Jurnal Elektronik Ilmu Komputer Udayana
Volume 9 No. 3, February 2021

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Location Named-Entity Recognition using Rule-Based Approach for Balinese Texts

Ni Putu Ayu Sherly Anggita Sugiarta[a1], Ngurah Agus Sanjaya ER[a2]

[a]Informatics Department, Udayana University
Bali, Indonesia
[1]nptayusherly@gmail.com
[2]agus_sanjaya@unud.ac.id

Abstract

In Natural Language Processing (NLP), Named Recognition Entity (NER) is a sub-discussion widely used for research. The NER's main task is to help identify and detect the entity-named in the sentence, such as personal names, locations, organizations, and many other entities. In this paper, we present a Location NER system for Balinese texts using a rule-based approach. NER in the Balinese document is an essential and challenging task because there is no research on this. The rule-based approach using human-made rules to extract entity name is one of the most famous ways to extract entity names as well as machine learning. The system aims to identify proper names in the corpus and classify them into locations class. Precision, recall, and F-measure used for the evaluation. Our results show that our proposed model is trustworthy enough, having average recall, precision, and f-measure values for the specific location entity, respectively, 0.93, 0.93, and 0.92. These results prove that our system is capable of recognizing named-entities of Balinese texts.

Keywords: *Location, Named-Entity Recognition, Rule-Based Approach, Balinese Texts*

1. Introduction

Natural Language Processing (NLP) is an Artificial Intelligence (AI) branch that focuses on natural language processing. Natural language is the language commonly used by humans in communicating with each other. NLP tries to make computers understand human language by giving computers knowledge of the human language. NLP is a computational technique for analyzing and representing natural text at one or more linguistic analysis levels to achieve human-like language processing for various tasks or applications. There are a lot of fields that apply NLP technologies such as Information Retrieval (IR), Information Extraction (IE), Question-Answering, etc. [1]. One of the sub-tasks of IE is to help the process to identify and extract such information called named-entity, and it is known as a Named-Entity Recognition (NER) process.

NER is a crucial component in many NLP applications, such as question answering, information extraction, clustering, topic tracking, and summarization [2][3]. Identification and classification are the aims of NER. The main issue of the NER is to identify proper names in text documents and to classify them in some of the predefined types, such as persons, organizations, locations, temporal expressions, numeric expressions, etc. which is very useful in the case of information extraction [3][4]. The studies' domain usually influences the NER algorithm's implementation for NLP and on different languages may require other techniques in recognizing the named-entity [3]. For example, detecting the entity type for a document written in English can quickly be done by detecting proper nouns. Proper nouns usually start with a capital letter. It is used to represent named-entities such as people, locations, organizations, etc. However, this method may not apply to documents written in Arabic [5].

Algorithms for NER systems can be classified into three categories; rule-based, machine learning, and hybrid [6]. A Rule-Based NER algorithm detects the named-entity by using a set of rules and a list of dictionaries manually predefined by humans [3]. The rule-based NER algorithm applies a

set of rules to extract patterns, and these rules are based on pattern base for location names, pattern base for organization name, etc. The patterns are mostly made up of grammatical (e.g., part of speech), syntactic (e.g., word precedence), and orthographic features (e.g., capitalization) in combination with dictionaries [6]. Rule-based methods are usually based on an existing lexicon of proper names and a local grammar that describes patterns to match NEs using internal evidence (gazetteers) and external evidence provided by the context in which the named-entity appear [7]. Systems based on the machine learning approach use stochastic techniques and learn specific knowledge of a massive learning corpus where the target named-entity is labeled. Nevertheless, this approach requires an enormous amount of learning data for its learning algorithm [8]. And the hybrid approach combines the two techniques mentioned above for their complementarity.

