applications*, *Journal of Statistical Computation and Simulation*, *Communications in Statistics-Computation and Simulation*, *Journal of Clinical and Translational Endocrinology*, and *Computers, Materials and Continua*.

Professor Dhiya Al-Jumeily OBE is a professor of Artificial Intelligence and the president of eSystems Engineering Society. His research focus is on developing AI analytics for improving healthcare and the environment fulfilling the United Nations SDGs. His research has been well-recognized and featured in 300+ peer-reviewed articles, 40+ books/book chapters, and attracted over £7.5M. He has successfully supervised 25+ Ph.D. students to completion and has been an external examiner in UK and global universities. He is actively involved as a member of the editorial board/review committee for numerous international journals. He is the founder/general series chair of the IEEE International Conference on Developments in eSystems Engineering since 2007 and DASET since 2022. He was promoted and appointed by The Queen to the Most Excellent Order of the British Empire, “OBE-Ordinary Officers of the Civil Division” of the said Most Excellent Order for the “Services to Scientific Research”.

Professor Michael W. Berry is the co-author and an editor of sixteen books covering topics in scientific computing, information retrieval, text/data mining, and data science. He is the co-editor of the Soft Computing in Data Science volumes from

2015 to 2021 and Data Science and Emerging Technologies 2022 proceedings by Springer. He is also the co-author of popular books published by Society for Industrial and Applied Mathematics: *Understanding Search Engines: Mathematical Modeling and Text Retrieval, Second Edition*, and *Computational Information Retrieval*. He has published over 115 refereed journals and conference publications. He is a member of SIAM, ACM, MAA, ASEE, and the IEEE Computer Society and is on the editorial board of *Foundations of Data Science* (AIMS) and the *SIAM Journal on Matrix Analysis and Applications* (SIAM). He is also a certified program evaluator for the Computing Accreditation Commission (CAC) of the Accreditation Board for Engineering and Technology, Inc. (ABET).

Artificial Intelligence

A Comparative Study of Lemmatization Approaches for Rojak Language



Liu Jun Yoon, Xuan Yi Tan, Khai Yin Lim, Chi Wee Tan, Ling Ern Cheng, and Jenny Tan

Abstract Lemmatization is an important preprocessing step in most natural language processing (NLP) applications where it extracts a valid and linguistically meaningful lemma from an inflectional word. This allows different inflected forms of a word to be grouped into a common root which is the base-form or dictionary-form of a word, known as lemma. Due to the rapid spread of code-mixing languages like the Rojak language that mixes English with Malay, a lemmatizer capable of lemmatizing the language is needed for NLP applications involving this language. Thus, this work proposes a Rojak language lemmatization approach that is able to handle both languages without requiring users to input texts in different language separately. Various methods including rule-based, corpus-based, machine learning, and deep learning-based were experimented and compared using the English Web Treebank (EWT) and Indonesian GSD corpora from the Universal Dependencies (UD) framework. Besides, the effect of POS tags on the performance of lemmatizers was also evaluated based on the accuracy of the train and test sets. From the experiments conducted, the corpus-based approach produced the best results with

L. J. Yoon · X. Y. Tan · K. Y. Lim (✉) · L. E. Cheng · J. Tan
Department of Computing and Information Technology, Penang Branch, Tunku Abdul Rahman University of Management and Technology, Tanjung Bungah, Malaysia
e-mail: limky@tarc.edu.my

L. J. Yoon
e-mail: yoonlyj-pm20@student.tarc.edu.my

X. Y. Tan
e-mail: tanxy-pm20@student.tarc.edu.my

L. E. Cheng
e-mail: chengle-pm20@student.tarc.edu.my

J. Tan
e-mail: tanj-pm20@student.tarc.edu.my

C. W. Tan
Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur Campus, Kuala Lumpur, Malaysia
e-mail: chiwee@tarc.edu.my

99.90% and 92.27% test set accuracy for Malay and English, respectively, whereas the deep learning-based with POS tag approach produced the worst results of 79.78 and 91.15%.

