Journal of Theoretical and Applied Information Technology
15th December 2020. Vol.98. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A REVIEW OF NAMED ENTITY RECOGNITION AND CLASSIFICATION ON UNSTRUCTURED MALAY DATA

[1]ROSMAYATI MOHEMAD, [2]NAZRATUL NAZIAH MOHD MUHAIT, [3]NOOR MAIZURA MOHAMAD NOOR, [4]ZULAIHA ALI OTHMAN

[1-3]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.

[4]Center for Artificial Intelligence Technology, Faculty Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.

Email: [1]rosmayati@umt.edu.my, [2]nazz1799@gmail.com, [3]maizura@umt.edu.my, [4]zao@ukm.edu.my

ABSTRACT

In recent years, due to the emergence of various social network platforms, a massive amount of data is continuously generated and shared. The majority of the data is unstructured, which contains information that might be crucial and valuable if analyzed. Effective use of these unstructured data is a tedious and labor-intensive task. Information extraction is one of the on-going research areas to extract potentially useful information out of voluminous data. Several different techniques and methods for information extraction have been proposed to understand the content and context of any available unstructured data at the low-level structure. However, there are limited studies conducted to investigate the challenges of Named Entity Recognition and Classification (NERC) on unstructured Malay data, which is known as one of the main subtasks in information extraction. Therefore, this paper addresses a comprehensive review of the existing NERC techniques for processing unstructured Malay data along with its limitations and challenges. The contributions of this paper are twofold. The primary contribution is it presents the overview of prior studies on NERC techniques of unstructured Malay data. Second, it scrutinizes the limitations and challenges of theses existing techniques due to the voluminous, dimensionality, and heterogeneity of unstructured Malay data. The findings show that most of the previous studies using a machine learning-based approach produce a satisfactory result rather than a rule-based approach. Furthermore, the challenges in terms of the different morphological of Malay language compared to resource-rich languages such as English, limitation of Malay corpus and annotated Malay text, and Malay text ambiguities could influence the performance of Malay NERC system efficiency, which should be carefully considered during the design of the systems.

Keywords: *Information Extraction, Malay Language, Named Entity Recognition and Classification, Unstructured Data*

1 INTRODUCTION

In this digital era, with most of the processes are moving towards digitization and the emergence of diverse social network platforms, huge volume, and variety of data either in text, image, audio or video is continuously generated and shared. According to the International Data Corporation (IDC), the growth of digital universe data is predicted to increase tremendously from 33 Zettabytes in 2018 to 175 Zettabytes by 2025 [1]. The vast majority of this never-ending data expansion is unstructured, which contributed as much as 95% of the global data in 2020 [2]. This type of data contains an abundant amount of information that might be crucial and valuable if analyzed to resolve issues in many domains such as medical informatics [3][4][5][6], geological [7], construction [8][9], consumer review and recommendation [10][11] and tender management [12].

The characteristics of unstructured data often appear in a non-standardization format, loads with heavy-text, and come from heterogeneous and diverse sources [2]. Due to the complexity of this unstructured data, extracting meaningful information from this type of data is one of the non-trivial tasks in text mining, and on-going researches are performed in many focus areas such as data mining, web mining, information retrieval, question

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

answering, and machine translation. Information extraction is defined as an organization of tasks between structured data from unstructured data and semi-structured data [13]. The ultimate goal of information extraction is to gain insight and obtain useful information hidden in the unstructured data and convert it into meaningful structured information so that it can be accessed and analyzed to help predict results, discover knowledge, and make informed decisions. This useful information can be categorized as entities, relations, facts, terms, patterns, and other types of information that could assist the data analysis.

The correct transformation and interpretation of these unstructured data in a well-processed form enhance the performance of information extraction. Named Entity Recognition and Classification (NERC) is a subtask in the domain of information extraction. NERC is characterized as the process of finding certain useful terms such as entity names (the names of persons, locations, organizations), temporal expressions (dates, times), and numerical expressions (monetary values, percentages), and classify them into rhetorical categories based on the surrounding context [14][15]. A set of predefined semantic categories is assigned or annotated to the words or phrases that are known as the recognized entity. The extracted terms of named entities, however, may vary, depending on the domain of interest and the language. Generally, a named entity is any real-world object that can be denoted as a word or combination of words or terms.

Various NERC techniques have been explored in prior studies for extracting diverse named entities, and the proper utilization of the extracted entities holds the utmost importance in text mining and natural language applications. NERC is language-dependent, which the approach adopted for most NER systems is domain-general, meaning they are built based on a language and not a particular targeted domain [16]. Till now, most of the studies are focused on the development of NERC in different domains and diverse resource-rich languages such as English [17][18], Indian [19][20][21], Arabic [22][23], and Turkish [24][25]. However, there are limited NERC studies on unstructured text written in Malay since Malay is an agglutinative language that may alter the semantic of its words with different morphology used, such as affixation, composition, reduplication.

The lack of credible Malay NERC tools [26] and very few publicly shared reference annotated corpora [14][27] have increased interest among the research community to develop more new approaches in NERC for improving the performance of Malay natural language applications. However, the researcher has faced some obstacles to developing an accurate Malay NERC system. The limitations are the lack of annotation of training data and also text ambiguity. To the extent of our knowledge, there are very few reviews of Malay NERC so far, and the issue has not been extensively explored in the reviews to identify the impact of unstructured Malay data on the existing NERC techniques. Arguably, the most established one was published by Morsidi et al. [14] in 2015. However, they did not include recent publications in the period from 2015 to 2020.

The motivation for conducting this review is to highlight the present status of NERC techniques developed from 2010 to 2020 for unstructured Malay data and to identify numerous issues and challenges in extracting unstructured Malay data. Furthermore, this work will contribute to the body of knowledge for the Malay NERC. The research question of this study is how NERC is carried out on unstructured Malay text from 2010-2020? The following complementary questions are constructed to define the research question:

Question 1: What are the main approaches used in NERC?

Question 2: Concerning the approaches, to what extent the experimental researches in NERC on unstructured Malay data are conducted in terms of the features, methodologies employed, and the evaluation of performance?

Question 3: Concerning the prior researches conducted in NERC on Unstructured Malay data, what are the limitations and challenges faced?

The remaining article is structured as follows. Section 2 presents the research methodology applied. Meanwhile, approaches in NERC are discussed in Section 3. Section 4 explains about Malay name entity recognition using rule-based approaches. Subsequently, Section 5 discusses on Malay NERC using machine learning-based approaches. The issues and challenges that are to be handled while designing the NERC system for unstructured Malay text are highlighted in Section 6. Finally, the last section concludes the article with future research directions to improve the research in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

this field.

2 RESEARCH METHODOLOGY

In this study, a systematic literature review is conducted. The systematic literature review is a process of aggregating, critically assessing, summarizing, and interpreting the available evidence concerning a clearly defined problem that requires further investigation. It consists of several steps. The steps are comprehensive search of relevant data source, selection of quality criteria, assessment of the quality of potential evidence, extraction of the evidence, synthesis of the evidence, and reporting. These steps follow the guidelines and the systematic review protocol, as proposed by Kitchenham and Charters [28].

