IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 5, October, 2023, pp. 1335-1342
Journal homepage: http://iieta.org/journals/isi

Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models

Arif Ridho Lubis[1*], Yuyun Yusnida Lase[1], Darwis Abdul Rahman[2], Deden Witarsyah[3]

[1] Department of Computer Engineering and Informatics, Politeknik Negeri Medan, Medan 20155, Indonesia
[2] Department of Mechanical Engineering, Politeknik Negeri Medan, Medan 20155, Indonesia
[3] School of Industrial and System Engineering, Telkom University, Bandung 40257, Indonesia

Corresponding Author Email: arifridho@polmed.ac.id

https://doi.org/10.18280/isi.280522

Received: 8 June 2023
Revised: 29 August 2023
Accepted: 11 October 2023
Available online: 31 October 2023

Keywords:
text preprocessing, Convolutional Neural Network (CNN), deep learning, performance, Bahasa Indonesia

ABSTRACT

Spell checking capabilities, crucial within the domain of natural language processing, often encounter limitations in the context of Bahasa Indonesia due to data irregularities and the scarcity of high-quality training data. This study aims to enhance spell checker performance through the implementation of various text preprocessing techniques, including case folding, tokenization, stemming, and the removal of stop words. A Convolutional Neural Network (CNN), a deep learning model, was employed in this research to facilitate the overall process. The study utilized data gathered from social media communities, comprising a total of 10,000 entries. This data was divided into two subsets; 80% (8,000 entries) was allocated for training and the remaining 20% (2,000 entries) was designated for testing. A series of tests were conducted on datasets subject to different preprocessing approaches: without case folding, without stop words removal, without stemming, and with all text preprocessing stages implemented. The evaluation metrics employed in this study included accuracy, recall, precision, and the F1 score. The results demonstrated notable improvements in spell checker performance with appropriate text preprocessing. Specifically, the accuracy reached 0.86 for the dataset without stemming, 0.74 for the dataset without stop words removal, 0.7 for the dataset without case folding, and 0.89 for the dataset where all preprocessing stages were applied. These findings suggest that a comprehensive text preprocessing approach, paired with deep learning models, can significantly enhance spell checker performance for Bahasa Indonesia.

1. INTRODUCTION

With its origin in Indonesia, the Indonesian language, or Bahasa Indonesia, boasts an extensive reach, being spoken by approximately 200 million individuals globally [1]. Despite this, the language is characterized by a complexity in its spelling rules-rules that dictate changes in word endings or the use of capital letters [2, 3]. This complexity necessitates the use of a spell checker to ensure standard and correct language usage, as spelling errors can significantly impact comprehension [4]. Given the intricate and often perplexing spelling rules in Indonesian, the utilization of an automatic spelling checker can substantially improve the quality of written Indonesian [5].

The evolution of technology has led to an increased reliance on automatic spelling checkers, particularly in word processing and natural language applications [6, 7]. Despite advancements in natural language processing technology, the spell checkers for Indonesian continue to present substantial shortcomings in detecting and correcting misspellings [8-10]. The primary factors contributing to these challenges include the quality of unstructured data, which complicates the application of deep learning models [11].

For optimal performance, deep learning models require clean and structured training data [12, 13]. However, natural language data is frequently unstructured and laden with noise, such as punctuation and irrelevant words. As such, it becomes imperative to apply text preprocessing techniques to natural language data prior to its processing by deep learning models [14, 15].

In addressing the problem of poor data quality and scarcity of training data in this study, text preprocessing techniques are employed to overcome the irregular properties of the text data, aiming to achieve low bias and high accuracy in spell checking. These techniques encompass several stages, including tokenization, stopword removal, stemming, and lemmatization [16]. Tokenization converts a sentence into a collection of tokens or words, which serve as the smallest units in natural language processing [17, 18].

Stopword removal eliminates frequently occurring words in sentences that offer no predictive value, such as "dan", "atau", or "yang mana". Stemming removes prefixes or suffixes from words to revert them to their base forms, while lemmatization transforms words into their basic forms so that variations of a word with the same meaning are standardized [19, 20]. The quality of the final model is influenced by the preprocessing technique used [21, 22].

One of the challenges in creating an Indonesian spell checker using deep learning techniques is the lack of a representative dataset and the limited amount of data. Therefore, the choice of an appropriate preprocessing technique can impact the quality of the model, especially when

dealing with unbalanced data with large variations.