Keywords Lemmatization · Rojak language · Natural language processing (NLP) · POS tag

1 Introduction

Lemmatizer plays an important role in the preprocessing part in most natural language processing (NLP) applications like information retrieval system (IRS) and machine translation system (MTS). This is because in most NLP applications, extraction of a valid and linguistically meaningful lemma from an inflectional word is required. For example, in MTS, the first phase involves analysis and testing the input text written in the source language for its grammatical correctness. Hence, a lemmatizer is required to perform text normalization to break down inflectional words into root and affixes [1, 2]. By performing lemmatization, different inflected forms of a word are grouped into a common root which is the base-form or dictionary-form of a word, known as lemma, that is a valid root and linguistically meaningful word. For example, if the words “*play*, *plays*, and *played*” are given to a lemmatizer, “*play*” will be returned by the lemmatizer as a lemma [2]. Since lemmatization is widely used in most preprocessing of NLP applications, a lemmatizer that is able to give a more accurate lemma has to be studied.

Code-mixing is a widespread phenomenon in Asia that mixes two or more languages. In Malaysia, English has been the focal mixing language practiced with Malay, resulting in the creation of Rojak language [3, 4]. For example, Malaysians tend to write “*That is such a cangguh gadget, I want one too*”. Due to the rapid spread of Rojak language, a lemmatizer capable of lemmatizing the Rojak language, namely a combination of Malay and English, is needed [5]. Besides, all existing stemmers and lemmatizers for Malay or Indonesian confuse roots with stems or lemmas. For instance, although the Sastrawi stemmer is a stemmer, it returns roots instead of stems, while Malaya also inherits this problem. This is because they do not consider the parts-of-speech (POS) tags when performing stemming. Conversely, MorphInd is currently the most widely used morphological analyzer for Indonesian regards roots as lemmas. No existing tools provide stems and lemmas [6]. Therefore, an approach capable of lemmatizing inflected Malay and English words, constituting the Rojak language, into its respective lemmas is needed. This study encompasses three primary objectives: (1) to introduce an approach capable of lemmatizing Rojak language, which is able to handle English and Malay simultaneously, without requiring the user to manually identify and separate the language; (2) to compare the performance of rule-based, corpus-based, machine learning, and deep learning-based lemmatization approaches; and (3) to evaluate the impact of POS tags on the performance of machine learning and deep learning-based lemmatizers.

This article is structured as follows. Section 2 analyzes related works. Section 3 describes the experimental set up and proposed methodology. Section 4 details the results and discussions. Section 5 concludes the work.

2 Literature Review

Recently, a lemmatizer that uses a novel dictionary lookup approach has been proposed for the Urdu language [7]. In the approach, the lemma returned depends on the corpus used to train the model. The lemma of a given word was only returned if both the word and its corresponding POS tag were present in the corpus. If the word and POS tag were not found, the lemmatizer would return the word as it was passed in, without any changes. Any differences caused by spelling mistakes or different blank spaces from the user input words would result in a no match in the corpus. The proposed lemmatizer was evaluated with and without POS tagging. The lemmatizer obtained the highest accuracy 76.44% when words without POS tags were lemmatized. On the other hand, the accuracy achieved by words that are passed in together with POS tags obtained a lower accuracy of 66.79%.

A lemmatizer for Icelandic known as Nefnir was proposed in 2019 [1]. Nefnir was developed from rules derived from a morphological database known as Database of Modern Icelandic Inflection (DMII). Suffix substitution rules were derived from the database and used to lemmatize tagged text. New rules were generated to minimize the number of errors in the training set until no further reduction in the error count. In Nefnir, it was assumed all word forms are identical to their lemma. A list of rules was generated for all the errors. The rule that minimizes the number of remaining errors was selected and applied to the training set until the number of errors does not reduce. The criteria for rule generation are that rules are generated only if the rule correctly lemmatizes at least two examples in the training set. The evaluation of Nefnir was performed to determine the accuracy of Nefnir in lemmatizing words with correct POS tags and words that are automatically tagged with POS tagger which was IceTagger. The accuracy achieved by Nefnir with correct POS tags was 99.55%, while for words that are tagged automatically with IceTagger, the accuracy was 96.88%. It was shown that Nefnir accuracy dropped when lemmatization was performed on words that were automatically tagged by POS tagger.