2.1 Searching evidence from the relevant data source

A comprehensive search of the English and Malay language literature was performed, incorporating both open and closed-access electronic data sources. The electronic search was performed using the Google Scholar engine for finding the refereed research literature of the open-access database. Meanwhile, for the subscription database, the searching was done in the IEEE Xplore, ISI Web of Knowledge, ScienceDirect, Scopus, and SpringerLink, which these databases are well-known to contain computer science archives. It deals with peer-reviewed and published information, including journal articles, book chapters, and conference proceedings.

All of these data sources were searched using the following keywords, as shown in Table 1. The search strings are formed by categorizing the keywords into groups based on their synonyms or similar semantic meaning or a different way of spelling to ensure comprehensive searching. For example, the term "named entity recognition and classification" could be represented as "named entity recognition", or "name entity recognition". Another example, the synonym term for "terminology-based" that could be designated as "dictionary-based" or "dictionary lookup".

Table 1. Searching keywords used

	Group 1	Group 2	Group 3
Keyword[1]	"named entity recognition and classification"	"named entity recognition"	"name entity recognition"
Keyword[2]	"Malay named entity recognition and classification"	"Malay named entity recognition"	"Malay name entity recognition"
Keyword[3]	review	state-of-the-art	survey
Keywrod[4]	rule-based	rule	pattern
Keyword[5]	"machine learning"	machine	
Keyword[6]	Malay		
Keyword[7]	"compound nouns"		

Implementing the search strategy was achieved by combining the keywords using the Boolean operator, AND. There are several search strategies, which was designed to answer the research questions, including;

1. state-of-the-art NERC in between 2010 and 2020
2. state-of-the-art NERC on unstructured Malay text in between 2010 and 2020
3. NERC on unstructured Malay text in between 2010 and 2020
4. NERC on unstructured Malay text using rule-based in between 2010 and 2020
5. NERC on unstructured Malay text using machine learning-based in between 2010 and 2020

The search was restricted to the articles published between 2010 and 2020. It is the time in which the Malay NERC work is starting to get the focus. Table 2 details the number of refereed documents obtained based on the defined searching strategies by combining different keywords. Initially, the search was performed based on the identified keywords on the document's title. Furthermore, the search was extended to look for keywords in the abstract or entire article if the negative outcome was obtained from the previous query. A total of 355 refereed documents were retrieved from the data source searches. As far as we are concerned, there are very few comprehensive reviews in Malay named entity recognition and classification, according to the findings.

2.2 Selection of relevant evidence

The documents obtained should be filtered since the searching evidence in the prior steps return the number of publications far larger than manageable. In this step, removing unrelated

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

documents was done according to the following exclusion criteria:

1. duplicate documents within the search documents
2. the same document published in different data sources
3. inaccessible documents due to the restrictions on the publisher side
4. documents published before 2010
5. documents are not written in English and Malay language

Table 2. Total number of publications from the comprehensive searching for each database

Searching Strategy	Combination Keywords	IEEE Xplore	ISI WOS	Science Direct	Scopus	Springer Link	Google Scholar
Strategy 1 (n = 229)	$Keyword^{1,1}$ AND $Keyword^{3,1}$	0	1	1	1	0	1
	$Keyword^{1,2}$ AND $Keyword^{3,1}$	2	4	1	9	82	21
	$Keyword^{1,3}$ AND $Keyword^{3,1}$	0	0	1	0	1	0
	$Keyword^{1,1}$ AND $Keyword^{3,2}$	0	0	0	0	0	1
	$Keyword^{1,2}$ AND $Keyword^{3,2}$	0	2	0	3	1	13
	$Keyword^{1,3}$ AND $Keyword^{3,2}$	0	0	0	0	2	0
	$Keyword^{1,1}$ AND $Keyword^{3,3}$	0	1	0	1	0	3
	$Keyword^{1,2}$ AND $Keyword^{3,3}$	5	7	1	11	3	43
	$Keyword^{1,3}$ AND $Keyword^{3,3}$	0	1	1	1	1	3
	TOTAL	**7**	**16**	**5**	**26**	**90**	**85**
Strategy 2 (n = 2)	$Keyword^{2,1}$ AND $Keyword^{3,1}$	0	0	0	0	0	0
	$Keyword^{2,2}$ AND $Keyword^{3,1}$	1	0	0	0	0	1
	$Keyword^{2,3}$ AND $Keyword^{3,1}$	0	0	0	0	0	0
	$Keyword^{2,1}$ AND $Keyword^{3,2}$	0	0	0	0	0	0
	$Keyword^{2,2}$ AND $Keyword^{3,2}$	0	0	0	0	0	0
	$Keyword^{2,3}$ AND $Keyword^{3,2}$	0	0	0	0	0	0
	$Keyword^{2,1}$ AND $Keyword^{3,3}$	0	0	0	0	0	0
	$Keyword^{2,2}$ AND $Keyword^{3,3}$	0	0	0	0	0	0
	$Keyword^{2,3}$ AND $Keyword^{3,3}$	0	0	0	0	0	0
	TOTAL	**1**	**0**	**0**	**0**	**0**	**1**
Strategy 3 (n = 60)	$Keyword^{1,1}$ AND $Keyword^{6,1}$	0	0	0	0	0	0
	$Keyword^{1,2}$ AND $Keyword^{6,1}$	2	4	0	6	5	14
	$Keyword^{1,3}$ AND $Keyword^{6,1}$	0	2	0	2	2	2
	$Keyword^{7,1}$ AND $Keyword^{6,1}$	3	2	0	6	0	10
	TOTAL	**5**	**8**	**0**	**14**	**7**	**26**
Strategy 4 (n = 41)	$Keyword^{1,1}$ AND $Keyword^{4,1}$ AND $Keyword^{6,1}$	0	0	0	0	0	0
	$Keyword^{1,2}$ AND $Keyword^{4,1}$ AND $Keyword^{6,1}$	0	0	0	4	5	3
	$Keyword^{1,3}$ AND $Keyword^{4,1}$ AND $Keyword^{6,1}$	0	0	0	1	2	0
	$Keyword^{1,1}$ AND $Keyword^{4,2}$ AND $Keyword^{6,1}$	0	0	0	0	0	0
	$Keyword^{1,2}$ AND $Keyword^{4,2}$ AND $Keyword^{6,1}$	0	0	0	7	5	3
	$Keyword^{1,3}$ AND $Keyword^{4,2}$ AND $Keyword^{6,1}$	0	0	0	2	2	0
	$Keyword^{1,1}$ AND $Keyword^{4,3}$ AND $Keyword^{6,1}$	0	0	0	0	0	0
	$Keyword^{1,2}$ AND $Keyword^{4,3}$ AND $Keyword^{6,1}$	0	0	0	2	3	0
	$Keyword^{1,3}$ AND $Keyword^{4,3}$ AND $Keyword^{6,1}$	0	0	0	1	1	0
	TOTAL	**0**	**0**	**0**	**17**	**18**	**6**
Strategy 5 (n = 23)	$Keyword^{1,1}$ AND $Keyword^{5,1}$ AND $Keyword^{6,1}$	0	0	0	0	0	0
	$Keyword^{1,2}$ AND $Keyword^{5,1}$ AND $Keyword^{6,1}$	2	0	0	2	5	0
	$Keyword^{1,3}$ AND $Keyword^{5,1}$ AND $Keyword^{6,1}$	0	0	0	0	2	0
	$Keyword^{1,1}$ AND $Keyword^{5,2}$ AND $Keyword^{6,1}$	0	0	0	0	0	0
	$Keyword^{1,2}$ AND $Keyword^{5,2}$ AND $Keyword^{6,1}$	2	0	0	3	5	0
	$Keyword^{1,3}$ AND $Keyword^{5,2}$ AND $Keyword^{6,1}$	0	0	0	0	2	0
	TOTAL	4	0	0	5	14	0

There were 215 documents excluded. The balance of 140 papers was evaluated by screening on the abstract and the full text based on the inclusion criteria and quality criteria. The final list of these 140 documents was downloaded for further analysis.

