Application of named entity recognition method for Indonesian datasets: a review

Indra Budi[1], **Ryan Randy Suryono**[2]
[1]Department of Information Systems, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2]Department of Information Systems, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia

Article Info

Article history:

Received Aug 7, 2022
Revised Oct 6, 2022
Accepted Oct 24, 2022

Keywords:

Bibliometric analysis
Named entity recognition
Online news
Social media
Systematic literature review

ABSTRACT

A name entity (NE) is a proper name that designates a person, location, or organization. For humans, named entity recognition (NER) is a straightforward process insofar as many named entities are self-names, and most of them have initial capital letters and can be easily recognized, but it is very difficult for machines. This study discusses research trends in the application of NER to Indonesian datasets, particularly as it concerns certain tasks, datasets, methods/techniques, and entity labels. By conducting a systematic literature review (SLR) and bibliometric analysis with VOSviewer, this article hopes to provide opportunities for adopting old methods, combining models from previous research, and even proposing new methods. In addition, the motivation for doing SLR at NER is to look for new strategies in the supervision of financial technology (Fintech). If machines can find illegal Fintech entities on social media and online news, it can help the government to block these illegal Fintech entities. To this end, this study provides an overview of research trends in applying the NER method to *Bahasa Indonesia* (Indonesian) datasets, including the extraction of news articles, the monitoring of floods, and traffic.

Corresponding Author:

Indra Budi
Department Information Systems, Faculty of Computer Science, Universitas Indonesia
Pondok Cina, 16424 Depok, West Java, Indonesia
Email: indra@cs.ui.ac.id

1. INTRODUCTION

Named entity (NE) was introduced at the sixth message understanding conference (MUC-6). With the introduction of NE, the MUC conference has helped to advance the field of information extraction [1]. NE refers to a proper name that designates a person, location, or organization. For example, there are three NE in the following sentence: “James is a doctoral student in the Faculty of Computer Science at the University of Indonesia.” James an NE insofar as it is the name of a person (P); Indonesia refers to a location (L); and the Faculty of Computer Science refers to the organization (O). Named entity recognition (NER) is a procedure that finds, extracts, and automatically classifies named entities from open domains and unstructured texts such as newspaper articles. It then categorizes these NE into predefined types [2]. There are four approaches to NER: i) a rule-based approach, which does not require annotated data because it relies on artificial rules; ii) an unsupervised learning approach; iii) a feature-based supervised learning approach that relies on supervised learning algorithms with careful feature engineering; and iv) a deep-learning-based approach, which automatically finds the required representation for detecting or classifying raw input in an end-to-end manner [3], [4]. NER is a straightforward process for humans because many named entities are self-names, and most of them have initial capital letters and can be easily recognized, but for machines, it is very difficult [5].

Information extraction often uses data available on social media, online news, and e-commerce [3]. Much information can thereby be obtained, including product reviews, analysis, and information extraction. For example, NER research is used for Indonesian news articles [6]. The use of NER is also carried out for the extraction of comments related to flood monitoring and traffic monitoring [7], [8]. On the other hand, the use of this method is also useful for quote identification [9]. The role of language in text-analysis often determines which model is used [10], because not all libraries are available for specific tasks [2].

NER has been applied to a wide variety of tasks [3], but a brief survey of the application of NER to texts in the Indonesian language reveals a total of only 241 documents (accessed December 2021). Meanwhile, the need to perform NER with Indonesian datasets is continuing to grow. Currently, there are libraries and tools available to facilitate machine learning (ML) as it pertains to the use of NE to extract information, but are there enough datasets? To what extent is NER used to extract information on social media and online news in an Indonesian-language context? Not all natural language processing (NLP) functions are available in Indonesian because, unlike in English, the functions that rely on the ML model mentioned above are not directly supported [11], [12].

In addition, another motivation for doing SLR is triggered by the emergence of illegal financial technology (Fintech) problems [13], [14]. Several previous studies on Fintech have been carried out and the main side that can be solved is by monitoring entities on social media [15]. But the challenge is that not all corpus (text set) is available in all languages.

This study looks at research trends in the application of NER to Indonesian datasets, including specific tasks, datasets, method/techniques, and entity labels. Therefore, this article will help facilitate the design of experiments to extract Fintech information on social media and online news. With the hope that it is not only a Fintech platform but can be a proposal for supervision of agencies or organizations based on social media data and online news.

2. METHOD

2.1. Systematic literature review

First, this article presents a SLR of the field of NER research. A SLR aims to collect all research on a particular topic, evaluates it critically, and reaches conclusions that synthesize that research. Then follows a discussion of how NER has been applied to Indonesian texts. SLR has been used in various research domains such as P2P lending [13], Fintech [14], Teaching and learning via webinars [16], supply chain management model [17], and software engineering [18].

A SLR was carried out in three stages: the planning stage, the implementation stage and the reporting stage (see Figure 1). In the first stage, the planning stage is carried out to identify the need for a systematic review of the use of the agile project management (APM) method. At this stage, a review protocol was also developed by setting research questions (RQ) and formulating a boolean search to determine search keywords. This study used the population, intervention, comparison, outcomes, and context (PICOC) strategy to determine the RQ, as shown in Table 1.