Another rule-based lemmatizer that uses the longest-affix-match approach was proposed for Kannada inflectional words [2]. In the approach, the input word that contains prefix or suffix would be applied with a set of linguistic rules to get the appropriate lemma. Prefixes and suffixes for Kannada inflected nouns and finite verbs were collected manually from Kannada grammar textbooks [8]. A root dictionary was created from Kannada dictionary “Kannada Rathnakosha”. The proposed lemmatizer searched for a lemma in the root dictionary, and if the lemma was not found, the lemmatizer would append the obtained lemma to the root dictionary. This further improved the performance of the lemmatizer. The proposed lemmatizer was tested on four datasets with lemmatization performed on official circulars that achieved

an accuracy of 85.72%, newspaper 95.80%, legal documents 97.08%, and All India Radio news 95.39%. The accuracy of the proposed lemmatizer achieved above 85% on different dataset.

In 2020, lemmatization of the Russian language based on machine learning algorithms was proposed [8]. Vectorized word forms obtained from open dictionaries were fitted into various machine learning regression models, which were decision tree, random forest, extra tree, and bagging. Decision tree produced the highest accuracy on the lemmatization of real-world corpora, ABBYY corpus and Open-Corpora corpus, with the accuracy of 75.61% and 70.88%, respectively.

Deep learning sequence-to-sequence approach was proposed in 2021 to perform the automatic Romanian lemmatization [9]. The encoder and the decoder in the sequence-to-sequence model for lemmatization of Romanian words contain a single long short-term memory (LSTM) layer. The encoder and the decoder were enriched with one or two additional LSTM layers to improve the system's accuracy. When the deep learning models were evaluated on Romanian Explicative Dictionary (DEX) dataset, one layer LSTM-based architecture achieved the highest accuracy at both word and character levels with an accuracy of 95.93% and 97.29%, respectively. When POS information was included, the system's accuracy increased by 3.39% at word level and by 2.14% at character level resulting in an accuracy of 99.32% and 99.43%, respectively. The model's accuracy improved when POS information was included.

3 Proposed Methodology

This section describes the proposed framework proposed and the workflow for developing a lemmatizer for Rojak language texts that mixes English and Malay. Figure 1 shows the overall framework of the general process design.

3.1 Data Acquisition

Two corpora, one for the English language and another for the Malay language, were utilized in constructing the lemmatizer designed for the Rojak language. In light of the findings [10] indicating a similarity of over 90% between the Malay and Indonesian lexicons, an Indonesian corpus was used due to its relatively greater availability of resources as compared to Malay language. In this study, the English Web Treebank (EWT) corpus and Indonesian GSD corpus were employed [11–13]. Both the corpora were split into train, development, and test sets in the CoNLL-U file format. The train and development sets were combined to form the train set. Each set consists of sentences made up of words, where various information is provided for each word such as ID, FORM, LEMMA, UPOS, and XPOS. For building the lemmatizers, only the FORM, LEMMA, and UPOS which are universal POS tags

3.2 Data Preprocessing

Lowercasing was performed in this stage to ensure all words can be mapped or lemmatized despite the different text casings in the corpus.

3.3 Feature Extraction

Feature extraction is required by machine learning-based and deep learning-based lemmatizers. In this process, each word and lemma was transformed into a sequence of numerical values, forming either the feature vector or word embedding [15]. A pre-trained model known as fastText [16] was used to generate word embedding of size 300 for each word in the corpus, which were then fed to the machine learning and deep learning models for training.

3.4 Lemmatizer Building

Lemmatizers for the English and Malay languages were individually constructed using distinct methods. Four different lemmatizers, consisting of rule-based, corpus-based, machine learning-based, and deep learning-based were developed. The training or development process for each method was replicated to create two lemmatizers of the same method for the distinct languages, namely English and Malay.