2.3 Assessment of the evidence quality

The identification of inclusion criteria and quality criteria is essential to obtain documents that are thematically relevant to address the research question. Table 3 lists the inclusion criteria and quality criteria applied for screening both abstract and full text. Firstly, the abstract and full text of the documents were screened based on the inclusion criteria. Subsequently, the full text of relevant documents reviewed. If the abstract was considered significant to be included in the review, then the full-text was examined to ensure the content was

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

relevant. As depicted in Figure 1, 68 documents were omitted after the abstract review, and 72 documents were assessed based on the inclusion criteria for their entire text. At this stage, 15 documents that were unrelated to the systematic review of named entity recognition, and Malay named entity recognition approaches were eliminated.

Then, the balance of 57 documents is a set of relevant studies that will be filtered out for quality criteria. After performing the full-text screening, only 31 documents were eligible for qualitative analysis which are conducted in the next stage. The main reason for rejection is the studies did not describe a rationale for conduction the review.

Table 3. Inclusion and quality criteria

Inclusion Criteria	Quality Criteria
IC^1: The study identifies the title as a review of NERC	QC^1: There is a clear aim of the research.
IC^2: The study identifies the title as a review of Malay NERC	QC^2: The study describes the rationale for the review
IC^3: The study focuses on Malay NERC using rule-based	QC^3: There is a clear review of NERC approaches
IC^4: The study focuses on Malay NERC using machine learning	QC^4: The study discusses the limitations and challenges of NERC
IC^5: The study focuses on Malay NERC using the hybrid approach	QC^5: The study specifies the Malay NERC approach and its performance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

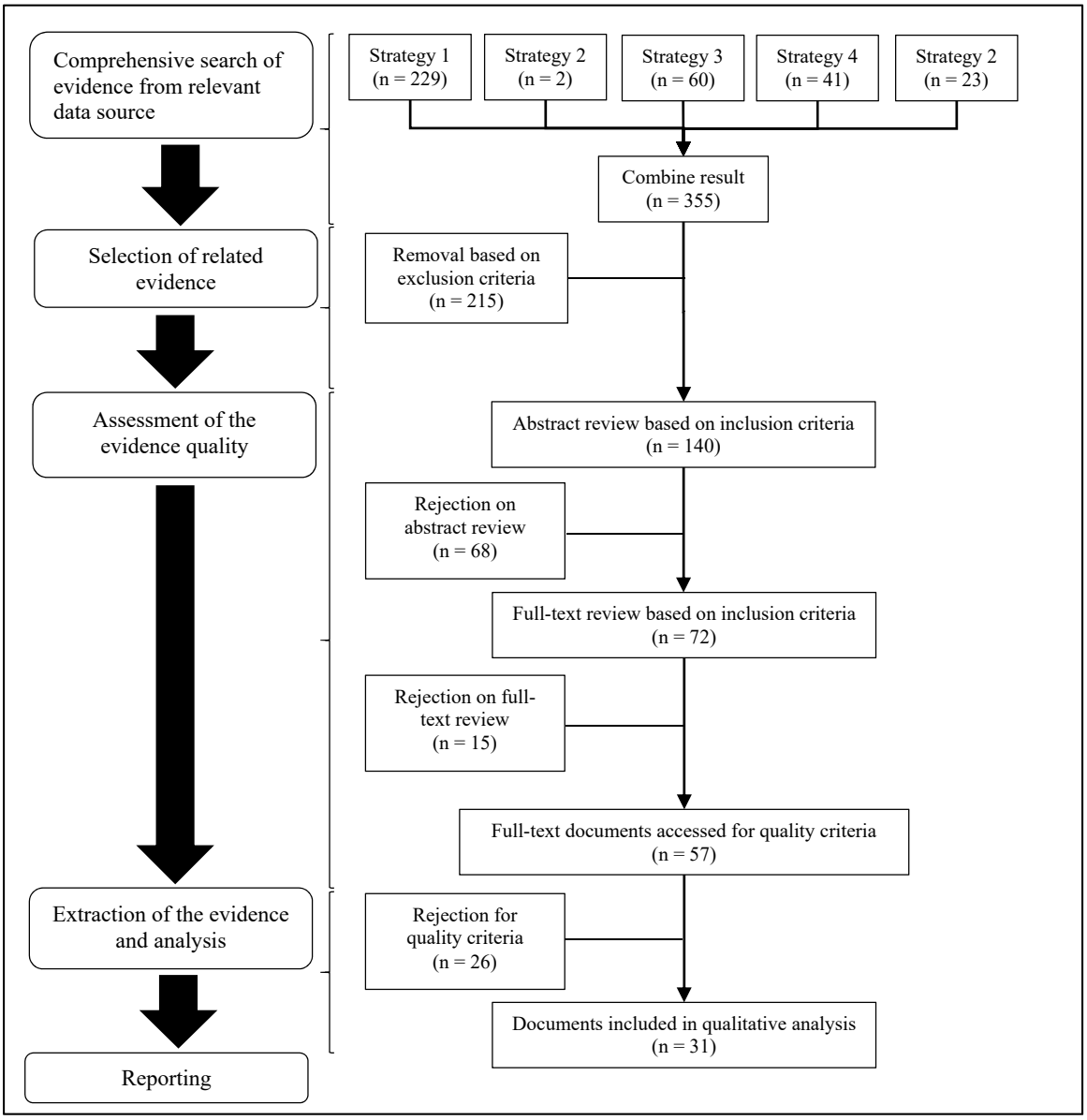

Figure 1: The flow of selecting relevant documents

2.4 Extraction of the evidence and analysis

In this step, an evidence extraction form, as shown in Table 4, is used to guide the qualitative analysis of documents. There are two types of study that involved in this paper; 1) review and 2) experimental.

Table 4. Evidence extraction form

Item	Data Extracted
Reference	Author, year, venue, type of publications
Form of study	Review: domain, language, NERC approach
	Experimental: NERC approach, performance results

In between 2011 and 2020, 30% (n=4) of refereed documents dealing with the review of NERC in English and Indian languages respectively, meanwhile 15% (n=2) of the publications are doing a systematic review of NERC in Turkish and Arabic and 8% (n=1) of the publication in Malay. The nine articles reporting on NERC had been published in the scientific journals, three book chapters and one conference proceedings.

Meanwhile, for the experimental studies conducted using Malay NERC approached, there are 18 articles were selected, which 45% (n=8) of the documents are dealing with Malay NERC approach using rule-based, 33% (n=6) of publications are using machine learning-based approach, and the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

balance of 20% (2=4) are studying on the hybrid system.

