JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Indonesian news classification application with named entity recognition approach

Nurchim[1,*], Nurmalitasari[2], Zalizah Awang Long[3]
[1,2]Universitas Duta Bangsa Surakarta
[3]Universiti Kuala Lumpur
[1,2]Jl. Bhayangkara, No. 55, Surakarta 57154, Indonesia
[3]Jl. Sultan Ismail 1016 Bandar Wawasan, Kuala Lumpur 50250, Malaysia
[*]Corresponding email: nurchim@udb.ac.id

Received 29 January 2023, Revised 27 February 2023, Accepted 11 May 2023

Abstract — Nowadays, many netizens search for news via search engines with countless amounts of information, so it is increasingly difficult to determine when the number of news articles that appear changes very quickly and dynamically. Thus, it is necessary to process the extraction of news information to display the core information of the news. Problems arise, especially in Indonesian, which has a structure of various noun phrase entities with shallow parsing or grammatical induction. Yet, it continues to confront the feature representation difficulty when analyzing from a formal lexical grammar, interpreting semantically, and extracting information. Named entity recognition (NER) has the opportunity to overcome this because it can extract news entities in depth, starting from proper nouns in text documents containing information search, machine translation, answering questions, and automatic summarization. This study offered the NER approach for constructing an application of classification Indonesian news material. The NER approach is distinct from others, which may be rule-based, dictionary-matching-based, machine learning-based, or a combination of them. Moreover, NER may be extended to recognize distinct entities based on their requirements. This study uses Design-Based Research whose process includes (1) pre-implementation, (2) design, (3) implementation and revision, and finally, (4) reflection and evaluation. This application was developed on the platform Python, streamlet, BeautifulSoup, news, and spacy library. The results of application accuracy testing have an F1-Score value of 89.69 % for all entities consisting of place, figure, day, date, and organization.

Keywords – classification, named entity recognition, news, web

I. Introduction

Digital news portals are internet users' most important news sources [1]. This is indicated by the interest in reading news on traditional web media or commercial portals, which is higher than new news media such as the Internet BBS [2]. Now, people prefer to search for news through search engines where the results can display a combination of several news portals according to the news keywords [3]. However, searching through search engines that can produce countless combinations of information sometimes takes much time to find the expected information [4]. Moreover, relevant news information will be even more challenging to determine when changes in the number of news articles that appear are swift and dynamic [5]. Due to the frequent millisecond updates to the storage, real-time data is enormous [6].

An information extraction process is needed on the news so that it can display the core information of the news quickly. The problem is that the structure of Indonesian has various noun phrase entities with shallow parsing or grammatical induction [7], so it takes action to classify text documents into correct classes based on techniques in machine learning [8]. To solve this issue, use named entity recognition (NER) to extract the core text ideas in ontology [9]. NER can classify proper nouns in text documents widely used in information retrieval, machine translation, answering questions, and automatic summarization [10]. The NER work process utilizes bags of words (BoW) that handle locations, people, organizations, or institutions [11]. In addition, NER's performance can extract each news entity in depth to support the ability to detect fake news [12].

This study proposed the NER method for developing an application for classifying Indonesian news information. NER is one method found in information extraction to detect and classify certain entities in a text [13]. NER extracts a person's name, organization, and location from a document's text. However, NER may be expanded to identify different things based on the needs. Research on NER has been carried out in various languages other than English and Indonesian, such as Chinese [14], [15], Arabic [16], Indian [17], Turkish [18], and Khmer [19]. Previous research also mentions using NER as a text analysis resource in Uyghurs [20]. Many aspects of Natural Language Processing necessitate the use of NER. Babych and Hartley [21] stated that NER could improve Machine Translation performance. Various research topics in Natural Language Processing, such as Quote attribution in Elson and McKeown [22], implement NER to detect the quote's author. Another function, NER, is used for the reference approach to detect text document entities [23].

II. Research Method

Text Classification is a classification method that involves assigning input from a text document to a set of pre-defined classes using a machine learning algorithm [24]. Automatic text classification utilizing machine learning classifiers such as naive Bayes, support vector machines, neural networks, and decision trees has improved performance in recent years [25]. However, it still faces the feature representation challenge in analyzing from a lexical a formal grammar, interpreting semantically, and extracting information [26]. The NER technique differs from the others, which may be rule-based, dictionary-matching-based, machine learning-based, or a mix [26]. Moreover, NER may be expanded to identify different entities based on their needs [9].