In the context of this study, text preprocessing ensures that the training data utilized for the Deep Learning model is clean and relevant, thereby enhancing the performance of the deep learning model in conducting spell checks. This improvement in performance is anticipated to aid Indonesian language users in writing correctly and correcting potential spelling errors. Moreover, the findings of this study can be incorporated into existing Indonesian spell checkers to improve their performance and quality. This study sets out to investigate and explore the use of text preprocessing techniques for deep learning models tasked with spell checking, contributing to the field of natural language processing.

2. LITERATURE REVIEW

2.1 Indonesian spell checker

Spell checking, a technique that meticulously processes natural language to identify and correct spelling errors in texts, is achieved through the validation of a given corpus, examining each word in a text document to produce structured sentences [23, 24]. The application of spelling correction in the field of natural language processing necessitates undergoing a text preprocessing process, particularly utilizing a machine learning approach [7]. Therefore, a spell checker plays an indispensable role in maintaining the accuracy and quality of Indonesian texts.

2.2 Text preprocessing

Text preprocessing is an initial, integral stage in the field of natural language processing that involves performing a series of operations on text data prior to further processing [25, 26]. The primary objective of text preprocessing is to cleanse, modify, or reduce text data, thereby enhancing its suitability for more efficient and accurate processing. The role of text preprocessing is pivotal in improving the quality of text data and facilitating subsequent analysis and processing [11].

Several key techniques of text preprocessing include:

(1) Tokenization: Text data is fragmented into smaller tokens or individual words within a sentence.

(2) Case Folding: All letters in the document are systematically converted to lowercase.

(3) Stopword Removal: Insignificant words, despite their high frequency of occurrence, are eliminated.

(4) Stemming: Words are standardized within the document.

The application of text preprocessing enables more efficient processing of text data and produces more accurate analysis results. This is critical in many NLP applications, including search engines, sentiment analysis, and named entity recognition.

2.3 Deep learning

Deep learning, a subset of machine learning, employs multiple data processing layers to produce accurate outputs autonomously [27, 28]. Its primary objective is to enable machines' independent learning capacity and solve complex tasks, including facial recognition, voice recognition, and decision-making based on the provided data [29]. Notably, the depth of the layers in a deep learning model directly correlates with the complexity of the data representations it can learn. This makes deep learning particularly effective when managing large and complex datasets, such as images, sounds, and text [30, 31].

Despite its advantages in solving complex tasks, processing vast datasets, and not being confined to structured data, deep learning presents several challenges [32]. Such challenges encompass the requirement of substantial computational resources and a large volume of training data for effective performance. As such, the adoption of deep learning should be carefully considered and implemented by trained experts.

2.4 Convolutional Neural Network

A Convolutional Neural Network (CNN) is a specialized type of neural network employed in computer vision for processing images and videos, designed specifically to recognize patterns and features within complex and abstract images [33, 34]. The architecture of a CNN involves four interconnected layers: the convolution layer, the pooling layer, and the fully connected layer, all working in unison to discern the most significant features. An illustration of the CNN architecture is provided in Figure 1.

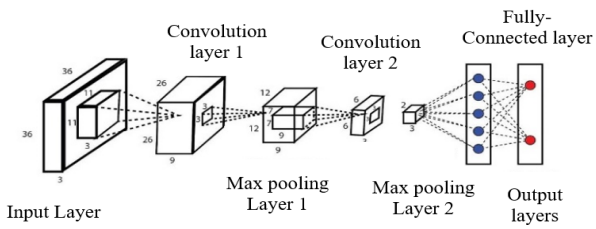

Figure 1. Architecture of a CNN

CNNs have been employed in various applications, including object recognition in images, face detection, handwriting recognition, and sign language translation. They have demonstrated exceptional performance in image and video processing, contributing significantly to advancements in the field of computer vision. CNN's ability to process images of varying sizes and recognize complex and abstract features stands out as one of its key strengths. However, similar to deep learning, CNN also faces challenges such as the requirement of substantial computational resources and a large volume of data for effective training [35].

2.5 Evaluation metrics

Evaluation metrics, essential tools in machine learning and data science, are utilized to assess the performance of a model or classification system in predicting classes within a specific dataset [36, 37]. These metrics provide critical insights into a model's performance, identifying areas that require optimization. Metrics commonly employed in this study encompass accuracy, precision, recall, and the F1-score. The formulas for calculating these metrics are as follows:

(1) Accuracy: Quantifies the frequency at which the model accurately predicts the correct class across all instances.

