JOURNAL OF DIGITAL SYSTEM DEVELOPMENT

https://e-journal.uum.edu.my/index.php/jdsd

How to cite this article:
Bakar, J. A., Ariffin, M. A., Harun, N. H., Rasli, R. M., Saad, N. H. M. & Lisnawita (2024). Malay Named Entity Recognition System Using Machine Learning for Tourism in Malaysia. *Journal of Digital System Developments, 2* (2), 46-63. https://doi.org/10.32890/jdsd2024.2.2.4

MALAY NAMED ENTITY RECOGNITION SYSTEM USING MACHINE LEARNING FOR TOURISM IN MALAYSIA

[1]Juhaida Abu Bakar, [2]Muhammad Asyraf Ariffin, [3]Nor Hazlyna Harun, [4]Ruziana Mohamad Rasli, [5]Nurul Huda Mohamad Saad, [6]Lisnawita

[123]Data Science Research Lab, School of Computing,
Universiti Utara Malaysia, Sintok, Kedah, Malaysia

[4] School of Multimedia and Communication,
Universiti Utara Malaysia, Sintok, Kedah, Malaysia

[5]Academy of Language Studies,
Universiti Teknologi MARA (UiTM), Alor Gajah, Melaka, Malaysia

[6]Faculty of Computer Science,
Universitas Lancang Kuning, Indonesia

[1]Corresponding author: juhaida.ab@uum.edu.my

Received: 05/08/2024 Revised: 19/08/2024 Accepted: 26/08/2024 Published: 20/10/2024

ABSTRACT

Analysing unstructured textual data has become increasingly common due to its rich informational value across various fields. Named Entity Recognition (NER) is crucial for identifying entities in open-domain text documents. Current NER techniques often rely on manually labelled documents, which are time-consuming and prone to inaccuracies. While methods such as Spacy and Polyglot have been used, more research is needed on applying machine learning techniques to this problem. This work addressed this gap by developing a Malay language NER system using machine learning. The system used available Malay corpus resources to identify, learn, tag, and store entities from Malay texts. It was designed to handle structured and unstructured data, extracting names of people, places, organisations, and other entities. The Malay NER System using Machine Learning was developed as a web-based application. It employed advanced machine learning models, specifically BERT and ALXLNET, to process and analyse data. This study shows good agreement among the respondents regarding the usability, perception, and feedback on the specific pages, with the lowest mean score being 76.67%. Regarding system functionalities, there is room for refinement to ensure more accurate and reliable output. The system featured a web interface allowing users to input Malay text and receive recognised entities as output. Performance was assessed using standard evaluation metrics. This work advanced natural language processing capabilities in Malay by creating a user-friendly, efficient tool for NER in Malay.

Keywords: Named Entity Recognition, Malay Text, Machine Learning, BERT, ALXLNET

INTRODUCTION

With the emergence of the internet and modern technology, vast amounts of data and information are now available and distributed in multiple languages in digital form. This information may be stored in either structured or unstructured formats. Structured data typically follows a specific format or schema, while unstructured data does not have a predefined format or organisation. Today, a vast amount of unstructured data is available online in various formats, including images, running text, audio, video clips, and time series data (Zadgaonkar & Agrawal, 2024). Such information must be processed and extracted through Natural Language Processing (NLP) tasks. Analysing unstructured textual data is becoming increasingly common, as this type offers valuable and helpful information across various fields. Natural language processing (NLP) is a significant area of study in computer science. Text analysis using various approaches and technologies is known as NLP (Sazali et al., 2016). Named Entity Recognition (NER) is one of the many tasks that utilise NLP in text analysis. NER is a method for identifying entities in open-domain text documents. Current NER techniques are often performed using documents that have been manually labelled, which can be time-consuming.

The primary objective of NER is to reduce the arduous and time-consuming manual annotation of named entities in texts by human annotators (Ulanganathan et al., 2017). However, machines must be trained to examine and comprehend the text's content to identify the listed entities. While many corpora are available for widely used languages such as English and Chinese, a new corpus must be created to train the Malay NER model due to the need for existing corpora for the Malay language. A NER-annotated corpus is necessary to develop a machine learning model to identify the appropriate entity types for new words and phrases based on context.

Named Entity Recognition (NER) is widely regarded as the backbone of Natural Language Processing (NLP) applications such as information extraction and retrieval, text mining, machine translation, and more. In recent years, there has been a dramatic increase in data generation and sharing due to the constant development of various social network platforms. Much of this data is unstructured, containing information that could be beneficial if properly analysed. This has led to the application of NER solutions in various fields, including education, healthcare, industry, business, and politics. Effectively utilising this unstructured data requires significant time and effort. Information extraction, one of the active research fields, aims to extract potentially relevant information from large amounts of data (Rosmayati et al., 2020). However, the implementation methods, particularly for the Malay language, are constrained for certain unstructured data types. A comprehensive textual analysis process is necessary to obtain crucial information for decision-making. Manual annotation of articles often results in inaccurate findings. Previous research has implemented methods such as Spacy (Honnibal, 2017) and Polyglot (Rami, 2015). However, there are limited studies that have implemented machine learning techniques. Therefore, an automated information extraction process using machine learning is specifically required.