The application of the NER algorithm for the classification of Indonesian language news is carried out using the design-based research (DBR) approach. The DBR approach is iterative in designing, implementing, evaluating, and improving problem-specific interventions by considering limited resources and technology [27]. Overall the stages of the DBR approach include (1) pre-implementation, (2) design, (3) implementation and revision, and finally, (4) reflection and evaluation [28]. The complete stages of the research will be shown in Fig. 1.

Finding news data includes looking for Indonesian news themes in the Google News library and scraping news articles. A website is used for packaging the application interface display design. Python and Streamlit are the platforms used for this experimentation. After that, put the model's accuracy, recall, and F1-Score to the test.

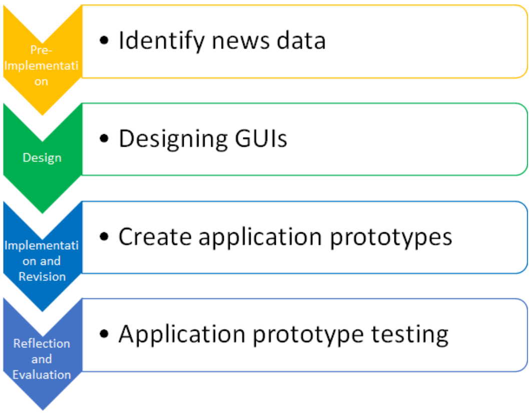

Pre-Implementation
• Identify news data
Design
• Designing GUIs
Implementation and Revision
• Create application prototypes
Reflection and Evaluation
• Application prototype testing

Fig. 1. Application development process.

III. Result

This section discusses identify news topics, scraping news content, classification using NER, and evaluation of custom NER model.

A. Identify News Topics

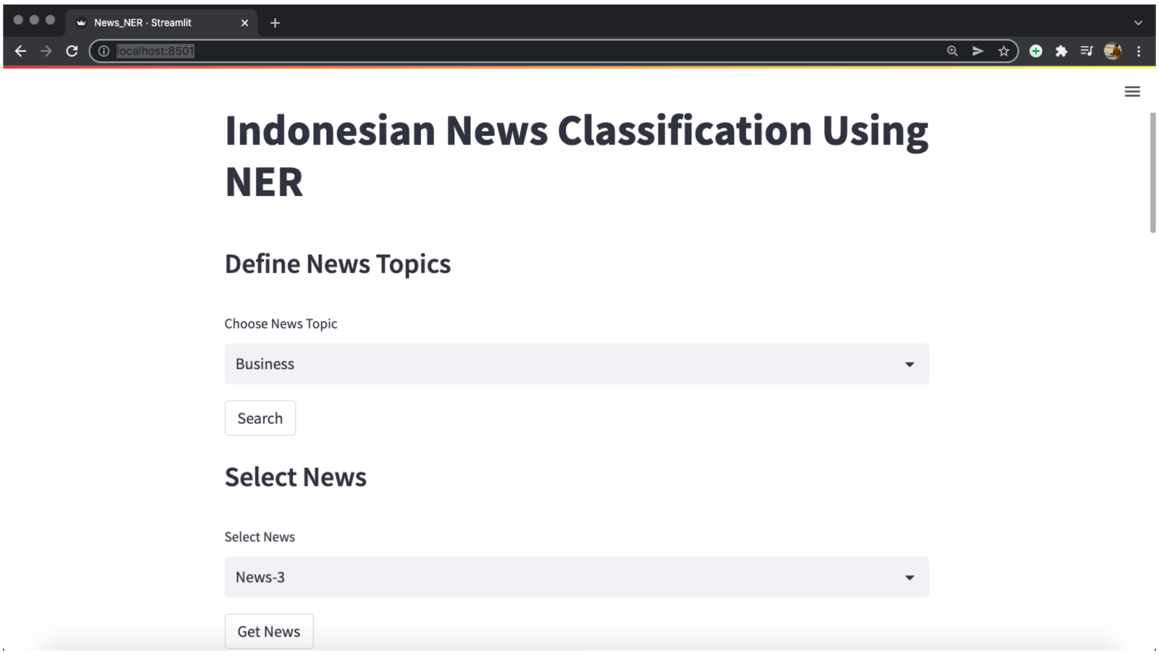
The Google News library, or news library in Python, is used while looking for news subjects. The features in this collection enable you to identify the most popular news headlines within a specific time frame and in a specified language [29]. In this research, only news in the Indonesian language is covered. The findings included a wide range of subjects, including world, nation, technology, entertainment, sports, science, and health shown in Fig. 2.