3.5 Rule-Based

Lowercasing was performed in rule-based lemmatizer on both the train and test sets. The rule-based lemmatizer utilizes information regarding word forms, lemmas, and POS tags to generate rules, as demonstrated in Table 2 for English and Table 3 for the Malay language, enabling it to carry out lemmatization. During the lemmatizer building stage, the initial step involves constructing a corpus that encompasses all word forms and lemmas. This corpus serves as a reference for the lemmatizer to determine whether a generated lemma represents a meaningful or dictionary-form of a word. Subsequently, rules for lemmatizing English and Malay words were formulated based on linguistic knowledge of the languages, observations from the corpora, and insights garnered from prior research. The overall workflow for the rule-based lemmatizer is outlined in Algorithm 1.

Table 2 Rules for English rule-based lemmatizer

POS tag	Rules	Example
NOUN/PROP	1. plural form → singular form	1. geese → goose 2. families → family 3. cacti → cactus
VERB	1. past tense → present tense 2. continuous tense → present tense 3. perfect tense → present tense 4. singular verb (with s) → plural verb (without s)	1. played → play 2. playing → play 3. seen → see 4. plays → play
ADJ	1. comparative → positive 2. superlative → positive	1. greater → great 2. greatest → great
ADV	1. better → well	
NUM	1. remove comma	
AUX	1. third person → first person 2. past tense → present tense 3. 'm, am, 's, is, 're, are, art, was, were, being, been → be 4. 've, has, had, having → have 5. does, did, done, doing → do 6. 'll → will 7. 'd → would	1. has → have 2. had → have
PART	1. n't → not	
PRON	1. object → subject 2. possessive (with 's) -> not possessive (without 's)	1. him → he 2. who's → whose

Algorithm 1: The execution steps for the rule-based lemmatizer

1	Input: User input text in Rojak language
2	Output: Lemmatized words in either English or Malay
3	Step 1: User inputs sequence of words in Rojak language
4	Step 2: Convert the received text to lowercase
5	Step 3: Detect language
6	Step 4: Check the received POS tag of the received word (perform on both English and Malay text)
7	Step 5: Perform lemmatization based on the received POS tag of the word where there will be 2 cases:
8	if the POS tag indicates that no lemmatization should be performed on the word, the word will be returned as the lemma
9	else the word will be lemmatized according to the rules defined for lemmatizing words with the specific POS tag
10	Step 6: Check if the output lemma exists in the corpus where there will also be 2 cases:
11	if the lemma exists in the corpus which means that it is a valid word, return the lemma

(continued)

(continued)

12	else the received word will be returned as the lemma
----	---

The rules for lemmatizing English words were derived from linguistic knowledge and observing the word-lemma pairs in the training set. As there are many irregular verbs and plural forms in English, some irregular words found from the training set are explicitly added to the list of rules for lemmatizing English words to increase coverage and improve accuracy. As for lemmatizing Malay words, the rules are mainly derived from the findings of previous studies supplemented by linguistic knowledge [17].

3.6 *Corpus-Based*

As for the corpus-based lemmatizer, the training dataset and development database from UD Treebanks for the two different languages were used as training data. Test set was used for evaluating the corpus-based lemmatizer to determine the accuracy of the lemmatizer on unseen dataset. The idea behind this approach is to build a corpus containing all the possible lemmas for a word according to the different POS tags, for each language. After the corpus has been built, it can then be used to look for a matching word form and POS tag to return its corresponding lemma, given an input word and POS tag.

Therefore, to build the corpus, lowercasing was first performed on the UD Treebanks data. The words and lemma from the training set were then used to build a dictionary corpus for the lemmatizer. Words in the training set that are not in the dictionary were added to the corpus together with the POS tag. At the same time, the POS tag of the word was checked if it exists in the corpus. In cases where the POS tag was not found in the corpus, it was appended to the corpus along with its lemma. Hence, to use the lemmatizer, it takes a word and its POS tag as input and then checks the corpus to determine whether the given word and POS tag exist in the corpus. If they exist, the corresponding lemma is returned; otherwise, the received word is returned.