This project uses machine learning methods to build a Named Entity Recognition (NER) system for the Malay language. Equipped with sufficient Malay corpus resources, the system can identify entities in Malay text. It can also learn new words, tag them, and store them in the system's corpus database. The system can also detect and extract information from structured and unstructured Malay texts, such as people's names, places, organisation names, and other entities in Malay text documents.

The three objectives achieved in this work are to identify the requirements for Malay Named Entity Recognition using a machine learning model, to develop a prototype of Malay Named Entity Recognition using a Machine Learning Model (Malay NER – ML), and to evaluate the Malay NER – ML model based on standard statistical measurements. This work established a baseline for developing NER systems using machine learning techniques specifically for Malay text. With advancements in technology, increased information, and enhanced methods, this work was anticipated to improve in accuracy and reliability.

RELATED WORKS

Research conducted by Rayner et al. (2014) implemented a rule-based algorithm to develop a Malay Named Entity Recognition (NER) system. The study found that the proposed Malay NER achieved results comparable to existing NER algorithms for other languages, with an F-measure of 89.47%, a recall rate of 94.44%, and an accuracy rate of 85%. The researchers concluded that the predefined rules and dictionaries used in the named entity recognition process need improvement for the Malay NER algorithm. They identified updating all utilised libraries as the biggest challenge in developing a functional Malay NER. Therefore, a practical approach is necessary to update the list of current dictionaries consistently.

In other research, Salleh et al. (2017) developed a Named Entity Recognition (NER) model using the Conditional Random Fields (CRF) method, which is one of the techniques for NER systems. The CRFsuite Python library and a dataset from Bernama News were used to implement the CRF method for the training and testing. The overall F1-score, or test accuracy, was 70.0% because the training model was developed using a very small dataset. The evaluation results were affected by the model being trained on a limited amount of data. Therefore, a larger dataset is required to improve the performance of a Malay NER model.

Current NER studies on the Malay language, conducted by Safwan et al. (2021), have developed the Malay Roman Corpus Annotation system, which can define and extract Malay Roman NER characteristics from unstructured Malay documents. The system, developed using Flask Python and powered by Polyglot, follows a five-phase method: sentence segmentation, tokenisation, part-of-speech tagging, entity recognition, and relationship recognition. The functionalities developed include managing users, managing text input, clearing text, viewing entity labels, analysing text, and managing results.

Research conducted by Asmai et al. (2018) developed a Malay NER model using a combination of fuzzy c-means and the K-Nearest Neighbours (k-NN) algorithm for crime analysis. Their system achieved an overall accuracy of 95.24% during k-NN classification. From an overall perspective of the evaluation process, the accuracy can be improved by increasing the training dataset for better results. Continuous focus on selecting appropriate features is necessary, as these features significantly affect the performance of the NER model, especially for the Malay language, which has a complex sentence structure. The researchers believe their system can be further improved by increasing the corpus references in Malay texts.

Continuing their work on NER studies, Salleh et al. (2018) implemented the fuzzy c-means method on the proposed Malay NER using RapidMiner software with a dataset extracted from Bernama Malay news. The analysis process involved three phases: preprocessing data, fuzzy c-means analysis, and evaluation of the results. The capitalisation feature was identified as a critical element for extracting named entities, as every named entity or proper noun in the Malay language's Part of Speech (POS) category begins with a capital letter. They discovered that using several features, such as POS tag, POS assigned value, character length, token position in the document, term frequency (t), capitalisation, and TF-IDF, improved the accuracy of results for Malay named entity recognition.

Research conducted by Mbouopda and Yonta (2019) proposed a named entity annotator for low-resource languages, specifically targeting Ewondo, a Bantu language of Cameroon. The model for this NER system is based on the POS tag projection model developed by Zennaki et al. (2015). They manually created a named entity-annotated English-Ewondo parallel corpus using a dataset extracted from an online Bible, which provides both English and Ewondo versions of each verse. The experiment results indicate that their model performs well on this dataset, achieving a precision of 76.69%, a recall of 66.33%, and an F1-score of 70.18% in detecting named entity tags from English to Ewondo.

Lexicon-based techniques link text sections with entity names using a lexicon or gazetteer derived from external knowledge sources. A study by Ma et al. (2020) simplified the usage of lexicons in Chinese NER, demonstrating how effectively incorporating word lexicons can enhance system performance. Their research developed a novel method to integrate lexicon information into character representations, achieving a high-performing Chinese NER system with quick inference times. Experiments on four benchmark Chinese NER datasets showed that their approach outperformed state-of-the-art methods in both inference speed and overall performance.