Many studies have used NER in various languages, including Indonesian, Portuguese, Turkish, Malaysian, Arabic, Persian, Indian, and Korean, with different methods [9]. Wulandari et al. [10] focused on Indonesian cell biology documents; they built the system using rule-based and Naïve Bayes Classifier. The highest average precision, recall, and f-measure with a micro average on rule-based is 85%. Dias et al. [11] proposed NER to handle a sensitive data discovery in Portuguese. They combine several techniques, such as rule-based/lexical-based models, machine learning algorithms, and neural networks. Rule-based and lexical-based approaches are used only for a specific set of classes. The Conditional Random Fields, Random Forest, and Bidirectional-LSTM method are used for the remaining entity classes. The best score was obtained using the Bidirectional-LSTM method, achieved a result of 83.01%. Another study by Alfred [3] developed a system by using Malay Part-Of-Speech (POS) tagging features and contextual features to construct a system Malay NER to handle three named-entities; person, location, and organizations. The experimental results show a good output of 89.47% for the F-Measure value. Mesfar [12] has developed a system for Arabic NER. He used the NooJ linguistic platform for building his system. The system contains a gazetteer, tokenizer, triggers, and morphological analyzer for recognizing proper names, dates, and temporal expressions used in Arabic text. The overall average accuracy of the F-measure is 87%.

The Balinese language is one of the Austronesian languages, the mother tongue for the Balinese people who live in Bali, Indonesia. As a language included in the top ten regional languages with the most speakers in Indonesia, this language is essential for its existence [13]. The Balinese language is used in government affairs, education, and other matters like a daily conversation. Nevertheless, until now, there has been no research related to the application of NER in Balinese. It's motivated us to take up NER in Balinese text as the proposed research area, especially for location type, because that type is general enough to be useful for many application domains [14]. We use a rule-based approach instead of a machine learning approach, considering that there are not as many text documents in Balinese as available in Indonesian or English. An example of implementing a rule-based NER can be seen in the following sentence, *"Tunyan semeng I Made mare teka uli Denpasar."* NER will detect *Denpasar* as a location marking entity.

This paper describes a rule-based Balinese NER system used to identify and classify the location named-entities in a Balinese text document. Rules-based NER is expected to provide adequate system performance. What follows are the details of the proposed research work. Section 2 discusses the detailed methodologies used in this paper. The results are evaluated and discussed in section 3. Section 4 shows our conclusion of the research.

2. Research Methods

In this study, the proposed method approach for detecting location marking entities in Balinese text documents uses rule-based NER. The proposed architecture in this study can be seen in Figure 1. A rule-based approach is used to find named-entities of each word in Balinese text documents. The rules used are obtained based on observations on the data, which these rules will be used in NER rule-based. The general flow of the research we will do begins with collecting dirty data from the internet. After that, the preprocessing stage will be carried out, including punctuation remove, normalization, and tokenization. Then enter the rule-based named-entity classification stage and continue with the evaluation process.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