3 NAMED ENTITY RECOGNITION AND CLASSIFICATION APPROACH

Named Entity Recognition and Classification is a vital subtask of information extraction. The term "named entity" was initially coined by Grishman and Sundheim in Message Understanding Conference-6 (MUC-6), which involves the process of identifying entity names of people, organizations, and geographical locations, as well as numeric expressions [29]. Since then, there has been a growing interest in named entity recognition and has been progressively studied among the research community in various languages such as English, Spanish, and Arabic.

NERC is the task of recognizing terms and phrases that mention named entities of interest from free text and classifying them into a group of predefined labels or categories such as a person, organization, locations, etc. [30]. NERC system aims to automatically identify and classify the proper nouns in the unstructured text to describe concepts of interests in a given domain, that afterwards could be linked to a knowledge base where these concepts are referred for further analytics applications [31][32]. There are different types of approaches that have been developed to identify entities in various domains. These approaches are terminology-based, rule-based learning, machine learning-based, hybrid and deep learning-based. Different languages may have different morphologies and thus require different NERC processes, which is very useful for increasing the performance of information extraction. Implementation of the NERC algorithm is usually influenced by the domain of the studies [2]. For example, a domain-specific NERC for crime may not be applicable for recognizing named entities on other specific domains such as a biomedical field. It is because the biomedical field has its scientific term to extract entities such as protein and gene names.

3.1 Terminology-based approach

Terminology-based, also called dictionary-based, is the most basic approach in named entity recognition, which it works based on the concept of one to one matching between words or phrases within unstructured text and terms that exist in the terminological glossaries or dictionaries [33]. The earlier terminology-based approach employs the simple dictionary lookup method, which necessitates a set of predetermined terminological lexicons within a specific domain as the input. Finding matches within the text phrases against the terms in the lexicon is done using strict string matching techniques [31]. It has been widely implemented in a domain-specific such as biomedical [34][35] and chemical [36]. Although this approach is very straightforward and precise, the reliance on lexicons could lead to low recall performance as it is unable to solve the disambiguation problem. Another disadvantage of this approach is the strict matching technique only capable of detecting entity as predetermined in the lexicons without considering the spelling mistake and unable to handle noisy data, hence loss the semantic matching.

3.2 Rule-based learning approach

Rule-based learning is another NERC approach, which uses the constructed rules based on the characteristics of the entities of interest, as well as the usage of information from dictionaries, thesauri references, or gazetteers [33]. The constructed rules depend on textual features such as orthographic and morphological, and context of the word in the text. It does not require annotated data. In rule-based NERC, the properties of language-related knowledge are exploited using regular expression patterns and syntactic-lexical patterns to detect the desired entities such as the name of the person, location, and organization. For example, the "Steve Job" query should not just check on the word "Steve" or "Job." The term "job" can lead to the searching of another similar word, such as "occupation." So instead of "Steve Job," the co-founder of Apple Inc., it will contribute to another purpose. The extracted named entities, however, could also include other concepts related within the domain of interest. This approach is highly efficient as it allows language grammatical and morphological to be expressed in the rules. It also employs domain-specific features to obtain sufficient accuracy [17]. However, the main disadvantages of this approach are the hand-crafted rules is time-consuming as the process is highly dependent on human experts, and domain dependency requires experts to involve closely in constructing rules. Since the experts manually craft the rules, it is non-applicable across domain and non-adaptable to recognize any potential new named entities categories.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3 Machine learning-based approaches

Machine learning is a statistical-based approach that uses algorithms on the training data and the number of selected textual features to automatically learn the complicated pattern to perform an identification process [37]. Features are words descriptors designed for the use of the algorithm such as orthographic (capitalization, decimal, digits, symbols), part-of-speech tags, morphological (affix, suffix, prefix), the word left/right, word length and others. Compared to the rule-based system, machine learning-based can be performed across the domain. The critical issue in the machine learning-based approach is that an enormous amount of data is required for training, as a lack of data can be one of the barriers to NERC system success. Besides, this approach relies on well-selected features [38]. There are three different types of machine learning approach, including supervised learning, semi-supervised learning, and unsupervised learning.

Supervised learning is the process of designing an adaptive model based on the labelled training dataset and the features associated with the data to recognize the correlation between the reliable output and input characteristics so that prediction of output values for new data can be inferred based on the relationships learned from the previous dataset [39]. It automatically constructs rules based on externally given examples and is thus used to predict future instances. Annotated corpora are prerequisites to the supervised learning algorithm during training phase [40], which requires a considerable effort of expert annotators that leads to time-consuming. Incomplete or missing annotated corpora affects the inferring process in the output. Under supervised setting, the machine learning-based methods are developed using supervised classifiers such as Hidden Markov Model, Maximum Entropy Markov Model, Conditional Random Fields, Support Vector Machine, Logistic Expression Classifier, and Decision Tree.

Semi-supervised learning is an intermediate approach between supervised learning and unsupervised learning which includes a small degree of supervision. A set of semi-supervised learning approaches aim to produce high-quality training data by using a small set of labelled corpus and a large set of unlabelled corpus [41]. This technique can provide a significant increase in learning accuracy. Bootstrap is an example method of semi-supervised learning.

Unsupervised learning relies on unsupervised algorithms to infer name entities based on the data that neither classified nor labelled during the training [36]. This algorithm acts without supervision during the computation of the lexical resources and lexical patterns to build representations from data. The data is organized based on similarities and information differences, although there is no classification in the data. The most common method for unsupervised learning is clustering and association rules.

3.4 Deep learning-based

Deep learning is an advanced model of machine learning that relies on the multiple processing layers of neural networks to learn representations of data from raw input with numerous levels of abstraction [18]. It can be classified into convolutional neural network and recurrent neural network [42]. The deep learning-based solution to NERC has begun to receive significant attention in recent years due to its improved performance. Compared to supervised machine learning, deep learning model does not need domain expertise to design NERC features since it can automatically learn complex, complicated and detail data features. Nevertheless, to the best of our knowledge, this approach has not been implemented to Malay NERC.

3.5 Hybrid approaches

The hybrid approach incorporates more than one approach of NERC, including any combination of a statistical-based, rule-based approach, and machine learning approach to optimize overall performance [43]. The hybrid approach has its advantages and disadvantages. Most of the studies that using a hybrid approach improve the performance in terms of accuracy [44]. It has a few modules based on the domain of the studies. Some studies might have only two modules, which are rule-based and machine-learning module only. But some studies are done by the combination of the machine learning algorithm and pattern matching rules to extract several entities such as a person, location, and organization.

One of the studies done by [32] proposed Malay-English Word Aligner or MEWA to align the Malay corpus with an English corpus by using a hybrid algorithm.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4 MALAY NAMED ENTITY RECOGNITION AND CLASSIFICATION USING RULE-BASED

The Malay language is one of the language fields that interest researchers in implementing the recognition task of the named entity. In Malay, it focuses on defining proper nouns. Like other languages, in the presentation of information, the Malay language also has its characteristics based on the order of sentences and the form of words that have specific meanings. Discussions on the implementation of the recognition of a named entity in the Malay language include orthography, morphology.