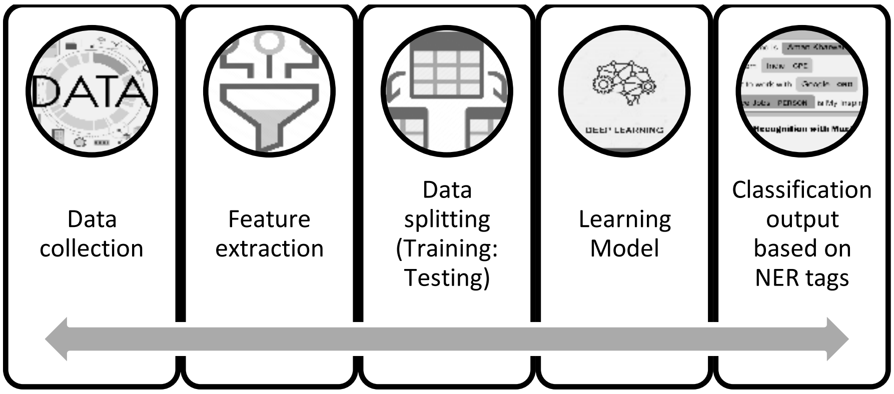
Based on the related works above, it is evident that the implementation of Machine Learning is minimal. Therefore, this study uses machine learning models such as BERT and ALBERT to extract relevant features and recognise named entities based on Zolkepli's (2018) software. This study also aims to develop a prototype system of Malay NER using machine learning.

METHODOLOGY

Five main phases are involved in developing the NER model: data collection, feature extraction, training-testing data splitting, model learning, and output classification based on NER tags, as shown in Figure 1.

Figure 1

The phases of the methodology


DATA
Data collection
Feature extraction
Data splitting (Training: Testing)
DEEP LEARNING
Learning Model
Classification output based on NER tags

First phase: Data collection

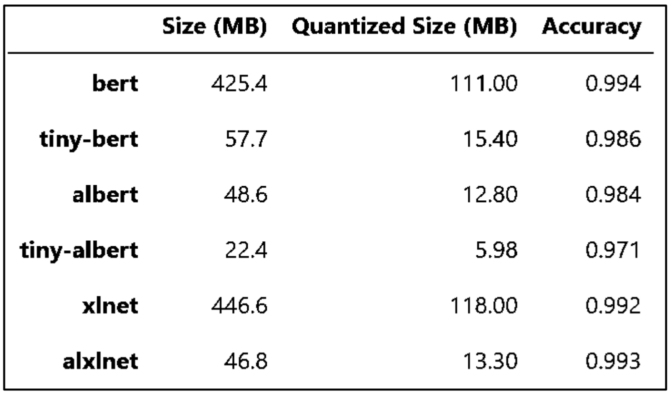
The first phase involves data collection. The raw dataset used in developing this model consists of articles about tourist attractions in Malaysia, extracted from various leading Malaysian newspapers such as Utusan Malaysia, Berita Harian, and Harian Metro. The dataset covers tourist attractions across all 13 states in Malaysia and includes articles written by different journalists, providing a mix of perspectives and writing styles. With over 100,000 articles published over several years, this dataset offers a substantial amount of text data, making it suitable for robust training of models, particularly for tasks like Natural Language Processing (NLP) and Named Entity Recognition (NER). In this work, we extracted only about 1% of the articles from the newspaper.

Second phase: Feature extraction

Preprocess the data using the BERT and XLNET models based on Zolkepli's (2018) software, the 'Malaya' package. BERT and XLNET were chosen for their high accuracy in NER tasks, with BERT achieving 99.4% accuracy and XLNET achieving 99.3% accuracy. Figure 2 shows the performance of these deep learning models in the Malaya package.

Figure 2

Available Transformer NER models (Source: Zolkepli, 2018)

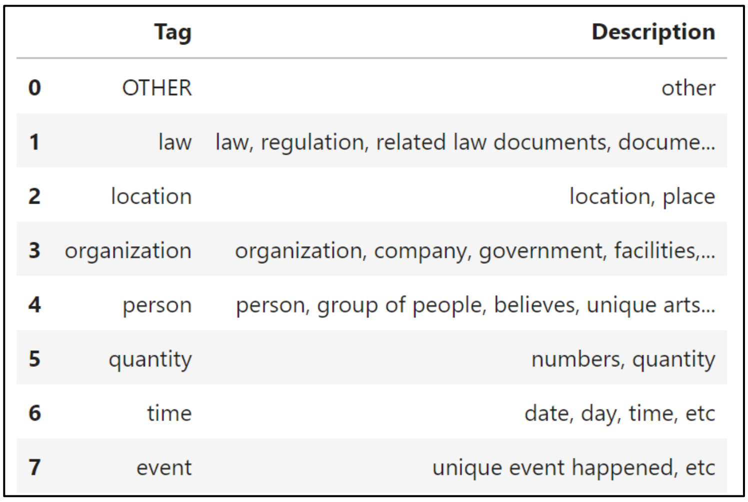


	Size (MB)	Quantized Size (MB)	Accuracy
bert	425.4	111.00	0.994
tiny-bert	57.7	15.40	0.986
albert	48.6	12.80	0.984
tiny-albert	22.4	5.98	0.971
xlnet	446.6	118.00	0.992
alxlnet	46.8	13.30	0.993

These models can classify entities by location, organisation, person, quantity, time, event, and law. Any words not corresponding to these NER tags are labelled 'OTHER'. BERT and other deep learning models can be utilised by passing them through `malaya.entity.transformer(model=model)`. These models extract relevant features and represent the Malay text in a suitable format for named entity recognition (NER). The output is then transformed into a CSV format with four columns: word, NER tags, model, and file. Figure 3 shows the NER tags extracted using the Malaya package, and Figure 4 shows the features extracted from the Terengganu articles along with the NER tag information.