(2) Precision: Measures the frequency at which the model accurately predicts the positive class.

(3) Recall or Sensitivity: Determines the frequency at which the model detects the positive class.

(4) F1-score: Represents the harmonic mean of precision and recall.

2.6 Previous research

The following research studies have contributed to the understanding of the impact of text preprocessing techniques on deep learning models, aiming to enhance the performance of the Indonesian spell checker:

(1) Maslej-Krešňáková et al. [38] conducted a comparison study between machine learning models and transformer models, utilizing TF-IDF. The research provided valuable insights into the impact of the model, reporting an accuracy of 91%.

(2) Zaky and Romadhony [2] implemented the Long Short Term Memory (LSTM) model, leveraging the part of speech feature on article data, to identify word errors. The study's evaluation results on the testing dataset revealed an accuracy rate of 83.76%.

(3) Rahutomo et al. [39] proposed a model to improve text by deploying sentence rules in the Indonesian language, addressing common errors in Indonesian writing due to misunderstandings. The proposed model was tested with 20 rules using a rule-based grammar, resulting in an accuracy of 85%.

(4) Toleu et al. [40] suggested a spell scheme utilizing automatic correction. The research demonstrated that the integration of reliability models and symmetrical patterns could provide a model capable of arranging spelling. The approach was evaluated through several trials, revealing promising results, particularly after integrating the context model. QazSpell has been proven to outperform similar products available in the market.

(5) Alharbi and Qamar [41] the research uses a deep learning approach using an algorithm (LSTM) which will be used for sentiment analysis on Arabic text data. This research uses 1371 data after text preprocessing. Text preprocessing is the initial stage which has the advantage of changing text from unstructured to structured, then the data resulting from text preprocessing will be applied using LSTM which is able to obtain 83% accuracy.

(6) Asqolani and Setiawan [42] identifying cyberbullying on the Twitter social media platform. In the process, the data will be processed including the stages of data labeling, data processing, and feature extraction with TF-IDF. Data processing will be beneficial for data originating from social media, so this research will use a deep learning approach by utilizing hybrid algorithms from CNN, LSTM, CNNLSTM, and LSTM-CNN.

3. METHODOLOGY

The methodology used in this study includes several stages. Such as data collection, text preprocessing, and creation or development of deep learning models that will use the Convolutional Neural Network (CNN) approach. This model is trained using data that has gone through the preprocessing stage before. Next, an evaluation is made of the model that has been trained using the test data that has been prepared. Evaluation is carried out by comparing the output results of the model with the actual test data to determine the accuracy of the model in conducting an Indonesia spell check. To be able to see the effect of the preprocessing process on the deep learning model, a test will be carried out by comparing the performance of the model before and after preprocessing. The data to be used for testing is the same data as the data used in implementing the model with deep learning algorithms. The results obtained are to see the effect of the preprocessing technique in conducting Indonesian spell check.

3.1 Dataset

The dataset used in this study is a community dataset on Twitter social media obtained through crawling techniques. The dataset consists of 10,000 tweets in Indonesian containing various types of text, including formal and informal texts, texts with general and special vocabulary, and texts with different spelling variations. The dataset then goes through a preprocessing stage, including punctuation removal, tokenization, stemming, and stopword removal. The following is the dataset used before preprocessing the text contained in Table 1.

The dataset that will be used amounts to 10,000 tweet data, implementing spell check will divide data for the training process and the testing process, for the data training process it uses 80% of the total data while for the testing process, it uses 20% of the amount of data to be processed in the model deep learning with the CNN approach.

Table 1. Data before text preprocessing

No	Tweet Data
1	kelar anjay senang hati
2	Hwhwhwh Berani Ngemal Tokoh Hidup Membeli
3	Cakung Banget Tidur Sore
4	Komika Cowok Anon Waras Jam Avatar Anugerah
5	Untung Dunia Fangirl Gosip Orang Suka Sastra Lihat Tersambar Pernyataan Berlangganan Indah Banget Dibikinin Daftar
6	hihi habis gangguan night ayg
7	julukan produksi jamkos cuy jam gelar sore
…	….
…	….
9998	tidak wadah teman cowok ngjak bareng tahu ntuh arti aph
9999	segar bahagia nugas selesai cevat bagoes
10000	wadah ayah ajak anak anak jalan jalan kuliner wekend habis sudah keluarga bangun

3.2 Research architecture

In this study, text preprocessing techniques used include text normalization, tokenization, and stopword removal. This research collects community tweet data that resides on Twitter social media. The data is then processed using the text preprocessing technique previously mentioned and used as training data for deep learning models. Next, an evaluation of the model performance is carried out using test data taken from the same source as the training data. The following is the research architecture contained in Figure 2.

