Future Computer Science Journal (FCSJ)
Vol. 1, No. 1, June 2023: xxx-xxx
http://asasijournal.id/index.php/fcsj
http://doi.org/10.xxxxx/fcsj.xxxx.xxx

BERT with Entity Recognition for Classifying Local and Import Products to Support Local MSMEs in the Global Market

Dea Annisayanti Putri[1*], Dwi Retnoningrum[2], Indra Budi[3], Aris Budi Santoso[4], Prabu Kresna Putra[5]
[1]Department of Information Technology Magister, Faculty of Engineering, Universitas Indonesia, Indonesia
[2]Department of Clinical Pathology, Faculty of Medicine, Universitas Diponegoro, Indonesia
[3,4,5]Master of Information Technology, Faculty of Computer Science, Universitas Indonesia, Indonesia

Abstract
The digital era and the COVID-19 pandemic have encouraged a shift toward online shopping. E-commerce has expanded the market by giving buyers access to a global market, resulting in increased cross-border transactions. As a direct challenge for local micro, small, and medium enterprises (MSMEs), the government has made regulations and campaigns to prioritize local products. This study presents a machine-learning model for classifying local products. We fine-tuned pre-trained Bidirectional Encoder Representations from Transformers (BERT) trained on Bahasa on product titles and used Entity Recognition to extract brand names as additional features. The model performs nearly perfectly with an accuracy of 97.79%. Adding brand name information provides an excellent signal to classify local products, indicated by the improvement after adding the brand name as a feature. With this implementation, the government and all e-commerce in Indonesia can collaborate to support the government campaign to encourage the competitiveness of MSMEs by prioritizing local products, reducing the number of imported products, and evaluating government programs specifically aimed at accelerating Indonesian MSMEs.

Keywords:
MSMEs; E-Commerce; Machine Learning; Text Classification; BERT;

Article History:
Received: June 14, 2023
Revised: June 20, 2023
Accepted: June 20, 2023
Published: June 26, 2023

Corresponding Author:
Dea Annisayanti Putri
Information Technology Magister Department, Universitas Indonesia, Indonesia
Email: deaannisayanti@gmail.com

INTRODUCTION

In the modern age, social activities have migrated to social media, while business transactions have transitioned to digital platforms through e-. According to Kinda [1], e-commerce is an online medium for conducting business transactions involving the buying and selling of products (or services), including their logistics processes. With the advancements in technology and the ubiquity of the internet, e-commerce offers individuals greater accessibility and convenience to more comprehensive and cost-effective products than traditional brick-and-mortar stores, as noted by Alagoz and Hekimoglu [2] and Hartono et al. [3].

One external factor that has contributed to the upsurge of e-commerce is the COVID-19 pandemic, which has had far-reaching implications for many sectors, as discussed by Bai [4], Erokhin and Gao [5], and Tudor and Sova [6]. The pandemic has hindered the physical economy by shuttering businesses and bolstering the shift towards digitalization by increasing the use of digital channels, especially e-commerce, as stated by Fletcher and Griffiths [7] and Priyono et al. [8]. Thus, digital transformation has accelerated the e-commerce era while the global economy is undergoing a downturn. This unprecedented phenomenon is an anomaly in history. The e-commerce sector plays a critical role in society, as it enables consumers to obtain goods in situations that preclude physical contact, thereby promoting public health, as per Lone et al. [9]. E-commerce has witnessed significant growth in several countries, and Indonesia is no exception. As reported by Situmorang [10], there has been a remarkable increase in e-commerce usage in Indonesia, rising from 54% in 2019 to 91% in 2020.

As previously noted by Kinda [1], e-commerce has provided opportunities that traditional shopping has only offered to a select group of sellers and buyers. With the existence of e-commerce, the market has expanded, granting buyers access to a broader array of products and services from both local and foreign sellers. However, this opportunity introduces a new challenge for local businesses, particularly small and

medium-sized local product manufacturers. They are directly challenged by the growth of the digital economy and the subsequent surge in cross-border transactions through e-commerce. Imported products often possess superior quality and sell well on e-commerce platforms, leading to increased competition for local businesses. Even in Europe, Cbcommerce [11] reports that 71% of e-shoppers purchase products from retailers and marketplaces abroad, further exacerbating the competition faced by local businesses.