Figure 3

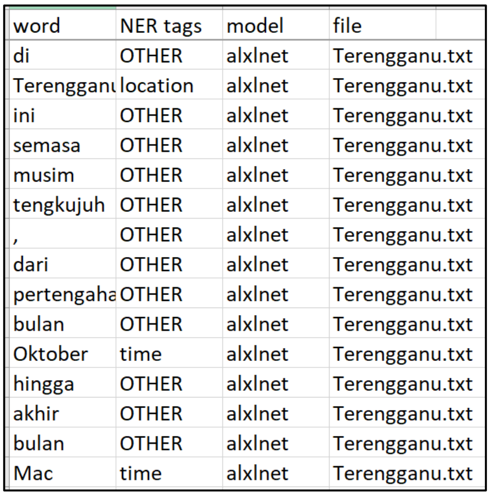
The description of supported entities (Source: Zolkepli, 2018)



	Tag	Description
0	OTHER	other
1	law	law, regulation, related law documents, docume...
2	location	location, place
3	organization	organization, company, government, facilities,...
4	person	person, group of people, believes, unique arts...
5	quantity	numbers, quantity
6	time	date, day, time, etc
7	event	unique event happened, etc

Figure 4

The features extracted in the Terengganu article



word	NER tags	model	file
di	OTHER	alxlnet	Terengganu.txt
Terenggani	location	alxlnet	Terengganu.txt
ini	OTHER	alxlnet	Terengganu.txt
semasa	OTHER	alxlnet	Terengganu.txt
musim	OTHER	alxlnet	Terengganu.txt
tengkujuh	OTHER	alxlnet	Terengganu.txt
,	OTHER	alxlnet	Terengganu.txt
dari	OTHER	alxlnet	Terengganu.txt
pertengaha	OTHER	alxlnet	Terengganu.txt
bulan	OTHER	alxlnet	Terengganu.txt
Oktober	time	alxlnet	Terengganu.txt
hingga	OTHER	alxlnet	Terengganu.txt
akhir	OTHER	alxlnet	Terengganu.txt
bulan	OTHER	alxlnet	Terengganu.txt
Mac	time	alxlnet	Terengganu.txt

Third phase: Training-Testing splitting

All features were then split into training and testing datasets for learning purposes. The dataset was divided with a ratio of 70% for training and 30% for testing.

Fourth phase: Learning model

The training data, consisting of text, is converted into numerical features using a CountVectorizer. This counts the occurrences of words in each text, creating a matrix of word counts. Then, a TfidfTransformer is applied to this matrix, which calculates the TF-IDF values for each word, giving more importance to informative words. The transformed matrix trains a Multinomial Naïve Bayes (Multinomial NB) classifier that learns patterns in the data. Multinomial NB is a commonly used algorithm in text classification tasks, including Named Entity Recognition (NER). Finally, the trained CountVectorizer and classifier models are saved as files for future use.

Fifth phase: Classification output based on NER tags

A trained model is created based on the compiled training dataset, which includes 1,000 articles. Using this trained model, saved as a ‘pickle’ file, the expected output can predict the named entity of the words that the user enters.

DESIGN AND DEVELOPMENT OF MALAY-NAMED ENTITY RECOGNITION USING MACHINE LEARNING

This section describes designing and developing a web-based Malay Named Entity Recognition (NER) system using machine learning algorithms. The development process follows the first six phases of the Waterfall development methodology. The NER system is designed to fulfil the requirements for Malay Named Entity Recognition by leveraging the power of machine learning. Requirements for the system were gathered through a thorough process that involved analysing relevant documents, conducting interviews, and researching existing NER systems. The gathered requirements are categorised into several vital functionalities: registration, login, data management, model management, visualisation, and contact information.

1. Requirement Gathering Process

During the requirement-gathering process, various sources were consulted, including academic papers, industry reports, and existing NER systems. Open-ended questions were used to gather information on the desired features and functionalities of the NER system. Additionally, interviews were conducted with domain experts and potential users to gain further insights into their requirements and expectations. The gathered requirements were carefully analysed and organised based on relevance and priority.

2. Requirements for Malay Named Entity Recognition System

The requirements obtained from the requirement-gathering process are presented in Table 1, along with their respective priority levels. These requirements reflect the critical functionalities of the Malay NER system, with a specific focus on utilising machine learning algorithms for accurate entity recognition.