Named entity recognition can be classified by using the rule-based approach according to dictionaries and gazetteer list [45]. Predefined dictionaries or set of rules can be used to define the name entity rule-based algorithms. These algorithms are applied to extract pattern for location names, organization names, etc. The recognition process could be sped up based on dictionary types. The type of dictionary used to affect the performance of the NERC system and result. Usually, the dictionaries include the list of countries, main cities, companies, first names and titles [45].

Most of the system develops for Malay is focusing on NERC [46][32][47]. The majority of the systems developed based on manually predefined dictionaries by a human. Alfred [46] proposed a NER rule-based for Malay text by using contextual rules to recognize person, location, and organization entities. The Malay NER framework is designed to detect the names of entities to perform a good result of retrieving Malay articles with more effective and efficient. The research compiled a few types of dictionaries lookup that probably exist in Malay text such as location, location prefix, person title, organization prefix, organization abbreviation, organization name, and organization suffix. The results show a reasonable output with the proposed algorithm. However, the performance can be improved by re-formulating the rules and having more complete dictionaries.

Ulanganathan et al. [48] developed a Malay Entity Recognition Engine (Mi-NER) by using a probabilistic approach. This system is used to identify person, location and organization entities. The result shows Mi-NER perform better of identifying person and organization identities by adding more different salutation for people name and variety of organization suffix. This strategy would be powerful to find person and organization specifically.

Sharum et al. [49] proposed a Name Indexer system to detect person names of 117 Malay newspapers' article. The system can extract the title and name found in Malay news article. The application can be used to improve the process of searching performance. The output of name indexer specifies all existing names, including the documents links. Based on this research, a few techniques were developed to recognize a person's names. Those of the techniques are scanning for common first name, scanning for known titles, scanning for a word with capitalized initials and also scanning with kinship. Overall, the application provides a precise result. However, the application requires more specific rules and patterns to recognize unknown names like foreign names.

Moreover, [50] presents a rule-based to detect person, organization, location and misc entities by using Stanford and Illinois NER. This research was interesting because they are using English NER system to extract the Malay Named Entity. As we know, most of the existing English NER is not suitable for another language, especially for the Malay language. The researcher is able to prove that Stanford NER and Illinois NER can recognize location and organization entities with a poor result. However, it is unable to classify person and other entities through the Malay documents due to the different morphology between English and Malay. Both of NER showed unsatisfied result and tended to produce more error during experiments. It's explained that Stanford and Illinois NER system are not compatible with recognizing Malay entities.

Additionally, [51] proposed an algorithm for extracting nouns from Malay classical documents. The objective of this research is to find the best way to extract noun by using Lookup List, Morphological Rules (Noun Affixes), Morphological Rules (Verb, Adjective and Noun Affixes) and Morphological Rules (Rayner's Rules). The experiment used 75 Malay classical documents extracted from local universities repository. Two language experts evaluated the outputs of Noun Extraction.

All of the aforementioned systems are using a rule-based approach. Most of the entities in their investigation are the same (person, location, organization) except for the noun extraction system [51] and name indexer [49]. Furthermore, there is no

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

complete resource, such as a dictionary or gazetteer in Malay. Some of the predefined terms in the dictionaries are manually hand-crafted. This task is very challenging because it consumes a lot of time. One of the above studies that be done by [52] shows that they obtain the poor error result due to the limited list of words in dictionaries.

Table 1 shows the evaluation result for the rule-based system. The usage of specific rules constructed for Malay NER indicates significant results in extracting person, location and organization within Malay text, compared to the use of English NER as proposed by Sulaiman et al. in [50].

Table 1: Summary of precision, recall, and f-measure for the rule-based systems.

System	Entity	Precision (%)	Recall (%)	F-measure (%)
Malay NER Recognition [46]	Person	85.00	94.44	89.47
	Location			
	Organization			
Mi-NER [48]	Person	80.95	72.12	76.28
	Location	96.64	69.37	80.06
	Organization	89.76	54.76	68.02
Name Indexer [49]	Name	92.00	54.00	68.00
Stanford NER [50]	Person	39.66		36.55
	Misc			
	Organization			
	Location			
Illinois NER [50]	Person		37.19	35.24
Noun extraction [51]	Verb,Adj,Noun Affixes			77.61
	Rayner's Rule			52.31
	Noun affixes			34.16
Noun+Verb Identification [53]	Compound nouns	93.5	27.5	
Malay Name Entity Recognition [54]	Person	96.00	99.00	97.00
	Location	97.00	98.00	99.00
	Organization	100.00	89.00	94.00
	Position	100.00	95.00	85.00

5 MALAY NAMED ENTITY RECOGNITION AND CLASSIFICATION USING MACHINE LEARNING-BASED.

On the other hand, there are also studies by using machine learning in Malay NERC. Salleh and his team [26] designed an automated Malay named entity recognition (Amner) by using a conditional random field method. Pos tagging was used in this work. The proposed NER system has three main phases includes pre-processing data, generating model, and evaluation. The training phase was tested from Bernama news by using a few features such as capitalization, lowercase, and part of speech (pos) tagging. Based on the studies, the features selected show the potential improvement for the result accuracy.

Furthermore, Ulanganathan et al. [31] created a system that can recognize the entity and non-entity from unstructured Malay text. This work implements a fuzzy c-means clustering method by using Bernama Malay news as a dataset. The studies need to go through several steps include data pre-processing, text features transformation, experimental, and evaluation. A rapid miner tool is used in these studies to analyze the entities.

Meanwhile, Asmai et al. [39] integrated two machine learning approach techniques (fuzzy c-means and K-Nearest Neighbours Algorithm) for identifying six entities of crime data in Malay such as organization, location, date, crime, person, and other. The experiment showed the improvement of the entities' recognition performance when there is a combination method between fuzzy c-means and K-Nearest Neighbours. However, appropriate features selection need to be considered to get the best performance. [55] also used the K-Nearest Neighbours algorithm in their studies to detect proper nouns with three different datasets. The dataset from Astro Awani news, Berita Harian news, and Bernama news. Regex identifier is adopted in this experiment to detect the proper noun effectively. However, human validation is still needed to process the data since regex only depends on the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

preconfigured rules.

Table 2 shows the evaluation results in the previous researches that have been done using the machine-learning approach. All of the machine-based NERC systems show the highest score in f-measure performance except for the noun detection system. It is due to the noun detection system is using unannotated Malay language news articles in their studies. It is noticeable that the machine-based learning approach needs an annotated corpus and large training datasets to gain the highest performance.

Table 2: Summary of precision, recall, and f-measure for machine learning systems.

System	Entity	Precision (%)	Recall (%)	F-measure (%)
AMNER [47]	Person	0.82	0.75	0.78
	Organization	0.93	0.71	0.81
	Facility	0.56	1	0.72
	location	0.81	0.81	0.81
Entity Recognizer [56]	Entity	100	88.97	
	Non-Entity	98.39	100	
Entity Extraction [39]	Organization	89.77	90.99	
	Location	80.83	72.39	
	Date	81.72	87.3	
	Crime	85.48	85.48	
	Person	89.27	89.96	
	other	97.88	98.09	
Noun detection [55]	AA dataset	0.52	0.53	0.504
	BH dataset	0.57	0.61	0.567
	BER dataset	0.47	0.46	0.450

Table 3 provides a comparison between rule-based systems and machine learning systems regarding the process, a form of corpus and whether they use gazetteers or not, part-of-speech and/or stemming. As we can see, all systems didn't use stemming process, but most of them were used POS tagging and annotated corpus especially for machine learning. There is no evaluation table for hybrid approach since this approach are not popular for Malay NERC research.