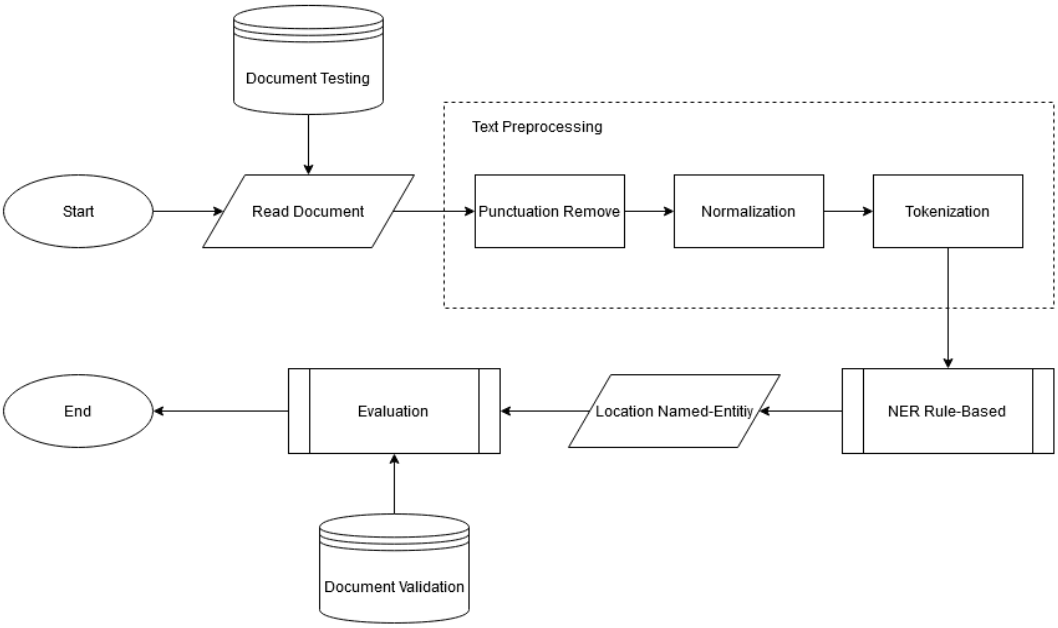

Figure 1. The NER Architecture

2.1 Data Collection

The data used in this research is secondary data. Secondary data is already available before we start the research; this data is related to the research. The data used were 70 Balinese text documents sourced from the internet. The dataset used in this study is a Balinese language document with the *.txt file format written in Latin letters. Table 1 is the sample Balinese language text with the location named-entities marked:

Table 1. Sample Balinese language text with the location named-entities marked

Bulan Bahasa Bali Pinakaa Pikamkam Ngelestariang Bahasa Bali
Beritabali.com, [LOC DENPASAR]. Bulan Bahasa Bali wantah silih tunggil program sane kapikamkam olih Pemerintah Provinsi Bali ri sajéroning utsaha ngrwérdiang, ngélastariang miwah ngélimbakang kawenténan basa, aksara miwah sastra Bali.
Pucak tawur agung panca wali krama miwah karya bhatara turun kabeh ring [LOC Pura Besakih], sampun kamargiang saha sane mangkin ngéranjing ring pamargin bakti pénganyar. Upacara puniki pinaka dudonan pamargin tawur agung panca wali krama wiadin karya bhatara turun kabeh ring [LOC Pura Penataran Agung Besakih], sadurung upacara pényinéban.

2.2 Data Preprocessing

Preprocessing is used to present text documents in a clear word format. The steps taken for preprocessing in this study were punctuation remove, normalization, and tokenization.

a. Punctuation removal

Punctuation removal is a process to remove symbols contained in text documents. The symbols that will be removed are \"#$%&()*+:;<=>@[\]^_`{|}~\n. Table 2 is an example of the results of the punctuation removal stage.

Table 2. Example of Punctuation Removal

Text Data	Punctuation Remove Results
$#Bulan Bahasa %%Bali wantah silih tunggil program sane kapikamkam olih Pemerintah Provinsi Bali ri $sajéroning utsaha&& ngrwérdiang, ngélastariang miwah ngélimbakang *+kawenténan basa, aksara miwah sastra Bali*+.	Bulan Bahasa Bali wantah silih tunggil program sane kapikamkam olih Pemerintah Provinsi Bali ri sajéroning utsaha ngrwérdiang, ngélastariang miwah ngélimbakang kawenténan basa, aksara miwah sastra Bali.

b. Normalization

Normalization is the process of converting text into standard forms. In this research, the character é will be changed to e. Table 3 is an example of the results of the tokenization stage.

Table 3. Example of Normalization

Text Data	Normalization Results
Bulan Bahasa Bali wantah silih tunggil program sane kapikamkam olih Pemerintah Provinsi Bali ri sajéroning utsaha ngrwérdiang, ngélastariang miwah ngélimbakang kawenténan basa, aksara miwah sastra Bali.	Bulan Bahasa Bali wantah silih tunggil program sane kapikamkam olih Pemerintah Provinsi Bali ri sajeroning utsaha ngrwerdiang, ngelastariang miwah ngelimbakang kawentenan basa, aksara miwah sastra Bali.

c. Tokenization

Tokenization is the process of decomposing a description from sentences into tokens, in this case, in the form of words. Table 4 is an example of the results of the tokenization stage.

Table 4. Example of Tokenization

Data Teks	Tokenization Results			
Bulan Bahasa Bali wantah silih tunggil program sane kapikamkam olih Pemerintah Provinsi Bali ri sajeroning utsaha ngrwerdiang, ngelastariang miwah ngelimbakang kawentenan basa, aksara miwah sastra Bali.	Bulan Bahasa Bali wantah silih tunggil program	sane kapikamkam olih Pemerintah Provinsi Bali ri	sajeroning utsaha ngrwerdiang , ngelastariang miwah ngelimbakang kawentenan	basa , aksara miwah sastra Bali .