As a form of support for the competitiveness of local MSMEs, the government makes regulations and campaigns that prioritize local products so that MSMEs and the Indonesian economy will prosper in their own country. The implementation of this support is in the form of implementing the "Gerakan Nasional Bangga Buatan Indonesia" (national movement to be proud of products made in Indonesia), which is strengthened by Presidential Decree No. 15 of 2021 [12]. In 2022, the government strengthened support for the local MSMEs by encouraging the use of local products for government procurement through Presidential Instruction number 2 of 2022 [13].

Government regulation is expected to be supported by various parties, the government, intermediaries such as e-commerce, and buyers. For e-commerce, this campaign requires them to create a particular campaign on their website to promote local products, as shown in Figure 1. Not only the private sector, as Lembaga Kebijakan Pengadaan Barang/Jasa Pemerintah (LKPP) or national public procurement agency reported in [14], that they bans 20,000 imported products from the official website of government procurement, e-katalog.

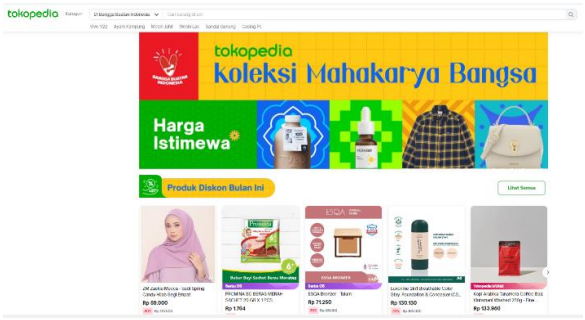

Figure 1. Indonesia E-commerce Page for Local Products

Natural Language Processing (NLP) has been widely applied for various implementations in the e-commerce domain, utilizing text information on products such as product titles, descriptions, or reviews. Examples of its application include product category classification with word embedding and Gated Recurrent Unit (GRU) by Kim et al. in [15], and using BERT by Zahera and Sherif in [16]. Product information on e-commerce can also be used to classify a product, whether it is a local or imported product. With this additional information, the government (especially the procurement department) and all e-commerce in Indonesia can collaborate to support the government campaign to encourage the competitiveness of MSMEs. The action can be by prioritizing local products or reducing the number of imported products. In addition, with this information on every product circulating in the community, the government can continually evaluate programs specifically aimed at accelerating Indonesian MSMEs.

By raising this opportunity, we formulate two questions: "What is the best model to classify local products?" and "What is the impact of brand name on the local product classification?".

LITERATURE REVIEW

Pre-processing

As discussed by Lunando and Purwarianti [17] and Robert and Gosselin [18], pre-processing is a critical step that aims to minimize irrelevant information in the text. In e-commerce, sellers often add extraneous words and characters to their product titles, such as using numbers to replace letters, repeating vowel characters, and using non-standard words, as noted by Mao et al. [19].

Case folding, which involves converting all the characters in the text to lowercase, is one pre-processing technique. Tokenization is another technique that involves dividing the text into specific parts, such as punctuation marks and words, as explained by Klampanos [20] and Nurdeni et al. [21]. Additionally, removing emojis and punctuation marks is another commonly used method for cleaning up text since these elements may not contain useful information for some tasks.

Bidirectional Encoder Representation from Transformers (BERT)

The Transformer architecture of neural networks, which is based on self-attention processes, has made significant progress in natural language processing, as noted by Vaswani et al. [22]. Building upon the benefits of the Transformer architecture, a novel model called Bidirectional Encoder Representation of Transformers (BERT) was created. BERT leverages word extension and contextualization to overcome the limitations of Recurrent Neural Network (RNN) and Long short-term memory (LSTM). The results of experiments conducted by Devlin et al. [23] show that BERT significantly enhances text classification performance.

One of the key benefits of using BERT-based models is that they require only a manageable amount of text data to train the models. However, given BERT's extensive pre-training, users must fine-tune it using sufficient training data. Moreover, the ease of access to the BERT repository and the BERT framework, such as the Huggingface framework, has accelerated model development, including BERT in Bahasa.