Table 3: A Comparison between the Malay NER Systems.

System	Method	Stemming	Gazetteer	POS	Annotated Corpus
Malay NER Recognition [46]	Rule-based	No	Yes	Yes	No
Mi-NER[48]	Rule-based	No	No	No	Yes
Name Indexer [49]	Rule-based	No	No	Yes	No
Illinois NER [50]	Rule-based	No	Yes	Yes	No
Illinois NER [50]	Rule-based	No	Yes	Yes	No
Noun extraction [51]	Rule-based	No	No	No	Yes
AMNER [47]	Conditional Random Field	No	No	Yes	Yes
Entity Recognizer [56]	Fuzzy C-means clustering	No	No	Yes	Yes
Noun Detection [55]	KNN Algorithm	No	No	Yes	Yes
Entity Extraction[39]	Fuzzy C-means, KNN	No	No	Yes	Yes

6 MALAY NAMED ENTITY RECOGNITION USING HYBRID APPROACH

Usually, the hybrid approach produces more significant results, but the linguist and domain expert needs a great deal of professional work [57]. For this reason, most of the recent work in Malay NER focuses only on the rule-based method and the techniques in machine-learning. There are presently limited studies in Malay using a hybrid approach. One studied that been done by [58] employed a hybrid approach to developing an automated text summarization system. This system is purposely designed for electronic documents that can help people to get a summary from Malay's news article. While there are many tools for academic and clinical text summarization, no studies are officially reported

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for the Malay language. They have been adopting some techniques from existing successful results includes pre-processing, text extraction, and sentence selection. Since Malay text summary corpora are minimal and do not publicly share, the corpus is required to be built manually.

7 CHALLENGE AND ISSUE OF MALAY NERC

First of all, the Malay text is limited to publicly available because most of the corpora are private for academic use [14]. [59] argued that Dewan Bahasa and Pustaka, Pangkalan Data Korpus (UKM-DBP) could be the most comprehensive corpus in Malaysia. Some scholars have no choice but to switch to their hand-crafted domain-specific corpus they need to construct the new corpus. Numerous morphological uses have hindered the adoption of English text processing techniques in Malay.

The use of NER in Natural Language Processing in a wide variety of applications appears promising. But the development of the field is hindered by:

7.1 The difference in morphology.

There are different morphological features between Malay, English, Arabic, and other languages. Every language has rich and complex morphological features which contribute to the difficulties of implementing a method to develop the correct NERC system. English NER application might not be suitable for recognizing Malay name-entities because the implementation of NER depends on the domain of studies and type of language. Many of the NER systems exist for other languages such as English, Indonesia, Arabic, and Urdu [60] [61] [62] [63]. It is also challenging to do a pre-processing for the Malay language because there are no existing pre-processing tools for Malay. If the researcher wants to use English pre-processing tools, they need to custom their coding first to get the result. This activity takes plenty of time because Malay morphology is quite complicated to analyze. One of the studies that been done by Alfred [46] found that morphologies for the Malay language are too complex and challenging to implement in the Malay NER algorithm. Hence, the powerful pre-processing techniques for the Malay language are required to improve the quality of unstructured data.

7.2 Limitation of Malay corpus.

A corpus can be characterized as a collection of common dialect information either in content or discourse frame. The development of the Malay corpus is mainly for the academic field, so the corpus does not provide to the public [64]. It Is not easy to construct the Malay corpus because it involves several phases such as data gathering include pre-processing activities, extraction, filtration, and association [26]. Designing the online corpus repository is a complex task [65] [66]. The different studies might use the diverse corpus to extract the information. It is built to the specific criteria based on the domain and the linguistic applications. One of the studies [67] has developed the Malaysian Corpus of Financial English (MaCFE). This type of corpus contains a variety of documents from all financial institutions in Malaysia.

7.3 Lack of annotated resources for the Malay language.

Lack of annotated corpus resources for the Malay language compared to English and another language [51]. This challenge most probably faces by a researcher who applies the rule-based method in their researcher because they can't use existing English corpus into the Malay language. Furthermore, it takes a long time to interpret the Malay text to be used for Malay entities' recognition.

Creating resourced for Malay NLP is exceptionally complicated and time-consuming, especially for the Malay language. This challenge might happen for other languages such as Persian and Arabic [68] [69] [70].

Lack of annotated resources for the Malay language will affect the machine learning method result. It is impossible to implement the NER method without having an extensive and comprehensive annotated data. These annotated corpora is the main requirement for the supervised machine learning method. Since there is a limitation of the Malay annotated data, the previous researcher would prefer to choose the unsupervised technique.

7.4 Ambiguities in text.

The ambiguities in human language are a significant obstacle for computers to understand and interpreted [71]. Text ambiguity occurs when

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

multiple words have the same meaning (synonym) or a word with different meanings (homonym). There is always an issue for text ambiguities, especially for the Malay languages due to multiple languages for certain words [72]. For example, the Malay word "sepak" has several meanings, such as to kick or to slap.

8 CONCLUSION

In this paper, the author has presented named entity recognition systems that were developed for the Malay language. These systems are split into two groups based on machine learning and rule-based approach. The objectives, methods, performance, strengths, and weaknesses are discussed for previous research. Besides, the comparison is deliberate between the rule-based approach and machine learning approach in Malay NERC systems. However, the hybrid approach is not included since the approach is quite complicated and less been used by previous researchers.

According to the findings obtained from the reviews, there still a lot of deficiency in previous studies. The future research in Malay NLP needs to make some improvement to have superb Malay NERC systems. Having complete dictionaries would improve the algorithm in NER studies. However, the NERC tools should be selected appropriately and compatible with the language of the studies because the different language has different morphology, which could affect the performance of information extraction. Other than that, having a large test collection will give the best and accurate result.

Named entity recognition is a field that is very popular among researchers and data scientists. This approach is very much beneficial for handling big data. Thus, this paper is expected to assist the researchers to raise some awareness on the important of the current state of the art and certain limitations during the development of NERC systems. It would allow other researchers to gain insight into the problem encountered in order to analyze and extract hidden features from the Malay documents.

ACKNOWLEDGEMENT

This research was supported by the Fundamental Research Grant Scheme (FRGS) with reference code of FRGS/1/2018/ICT04/UMT/02/3 and vote number 59541 under the Malaysia Ministry of Education (FRGS Phase 1/2018).

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, "The Digitization of the World From Edge to Core," Framingham, USA, 2018.

[2] K. Adnan and R. Akbar, *An analytical study of information extraction from unstructured and multidimensional big data*, vol. 6, no. 1. Springer International Publishing, 2019.

[3] E. Yehia, H. Boshnak, S. AbdelGaber, A. Abdo, and D. S. Elzanfaly, "Ontology-based clinical information extraction from physician's free-text notes," *J. Biomed. Inform.*, vol. 98, p. 103276, 2019.