B. Scraping News Content

After choosing a news topic, the top 3 stories currently trending in Indonesia will be shown along with their respective news titles and URLs (shown in Fig. 3). Additionally, web scraping is used to get news data content, as in Fig. 4. Data collection from potentially limited access platforms will be simpler with web scraping [30]. Among these is BeautifulSoup, a Python package created by Leonard Richardson and many other programmers that enables the extraction of structured data from web pages by parsing XML and HTML [31].

C. Classification using NER

The named entity recognition (NER) technique is used to classify news articles. NER is a natural language processing activity that extracts specified entity words or phrases from unstructured text data and categorizes them (entity type, time type, and number type) [32]. This task typically targets well-known things such as People, Organizations, Dates, *etc.* [33]. This study employs a unique NER model created with the Spacy framework version 3. Spacy allows it to operate on words and subwords rather than tokens. The research entities include place, figure, day, date, and organization, as in Fig. 5.



Indonesian News Classification Using NER

Define News Topics

Choose News Topic

Business

Search

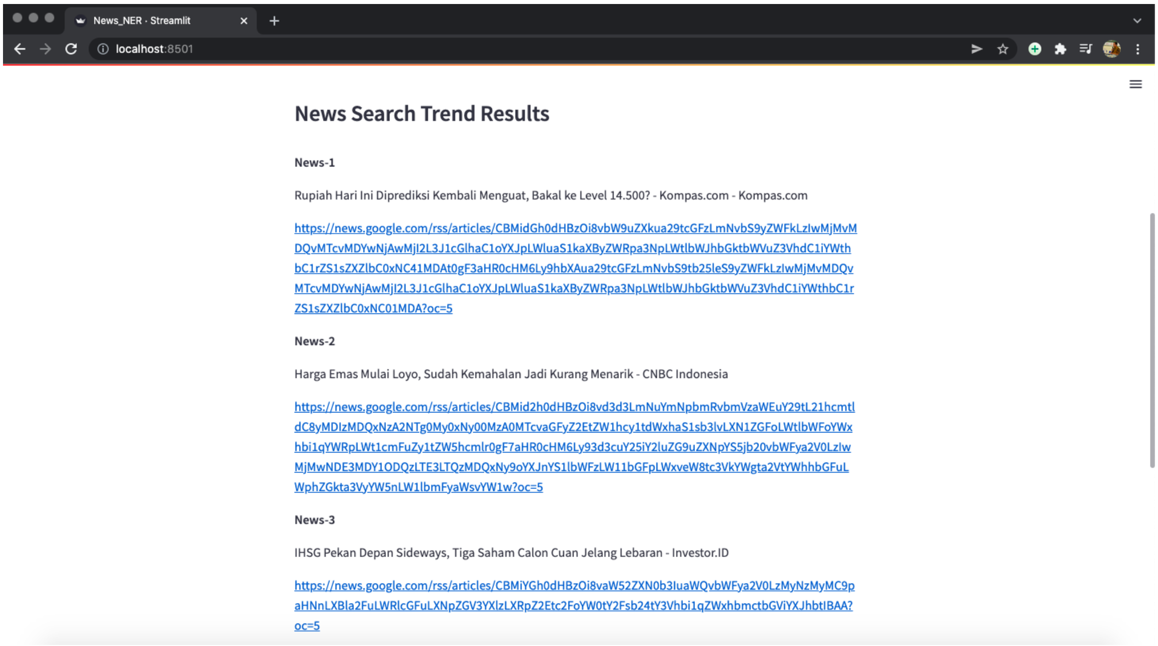
Select News

Select News

News-3

Get News

Fig. 2. Identify new topics.