Table 1

List of Requirements for Malay Named Entity Recognition System

Requirements ID	Description	Priority
MNM_01	*Register*	
MNM_01_01	Users and admins enter registration details	M
MNM_01_02	The system stores the user and admin details	M
MNM_02	*Login and Logout*	
MNM_02_01	The system allows admin registration	M
MNM_02_02	The system allows user registration	D
MNM_02_03	The system allows users to log in	M
MNM_02_04	The system displays a success message for registration and login	O
MNM_03	*Manage Corpus*	
MNM_03_01	The admin can view the corpus	D
MNM_03_02	The admin can edit the corpus	D
MNM_03_02	The admin should be able to add data to the corpus	D
MNM_03_04	The system should be able to store training data as the initial corpus	D
MNM_04	*Manage Model*	
MNM_04_01	The system should be able to detect the Malay language	M
MNM_04_02	The system should be able to identify Malay entities based on POS tagging	M
MNM_04_03	Users and admins should be able to	M

	enter Malay text	
MNM_05	*Reliability Issues*	
MNM_05_01	The system should run normally with the provided dataset	M
MNM_06	*Usability Issues*	
MNM_06_01	The system should be able to detect Malay entities in user-entered text	M

The requirements presented in Table 1 were translated into the computer system functionality.

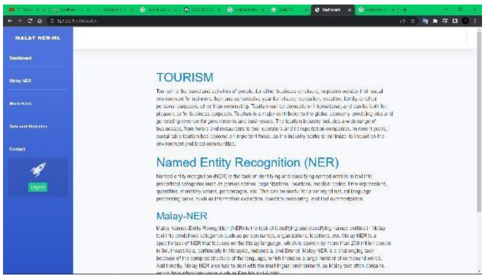
3. The Development of a Web-Based Interface for the Malay Named Entity Recognition System

Developing the web-based interface for the Malay Named Entity Recognition (NER) system involved utilising preprocessed data from BERT and ALXLNET models. The system was trained on tourism attraction articles from the 13 states in Malaysia. The goal was to create a user-friendly webpage with several pages dedicated to providing information and functionalities related to NER and tourism. Below are the critical components of the web-based interface.

A. Homepage: The homepage (Figure 5) serves as the system's main landing page. It provides information about the NER system, its capabilities, and the relevance of named entity recognition in tourism. The homepage also highlights the key features and benefits of the system.

Figure 5

The interface of the main page of the Malay Named Entity Recognition System



TOURISM

Tourism is the travel and activities of people, for either business or leisure, to places outside their usual environment for not more than one consecutive year for leisure, recreation, vacation, family, or other personal purposes, other than commuting. Tourism can be domestic or international, and can be both for pleasure or for business purposes. Tourism is a major contributor to the global economy, providing jobs and generating revenue for governments and businesses. The tourism industry includes a wide range of businesses, from hotels and restaurants to tour operators and transportation companies. In recent years, sustainable tourism has become an important focus, as the industry works to minimize its impact on the environment and local communities.

Named Entity Recognition (NER)

Named entity recognition (NER) is the task of identifying and classifying named entities in text into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. This can be useful for a variety of natural language processing tasks, such as information extraction, question answering, and text summarization.

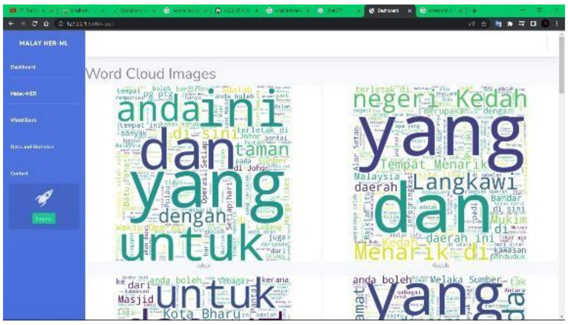
Malay-NER

Malay Named-Entity Recognition (NER) is the task of identifying and classifying named entities in Malay text into predefined categories such as person names, organizations, locations, etc. Malay NER is a specific task of NER that focuses on the Malay language, which is spoken by more than 200 million people in Southeast Asia, particularly in Malaysia, Indonesia, and Brunei. Malay NER is a challenging task because of the complex structure of the language, which includes a large number of compound words. Additionally, Malay NER also has to deal with the multilingual environment, as Malay text often contains

B. Wordbank Page: The Wordbank page (Figure 6) highlights word clouds based on the 13 states in Malaysia. These word clouds provide a visual representation of the frequently occurring words in the tourism attraction articles from each state. Users can explore the word clouds to gain insights into the popular tourism-related entities and themes in different states.

Figure 6

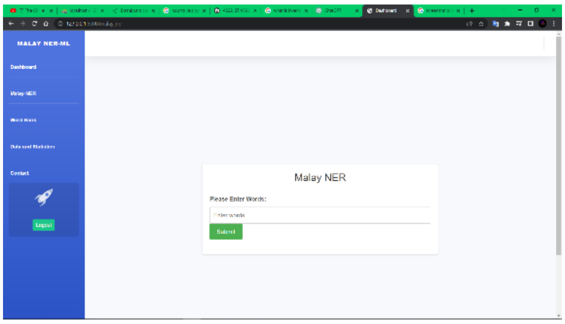
The interface of the Word cloud page of the Malay Named Entity Recognition System



C. Malay NER Page: The Malay NER page (Figure 7) is the core functionality of the web-based interface. Users can input Malay text into the system, and the NER system, powered by BERT and ALXLNET models, performs entity recognition on the text. The recognised named entities are displayed to the users, enabling them to identify and extract meaningful information from the text.