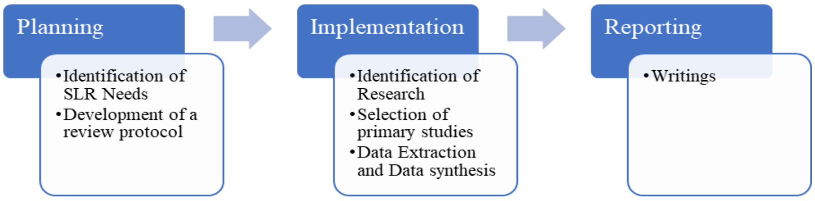

Planning
•Identification of SLR Needs
•Development of a review protocol
Implementation
•Identification of Research
•Selection of primary studies
•Data Extraction and Data synthesis
Reporting
•Writings

Figure 1. Steps for a SLR

Table 1. Criteria of research question

Population (P)	Name-entity recognition, NER, name entity recognition
Intervention (I)	Online news, social media
Comparison (C)	n/a
Outcomes (O)	Trend and application of NER in Indonesian context
Context (C)	Bahasa, Indonesia

There follows the RQ that guided the following analysis:
RQ. “What are the trends in the application of NER to extract information from Indonesian online news and social media?”

In this study, the search string is ("named-entity recognition" OR NER OR "named entity recognition") AND ("online news" OR "social media") AND (Indonesia* OR Bahasa). According to the research question, the criteria for inclusion and exclusion in Table 2 were used to define the results. In the second stage, this research defines a search strategy, namely selecting a publication database, selection results for research, data extraction and the synthesis process. These processes are sequential processes where each process aims to find the right study to be used in this research. The search and selection process are an elimination process based on the criteria specified in each process.

The authors collected papers from relevant electronic databases such as SCOPUS, ACM, IEEEXplore, and Science Direct, then used Mendeley software to organize the data. Some irrelevant papers were omitted in the first stage of collection based on the title and abstract. The second stage of selection articles is a full-text selection. Figure 2 illustrates the procedure of text-selection. The total number of papers obtained from the four databases was initially 241. Upon completion of the selection procedure, however, only 20 papers remained. The low number of papers is both a challenge to and an opportunity for NER research in an Indonesian context, as few studies have used the "*Bahasa*" dataset. The third stage is reporting the results and analyzing the results of this review. We mapped research results from previous studies and examined how the experimental process in NER was, what libraries could be used for Indonesian language datasets, how to approach NER, and proposed future research.

Table 2. Criteria of selection studies process

Inclusion criteria	Exclusion criteria
The paper studied about NER	The paper is not using English
Studies published in the last 5 years, between 2016-2021	Not full-text paper
The paper being studied is in the form of a journal or proceedings/conference	Same papers from different database
	Papers discussing NER but not the Indonesian text dataset

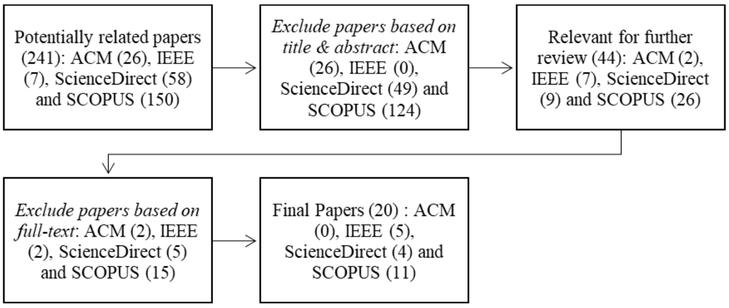

Potentially related papers (241): ACM (26), IEEE (7), ScienceDirect (58) and SCOPUS (150)
Exclude papers based on title & abstract: ACM (26), IEEE (0), ScienceDirect (49) and SCOPUS (124)
Relevant for further review (44): ACM (2), IEEE (7), ScienceDirect (9) and SCOPUS (26)
Exclude papers based on full-text: ACM (2), IEEE (2), ScienceDirect (5) and SCOPUS (15)
Final Papers (20) : ACM (0), IEEE (5), ScienceDirect (4) and SCOPUS (11)