IndoBERT is a new pre-trained language model for Bahasa, as reported by Koto et al. [24]. The model was evaluated using the IndoLEM dataset and was trained on a dataset comprising more than 220 million words collected from three primary sources: Indonesian Wikipedia, news articles, and Indonesian online corpus.

Feature Extraction

As highlighted by Li et al. [25], a crucial aspect of text processing is converting textual values into numeric values, such as vectors or other representations. The determination of features for machine learning approaches is adjusted to the categorization of text. Feature extraction methods can reduce the original features by removing irrelevant features, thus increasing accuracy and reducing machine learning processing time.

In addition to its use as a classification model, BERT is also famous for generating text representation. BERT can generate vectors from sentences by adding pooling on top of the model or using other processes like Sentence-BERT, as proposed by Reimers and Gurevych [26]. Sentence-BERT (SBERT) is a modification of the previously trained BERT network that uses Siamese and triplet network structures to obtain semantically meaningful.

Text Classification

Text classification is a process of extracting insights from textual information and organizing the information, as discussed by Shah et al. [27]. There are several algorithms in machine learning used for text classification, such as Decision Trees (DT), Support Vector Machines (SVM), and Neural Networks.

DT is a powerful method widely used in machine learning, image processing, and pattern recognition, as discussed by Damanik et al. [28]. It is a sequential model that efficiently and consistently combines a series of basic tests, comparing numerical properties to threshold values for each test. SVMs are another popular technique for classifying documents using discriminative classifiers. This technique can be used in all areas of data mining, such as text, images, and videos, as highlighted by Kowsari et al. [29]. SVM was initially used for binary classification tasks, but it has recently been applied to multi-class problems.

Neural networks are designed to learn through multiple connections in layers. Each layer only gets connections from the previous layer and only provides connections to the next layer in the remote part.

As an end-to-end classifier model, BERT undergoes two processes, as shown in Figure 2. First, feature transformations are done by applying sub-word tokenization on the text using a pre-trained tokenizer of BERT. The second process is to obtain the label, which will have its token id from each word and feed it to transformer layers. Every input example begins with a special token represented by CLS, where C is the final hidden vector of the special CLS token, which contains predicted class labels.

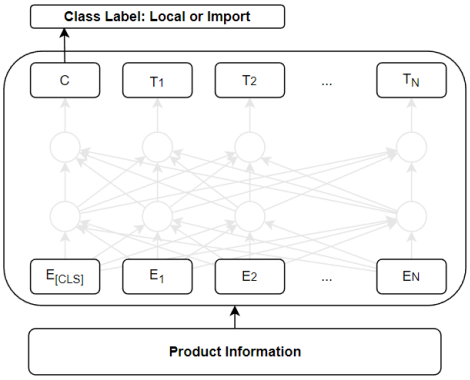

Figure 2. BERT as End-to-end Classifier

METHOD

Data Collection

The process of gathering product information from e-commerce websites involves scraping, which is an information extraction method utilizing Hypertext Transfer Protocol (HTTP). In this study, we employed this method to collect 4,800 product titles from three prominent Indonesian e-commerce platforms: Tokopedia, Shopee, and Ocistok. The products were classified into two categories based on their label: local and imported. Local products were scraped from e-commerce pages dedicated to showcasing Indonesia's locally-produced products, while imported products were gathered from Indonesia's e-commerce platforms that exclusively sell imported products. The collected products were well-distributed across eight categories: fashion, mom & baby, beauty, food & drinks, sport, automotive, household, and others (which contained uncategorized products). The distribution of each category and label can be observed in Table 1.

Table 1. Collected Data Distribution

Category	Local	Import
Fashion	818	1.224
Mom & Baby	246	26
Beauty	397	1
Food and Drink	351	0
Sport	28	0
Others	246	753
Automotives	51	0
Households	390	281
Total	2.527	2.285

Scraping as a data collection method has certain limitations, primarily concerning the number of iterations required to access a website before it is flagged as suspicious bot activity. To mitigate this risk, the scraping process in this study was conducted in bulk, with a focus solely on products listed on the Search Result Page (SRP) to avoid detection. However, this approach did have an impact on the number of features obtained, as only the product title was used as a source of information. Consequently, other information, such as product description or image, was not included in our data collection process.