News Search Trend Results

News-1

Rupiah Hari Ini Diprediksi Kembali Menguat, Bakal ke Level 14.500? - Kompas.com - Kompas.com

https://news.google.com/rss/articles/CBMidGh0dHBzOi8vbW9uZXkua29tcGFzLmNvbS9yZWFkLzIwMjMvMDQvMTcvMDYwNjAwMjI2L3J1cGlhaC1oYXJpLWluaS1kaXByZWRpa3NpLWtlbWJhbGktbWVuZ3VhdC1iYWthbC1rZS1sZXZlbC0xNC41MDAt0gF3aHR0cHM6Ly9hbXAua29tcGFzLmNvbS9tb25leS9yZWFkLzIwMjMvMDQvMTcvMDYwNjAwMjI2L3J1cGlhaC1oYXJpLWluaS1kaXByZWRpa3NpLWtlbWJhbGktbWVuZ3VhdC1iYWthbC1rZS1sZXZlbC0xNC01MDA?oc=5

News-2

Harga Emas Mulai Loyo, Sudah Kemahalan Jadi Kurang Menarik - CNBC Indonesia

https://news.google.com/rss/articles/CBMid2h0dHBzOi8vd3d3LmNuYmNpbmRvbmVzaWEuY29tL21hcmtldC8yMDIzMDQxNzA2NTg0My0xNy00MzA0MTcvaGFyZ2EtZW1hcy1tdWxhaS1sb3lvLXN1ZGFoLWtlbWFoYWxhbi1qYWRpLWt1cmFuZy1tZW5hcmlr0gF7aHR0cHM6Ly93d3cuY25iY2luZG9uZXNpYS5jb20vbWFya2V0LzIwMjMwNDE3MDY1ODQzLTE3LTQzMDQxNy9oYXJnYS1lbWFzLW11bGFpLWxveW8tc3VkYWgta2VtYWhhbGFuLWphZGkta3VyYW5nLW1lbmFyaWsvYW1w?oc=5

News-3

IHSG Pekan Depan Sideways, Tiga Saham Calon Cuan Jelang Lebaran - Investor.ID

https://news.google.com/rss/articles/CBMiYGh0dHBzOi8vaW52ZXN0b3IuaWQvbWFya2V0LzMyNzMyMC9paHNnLXBla2FuLWRlcGFuLXNpZGV3YXlzLXRpZ2Etc2FoYW0tY2Fsb24tY3Vhbi1qZWxhbmctbGViYXJhbtIBAA?oc=5

Fig. 3. Choosing news topics.

D. Evaluation of Custom NER model

The NER model was tested using the SpaCy library, and its performance was evaluated by calculating precision, recall, and F1-Score. Table 1 is the assessment metrics findings for all entities.