The spell-checker model works in three stages. The first stage involves text preprocessing techniques to make the data structured, then the model training stage is carried out using datasets that have passed the text preprocessing process, then tests are carried out using test datasets, the spelling checker research architecture using a deep learning approach can be explained in the figure. The following describes the stages of the research architecture:

(1) Preprocessing text to data

Text preprocessing is the process of converting raw text data into a format that is easier for machines to process and understand. Some common steps in text preprocessing include:

1) Eliminating non-alphanumeric characters: non-

alphanumeric characters such as punctuation marks, double spaces, and other special characters are often unnecessary in text analysis. Therefore, the first step in text preprocessing is to remove these characters.

2) Lowercasing: Usually, the text in the document is inconsistent in the use of uppercase and lowercase letters. To facilitate the text analysis process, all text is generally converted to lowercase.

3) Tokenization: The text is then broken down into smaller units called tokens. Tokens can be words, phrases, or characters. Tokenization allows machines to understand the structure of a text and perform further analysis.

4) Stop word removal: Stop words are words that are very common and often appear in texts, such as "yang", "is", "tdk", and "kw". These words do not provide much information about the text being analysed and are often removed from documents during preprocessing.

5) Stemming or Lemmatization: Stemming is the process of cutting the word endings to get the basic form. Lemmatization also has the same goal, which is to change words to their basic forms. This process is done to reduce dimensions and avoid problems such as overfitting and word redundancy in the model to be built.

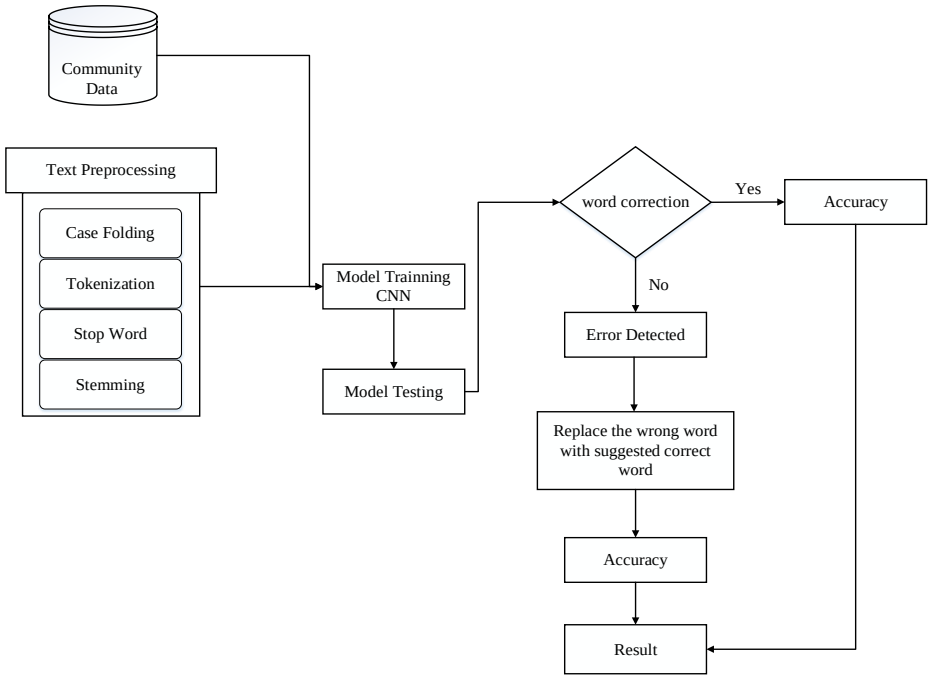

Figure 2. Architecture research spell checker

In Table 2 the dataset has not been preprocessed, then steps such as case folding, tokenization, stop word and removal are carried out so as to produce data as in Table 3.