2.3 NER Rule-Based Approach

A rule-based approach will help find each word's location marking entities in the Balinese text document. Rule-based is a method in which the rules in the system are made based on linguistic knowledge. As a rule-based approach is a domain-specific, rules define one language will not apply to other languages. The analysis carried out at the syntactic and semantic levels in more depth is an advantage of this method. The steps are as follows:

a. Step - 1 Read the Balinese text document.
b. Step - 2 The data will go through the preprocessing stage.
c. Step - 3 NER will detect a token, a location marker according to the rules used.
d. Step - 4 Repeat steps 1 - 3 for all data.

Rules are made with due regard to the morphological and contextual structures. Table 5 and Table 6 show a list of the features we use.

Table 5. List of contextual features

Feature Name	Explanation	Example
locPrefix	Location prefix	Gunung, Tukad, Pante, Desa, Kecamatan, Kota, Propinsi
locSufix	Location sufix	Utara, Timur, Tenggara, Selatan, Barat, Tengah Kaja, Kangin, Kauh, Kelod
preposition	Prepositions that are usually followed by location name	ring, saking, uli, ka
locArea	Denotes the area of a location	wewidangan, gumi, wawengkon, jagat
conjunction	After conjunction is the location if before the conjunction is location	lan, tur, miwah, sareng, utawi
date	Format date	12 Juli 1967

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Table 6. List of morphological features

Feature Name	Feature Name	Feature Name
titleCase	Begin with an uppercase letter and followed by all lowercase letter	Tanah Lot, Danu Bratan, Taman Sukasada
upperCase	All uppercase letter	GWK, MBG
lowerCase	All lowercase letter	ring, wewidangan
digit	All number	12
digitSlash	Number with slash	12/7

The following are the rules used in our system.

a. **IF** ([token] = 'locPrefix' morphological type = *titleCase*)
THEN (for [**next token**] morphological type = *titleCase* or *digit*) → location name
Example: [LOC Gunung Agung] punika gunung sane paling tegeh ring Bali.

b. **IF** ([token] = 'preposition' morphological type = *titleCase* or *lowerCase*)
THEN (for [**next token**] morphological type = *titleCase* or *upperCase*) → location name
Example: Wenten 8 kelurahan ring [LOC Karangasem].

c. **IF** ([token] = 'locSufix' morphological type = *titleCase*)
THEN (for [**previous token**] morphological type = *titleCase*) → location name
Example: [LOC Nusa Tenggara Barat] magenah ring sisi kauh Pulo Baline.

d. **IF** ([token] = 'date')
THEN (for [**previous token**] morphological type = *titleCase*) → location name
Example: [LOC Jimbaran], 12 Juli 1967

e. **IF** ([token] morphological type = *digitSlash*)
THEN (for [**previous token**] morphological type = *titleCase*) → location name
Example: [LOC Denpasar] (12/7) Lomba layangan sane wenten ring Panjer sampun usan.

f. **IF** ([token] = 'conjunction' and [token-1] = location)
THEN (for [**next token**] morphological type = *titleCase*) → location name
Example: Kabupaten Karangasem lan [LOC Klungkung] magenah ring Provinsi Bali.

g. **IF** ([token] = 'locArea')
THEN (for [**next token**] morphological type = *FirstCap*) → location name
Example: Ring wewidangan [LOC Desa Selat], wenten Pura Dalem sane ngeluanin Pura Desa.

2.4 Evaluation

Several measures have been defined to evaluate the quality of a NER system's output. The usual measures are called precision, recall, and F1-measure. The precision is the ratio between the correct positive class predictions' total results with the total data predicted as the positive class. The recall compares the total results of the correct positive class predictions with the total data that is truly positive. However, several issues remain in just how to calculate those values. This study's evaluation will involve expert data containing the location marking entities included in the data used. A scoring model developed for the Message Understanding Conference (MUC) and Multilingual Entity Task (MET) evaluations measures both precision (P) and recall (R), terms borrowed from the information-retrieval community, where [6]:

$$P = \frac{number\ of\ correct\ responses}{number\ of\ responses} \quad (1)$$

And

$$R = \frac{number\ of\ correct\ responses}{number\ correct\ in\ key} \quad (2)$$

F-Measure is one of the evaluation calculations in information retrieval that combines precision and recall:

$$F = \frac{R * P}{0.5 * (R + P)} \quad (3)$$

The term response denotes “answer delivered by the system”; the term key is used to mean “an annotated file containing correct answers”. In MUC-7, a correct answer from a NER is where the label and boundaries are correct.