Table 1. Evaluation Metric

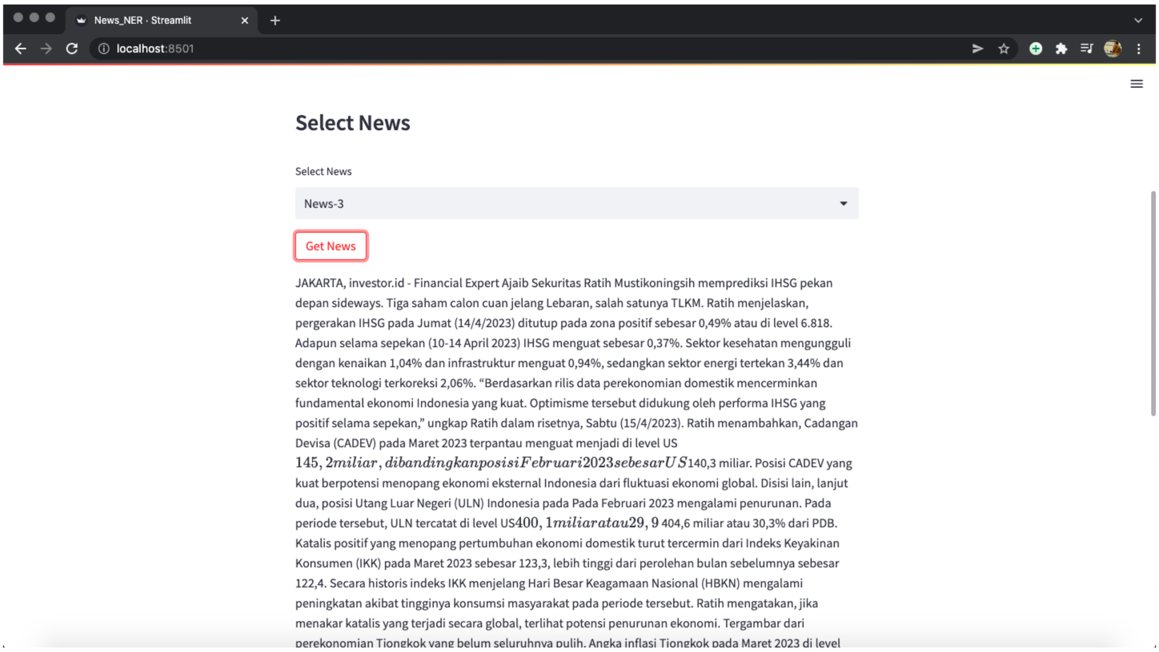
Evaluation Metric	
Precision (%)	88.21
Recall (%)	85.71
F1-Score (%)	86.96

IV. Discussion

This study's findings are likely to serve as the foundation for subsequent research on the usage of NER in Indonesian. The F1-score results obtained an accuracy of 86.96 %, tending to be better than previous studies by AI-Ash *et al.* [34] at 76 % and Wintaka *et al.* [35] at 84.11 %. This is based on the fact that Indonesian has an extensive vocabulary, and the number of word categorization class entities will be significant. As a natural language processing package, Spacy enables the creation of custom NER models with entities that meet certain specifications. Of course, more datasets and word-labeling annotations are required for custom NER model development.

V. Conclusion

This Indonesian news classification application was developed using the Python language, streamlet, Beau-