Table 2. Data before text preprocessing

No	Tweet Data
1	kelar anjay senang hati
2	Hwhwhwh Berani Ngemal Tokoh Hidup Membeli
3	Cakung Banget Tidur Sore
4	Komika Cowok Anon Waras Jam Avatar Anugerah
5	Untung Dunia Fangirl Gosip Orang Suka Sastra Lihat Tersambar Pernyataan Berlangganan Indah Banget Dibikinin Daftar
6	hihi habis gangguan night ayg
7	julukan produksi jamkos cuy jam gelar sore
…	….
…	….
9998	tidak wadah teman cowok ngjak bareng tahu ntuh arti aph
9999	segar bahagia nugas selesai cevat bagoes
10000	wadah ayah ajak anak anak jalan jalan kuliner wekend habis sudah keluarga bangun

The CNN model consists of several convolutional and pooling layers which will process input text data with a maximum length of max_len and embedding_size dimensions. Then, dropout layers will be applied to each convolutional and pooling layer to prevent overfitting of the model. Next, a fully connected layer will be added using the Relu activation function to classify the data. Layer dropouts are also applied to fully connected layers to prevent overfitting the model. Finally, the model will be compiled using the loss function categorical_crossentropy and the Adam optimizer to train the model with accuracy metrics. At this stage, the model that has been built will be trained using training data. This training process is carried out by compiling it first, namely by determining the optimizer, loss function, and metrics to be used to evaluate model performance. Next, the model will be trained using training data with a certain batch size and a number of epochs. After the training process is complete, the model will be evaluated using validation data. Finally, the model will be saved with a specific name so that it can be used at a later stage.

(2) The process of checking the spelling of the CNN algorithm

The model to be processed will use data from text preprocessing results and use data without text preprocessing results, then the CNN algorithm will process it, enter training data with correct spelling data and incorrect spelling data, then every word, every character will be broken into pieces small to process. In the process, the CNN network will include an input layer in the form of characters in the dataset, then will use several convolution layers for each feature, the polling

layer will reduce the feature space, and then the fully connected layer will be used to produce spell check output. The hyperparameter process will configure the number of hidden layers, batch size, and number of epochs.

1. The process of checking the spelling of the CNN algorithm

The results of the CNN algorithm process will be evaluated by measuring the accuracy produced by measuring predictions that produce the correct value of the whole.

Table 3. Data after text preprocessing

No	Tweet Data
1	selesai hati senang
2	berani jalan ke plaza sambil membeli
3	tanggung banget tidur sore
4	komika cowok tidak waras jam avatar anugerah
5	untung dunia perempuan gosip orang suka sastra lihat tersambar pernyataan berlangganan indah banget dibikinin daftar
6	habis gangguan malam sayang
7	julukan produksi jam bro jam gelar sore
….	….
….	….
9998	tidak wadah teman cowok ngajak bareng tahu arti nya
9999	segar bahagia tugas selesai cepat bagus
10000	wadah ayah ajak anak anak jalan jalan kuliner sabtu minggu habis sudah keluarga bangun

4. DISCUSSION AND RESULT

This discussion, this research will discuss the effect of text preprocessing techniques on deep learning models to improve the performance of Indonesian spell checkers. Text preprocessing techniques are techniques for cleaning, normalizing, and transforming text so that it can be further processed by deep learning models. In this study, experiments will be carried out using several different text preprocessing techniques to see the effect on the model's performance in checking Indonesian spelling.

4.1 CNN model

The model used in this study uses the CNN model with several convolutional and pooling layers, as well as dropout and fully connected layers. The implementation uses the Python language TensorFlow library which is platform-independent but model development and verification is done on the Windows platform. The test dataset uses community data originating from Twitter. This CNN model is trained on a dataset that utilizes the use of text preprocessing stages and is evaluated using all stages of text preprocessing. Model optimization in the CNN model is also quite a difficult task because it involves hyperparameters that affect the behavior of the model. In this study, manual selection is used to optimize the hyperparameters so that training and testing losses are minimum.

4.2 Evaluation metrics

In this study, several evaluation metrics are used to evaluate the performance of the proposed deep learning model after applying text preprocessing techniques. These metrics include precision, recall, f1-score, and accuracy. Precision and recall are used to measure the accuracy and completeness of the model in predicting the correct spelling of words, while the f1-score is the harmonic mean of precision and recall. Accuracy is used to measure how much the model's predictions are correct from the total predictions made. The results of the evaluation of this deep learning model will be used to determine whether the applied text preprocessing technique can improve the performance of the Indonesian spell checker or not. The following are the results of the evaluation metrics obtained from testing the deep learning model when performing a spell checker with CNN.