Figure 2. The selection procedure for final papers

2.2. Bibliometrics analysis

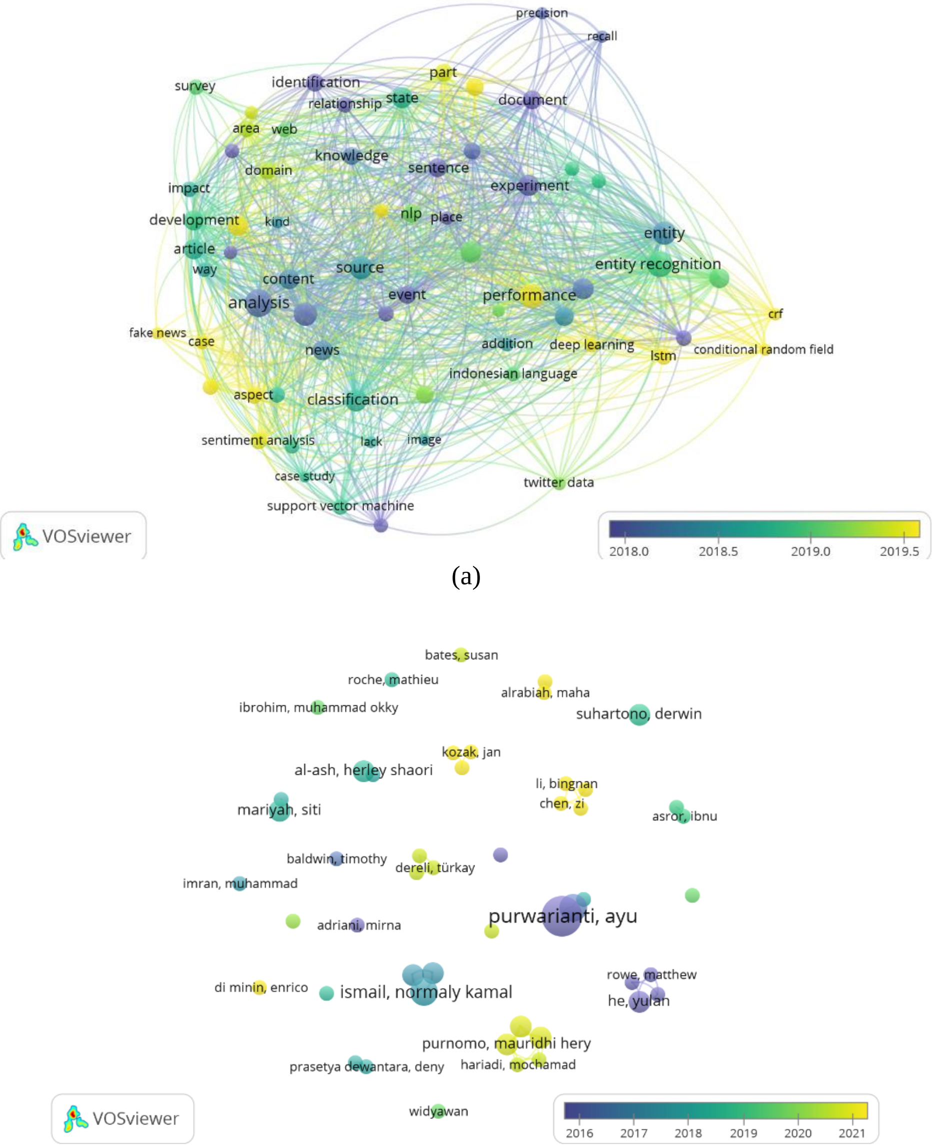
This study also presents a bibliometric analysis of the document results at the initial stage of selection. Bibliometrics is one way to perform statistical analysis of books, articles, or other publications. This analysis is carried out using data on the number and authors of scientific publications as well as articles and citations in them which aims to measure the outcomes of individuals or research teams, institutions, and countries, identify national and international networks and map the development of new fields of science and technology. The VOSviewer tool views keyword clusters and authors in the NER field and thereby helps to expand the scope of NER research.

3. RESULTS AND DISCUSSION

3.1. Bibliometrics analysis results

VOSViewer software helps to visualize research trends by putting the keywords of articles into clusters and constructing diagrams from them. From Figure 3(a) that NER research began to develop in early 2018 (see blue cluster) with content analysis tasks and experiments to identify documents and sentences. Some of these tasks included the comparison of precision-recall with simple ML model approaches such as conditional random fields (CRF) and support vector machines (SVM). It was not until the beginning of 2019 (see green cluster), however, that research using Indonesian datasets began. It was also around that time that several other classification tasks also began to develop. Research data does not only come from scholarly articles but is also available in the form of data and images on the web and on platforms such as Twitter. On the other hand, if we look at 2019-2021 (see yellow cluster), NER research has started working on fake news, conducting aspect-based sentiment analysis, and measuring the model's performance. This is where the NER approach with deep learning (DL) begins to emerge. DL is one of the implementation methods of ML which aims to imitate the workings of the human brain using an artificial neural network or artificial reasoning network. The algorithm results are naturally expected to improve the performance of ML.

Additionally, we conducted a VOSViewer analysis with the co-authorship feature to see which authors were actively researching NER topics. Of 684 authors, 49 met the threshold; however, the results show that researchers are not connected by any network. This shows that each NER experiment has its own research goals, dataset, methods/techniques, as well as part of speech (PoS) tagging process. It can be seen below that the one of the most active researchers in the field is Purwanti, see Figure 3(b).



(a)

(b)

Figure 3. Co-occurrence network of author keywords (a) co-authorship network and (b) from 2016 to 2021

3.2. SLR results

After investigating the VOSViewer results, we examined 20 articles collected from the ACM, IEEE, ScienceDirect, and SCOPUS portals. The articles were then extracted and mapped according on author, task, dataset, and method/technique (see Table 3). It is clear from the table above that several NER studies with Indonesian datasets have been carried out for the following tasks: complaint classification [19], quote identification [9], [20], flood monitoring extraction [7], traffic monitoring [8], [21], tourist attractions [22], zakat [23], lipstick product reviews [24], and various model combination tests for twitter [25]–[28], online news [6], [29]–[31], and Wikipedia [32], [33].