3. Result and Discussion

In this study, a rule-based NER application to find location marking entities in Balinese text documents is implemented using the python programming language. This study will involve data validation, which contains location entities in the data used. Precision, recall, and F1-measure will be searched by the formula described in equation (1) for precision calculations, equation (2) for recall calculations, and equation (3) for F1-measure calculations.

For now, the rules used in the system we built can only cover a specific location entity in the form of a proper noun. Proper nouns usually start with a capital letter, for example, Mount Indrakila, Pantai Pasir Putih, Tukad Unda, etc. For a nonspecific or general location entity in the form of a common noun, such as peken, carik, alase, etc., our system has not detected it. That is a weakness in our system. Therefore, in our experiment, we carried out two test scenarios: validation data containing only the specific location entity and validation data containing all location data, including the specific location entity and general location entity. That is to see how far the difference is from the results obtained. Table 5 shows the results of precision, recall, and F-measure, which is obtained.

Table 5. Ruled-Based Test Results for Location Named-Entity

Named-Entity	Recall	Precision	F-Measure
Specific location	0.9358732105261144	0.9366584209441352	0.9207823038481747
All location (specific and general location)	0.775722336807337	0.9355595198452339	0.8122392440626851

From the experiments that have been conducted, the results of the evaluation of the average recall, precision, and f1-measure, for the specific location entity respectively are 0.935, 0.936, and 0.920. And the results of the evaluation of the average recall, precision, and f1-measure for the all location entity are 0.775, 0.935, and 0.812. The visualization of the comparison results obtained from the two test scenarios performed is shown in Figure 2.

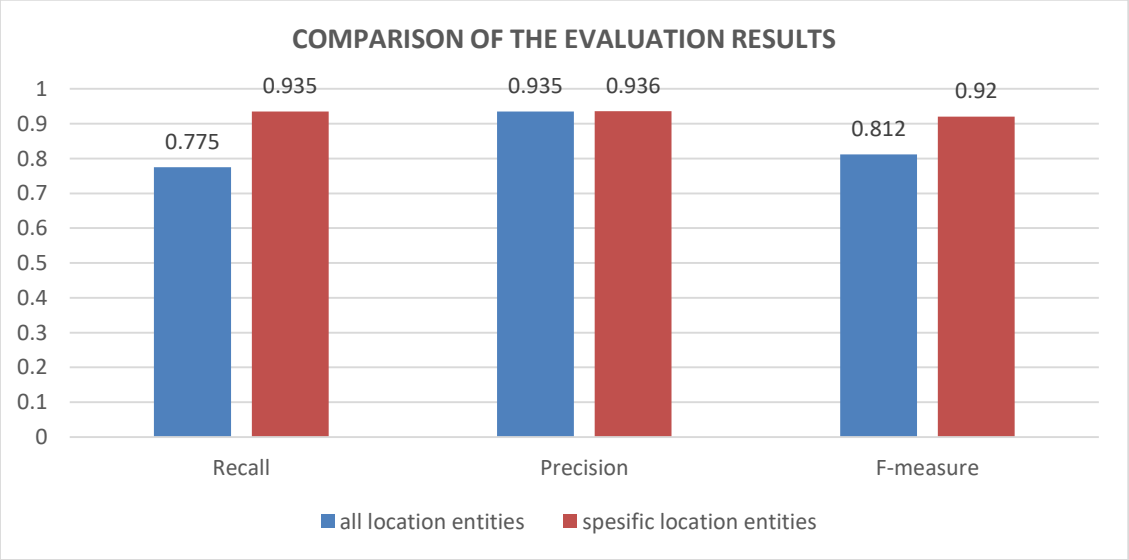

Figure 2. Comparison of The Evaluation Results

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Several factors cause the study's results for the specific location entities to be less than optimal. One of which is the error in writing the location's name, namely not using proper nouns, the absence of consistency in capital letters, for example, a word is written in all capital letters or without using capital letters. And some entities have not been recognized by the NER due to undefined rules. The evaluation value in all locations has decreased because there are no rules that cover nonspecific location entities. Nonspecific location entities are common nouns marked with lowercase letters; this makes it difficult for us to create rules that can identify nonspecific location entities.