3.7 *Machine Learning-Based*

In machine learning-based lemmatizer, the train and development sets were combined to form a train set for training the lemmatizer. Features extraction was then performed on the words and lemma through fastText word embedding. In this study, decision tree regression model (DTR) was selected as it managed to yield the highest accuracy during the lemmatization process on the real-world corpora for the Russian language [8]. Thus, DTR was expected to be able to produce outstanding results in Rojak

Table 3 Rules for Malay rule-based lemmatizer

POS tag	Rules	Example
NOUN	<ol style="list-style-type: none"> 1. anti+ 2. peN+ 3. +an 4. antar+ 5. ke+ 6. +wan 7. per+ 8. +wati 9. ke + tidak+ 	<ol style="list-style-type: none"> 1. olahragawan → olahraga 2. pengiriman → kirim 3. perbaikan → baik
ADJ	<ol style="list-style-type: none"> 1. non+ 2. ter+ 3. +an 4. ke+ 5. +nya 6. se+ 	<ol style="list-style-type: none"> 1. tercantik → cantik 2. kecantikan → cantik 3. secantik → cantik
VERB	<ol style="list-style-type: none"> 1. meN+ 2. per+ 3. +kan 4. di+ 5. + i 6. ber+ 	<ol style="list-style-type: none"> 1. mengambil → ambil 2. ber duduk → duduk 3. diletak → letak
NUM	<ol style="list-style-type: none"> 1. nya+ 2. ber+ 	<ol style="list-style-type: none"> 1. bertujuh → tujuh 2. tujuhnya → tujuh
PRON	<ol style="list-style-type: none"> 1. nya → dia 2. ku → aku 3. mu → kamu 4. kau → kamu 	

language lemmatizer. The DTR takes the word embedding of the inflected word as the input and learns to predict its corresponding output, which is the word embedding of the lemma of the input word. Depending on the type of approach, it can also take the encoded POS tag of the inflected word as input, allowing the DTR to learn from both the word and POS tag.

The DTR is a model based on the decision tree (DT) algorithm that learns using a tree structure that contains a root node, decision nodes, edges, and leaf nodes. It starts from the root node by selecting an attribute and splitting values as its starting point and continuously repeats this process to generate decision nodes, until it reaches its leaf nodes. The leaf nodes contain the values of the final prediction generated by the DT. In this case, the DTR uses the DT algorithm to perform lemmatization that is posed as a regression problem, whereby the predicted lemmas are represented by their feature vectors, which are sequences of continuous values. Using the generated tree, a path leading to the leaf nodes can be found to obtain the predicted values, which can also be expressed in rule form. Hence, the DTR is able to perform predictions and generate results for performing lemmatization [18].

Table 4 Parameters of the LSTM

Layers	Parameters	
	With POS tag	Without POS tag
Masking layer	mask_value = 0	mask_value = 0
LSTM layer	units = 317 return_sequences = true	units = 300 return_sequences = true
Dropout layer	rate = 0.2	rate = 0.2
Time distributed layer	layer = Dense units = 300 activation = 'linear'	layer = Dense units = 300 activation = 'linear'

3.8 Deep Learning-Based

Similar processes to the machine learning-based lemmatizer were performed. LSTM was employed for the lemmatization task. Similar to the machine learning-based approach, the LSTM also takes word embedding of the inflected word and optionally, encoded POS tags, and lemma as inputs and output respectively for training. With this, it is then able to predict the embedding of the lemma of a given word based on its embedding and optionally, its encoded POS tag.

LSTM is a variant of recurrent neural network (RNN) that retains its chain-like structure. However, the difference between the two is that RNN has difficulties dealing with long term dependencies, which are sequential data that require previous data or context to be retained for a longer period due to the vanishing gradient problem. LSTM overcomes this problem with the use of cell state, which differentiates its recurrent unit architecture from RNNs. It uses three types of gates, namely the forget gate, input gate, and output gate, in the cell state to control addition and deletion of information from the recurrent unit, while allowing information to flow through it using the door mechanism [19]. The parameters of the layers in the LSTM with POS and without POS tags are as shown in Table 4. About 100 epochs with a batch size of 4 were used to train the model. The chosen loss function was cosine_similarity, and the optimization technique utilized was rmsprop.

3.9 Evaluation

Both the English and Malay lemmatizers built on the different methods were evaluated using the accuracy (Eq. 1). To perform the evaluation, the test sets of the two corpora were used.

$$\text{Accuracy} = \frac{\text{Number of Words Correctly Lemmatized}}{\text{Total Number of Words in the Test Set}} \quad (1)$$