Table 3. SLR results

Author	Tasks	Dataset	Methods/techniques
Purwanti *et al.* [19]	Using InaNLP for complaint tweet classification	7,440 Twitter data	InaNLP: Indonesia natural language processing toolkit
Wibawa *et al.* [29]	Build Indonesian NER for newspaper articles with 15 classes	457 online news data from: Detiknews.com, Kompas.com, Mediaindonesia.com	Supervised machine learning in the NER (Naïve Bayes, SVM, and simple logistic)
Syaifudin *et al.* [9]	Identify quotes from Indonesian online news texts	2506 sentences from: kompas.com, tempo.co, and tribunnews.com	Rule-based method
Taufik *et al.* [25]	Modelling NER on Indonesian microblog messages	600 Twitter data	Rule-based method
Munarko *et al.* [26]	Grouping formal and informal Twitter	8,000 Twitter data	CRF
Gunawan *et al.* [32]	Using deep learning to identify Indonesian-language entities	4139 sentences from Wikipedia	Hybrid bidirectional long short-term memory (BLSTM) and convolutional neural network (CNN)
Herwanto *et al.* [21]	Propose an information extraction method to map traffic conditions from tweets	3,013 Twitter data	Rule-based method
Alifi *et al.* [8]	Designing model architecture for traffic information extraction	44,102 Twitter data	Bidirectional LSTM and CNN
Azarine *et al.* [27]	Build NER on Indonesian-language tweets with hidden Markov model (HMM) and add POS tagging feature extraction	500 Twitter data	HMM
Leonandya *et al.* [33]	Apply and evaluate the latest transfer learning techniques	Online news from kompas.com and tempo.co and Indonesian Wikipedia	Deep bidirectional language models and transfer learning
Wintaka *et al.* [28]	This study builds a model using a combination of deep learning and machine learning approaches for Twitter data.	250 formal tweets dan 350 Informal tweets	BLSTM dan CRF
Emcha *et al.* [20]	Extracting quotation on Indonesian online news	503 standard sentences and 395 indirect quotation sentences from kompas.com, detik.com, tempo.co, tribunnews.com, and antaranews.com	SVM algorithm
Rosyiq *et al.* [22]	Information Extraction of Indonesian Tourist Attractions	800 Twitter data	DBpedia ontology
Santoso *et al.* [6]	Propose a hybrid approach for named entity recognition	51.241 entities from Indonesian Online News	Hybrid CRFand K-Means
Azzahra *et al.* [34]	Conducting NER research for unstructured text format datasets in Indonesian using a deep learning approach	500 Twitter data	HMM
Yohanes *et al.* [35]	Building a framework to build a corpus for extracting Indonesian public figure quotes	Indonesian online newspaper (Kompas Daily)	GloboQuotes, PARC 3.0, PolNeAR and Quootstrap
Putra *et al.* [7]	Utilization of social media data to serve as flood monitoring data	72,212 Twitter data	Naive Bayes, random forest, SVM, logistic regression, and CRF
Sukmana *et al.* [23]	Building a knowledge graph for zakat involves data acquisition, extracting entities and their relationships, mapping to ontologies, and applying knowledge graphs and visualizations.	The entire documents are 24 documents with 15,979 words from online sources and four offline data documents.	Framework Indonesian-open domain information extractor for processing entity-relationship identification, mapping to ontology, and deploying knowledge graphs
Indarta *et al.* [24]	Extraction of aspects and opinions on lipstick product reviews	591 sentence reviews and 8,574	CRF and HMM
Santoso *et al.* [30]	Extracting the ontology building concept automatically with NER	29.587 Indonesian online news articles collected from CNN Indonesia	End-to-end model deployment using BLSTM

3.3. Discussion

The internet and especially social media are a strategic tool for disseminating information to the public. Techniques have recently emerged that allow one to extract information on a targeted topic from the internet and then to examine the relationship between the words associated with that topic. Moreover, these techniques allow one to map out the relationship between the chief exponents of that topic and perhaps even locate them by charting their movements. One technique, namely text mining, provides a set of methodologies and tools for finding, visualizing, and evaluating information from extensive collections of text data [36]. Four processes need to be executed in text mining (see Figure 4). There are two ways of collecting data from social media and online news: i) web crawling using an API or BOT automatically; ii) web scraping by inserting HTML or XML elements using the HTTP protocol. After the data is collected and cleaned, the next stage is pre-processing, which can be done with a tokenizer, by removing stopwords, or by stemming.

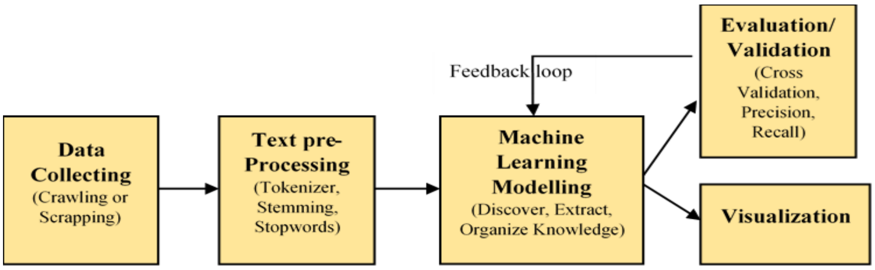

Data Collecting (Crawling or Scrapping)
Text pre-Processing (Tokenizer, Stemming, Stopwords)
Machine Learning Modelling (Discover, Extract, Organize Knowledge)
Evaluation/ Validation (Cross Validation, Precision, Recall)
Visualization
Feedback loop

Figure 4. The basic process of text mining

At that point a machine-learning approach to modeling performs a looping procedure for a final evaluation and validation. Finally, presentations of the data help to visualize the results of modeling after the tasks of categorization, recommendation, spam detection, and summarization have been completed. As explained in the introduction, the toolkit for analyzing languages, especially the natural language toolkit (NLTK), is intended for English. Each country needs to rely on other tools and cannot fully use NLTK. NLTK is a library and program for NLP written in the Python programming language. NLTK supports tokenization classification, stemming, tagging, parsing, and semantic reasoning functions. Some Indonesian language libraries have InaNLP, kateglo, BimaNLP, Indonesian Stemmer, Sastrawi, PySastrawi, and SentiStrengthID. Table 4 describes the most frequently used Indonesian language libraries. In addition, some tools and libraries for NER include SpaCy, GATE, OpenNLP, CoreNLP, NLTK, and CogcompNLP.