Brand Detection

Given the limitations concerning the number of features in our data collection process, we conducted additional activities to enhance our dataset, specifically the extraction of brands from product titles. For this purpose, we used product and brand data from eBay, which was translated into Indonesian. Named Entity

Recognition (NER), a widely used technique in natural language processing, was implemented using Python's SpaCy library [30] to detect brands.

To increase precision and reduce noise, we used frequency-based post-processing methods on the predicted brand data. Firstly, all the predicted brands were compiled and sorted by frequency. Next, we manually validated all the brands with a minimum of 20 occurrences, and only those included in the validated list were defined as valid and used as features. Through this extraction and cleaning process, we extracted 1,505 brands from the product titles, which comprised only 31% of all products. The low coverage of brand extraction is due to two reasons: some products do not have a brand, and the NER model's incorrect prediction. The results of our brand extraction process can be found in Table 2.

Table 2. Brand Detection Results

Product Title	**Predicted Brand**
ZM Zaskia Mecca - Yumi Maroon Blouse Romansa Khatulistiwa - bungaasoka - XL	ZM Zaskia Mecca
Bohopanna - oversized bodysuit - jumper bayi - angklung, 0-6m	Bohopanna
Wayang Kulit Mini 30cm	-
Baju impor Uniqlo	-

Data Preprocessing

The text utilized in this research is derived from product titles, which can be noisy and unclean due to sellers' freedom to give titles as they see fit. As highlighted by Mao et al. [19], product titles on Indonesian e-commerce platforms often contain irrelevant words and symbols. To address this issue, we conducted text pre-processing, which involved cleaning non-alphanumeric characters, lowercasing all text, and tokenization. This pre-processing step was critical in minimizing irrelevant information and preparing the text for further analysis.

Modeling

The proposed model in this study consists of two parts: feature extraction and an end-to-end model. For feature extraction, we utilized two pre-trained BERT models in Bahasa, namely indobert-base-p1 and indobert-base-p2 [24], and used Python's SentenceBERT NLP library to obtain vector representations of the product titles. We employed three classifiers, namely Decision Tree, SVM, and Neural Network, using Python's scikit-learn library.

For the end-to-end model, we fine-tuned the pre-trained BERT models using Python's HuggingFace library, and trained them for the specific task of local product classification. This approach allowed us to leverage the strengths of pre-trained models and adapt them to the Indonesian e-commerce context. By using a combination of feature extraction and end-to-end models, we aimed to improve the accuracy and effectiveness of our classification approach for local products..

RESULTS AND DISCUSSION

Model Performance

In this study, we utilized BERT both as a text representation model and as a classifier. The performance of BERT as a text representation model and as a classifier was evaluated and compared in Table 3. The results showed that the average performance of BERT as a text representation model was approximately 88.35% on weighted F1, while as a classifier, it was approximately 97.79%. This indicates a significant difference in performance when BERT is used as a classifier.

The representation of a sentence (in this case, the product title) obtained with BERT comes from an aggregation process, such as mean pooling, max-pooling, or CLS token, which reduces the semantic meaning value of every word. In contrast, when BERT is used as an end-to-end model, the text representation is obtained from the sub-words (word pieces) and is directly processed in the classification task. Based on these results, it can be concluded that BERT as an end-to-end model performs better than BERT combined with other classifiers, such as Decision Trees, SVM, and neural networks..

From the F1 results for both labels, it can be concluded that classifiers other than BERT are always better at predicting local than import labels. In contrast, the BERT classifier is slightly better at predicting imported products. In this study, we also compared the IndoBERT base-p1 and base-p2, while in this classification task, both models give similar results, even though base-p1 has 0.1% better predicting import labels.