Figure 7

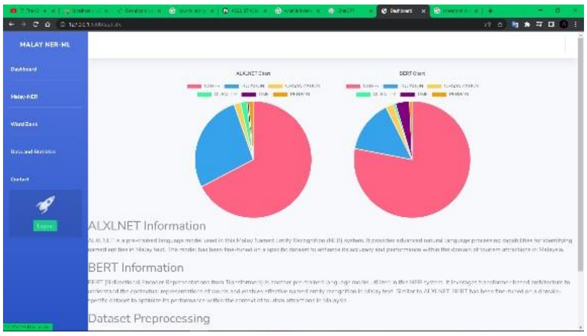
The interface of the Malay NER page of the Malay Named Entity Recognition System



D. Data and Statistics Pages: The Data and Statistics page (Figure 8) contains the models used to preprocess the dataset. Users can access details about the BERT and ALXLNET models, including their architecture, training process, and performance metrics. This Section provides transparency and insights into the underlying models used in the NER system.

Figure 8

The interface of the Data and Statistics page of the Malay Named Entity Recognition System



ALXLNET Information

ALXLNET is a pre-trained language model used in this Malay Named Entity Recognition (NER) system. It provides advanced natural language processing capabilities for identifying named entities in Malay text. The model has been fine-tuned on a specific dataset to enhance its accuracy and performance within the domain of tourism attractions in Malaysia.

BERT Information

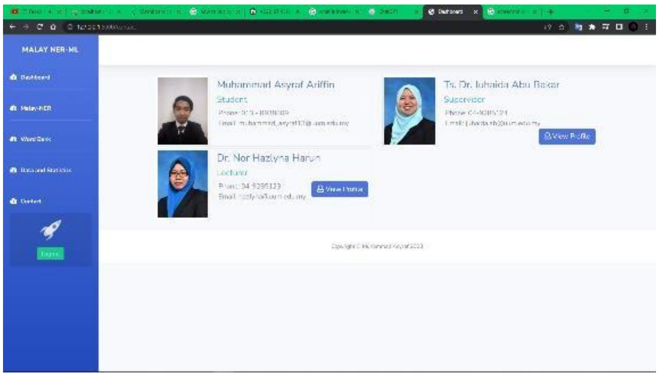
BERT (Bidirectional Encoder Representations from Transformers) is another pre-trained language model utilized in the NER system. It leverages transformer-based architecture to understand the contextual representations of words and enables effective named entity recognition in Malay text. Similar to ALXLNET, BERT has been fine-tuned on a domain-specific dataset to optimize its performance within the context of tourism attractions in Malaysia.

Dataset Preprocessing

E. Contact Page: The Contact page (Figure 9) includes contact information for the system developer, supervisor, and lecturer. Users can reach out for inquiries, feedback, or collaboration opportunities. This page enhances communication and fosters engagement between the system developer and users.

Figure 9

The interface of the Contact page of the Malay Named Entity Recognition System



Muhammad Asyraf Ariffin
Student
Phone: [illegible]
Email: [illegible]

Ts. Dr. Juhaida Abu Bakar
Supervisor
Phone: [illegible]
Email: [illegible]

Dr. Nor Hazlyna Harun
Lecturer
Phone: 04-9285123
Email: [illegible]

Copyright © Muhammad Asyraf 2023

The development of the web-based interface for the Malay Named Entity Recognition System leveraged preprocessed data from BERT and ALXLNET models trained on tourism attraction articles from the 13 states in Malaysia. The interface consists of multiple pages, including the homepage, Wordbank page, Malay NER page, Data and Statistics pages, and Contact page. This comprehensive web-based interface aims to provide users with valuable insights into tourism entities in Malaysia and facilitate efficient named entity recognition in Malay text.

4. Evaluation of Malay Text Simplification System

A. The Evaluation Setting

The respondents recruited for the usability test are students from Universiti Utara Malaysia. A total of 30 students participated in this usability test. The participants were recruited from the states where the students were initially from. The type of evaluation was conducted, and face-to-face usability tests were performed. Usability testing involves assessing the functionality of a website, app, or other digital product by observing real users as they attempt to complete tasks. Field testing was conducted to evaluate the Malay Named Entity Recognition System's significance, achievement of work objectives, development, effectiveness, efficiency, impact, and sustainability. Participants were asked to use the web-based system and complete a Google Form questionnaire to evaluate the application. The questionnaire consisted of five sections: Section A - Demographic information, Section B - System usability, Section C - Respondents' perception of tourism in Malaysia, Section D - Feedback on specific pages and Section E - Overall feedback on the Malay Named Entity Recognition System. Most sections used a four-point Likert scale, including Strongly Disagree, Disagree, Agree, and Strongly Agree. The respondents' information, responses, and feedback were kept private and used only for research.

B. The Respondents' Demographic Information

There are a total of 30 participants for this usability test. The respondents for the test are recruited among two or three people from each state in Malaysia. Analysis of the respondents 'demographic information also revealed that most respondents do not need to become more familiar with the Malay NER concept. 70% of 21 respondents must become more familiar with the Malay NER concept.