Table 4. Indonesian libraries

Libraries	Description
InaNLP [19]	An interface for InaNLP and Deeplearning4j's Word2Vec for Indonesian (*Bahasa Indonesia*) in the form of REST API.
Kateglo	The Indonesian thesaurus and glossary dictionary with 72253 dictionary entries, 191200 glossary entries, 2012 proverb entries, and 3423 abbreviations and acronyms.
BimaNLP	Repository for Python codes supporting NLP tutorials in Indonesian
Indonesian Stemmer [37]	Stemming Effect Study on Information Search in Indonesian based on Porter Stemmer
Sastrawi	High-quality stemmer library for Indonesian Language (Bahasa)
PySastrawi	Ported from Sastrawi project in PHP to Python
SentiStrengthID [38]	Sentiment Strength Detection in Bahasa Indonesia

NER is one of the first steps toward information extraction that seeks to find entities mentioned in a text and classify them into predefined categories such as the person's name, organization, location, time, value, and percentage [3]. NER is used in many NLP fields and can help address many needs [25], [39], [40]. NER is a critical pre-processing tool for various downstream applications such as information recovery, query answering, and machine translation. Recognition of named entities in search queries will help understand user intent better, thus providing better search results [41].

It is important to classify the various approaches that NER employs. Even though they both carry out classification functions, various other approaches to NER continue to develop. Figure 5 illustrates the NER approach. The NER approach to the non-ML algorithm consists of four steps. First, the rule-based method identifies the rules in the system that are made by themselves based on linguistic knowledge [9]. Second, the lexicon-based method works by first making a dictionary of opinion words (lexicon). Third, statistical based using probabilistic. For example, the CRF and HMM algorithms [24]. A CRF is a framework for building discriminative probabilistic models for segmenting and labelling sequential data. At the same time, HMM is the primary technique for POS tagging in NLP. HMM models observations using a Markovian process with a state that is not directly observed (hidden). The main idea of HMM is to solve the problem of sequence tagging. Fourth is ontology-based NER such as a machine-learning approach. This method can identify known terms and concepts in the unstructured or semi-structured text, but at the same time it also relies on updating. The ontology approach provides additional advantages in terms of making further reasoning and knowledge acquisition for the extracted concepts [23], [30].

In the field of NLP, researchers are interested in identifying the word class for each word in each sentence. For example, the sentence *Ryan menendang bola* ('Ryan kicks the ball'). After the POS tagging process, the classification is "Ryan/noun menendang/verb bola/noun." This is useful for choosing nouns in sentences. Word classes are referred to as syntactic categories. POS tagging is a form of sequential job classification.

There also exist several schemes to annotate NER data. Widely used tagging schemes include inside-outside (IO), inside-outside-beginning (IOB), and beginning-inside-last-outside-unit (BILOU). If two tags appear consecutively, IO cannot distinguish between their boundaries. However, IOB and BILOU can incorporate boundary information but differ concerning their respective abilities to model more acceptable context information [41].

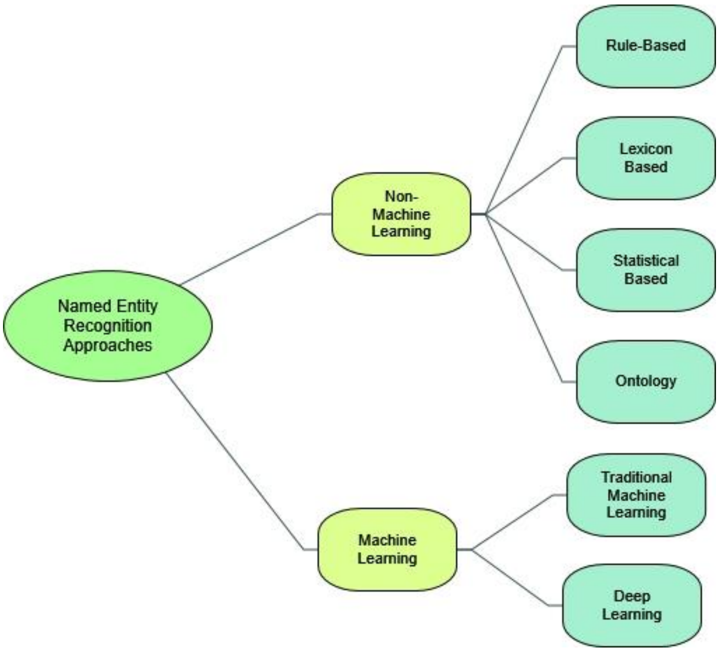


Figure 5. NER approaches

In conducting the NER experiment, the lack of datasets in Indonesian provides an opportunity for further research to build new datasets. The important thing in building this dataset is how to conduct crawling and scrapping. If we conduct scrapping manually, it may be necessary to spend time copying and pasting data. So, the suggestion for scrapping is to use coding, applications, and or browser extensions. HTML parsing techniques can also be performed via JavaScript and target linear and branching HTML pages. This method is more efficient in identifying HTML scripts from websites which are then used to extract text, links, and data. There is no one hundred percent effective scrapping technique because the data obtained are not always neat, and this depends on the structure of the page. So, understanding the structure of website pages is essential.