Table 3. Model's Performance

Embedding	Classifier	Accuracy (Import)	Accuracy (Local)	Accuracy
Indobert-base-p1	Decision Tree	80.12%	85.10%	82.74%
Indobert-base-p1	SVM	88.73%	91.50%	90.18%
Indobert-base-p1	Neural Network	91.30%	92.88%	92.13%
Average		**86.72%**	**89.83%**	**88.35%**
Indobert-base-p1	Indobert-base-p1	97.86%	97.72%	97.79%
Indobert-base-p2	Indobert-base-p2	97.85%	97.71%	97.78%
Average		**97.86%**	**97.72%**	**97.79%**

Feature Selection

Based on the evaluation results in Table 4, there is a positive effect with the presence of brand name as an additional feature. The average F1 with just a product title as a feature is 92.71%, while with the addition of a brand, it improves to 93.18%. Even though the coverage of brand extraction is relatively low (only 31%), brand names can provide a good signal as an indicator of local products.

Table 4. Model's Performance

Embedding	Classifier	Accuracy (Title)	Accuracy (Brand & Title)
Indobert-base-p1	Decision Tree	84.94%	87.20%
Indobert-base-p1	SVM	86.98%	88.81%
Indobert-base-p1	Neural Network	93.62%	92.56%
Indobert-base-p1	Indobert-base-p1	99.33%	99.00%
Indobert-base-p2	Indobert-base-p2	98.66%	98.31%
Average		**92.71%**	**93.18%**

CONCLUSION

Based on the experimental results, the end-to-end BERT-based model is the best for local product classification. Pretrained BERT Bahasa, used as the base model and fine-tuned on the local product classification, performs nearly perfectly with an evaluation result of F1 97.92%. This indicates that BERT-based models are suitable for classifying local products in Indonesia, and the end-to-end model outperforms the models combined with other classifiers.

Moreover, brand names can provide a good signal as an indicator of local products, despite the relatively low coverage of brand extraction (only 31% from all the data). The experiment results showed that the addition of brand names as an additional feature improved the classification performance by 0.5%. Therefore,

incorporating brand information as an additional feature can be a helpful strategy for improving the classification performance of local products.

With this information, the government (especially e-katalog) and all e-commerce in Indonesia can collaborate to support the government campaign to encourage the competitiveness of MSMEs. For example, they can prioritize local products or reduce the number of imported products. This can help increase the visibility of local products and improve the competitiveness of MSMEs in Indonesia. In addition, with the information on every product circulating in the community, the government can constantly evaluate programs specifically aimed at accelerating Indonesian MSMEs. This can lead to more effective programs and policies that support the development of local MSMEs in Indonesia.

In the future, it is recommended to expand the scope of this research by increasing the number of categories in the data and conducting a more detailed exploration of the evaluation results by product category. To make the data more relevant to government needs, specific product data from e-katalog can be obtained. Other features such as descriptions, prices, and images can also be utilized to improve the model's performance. Furthermore, brand extraction can be further improved by exploring better approaches. Additionally, it is suggested to investigate the impact of external factors, such as product reviews and ratings, on the classification performance. Finally, the developed model can be implemented in the e-commerce system to assist consumers in finding and purchasing local products, which can help promote the competitiveness of Indonesian MSMEs.

ACKNOWLEDGMENT

This work was supported by Information Technology Magister, Universitas Indonesia, Indonesia.

REFERENCES

[1] M. T. Kinda, *E-commerce as a Potential New Engine for Growth in Asia*. International Monetary Fund, 2019.

[2] S. M. Alagoz and H. Hekimoglu, "A study on tam: analysis of customer attitudes in online food ordering system," *Procedia-Social Behav. Sci.*, vol. 62, pp. 1138–1143, 2012.

[3] E. Hartono, C. W. Holsapple, K.-Y. Kim, K.-S. Na, and J. T. Simpson, "Measuring perceived security in B2C electronic commerce website usage: A respecification and validation," *Decis. Support Syst.*, vol. 62, pp. 11–21, 2014.

[4] H. M. Bai, A. Zaid, S. Catrin, K. Ahmed, and A. Ahmed, "The socio-economic implications of the coronavirus pandemic (COVID-19): A review," *Int. J. Surg*, vol. 8, no. 4, pp. 8–17, 2020.

[5] V. Erokhin and T. Gao, "Impacts of COVID-19 on trade and economic aspects of food security: Evidence from 45 developing countries," *Int. J. Environ. Res. Public Health*, vol. 17, no. 16, p. 5775, 2020.