[4] S. Malmasi, W. Ge, N. Hosomura, and A. Turchin, "Comparing information extraction techniques for low-prevalence concepts: The case of insulin rejection by patients," *J. Biomed. Inform.*, vol. 99, p. 103306, 2019.

[5] V. Suárez-Paniagua, R. M. Rivera Zavala, I. Segura-Bedmar, and P. Martínez, "A two-stage deep learning approach for extracting entities and relationships from medical texts," *J. Biomed. Inform.*, vol. 99, p. 103285, 2019.

[6] M. Abulaish, M. A. Parwez, and Jahiruddin, "DiseaSE: A biomedical text analytics system for disease symptom extraction and characterization," *J. Biomed. Inform.*, vol. 100, p. 103324, 2019.

[7] A. Rauch, M. Sartori, E. Rossi, P. Baland, and S. Castelltort, "Trace Information Extraction (TIE): A new approach to extract structural information from traces in geological maps," *J. Struct. Geol.*, vol. 126, pp. 286–300, 2019.

[8] M. Marzouk and M. Enaba, "Text analytics to analyze and monitor construction project contract and correspondence," *Autom. Constr.*, vol. 98, pp. 265–274, 2019.

[9] J. Sun *et al.*, "Text visualization for construction document information management," *Autom. Constr.*, vol. 111, p. 103048, 2020.

[10] M. Siering, A. V Deokar, and C. Janze, "Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews," *Decis. Support Syst.*, vol. 107, pp. 52–63, 2018.

[11] S. Moon and W. A. Kamakura, "A picture is worth a thousand words: Translating product reviews into a product positioning map," *Int. J. Res. Mark.*, vol. 34, no. 1, pp. 265–285, 2017.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[12] R. K., H. Srinivas, and S. S., "Industrial information extraction through multi-phase classification using ontology for unstructured documents," *Comput. Ind.*, vol. 100, pp. 137–147, 2018.

[13] T. M. Alam and M. J. Awan, "Domain Analysis of Information Extraction Techniques," *Int. J. Multidiscip. Sci. Eng.*, vol. 9, no. 6, 2018.

[14] F. Morsidi, S. Sarkawi, S. Sulaiman, M. Siti Asma, and A. W. Rohaizah, "Malay named entity recognition: a review," *J. ICT Educ.*, vol. 2, no. 1, pp. 1–14, 2015.

[15] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Information extraction from scientific articles: a survey," *Scientometrics*, vol. 117, no. 3, pp. 1931–1990, 2018.

[16] C. Y. Lim, I. K. T. Tan, and B. Selvaretnam, "Domain-General Versus Domain-Specific Named Entity Recognition: A Case Study Using TEXT BT - Multi-disciplinary Trends in Artificial Intelligence," 2019, pp. 238–246.

[17] A. Goyal, V. Gupta, and M. Kumar, "Recent Named Entity Recognition and Classification techniques: A systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–43, 2018.

[18] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans. Knowl. Data Eng.*, p. 1, 2020.

[19] G. Talukdar, P. P. Borah, and A. Baruah, "A Survey of Named Entity Recognition in Assamese and Other Indian Languages," *Int. J. Nat. Lang. Comput.*, vol. 3, pp. 105–112, 2014.

[20] K. Bhattacharjee *et al.*, "Named Entity Recognition: A Survey for Indian Languages," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2019, vol. 1, pp. 217–220.

[21] A. Jain, D. K. Tayal, D. Yadav, and A. Arora, "Research Trends for Named Entity Recognition in Hindi Language BT - Data Visualization and Knowledge Engineering: Spotting Data Points with Artificial Intelligence," J. Hemanth, M. Bhatia, and O. Geman, Eds. Cham: Springer International Publishing, 2020, pp. 223–248.

[22] B. A. Ben Ali, S. Mihi, I. El Bazi, and N. Laachfoubi, "A recent survey of Arabic named entity recognition on social media," *Rev. d'Intelligence Artif.*, vol. 34, no. 2, pp. 125–135, 2020.

[23] K. Shaalan, "A Survey of Arabic Named Entity Recognition and Classification," *Comput. Linguist.*, vol. 40, no. 2, pp. 469–510, 2014.

[24] D. Küçük, N. Arıcı, and D. Küçük, "Named Entity Recognition in Turkish: Approaches and Issues BT - Natural Language Processing and Information Systems," 2017, pp. 176–181.

[25] R. Yeniterzi, G. Tür, and K. Oflazer, "Turkish Named-Entity Recognition BT - Turkish Natural Language Processing," K. Oflazer and M. Saraçlar, Eds. Cham: Springer International Publishing, 2018, pp. 115–132.

[26] N. Omar, A. F. Hamsani, N. A. S. Abdullah, and S. Z. Z. Abidin, "Construction of Malay abbreviation corpus based on social media data," *Journal of Engineering and Applied Sciences*, vol. 12, no. 3. pp. 468–474, 2017.

[27] S. S. Sazali, N. A. Rahman, and Z. A. Bakar, "Characteristics of Malay translated hadith corpus," *J. King Saud Univ. - Comput. Inf. Sci.*, 2020.

[28] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering. In: Technical report EBSE 2007–001, Keele University and Durham University Joint Report," 2007.

[29] R. Grishman and B. Sundheim, "Message Understanding Conference - 6: A Brief History," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996, pp. 466–471.

[30] M. Asharef, N. Omar, and M. Albared, "Arabic named entity recognition in crime documents," *J. Theor. Appl. Inf. Technol.*, vol. 44, no. 1, pp. 1–6, 2012.

[31] F. M. Couto and A. Lamurias, "MER: a shell script and annotation server for minimal named entity recognition and linking," *J. Cheminform.*, vol. 10, no. 1, p. 58, 2018.

[32] M. L. Gavrilova, O. Gervasi, and B. O. Apduhan, "Name Entity Recognition for Malay Texts Using Cross-Lingual Annotation Projection Approach," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9156, pp. 242–256, 2015.

[33] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extraction of evidence-based dietary recommendations," *PLoS One*, vol. 12, no. 6, pp. 1–32, 2017.

[34] X. Wang, C. Yang, and R. Guan, "A comparative study for biomedical named entity recognition," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 3, pp. 373–382, 2018.

[35] R. R. V. Goulart, V. L. Strube de Lima, and C. C. Xavier, "A systematic review of named entity recognition in biomedical texts," *J. Brazilian Comput. Soc.*, vol. 17, no. 2, pp. 103–116, 2011.

[36] S. Eltyeb and N. Salim, "Chemical named entities recognition: a review on approaches and applications," *J. Cheminform.*, vol. 6, p. 17, Apr. 2014.

[37] Y. Sari, M. F. Hassan, and N. Zamin, "Rule-based pattern extractor and Named Entity Recognition: A hybrid approach," *Proc. 2010 Int. Symp. Inf. Technol. - Eng. Technol. ITSim'10*, vol. 2, pp. 563–568, 2010.

[38] Y. Wen, C. Fan, G. Chen, X. Chen, and M. Chen, "A Survey on Named Entity Recognition BT - Communications, Signal Processing, and Systems," 2020, pp. 1803–1810.

[39] S. A. Asmai, M. S. Salleh, H. Basiron, and S. Ahmad, "An enhanced Malay named entity recognition using combination approach for crime textual data analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 9, pp. 474–483, 2018.