Second, after getting the dataset, we need to understand the data cleansing approach, including tokenizer, stemming, and stopwords. Several features to remove punctuation marks, numbers, and emoticons are used so that text data are of a high quality before being used during data analysis. Text preprocessing prepares unstructured text into good data so that they are ready to be processed.

Third, the prepared dataset is generally divided into training data, development sets, and testing sets in building ML models. Training is the process of building a data model, and testing is testing the performance of the learning model. Development sets are generally not used when the data set is small. For example, 80% training data and 20% testing data or 70% training data and 30% testing data. the right approach must be selected to carry out NER carefully. Several studies demand a high level of accuracy and a high percentage of F1 scores.

3.4. Recommendations for illegal Fintech supervision strategies with the NER approach based on social media data and online news

Based on the SLR, new ideas emerge to utilize this method in the era of technological and social media transformation. The digital economy can change society and business's economic activities, from what was originally manual to fully automated. This impacts the provision of financial services by startups and Fintech companies. Currently, Fintech practices in Indonesia are very developed, starting from payments, funding, and Robo-advisors. However, in its implementation, Fintech lending (online lending) received special attention because it caused several problems, namely the emergence of illegal fintech. Unreasonable billing processes, issues of personal data protection, and even moral hazards are the focus of the supervision. In Indonesia, a government website channel is available for illegal Fintech complaints, but people tend to use social media to submit their complaints [15]. With the NER concept described in the previous section and the basics of libraries, POS tagging, and named entities, this research becomes the basis for developing ML models in the early identification of platform names on social media. Figure 6 is our proposed Fintech supervision model with social media data and online news that can be used for further research.

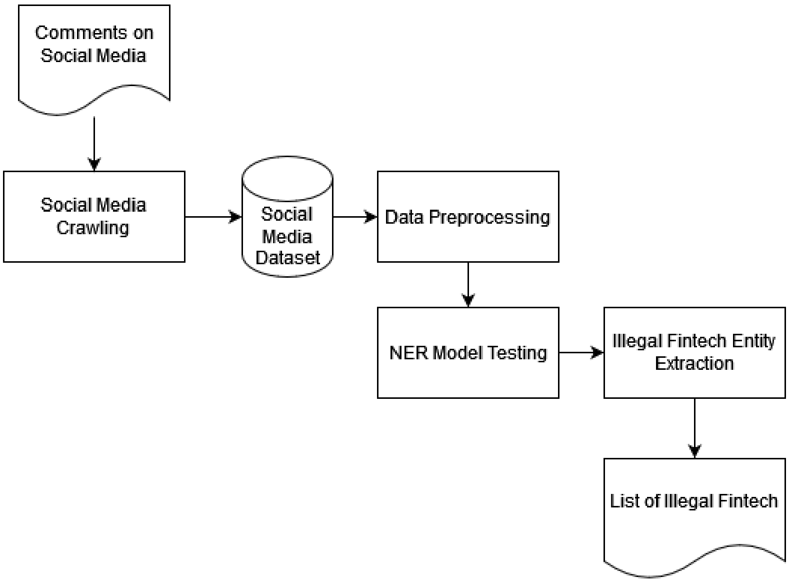


Figure 6. Proposed illegal Fintech supervision model with the NER approach

4. CONCLUSION

In conclusion, this study has provided an overview of research trends in applying the NER method to Indonesian datasets, including extracting news articles, flood monitoring, traffic monitoring, and quotation identification. Other areas of research to consider are data collection, building data sets, cleaning data, and selecting ML algorithm models for NER tasks. The theoretical implication of this research is to obtain the concept of NER and its application. This includes finding researchers and comparing the NER methods used. At the same time, the practical implication is that this NER approach can be used to extract social media comments for platform entity detection. As has been proposed, what is interesting is developing an Illegal Fintech supervision model from social media data.

This survey has limitations because the number of articles reviewed is low due to the lack of research using Indonesian datasets. This is an opportunity for further research in developing models and libraries that use Indonesian datasets. In the field of computer linguistics, the grammatical structure of each country will be a consideration and a challenge that can be explored for future research.

ACKNOWLEDGEMENTS

This research was funded by PUTI Pascasarjana Universitas Indonesia, grant number NKB-104/UN2.RST/HKP.05.00/2022.

REFERENCES

[1] A. Mansouri, L. S. Affendey, and A. Mamat, "Named Entity Recognition Approaches," *Journal of Computer Science*, vol. 8, no. 2, pp. 339–344, 2008.

[2] W. Etaiwi, A. Awajan, and D. Suleiman, "Statistical Arabic Name Entity Recognition Approaches: A Survey," *Procedia Computer Science*, vol. 113, pp. 57–64, 2017, doi: 10.1016/j.procs.2017.08.288.

[3] A. Goyal, V. Gupta, and M. Kumar, "Recent Named Entity Recognition and Classification techniques: A systematic review," *Computer Science Review*, vol. 29, pp. 21–43, 2018, doi: 10.1016/j.cosrev.2018.06.001.