C. The Usability of Malay Named Entity Recognition System

An analysis was conducted on the respondents' answers in Section B of the post-task questionnaire, as shown in Table 2. This Section measures the respondents' perceptions of the usefulness and ease of use of the web-based system's functionality and their satisfaction with the system. It consists of five statements. Based on the responses, the statement "The homepage is visually appealing and informative", and the statement "The instructions for using the NER function are clear and easy to follow" each received a mean score of 96.67% for "Agree" and "Strongly Agree" responses.

D. Perception of Tourism in Malaysia

An analysis was conducted on the responses from the respondents in Section B of the post-task questionnaire, as in Table 3, which aimed to measure their perceptions regarding the web-based system's usefulness and ease of use. This Section consisted of five statements related to the Malay Named Entity System.

The mean scores for each statement provide insight into the respondents' satisfaction levels. Most respondents agreed with each statement. In particular, the statements "The use of our website and its NER system positively influenced my perception of tourism in Malaysia" and "I am likely to revisit our website in the future to explore more about tourism in Malaysia" were chosen as the most agreed-upon by the respondents.

E. Feedback on Specific pages

The feedback on the "Specific Pages" section of the post-task questionnaire provides insights into the respondents' perceptions and experiences regarding the different pages of the NER system, as shown in Table 4. For the NER page, the mean score for the statement "The NER page helps identify Malay named entities" was 80% for "Agree" and "Strongly Agree". This indicates a good level of agreement among the respondents, suggesting that the respondents found the NER page helpful for identifying Malay-named entities. Similarly, for the statement, "The NER page helps identify attractions in Malaysia (location, organisation, people)," the mean score was 73.33%. This indicates a slightly lower agreement level than the first statement among the respondents.

In terms of the accuracy and reliability of the NER function, the respondents' mean score for the statement "The results provided by the NER function are accurate and reliable" was 56.67% for "Agree" and "Strongly Agree". This indicates a relatively lower level of agreement, suggesting that there may be room for improvement in the accuracy and reliability of the NER function.

Moving on to the word cloud page, the respondents' mean score for the statement "The word cloud page is visually appealing and informative" was 96.67% for "Agree" and "Strongly Agree." This indicates a high level of agreement, suggesting that the respondents found the word cloud page visually appealing and informative.

Similarly, for the statement "I did not encounter any difficulties in interpreting the word cloud," the respondents' mean score was also 96.67% for "Agree" and "Strongly Agree." This indicates a high level of agreement, suggesting that the respondents did not experience any difficulties in interpreting the word cloud.

Overall, the feedback from the respondents provides valuable insights into their perceptions and experiences with the NER system. The feedback highlights areas of strength, such as the visual appeal and informativeness of the word cloud page, while also indicating areas for improvement, particularly regarding the accuracy and reliability of the NER function.

Table 2

The Respondents' Responses on the System Usability

	Mean, n=30			
Statement	Strongly Disagree	Disagree	Agree	Strongly Agree
The overall interface of our website is user-friendly and visually appealing.	0.00	0.00	0.60	0.40
I was able to navigate through the different pages easily	0.00	0.00	0.57	0.43
The homepage is visually appealing and informative	0.00	0.03	0.73	0.23
I did not encounter any issues or errors while using the NER functionality	0.00	0.07	0.70	0.23
The instructions for using the NER function are clear and easy to follow	0.00	0.03	0.73	0.23

Table 3

The Respondents' Responses on the Perception of Tourism in Malaysia

	Mean, n=30			
Statement	Strongly Disagree	Disagree	Agree	Strongly Agree
Using our website and its NER system positively influenced my perception of tourism in Malaysia.	0.00	0.20	0.60	0.20
The NER system on our website helped me discover new tourist attractions or activities in Malaysia.	0.00	0.23	0.57	0.20
The NER system and the information provided on our website enhanced my engagement with Malaysia's tourism offerings.	0.00	0.27	0.53	0.20
I plan to revisit our website to explore tourism in Malaysia more.	0.00	0.20	0.60	0.20

Our website increased my confidence in planning a trip to Malaysia by providing relevant tourism information.	0.00	0.23	0.57	0.20

Table 4

The Respondents' Responses on the Feedback on the Specific Pages in the System

	Mean, n=30			
Statement	Strongly Disagree	Disagree	Agree	Strongly Agree
The NER page helps identify Malay-named entities	0.00	0.20	0.73	0.07
The NER page helps identify attraction in Malaysia entities (location, organisation, people)	0.00	0.27	0.70	0.03
The results provided by the NER function are accurate and reliable	0.00	0.43	0.53	0.03
The Word Cloud page is visually appealing and informative	0.00	0.03	0.73	0.23
I did not encounter any difficulties in interpreting the word cloud	0.00	0.03	0.73	0.23

The overall feedback from the respondents suggests that there are both positive and areas for improvement in the Malay Named Entity Recognition (NER) system. The usability analysis indicated that respondents found certain system aspects valuable and easy to use. However, some aspects could be further enhanced to improve the user experience.