Select News

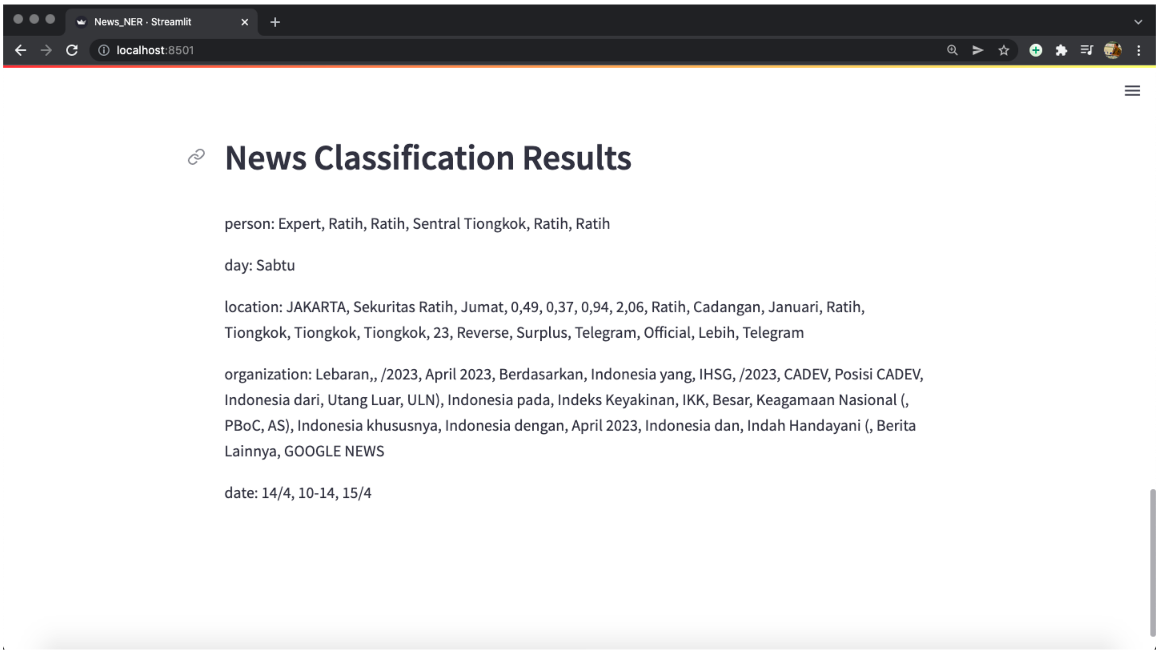
Select News

News-3

Get News

JAKARTA, investor.id - Financial Expert Ajaib Sekuritas Ratih Mustikoningsih memprediksi IHSG pekan depan sideways. Tiga saham calon cuan jelang Lebaran, salah satunya TLKM. Ratih menjelaskan, pergerakan IHSG pada Jumat (14/4/2023) ditutup pada zona positif sebesar 0,49% atau di level 6.818. Adapun selama sepekan (10-14 April 2023) IHSG menguat sebesar 0,37%. Sektor kesehatan mengungguli dengan kenaikan 1,04% dan infrastruktur menguat 0,94%, sedangkan sektor energi tertekan 3,44% dan sektor teknologi terkoreksi 2,06%. “Berdasarkan rilis data perekonomian domestik mencerminkan fundamental ekonomi Indonesia yang kuat. Optimisme tersebut didukung oleh performa IHSG yang positif selama sepekan,” ungkap Ratih dalam risetnya, Sabtu (15/4/2023). Ratih menambahkan, Cadangan Devisa (CADEV) pada Maret 2023 terpantau menguat menjadi di level US
$145, 2miliar, dibandingkanposisiFebruari2023sebesarUS$140,3 miliar. Posisi CADEV yang kuat berpotensi menopang ekonomi eksternal Indonesia dari fluktuasi ekonomi global. Disisi lain, lanjut dua, posisi Utang Luar Negeri (ULN) Indonesia pada Pada Februari 2023 mengalami penurunan. Pada periode tersebut, ULN tercatat di level US$400, 1miliaratau29, 9$ 404,6 miliar atau 30,3% dari PDB. Katalis positif yang menopang pertumbuhan ekonomi domestik turut tercermin dari Indeks Keyakinan Konsumen (IKK) pada Maret 2023 sebesar 123,3, lebih tinggi dari perolehan bulan sebelumnya sebesar 122,4. Secara historis indeks IKK menjelang Hari Besar Keagamaan Nasional (HBKN) mengalami peningkatan akibat tingginya konsumsi masyarakat pada periode tersebut. Ratih mengatakan, jika menakar katalis yang terjadi secara global, terlihat potensi penurunan ekonomi. Tergambar dari perekonomian Tiongkok yang belum seluruhnya pulih. Angka inflasi Tiongkok pada Maret 2023 di level

Fig. 4. Scraping news content.



News Classification Results

person: Expert, Ratih, Ratih, Sentral Tiongkok, Ratih, Ratih

day: Sabtu

location: JAKARTA, Sekuritas Ratih, Jumat, 0,49, 0,37, 0,94, 2,06, Ratih, Cadangan, Januari, Ratih, Tiongkok, Tiongkok, Tiongkok, 23, Reverse, Surplus, Telegram, Official, Lebih, Telegram

organization: Lebaran,, /2023, April 2023, Berdasarkan, Indonesia yang, IHSG, /2023, CADEV, Posisi CADEV, Indonesia dari, Utang Luar, ULN), Indonesia pada, Indeks Keyakinan, IKK, Besar, Keagamaan Nasional (, PBoC, AS), Indonesia khususnya, Indonesia dengan, April 2023, Indonesia dan, Indah Handayani (, Berita Lainnya, GOOGLE NEWS

date: 14/4, 10-14, 15/4

Fig. 5. Classification result of news topics with NER.

tifulSoup, news, and Spacy library. Streamlit is used to support making GUI applications in web form. BeautifulSoup is a library for scraping trending news content data through the news library. The classification model was custom developed using the Spacy v3 library consisting of place, figure, day, date, and organization entities. The evaluation value of the model has an F1-Score of 86.96 % for all entities. For the model accuracy value to be better, further research can add feature selection to the data preprocessing process.

ACKNOWLEDGMENT

This research was sponsored through the internal research program of Universitas Duta Bangsa Surakarta in 2022.