Table 4. Performance results of deep learning model with CNN

Dataset	Precision	Recall	f-measure	Accuracy
Datasets without stemming	0.882	0.938	0.909	0.86
Datasets without stopwords	0.7941	0.9310	0.8579	0.74
Dataset without case folding	0.75	0.8571	0.7992	0.7
Dataset with full text preprocessing	0.9184	0.9713	0.944	0.89

Based on Table 4, the process of obtaining value with evaluation metrics will be explained.

(1) Datasets without stemming, in testing the model using a dataset without the stemming process in text processing, the performance of the model is shown in Figure 3.

Figure 3 will explain the calculation of precision, recall, F-measure and accuracy as follows:

TP = 200, Precision = 0.882, Recall = 0.938, F-measure = 0.909, Accuracy = 0.869

(2) Dataset without stopword processing. In testing the model using a dataset without stopword processing in text processing, the performance of the model is shown in Figure 4.

Figure 4 will explain the calculation of precision, recall, F-measure and accuracy as follows:

TP = 350, Precision = 0.7941, Recalls = 0.9310 , F-measure = 0.8579, Accuracy = 0.7455

(3) Dataset without case folding process. In testing the model using datasets without case folding process in the preprocessing test, the performance of the model is as shown in Figure 5.

Figure 5 will explain the calculation of precision, recall, F-measure and accuracy as follows:

TP = 400, Precision = 0.75, Recall = 0.8571, F-measure = 0.7992, Accuracy = 0.7

(4) Dataset with text preprocessing as a whole. In testing the model using a dataset with a text processing process as a whole, the performance of the model is shown in Figure 6.

Figure 6 will explain the calculation of precision, recall, F-measure and accuracy as follows:

TP = 400, Precision = 0.9184, Recall = 0.9713, F-measure = 0.9441, Accuracy = 0.8933

The results of the study show that the use of appropriate text preprocessing techniques can improve the model's

performance in performing Indonesian spell checks. Some of the techniques applied, such as stemming, stopword removal, and tokenization, were able to increase the accuracy and F1-score of the built model. The following is the comparison accuracy of the model applied to the text preprocessing technique in Figure 7.