Regarding system usability, respondents generally agreed that the system's functionalities were helpful, but they suggested improvements regarding the accuracy and reliability of the NER results. While the system effectively identifies Malay-named entities, there is room for refinement to ensure more accurate and reliable output.

Furthermore, it was observed that the respondents needed to be more familiar with the concept of Malay NER. This suggests that users may need to provide more information or guidance to help them better understand and utilise the system effectively.

Regarding the perception of tourism in Malaysia, the respondents generally had a positive experience with the system. They agreed that the system positively influenced their perception of tourism, helped them discover new attractions, enhanced their engagement with Malaysia's tourism offerings, and increased their confidence in trip planning. This indicates that the system has the potential to contribute positively to users' perceptions and experiences of tourism in Malaysia.

Regarding specific pages, the NER page was found to help identify Malay-named entities, but there was a slightly lower level of agreement when identifying attraction entities. This suggests there may be areas for improvement in accurately identifying specific types of attractions-related entities.

On the other hand, the word cloud page received positive feedback, with respondents finding it visually appealing and informative. They also reported no difficulties interpreting the word cloud, indicating a user-friendly and intuitive design.

Overall, the feedback suggests that while the system has valuable features and has the potential to positively influence users' perception of tourism in Malaysia, there are areas that require further development. Improving the accuracy and reliability of the NER results, providing more information and guidance on the Malay NER concept, and addressing any limitations in identifying attraction entities can help enhance the overall user experience and satisfaction with the system.

CONCLUSION AND FUTURE WORK

In conclusion, the development of the Malay Named Entity Recognition (NER) system has provided valuable insights into Malaysia's usability and perception of tourism. The system demonstrated usefulness in identifying Malay-named entities and contributing to users' understanding and engagement with tourism offerings. However, there are areas for improvement, particularly regarding the accuracy and reliability of the NER results and the identification of attraction entities.

For future work, continuously updating and expanding the tourism information on the system can provide users with a broader range of attractions, locations, and activities to explore. Collaborations with tourism authorities, local communities, and relevant stakeholders can facilitate the acquisition of up-to-date and comprehensive tourism data.

By addressing this area, the Malay NER system can evolve into a more accurate, reliable, and user-friendly tool for identifying Malay-named entities and enhancing users' perception and engagement with tourism in Malaysia.

ACKNOWLEDGMENT

This research has been carried out under Geran UUM Press (SO Code: 21246) provided by Universiti Utara Malaysia (UUM), Malaysia.

REFERENCES

Honnibal, M., & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. GitHub.

Salleh, M. S., Asmai, S. A., Basiron, H., & Ahmad, S. (2017). A Malay-named entity recognition using conditional random fields. In *2017 5th International Conference on information and Communication Technology (ICOIC7)* (pp. 1-6). IEEE.

Mbouopda, M. F., & Yonta, P. M. (2019). A Word Representation to Improve Named Entity Recognition in Low-Resource Languages. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 333-337). IEEE.

Salleh, M. S., Asmai, S. A., Basiron, H., & Ahmad, S. (2018). Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(2-7), 121-126.

Rosmayati, M., Nazratul Naziah, M. M., Noor Maizura, M. N., & Zulaiha, A. O. (2020, December 15). A Review of Named Entity Recognition and Classification on Unstructured Malay Data.

Rami, A.-R. (2015). Named Entity Extraction — polyglot 16.07.04 documentation. https://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.htm l

Ma, R., Peng, M., Zhang, Q., & Huang, X. (2019). Simplify the usage of the lexicon in Chinese NER—*arXiv preprint arXiv:1908.05969*.

Alfred, R., Leong, L. C., On, C. K., & Anthony, P. (2014). Malay named entity recognition based on a rule-based approach.

Sazali, S. S., Rahman, N. A., & Bakar, Z. A. (2016). Information extraction: Evaluating named entity recognition from classical Malay documents. In *2016, the third international conference on Information Retrieval and knowledge management (CAMP)* (pp. 48-53). IEEE.

Asmai, S. A., Salleh, M. S., Basiron, H., & Ahmad, S. (2018). An enhanced Malay named entity recognition using a combination approach for crime textual data analysis. *International Journal of Advanced Computer Science and Applications*, *9*(9), 474-483.

Chang, S. S., Bakar, J. A., & Katuk, N. (2021). Malay Roman Corpus Annotation System. Multidisciplinary Applied Research and Innovation, 2(3), 001–004.

Ulanganathan, T., Ebrahim, A., Xian, B. C. M., Bouzekri, K., Mahmud, R., & Hoe, O. H. (2017). Benchmarking Mi-NER: Malay entity recognition engine. In *9th International Conference on information, process, and knowledge management* (pp. 52-58).

Zadgaonkar, A., & Agrawal, A. J. (2024). An Approach for analysing unstructured text data using topic modelling techniques for efficient information extraction. *New Generation Computing*, *42*(1), 109-134.

Zennaki, O., Semmar, N., & Besacier, L. (2015). Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle. In *TALN 2015*.

Zolkepli, H. (2018). Malaya, a natural-language-toolkit library for Bahasa Malaysia, powered by Pytorch. https://github.com/huseinzol05/malaya.