REFERENCES

[1] N. B. C. E. Jamil, I. B. Ishak, F. Sidi, L. S. Affendey, and A. Mamat, “A systematic review on the profiling of digital news portal for big data veracity,” in *Procedia Comput. Sci.*, vol. 72, pp. 390–397, 2015, doi: 10.1016/j.procs.2015.12.154.

[2] C. Hu, Y. Li, Y. Wang, and L. Wu, “Analysis of hot news based on big data,” in *Proc. - 17th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2018*, pp. 678–681, 2018, doi: 10.1109/ICIS.2018.8466427.

[3] M. U. Bokhari and M. K. Adhami, “How well they retrieve fresh news items: News search engine perspective,” *Perspect. Sci.*, vol. 8, pp. 469–471, 2016, doi: 10.1016/j.pisc.2016.06.002.

[4] M. Asenova and C. Chrysoulas, “Personalized microservice recommendation system for online news,” in *Procedia Comput. Sci.*, vol. 160, pp. 610–615, 2019, doi: 10.1016/j.procs.2019.11.039.

[5] C. Feng, M. Khan, A. U. Rahman, and A. Ahmad, “News recommendation systems-accomplishments, challenges future directions,” *IEEE Access*, vol. 8, pp. 16702–16725, 2020, doi: 10.1109/ACCESS.2020.2967792.

[6] S. Farhana, A. Lajis, Z. A. Long, and H. Nasir, “Impact of big data congestion in IT: An adaptive knowledge-based Bayesian network,” *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 2031–2036, 2020, doi: 10.11591/ijece.v10i2.pp2031-2036.

[7] A. T. Hermawan, Gunawan, and J. Santoso, “Natural language grammar induction of Indonesian language corpora using genetic algorithm,” in *Proc. - 2011 Int. Conf. Asian Lang. Process. IALP 2011*, pp. 15–18, 2011, doi: 10.1109/IALP.2011.58.

[8] N. R. Sulaiman and M. Md Siraj, “South China Sea conflicts classification using named entity recognition (NER) and part-of-speech (POS) tagging,” *Int. J. Innov. Comput.*, vol. 10, no. 1, pp. 41–48, 2020, doi: 10.11113/ijic.v10n1.255.

[9] 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 Syst. Appl.*, vol. 176, no. February 2020, p. 114856, 2021, doi: 10.1016/j.eswa.2021.114856.

[10] S. Vychegzhanin and E. Kotelnikov, “Comparison of named entity recognition tools applied to news articles,” in *Proc. - 2019 Ivannikov Ispras Open Conf. ISPRAS 2019*, no. February, pp. 72–77, 2019, doi: 10.1109/ISPRAS47671.2019.00017.

[11] K. Englmeier, “The role of text mining in mitigating the threats from fake news and misinformation in times of corona,” in *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 149–156, 2021, doi: 10.1016/j.procs.2021.01.115.

[12] W. Shishah, “JointBert for detecting Arabic fake news,” *IEEE Access*, vol. 10, no. June, pp. 1–1, 2022, doi: 10.1109/access.2022.3185083.

[13] K. Adnan and R. Akbar, ”An analytical study of information extraction from unstructured and multidimensional big data,” *Journal of Big Data*, vol. 6, no. 1, 2019.

[14] J. Gao, A. Wu, M. Li, and C. N. Huang, “Chinese word segmentation and named entity recognition: A pragmatic approach,” *Comput. Linguist.*, vol. 31, no. 4, pp. 531–574, 2005, doi: 10.1162/089120105775299177.

[15] G. Fu, “Chinese named entity recognition using a morpheme-based chunking tagger,” in *2009 Int. Conf. Asian Lang. Process. Recent Adv. Asian Lang. Process. IALP 2009*, pp. 289–292, 2009, doi: 10.1109/IALP.2009.68.

[16] K. Shaalan and H. Raza, “Arabic named entity recognition from diverse text types,” in *Advances in Natural Language Processing. GoTAL 2008. Lecture Notes in Computer Science, vol 5221.*, pp. 440–451, 2008, doi: 10.1115/OMAE2011-50009.