[6] C. Tudor and R. Sova, "Infodemiological study on the impact of the COVID-19 pandemic on increased headache incidences at the world level," *Sci. Rep.*, vol. 12, no. 1, p. 10253, 2022.

[7] G. Fletcher and M. Griffiths, "Digital transformation during a lockdown," *Int. J. Inf. Manage.*, vol. 55, p. 102185, 2020.

[8] A. Priyono, A. Moin, and V. N. A. O. Putri, "Identifying digital transformation paths in the business model of SMEs during the COVID-19 pandemic," *J. Open Innov. Technol. Mark. Complex.*, vol. 6, no. 4, p. 104, 2020.

[9] S. Lone, N. Harboul, and J. W. J. Weltevreden, "2021 european e-commerce report," 2021.

[10] A. Situmorang, "Pertumbuhan E-Commerce Tahun Ini Meningkat Tajam di Indonesia," 2020.

[11] Cbcommerce, "Top 500 EU cross-border analysis report 2020," 2020. https://www.cbcommerce.eu (accessed Feb. 05, 2023).

[12] *Keputusan Presiden Republik Indonesia Nomor 15 Tahun 2021 Tentang Tim Gerakan Nasional Bangga Buatan Indonesia*. Indonesia, 2021.

[13] *Keputusan Presiden Republik Indonesia Nomor 2 Tahun 2022 Tentang Percepatan Peningkatan Penggunaan Produk Dalam Negeri Dan Produk Usaha Mikro, Usaha Kecil, Dan Koperasi Dalam Rangka Menyukseskan Gerakan Nasional Bangga Buatan Indonesia Pada Pelaksanaan Pe*. Indonesia, 2022.

[14] LKPP, "LKPP Kawal Transaksi Produk Dalam Negeri, Produk Impor Dalam PBJP Dibatasi," 2022. http://www.lkpp.go.id/v3/#/read/6757

[15] H. I. H. Kim, G. Joo, "Product Category Classification using Word Embedding and GRUs," *J. Korean Inst. Inf. Technol.*, vol. 19, no. 4, pp. 11–18.

[16] M. A. S. H. M. Zahera, "ProBERT: Product Data Classification with Fine-tuning BERT Model," *MWPD@ ISWC*, 2020.

[17] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," in

2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2013, pp. 195–198.

[18] G. Robert and R. Gosselin, "Evaluating the impact of NIR pre-processing methods via multiblock partial least-squares," *Anal. Chim. Acta*, vol. 1189, p. 339255, 2022.

[19] H. Mao, A. Yusup, Y. Ge, and D. Chen, "Named Entity Recognition in Chinese E-commerce Domain Based on Multi-Head Attention," in *2022 9th International Conference on Dependable Systems and Their Applications (DSA)*, 2022, pp. 576–580.

[20] I. A. Klampanos, "Manning Christopher, Prabhakar Raghavan, Hinrich Schütze: Introduction to information retrieval: Cambridge University Press, Cambridge, 2008, 478 pp, Price 60, ISBN 97805218657515." Springer, 2009.

[21] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment analysis on Covid19 vaccines in Indonesia: from the perspective of Sinovac and Pfizer," in *2021 3rd East Indonesia conference on computer and information technology (EIConCIT)*, 2021, pp. 122–127.

[22] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv Prepr. arXiv1810.04805*, 2018.

[24] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," *arXiv Prepr. arXiv2011.00677*, 2020.

[25] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang, and H. Wan, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," *Expert Syst. Appl.*, vol. 150, p. 113277, 2020.

[26] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv Prepr. arXiv1908.10084*, 2019.

[27] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augment. Hum. Res.*, vol. 5, pp. 1–16, 2020.

[28] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, "Decision tree optimization in C4. 5 algorithm using genetic algorithm," in *Journal of Physics: Conference Series*, 2019, vol. 1255, no. 1, p. 12012.

[29] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[30] H. Shelar, G. Kaur, N. Heda, and P. Agrawal, "Named entity recognition approaches and their comparison for custom ner model," *Sci. Technol. Libr.*, vol. 39, no. 3, pp. 324–337, 2020.