[4] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2022, doi: 10.1109/TKDE.2020.2981314.

[5] M. Humbel, J. Nyhan, A. Vlachidis, K. Sloan, and A. Ortolja-Baird, "Named-entity recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future," *Journal of Documentation*, vol. 77, no. 6, pp. 1223–1247, 2021, doi: 10.1108/JD-02-2021-0032.

[6] J. Santoso, E. I. Setiawan, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Hybrid conditional random fields and k-means for named entity recognition on indonesian news documents," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 3, pp. 233–245, 2020, doi: 10.22266/IJIES2020.0630.22.

[7] P. K. Putra, D. B. Sencaki, G. P. Dinanta, F. Alhasanah, and R. Ramadhan, "Flood Monitoring with Information Extraction Approach from Social Media Data," in *Proceeding - AGERS 2020: IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology: Understanding the Interaction of Land, Ocean and Atmosphere: Disaster Mitigation and Regional Resillience*, 2020, pp. 113–119, doi: 10.1109/AGERS51788.2020.9452770.

[8] M. Riza Alifi and S. H. Supangkat, "Information extraction of traffic condition from social media using bidirectional LSTM-CNN," in *2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018*, 2018, pp. 637–640, doi: 10.1109/ISRITI.2018.8864265.

[9] Y. Syaifudin and A. Nurwidyantoro, "Quotations identification from Indonesian online news using rule-based method," in *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2016, pp. 187–194, doi: 10.1109/ISITIA.2016.7828656.

[10] J. Cheng, J. Liu, X. Xu, D. Xia, L. Liu, and V. S. Sheng, "A review of Chinese named entity recognition," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 6, pp. 2012–2030, 2021, doi: 10.3837/tiis.2021.06.004.

[11] R. R. V. Goulart, V. L. Strube de Lima, and C. C. Xavier, "A systematic review of named entity recognition in biomedical texts," *Journal of the Brazilian Computer Society*, vol. 17, no. 2, pp. 103–116, 2011, doi: 10.1007/s13173-011-0031-9.

[12] G. Simoes, H. Galhardas, and L. Coheur, "Information extraction tasks : a survey," *Simpósio de Informática*, vol. 540, pp. 1–550, 2009.

[13] R. R. Suryono, B. Purwandari, and I. Budi, "Peer to peer (P2P) lending problems and potential solutions: A systematic literature review," *Procedia Computer Science*, vol. 161, pp. 204–214, 2019, doi: 10.1016/j.procs.2019.11.116.

[14] R. R. Suryono, I. Budi, and B. Purwandari, "Challenges and trends of financial technology (Fintech): A systematic literature review," *Information (Switzerland)*, vol. 11, no. 12, pp. 1–20, 2020, doi: 10.3390/info11120590.

[15] R. R. Suryono, I. Budi, and B. Purwandari, "Detection of fintech P2P lending issues in Indonesia," *Heliyon*, vol. 7, no. 4, Apr. 2021, doi: 10.1016/j.heliyon.2021.e06782.

[16] A. Gegenfurtner and C. Ebner, "Webinars in higher education and professional training: A meta-analysis and systematic review of randomized controlled trials," *Educational Research Review*, vol. 28, no. November 2018, p. 100293, 2019, doi: 10.1016/j.edurev.2019.100293.

[17] P. A. Wibowo Putro, E. K. Purwaningsih, D. I. Sensuse, R. R. Suryono, and Kautsarina, "Model and implementation of rice supply chain management: A literature review," *Procedia Computer Science*, vol. 197, no. 2021, pp. 453–460, 2022, doi: 10.1016/j.procs.2021.12.161.

[18] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[19] A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, and F. Ferdian, "InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification," in *4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, ICAICTA 2016*, 2016, doi: 10.1109/ICAICTA.2016.7803103.

[20] A. C. Emcha, Widyawan, and T. B. Adji, "Quotation extraction from Indonesian online news," *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, pp. 408–412, 2019, doi: 10.1109/ICOIACT46704.2019.8938558.

[21] G. B. Herwanto and D. Prasetya Dewantara, "Traffic Condition Information Extraction from Twitter Data," *Proceedings - 2nd 2018 International Conference on Electrical Engineering and Informatics, ICELTICs 2018*, pp. 95–100, 2018, doi: 10.1109/ICELTICS.2018.8548921.

[22] A. Rosyiq, A. R. Hayah, A. N. Hidayanto, M. Naisuty, A. Suhanto, and N. F. Avuning Budi, "Information Extraction from Twitter Using DBpedia Ontology: Indonesia Tourism Places," in *Proceedings - 1st International Conference on Informatics, Multimedia, Cyber and Information System, ICIMCIS 2019*, 2019, pp. 91–96, doi: 10.1109/ICIMCIS48181.2019.8985194.

[23] H. T. Sukmana, J. M. Muslimin, A. F. Firmansyah, and L. K. Oh, "Building the Knowledge Graph for Zakat (KGZ) in Indonesian Language," *ASM Science Journal*, vol. 16, pp. 1–10, 2021, doi: 10.32802/asmscj.2021.758.

[24] D. Kun Indarta and A. Romadhony, "Aspect and Opinion Extraction of Indonesian Lipsticks Product Reviews using Conditional Random Field (CRF)," *KST 2021 - 2021 13th International Conference Knowledge and Smart Technology*, pp. 113–117, 2021, doi: 10.1109/KST51265.2021.9415829.