[17] A. Ekbal and S. Saha, “A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies,” *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14760–14772, 2011, doi: 10.1016/j.eswa.2011.05.004.

[18] D. Küük and A. Yazici, “A hybrid named entity recognizer for Turkish,” *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2733–2742, 2012, doi: 10.1016/j.eswa.2011.08.131.

[19] H. Pan, X. Yan, Z. Yu, and J. Guo, “A Khmer named entity recognition method by fusing language characteristics,” in *26th Chinese Control Decis. Conf. CCDC 2014*, pp. 4003–4007, 2014, doi: 10.1109/CCDC.2014.6852881.

[20] J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell, “Neural cross-lingual named entity recognition with minimal resources,” *arXiv Prepr. arXiv1808.09861*, pp. 369–379, 2018, doi: 10.18653/v1/d18-1034.

[21] A. Hartley and B. Babych, “Improving machine translation quality with automatic named entity eecognition,” in *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, 2003, p. 8.

[22] D. K. Elson and K. R. McKeown, “Automatic attribution of quoted speech in literary narrative,” in *Proc. Natl. Conf. Artif. Intell.*, vol. 2, no. 2009, pp. 1013–1019, 2010.

[23] S. Žitnik, L. Šubelj, and M. Bajec, “SkipCor: Skip-mention coreference resolution using linear-chain conditional random fields,” *PLoS One*, vol. 9, no. 6, 2014, doi: 10.1371/journal.pone.0100101.

[24] X. Luo, “Efficient English text classification using selected Machine Learning Techniques,” *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.

[25] M. N. Asim, M. U. G. Khan, M. I. Malik, A. Dengel, and S. Ahmed, “A robust hybrid approach for textual document classification,” in *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 1390–1396, 2019, doi: 10.1109/ICDAR.2019.00224.

[26] L. Zhang, X. Nie, M. Zhang, M. Gu, V. Geissen, C. J. Ritsema, D. Niu, and H. Zhang, “Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: A Deep learning approach,” *Front. Plant Sci.*, vol. 13, no. November, pp. 1–16, 2022, doi: 10.3389/fpls.2022.1053449.

[27] M. D. Wolcott, N. G. Lobczowski, K. Lyons, and J. E. Mclaughlin, “Design-based research: Connecting theory and practice in pharmacy educational intervention research,” *Curr. Pharm. Teach. Learn.*, vol. 11, no. 3, pp. 309–318, 2019, doi: 10.1016/j.cptl.2018.12.002.

[28] C. J. Cavallaro and S. F. Sembiante, “Facilitating culturally sustaining, functional literacy practices in a middle school ESOL reading program: A design-based research study,” *Lang. Educ.*, vol. 35, no. 2, pp. 160–179, 2020, doi: 10.1080/09500782.2020.1775244.

[29] B. Lyamin, E. Konnikov, A. Chernikova, and K. Shadrov, “Information logistics as a driver for the development of the rocket and space industry,” *Transp. Res. Procedia*, vol. 63, pp. 887–895, 2022, doi: 10.1016/j.trpro.2022.06.086.

[30] C. Fink, T. Toivonen, R. A. Correia, and E. Di Minin, “Mapping the online songbird trade in Indonesia,” *Appl. Geogr.*, vol. 134, no. June, p. 102505, 2021, doi: 10.1016/j.apgeog.2021.102505.

[31] M. A. Khder, ”Web scraping or web crawling: State of the art, techniques, approaches, and application,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[32] J. Sun, Y. Liu, J. Cui, and H. He, “Deep learning-based methods for natural hazard named entity recognition,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-08667-2.

[33] N. Jofche, K. Mishev, R. Stojanov, M. Jovanovik, E. Zdravevski, and D. Trajanov, “Named entity recognition and knowledge extraction from pharmaceutical texts using transfer learning,” in *Procedia Comput. Sci.*, vol. 203, no. 2019, pp. 721–726, 2022, doi: 10.1016/j.procs.2022.07.107.

[34] H. S. Al-Ash, I. Fanany, and A. Bustamam, “Indonesian protected health information removal using named entity recognition,” in *Proc. 2019 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2019*, pp. 258–263, 2019, doi: 10.1109/ICTS.2019.8850995.

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