4. Conclusion

Based on the results of NER's study on Balinese text documents using a Rule-Based approach, the following conclusions were obtained:

a. A rule-based approach can be used in NER for Balinese text documents. The average recall, precision, and f-measure values are obtained for the specific location entity, respectively, 0.93, 0.93, and 0.92. Meanwhile, the average recall, precision, and f-measure values are obtained for all location entities, each 0.77, 0.93, and 0.81.
b. Building a NER system in the Balinese language using a rule-based approach can help many significant applications. This finding is expected to be an initial approach in the development of NER for Balinese. In the future, perhaps the application of NER in Balinese text documents can be further expanded so that it can cover all locations in the document text, especially general location entities. Or maybe use a machine learning or hybrid approach to obtain better precision, recall, and F-measure results. In particular, NER's application in Balinese is expected to facilitate and significantly impact social work.

References

[1] E. D. Liddy, "Natural Language Processing," in *Encyclopedia of Library and Information Science*, 2nd ed., New York: Marcel Decker, Inc., 2001.

[2] S. Kumova Metin, "Named Entity Recognition in Turkish Using Association Measures," *ACIJ*, vol. 3, no. 4, pp. 43–49, Jul. 2012, doi: 10.5121/acij.2012.3406.

[3] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach," *IJMLC*, vol. 4, no. 3, pp. 300–306, Jun. 2014, doi: 10.7763/IJMLC.2014.V4.428.

[4] S. Morwal, "Named Entity Recognition using Hidden Markov Model (HMM)," *IJNLC*, vol. 1, no. 4, pp. 15–23, Dec. 2012, doi: 10.5121/ijnlc.2012.1402.

[5] A. M. Saif and M. J. A. Aziz, "An Automatic Collocation Extraction from Arabic Corpus Abdulgabbar," *J. Comput. Sci.*, vol. 7, no. 1, pp. 6–11, 2011, doi: 10.3844/jcssp.2011.6.11.

[6] A. Mansouri, L. S. Affendey, and A. Mamat, "A survey of named entity recognition and classification," *Lingvisticae Investig.*, vol. 30, no. 1, pp. 3–26, Aug. 2007, doi: 10.1075/li.30.1.03nad.

[7] W. Zaghouani, "RENAR: A Rule-Based Arabic Named Entity Recognition System," *ACM Trans. Asian Lang. Inf. Process.*, vol. 11, no. 1, pp. 1–13, Mar. 2012, doi: 10.1145/2090176.2090178.

[8] L. Chahira, Z. Anis, and Z. Mounir, "A Rule-based Named Entity Extraction Method and Syntactico-Semantic Annotation for Arabic Language," in *ALLDATA 2017: The Third International Conference on Big Data, Small Data, Linked Data and Open Data (includes KESA 2017)*, 2017, pp. 63–69.

[9] Rusliani and K. K. Purnamasari, "Named Entity Recognition Pada Teks Berbahasa Indonesia Untuk Pembangkit Pertanyaan Otomatis," Universitas Komputer Indonesia, 2017.

[10] D. W. Wulandari, P. P. Adikara, and S. Adinugroho, "Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier," *J.*

Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 11, pp. 4555–4563, 2018, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2988.

[11] M. Dias, J. Boné, J. C. Ferreira, R. Ribeiro, and R. Maia, “Named Entity Recognition for Sensitive Data Discovery in Portuguese,” *Appl. Sci.*, vol. 10, no. 7, p. 2303, Mar. 2020, doi: 10.3390/app10072303.

[12] S. Mesfar, “Named Entity Recognition for Arabic Using Syntactic Grammars,” pp. 305–306.

[13] I. G. B. W. B. Temaja, “POLA REDUPLIKASI BAHASA BALI: PERBANDINGANNYA DENGAN POLA REDUPLIKASI BAHASA-BAHASA AUSTRONESIA,” *PRASASTI J. Linguist.*, vol. 3, no. 2, p. 190, Nov. 2018, doi: 10.20961/prasasti.v3i2.17520.

[14] J. Jiang, “Information Extraction from Text,” in *Mining Text Data*, Research C., C. C. Aggarwal and C.-X. Zhai, Eds. Boston, MA: Springer US, 2012, pp. 11–41.