[25] N. Taufik, A. F. Wicaksono, and M. Adriani, "Named entity recognition on Indonesian microblog messages," in *2016 International Conference on Asian Language Processing (IALP)*, 2016, pp. 358–361, doi: 10.1109/IALP.2016.7876005.

[26] Y. Munarko, M. S. Sutrisno, W. A. I. Mahardika, I. Nuryasin, and Y. Azhar, "Named entity recognition model for Indonesian tweet using CRF classifier," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 403, no. 1, doi: 10.1088/1757-899X/403/1/012067.

[27] I. S. Azarine, M. A. Bijaksana, and I. Asror, "Named entity recognition on Indonesian tweets using hidden markov model," *2019*

7th International Conference on Information and Communication Technology, ICoICT 2019, 2019, doi: 10.1109/ICoICT.2019.8835277.

[28] D. C. Wintaka, M. A. Bijaksana, and I. Asror, "Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF," *Procedia Computer Science*, vol. 157, pp. 221–228, 2019, doi: 10.1016/j.procs.2019.08.161.

[29] A. S. Wibawa and A. Purwarianti, "Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning," *Procedia Computer Science*, vol. 81, pp. 221–228, 2016, doi: 10.1016/j.procs.2016.04.053.

[30] J. Santoso, E. I. Setiawan, C. N. Purwanto, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory," *Expert Systems with Applications*, vol. 176, p. 114856, 2021, doi: 10.1016/j.eswa.2021.114856.

[31] P. R. Togatorop, R. Siagian, Y. Nainggolan, and K. Simanungkalit, "Implementation of ontology-based on Word2Vec and DBSCAN for part-of-speech," *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, pp. 51–56, 2020, doi: 10.1145/3427423.3427431.

[32] W. Gunawan, D. Suhartono, F. Purnomo, and A. Ongko, "Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs," in *Procedia Computer Science*, 2018, vol. 135, pp. 425–432, doi: 10.1016/j.procs.2018.08.193.

[33] R. Leonandya and F. Ikhwantri, "Pretrained language model transfer on neural named entity recognition in Indonesian conversational texts," in *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation, PACLIC 2019*, 2019, pp. 104–113.

[34] N. S. Azzahra, M. O. Ibrohim, J. Fahmi, B. F. Apriyanto, and O. Riandi, "Developing name entity recognition for structured and unstructured text formatting dataset," *2020 5th International Conference on Informatics and Computing, ICIC 2020*, 2020, doi: 10.1109/ICIC50835.2020.9288566.

[35] Y. S. Yohanes, Y. J. Kumar, and N. Z. Zulkarnain, "Understanding quotation extraction and attribution: towards automatic extraction of public figure's statements for journalism in Indonesia," *Global Knowledge, Memory and Communication*, vol. 70, no. 6–7, pp. 655–671, 2020, doi: 10.1108/GKMC-07-2020-0098.

[36] N. A. Ghani, S. Hamid, I. A. Targio Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior*, vol. 101, no. July 2018, pp. 417–428, 2019, doi: 10.1016/j.chb.2018.08.039.

[37] F. Z. Tala, "A study of stemming effects on information retrieval in bahasa indonesia," Universiteit van Amsterdam, 2003.

[38] R. R. Suryono and I. Budi, "P2P Lending Sentiment Analysis in Indonesian Online News," in *Sriwijaya International Conference International Conference of Information Technology and its Applications*, 2019, vol. 172, no. Siconian 2019, pp. 39–44.

[39] B. Aryoyudanta, T. B. Adji, and I. Hidayah, "Semi-supervised learning approach for Indonesian Named Entity Recognition (NER) using co-training algorithm," *Proceeding - 2016 International Seminar on Intelligent Technology and Its Application, ISITIA 2016: Recent Trends in Intelligent Computational Technologies for Sustainable Energy*, pp. 7–12, 2017, doi: 10.1109/ISITIA.2016.7828624.

[40] B. S. Jati, S. T. Widyawan, and S. T. Muhammad Nur Rizal, "Multilingual Named Entity Recognition Model for Indonesian Health Insurance Question Answering System," *2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020*, pp. 180–184, 2020, doi: 10.1109/ICOIACT50329.2020.9332027.

[41] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named Entity Recognition and Relation Extraction: State-of-The-Art," *ACM Computing Surveys*, vol. 54, no. 1, 2021.

BIOGRAPHIES OF AUTHORS

Indra Budi is a lecturer in computer science and information systems at the Faculty of Computer Science, Universitas Indonesia. He is also a head of the information retrieval and natural language processing (IR-NLP) Laboratory (ir.cs.ui.ac.id/new). His research fields include information extraction, text mining, e-commerce, sentiment analysis, and social network analysis. He can be contacted at email: indra@cs.ui.ac.id.

Ryan Randy Suryono is a doctoral student at the Faculty of Computer Science, Universitas Indonesia. He is a member of the IR-NLP Laboratory, E-government and E-business Laboratory in the Faculty of Computer Science, Universitas Indonesia. He is also a lecturer at Universitas Teknokrat Indonesia, Bandar Lampung. His research interests include information systems, financial technology, and text analysis. He can be contacted at email: ryan@teknokrat.ac.id.