[40] R. E. Salah, L. Qadri, and Zakaria, "A Comparative Review of Machine Learning for Arabic Named Entity Recognition," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 2, pp. 511–518, 2017.

[41] U. Kanimozhi and D. Manjula, "A Systematic Review on Biomedical Named Entity Recognition," in *Data Science Analytics and Applications*, 2018, pp. 19–37.

[42] A. Thomas and S. Sangeetha, "Deep Learning Architectures for Named Entity Recognition: A Survey BT - Advanced Computing and Intelligent Engineering," 2020, pp. 215–225.

[43] C. Kiefer, P. Reimann, and B. Mitschang, "A hybrid information extraction approach exploiting structured data within a text mining process," *Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform.*, vol. P-289, no. Btw, pp. 149–168, 2019.

[44] K. Shaalan and M. Oudah, "A hybrid approach to Arabic named entity recognition," *J. Inf. Sci.*, vol. 40, no. 1, pp. 67–87, 2014.

[45] M. Ikhwan Syafiq, M. Shukor Talib, N. Salim, H. Haron, and R. Alwee, "A Concise Review of Named Entity Recognition System: Methods and Features," in *IOP Conference Series: Materials Science and Engineering*, 2019.

[46] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 3, pp. 300–306, 2014.

[47] H. B. and S. A. Muhammad Salleh, Siti Azirah, "A Malay Named Entity Recognition Using Conditional Random Fields," *Int. Conf. Inf. Commun. Technol.*, 2017.

[48] T. Ulanganathan, A. Ebrahimi, B. C. M. Xian, K. Bouzekri, R. Mahmud, and O. H. Hoe, "Benchmarking Mi-NER: Malay Entity Recognition Engine," *Ninth Int. Conf. Information, Process Knowl. Manag. (eKNOW 2017)*, no. c, pp. 52–58, 2017.

[49] M. Y. Sharum, M. T. Abdullah, M. N. Sulaiman, M. A. A. Murad, and Z. A. Z. Hamzah, "Name extraction for unstructured Malay text," *Isc. 2011 - 2011 IEEE Symp. Comput. Informatics*, pp. 787–791, 2011.

[50] S. Sulaiman, R. A. Wahid, S. Sarkawi, and N. Omar, "Using Stanford NER and Illinois NER to Detect Malay Named Entity Recognition," *Int. J. Comput. Theory Eng.*, vol. 9, no. 2, pp. 147–150, 2017.

[51] S. S. Sazali, N. A. Rahman, and Z. A. Bakar, "Information extraction: Evaluating named entity recognition from classical Malay documents," *2016 3rd Int. Conf. Inf. Retr. Knowl. Manag. CAMP 2016 - Conf. Proc.*, pp. 48–53, 2017.

[52] R. Alfred *et al.*, "A Rule-Based Named-Entity for Malay Articles," *Conf. Pap. Lect. Notes Comput. Sci.*, vol. 8346, no. PART 1, pp. 287–299, 2013.

[53] Z. A. Bakar, N. K. Ismail, and M. I. M. Rawi, "Identification of Noun + Verb Compound Nouns in Malay Standard document based on rule based," in *2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS)*, 2017, pp. 1–6.

[54] N. M. Noor, J. Sulaiman, and S. A. Noah, "Malay Name Entity Recognition Using Limited Resources," in *International Symposium of Information and Internet*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Technology, 2016.

[55] S. Sulaiman, R. A. Wahid, and F. Morsidi, "Feature extraction using regular expression in detecting proper noun for Malay news articles based on KNN algorithm," *J. Fundam. Appl. Sci.*, vol. 9, no. 5S, p. 210, 2018.

[56] M. S. Salleh, S. A. Asmai, H. Basiron, and S. Ahmad, "Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis," *J. Telecommun. Electron. Comput. Eng.*, vol. 10, no. 2–7, pp. 121–126, 2018.

[57] M. Fresko, B. Rosenfeld, and R. Feldman, "A hybrid approach to NER by MEMM and manual rules," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. 0, pp. 361–362, 2005.

[58] N. Zamin and A. Ghani, "A Hybrid Approach for Malay Text Summarizer," in *Proc. Int. Multi-Conference Eng. Technol. Innov*, 2010.

[59] H. Hasmy, Z. A. Bakar, and F. Ahmad, "Construction of Computational Lexicon for Malay Language," vol. 9429, pp. 257–268, 2015.

[60] H. K. Chih, A. Iriberri, and G. Leroy, "Crime information extraction from police and witness narrative reports," *2008 IEEE Int. Conf. Technol. Homel. Secur. HST'08*, pp. 193–198, 2008.

[61] I. Budi, S. Bressan, G. Wahyudi, Z. A. Hasibuan, and B. A. A. Nazief, "Named Entity Recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3735 LNAI, no. November 2015, pp. 57–69, 2005.

[62] K. Shaalan and H. Raza, "Person name entity recognition for Arabic," no. June, p. 17, 2007.

[63] U. P. Singh, V. Goyal, and G. S. Lehal, "Named entity recognition system for Urdu," *24th Int. Conf. Comput. Linguist. - Proc. COLING 2012 Tech. Pap.*, vol. 1, no. December 2012, pp. 2507–2518, 2012.

[64] S. Alias, S. K. Mohammad, G. K. Hoon, Eng, and T. T. Ping, "A Malay Text Corpus Analysis For Sentence Compression Using Pattern Growth Method.," *J. Teknol.*, vol. 8, pp. 197–206, 2016.

[65] R. Manurung, B. Distiawan, and D. D. Putra, "Developing an online Indonesian corpora repository," *PACLIC 24 - Proc. 24th Pacific Asia Conf. Lang. Inf. Comput.*, pp. 243–249, 2010.

[66] M.M.Din, N. H. H. Hashim, and M. M. Siraj, "Comparative Study on Corpus Development for Malay Investment Fraud Detection in Website," *Jsournal Fundam. Appl. Sci.*, vol. 9, no. 6s, pp. 828–838, 2017.

[67] R. Sadjirin, R. A. Aziz, N. M. Nordin, M. R. Ismail, and N. D. Baharum, "The development of malaysian corpus of financial english (MaCFE)," *GEMA Online J. Lang. Stud.*, vol. 18, no. 3, pp. 73–100, 2018.

[68] K. Dashtipour, M. Gogate, A. Adeel, A. Algarafi, N. Howard, and A. Hussain, "Persian Named Entity Recognition," *Proc. 2017 IEEE 16th Int. Conf. Cogn. Informatics Cogn. Comput. ICCI*CC 2017*, pp. 79–83, 2017.

[69] S. Larabi Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic natural language processing and machine learning-based systems," *IEEE Access*, vol. 7, pp. 7011–7020, 2019.

[70] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2019.

[71] M. Mulins, "Information extraction in text mining," *Comput. Sci. Grad. Undergrad. Student Scholarsh.*, 2008.

[72] M. F. Yahaya, N. A. Rahman, Z. A. Bakar, and H. Hasmy, "Evaluation on knowledge extraction and machine learning in resolving Malay word ambiguity," *J. Fundam. Appl. Sci.*, vol. 9, no. 5S, p. 115, 2018.