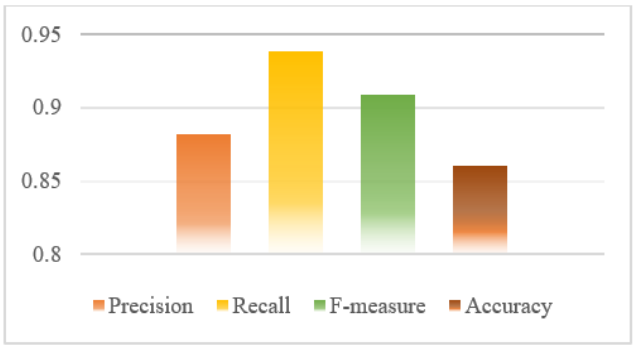

Figure 3. Performances metrik evaluasi Model 1

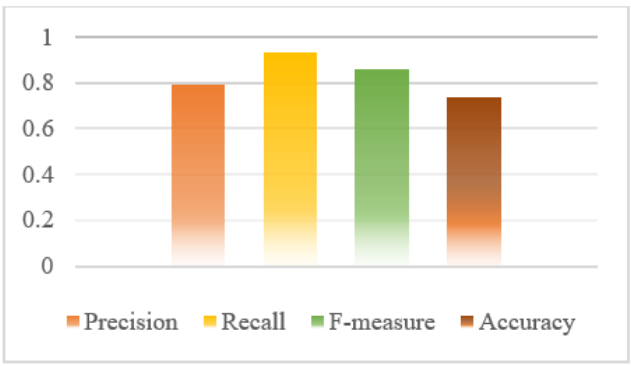

Figure 4. Performances metrik evaluasi Model 2

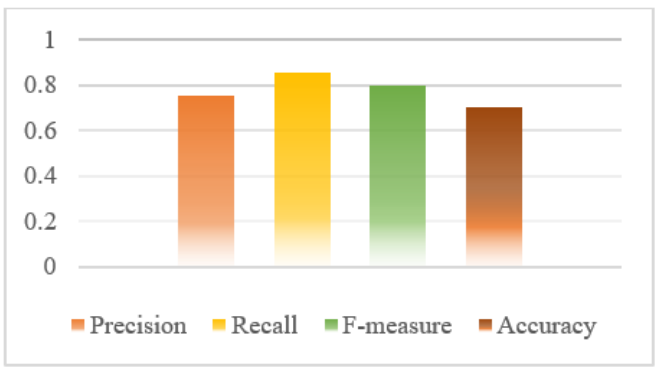

Figure 5. Performances metrik evaluasi Model 3

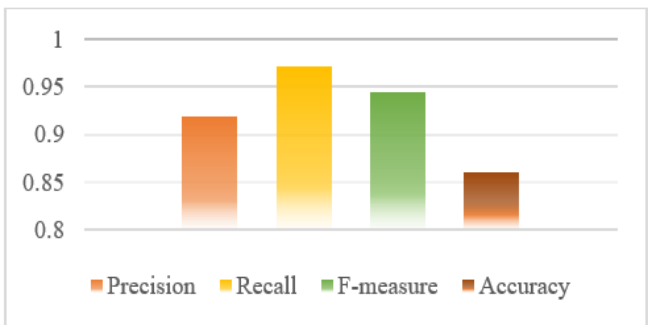

Figure 6. Performances metrik evaluasi Model 4

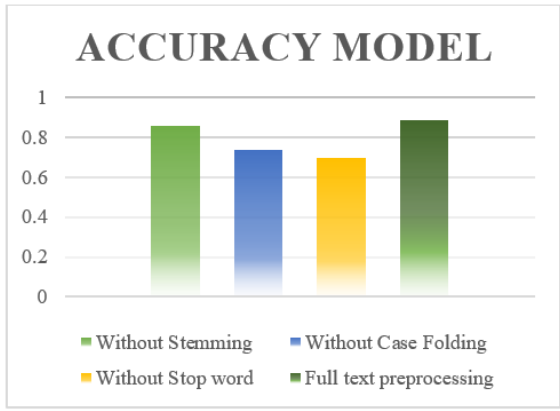

Figure 7. Performances accuracy using text preprocessing

In addition, the use of text preprocessing techniques can also speed up model training time and reduce the dimensions of the features used by the model. The use of text preprocessing it must be done for unstructured text data found on social media platforms such as Twitter, text preprocessing has several advantages such as text normalization, reducing multiple data dimensions so that the stages of text preprocessing can make changes to unstructured data.

4.3 Error analysis

Error analysis in the spell check model to be able to understand the performance of the model in the form of weaknesses such as errors in words that are not in the dataset, errors in words that are visually similar, and errors in words with abbreviations. The following in Table 5 is the data error in the spell-check model .

Table 5. Analysis of spelling errors in the model

No	Spelling Error	Data
1	Manga	Mangga
2	Merakah	mereka
3	Menyiruh	Meniru
4	Es krim	Eskrim

5. CONCLUSIONS

In this study, a deep learning model approach will be used by using CNN for the purpose of checking spelling on text data. In this study, the preprocessing technique will be used for data originating from social media with unstructured data properties. punctuation marks, emoticons, so that text data can be analyzed. The results of this study with the text preprocessing technique can improve the spelling check model, this is evidenced by the matrix evaluation. In the use of preprocessing text, each stage such as case folding, stemming and stop words will be analyzed and applied to the spell check model using the CNN approach. So this research will produce an accuracy of 0.7 for the CNN model without using case folding, an accuracy of 0.74 for the CNN model without using stop words, an accuracy of 0.86 for the model without using stemming while for the model using the whole stages of text preprocessing it will produce an accuracy of 0.89. so that this research can see differences in accuracy from using the stages of text preprocessing for the CNN model in conducting spell checks. The limitations of this model are spelling errors such

as errors in words that are not in the dataset, errors in words that are visually similar, and errors in words with abbreviations.

REFERENCES

[1] Sakhiyya, Z., Martin-Anatias, N. (2023). Reviving the language at risk: A social semiotic analysis of the linguistic landscape of three cities in Indonesia. International Journal of Multilingualism, 20(2): 290-307. https://doi.org/10.1080/14790718.2020.1850737

[2] Zaky, D., Romadhony, A. (2019). An LSTM-based spell checker for indonesian text. In 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Yogyakarta, Indonesia, pp. 1-6. https://doi.org/10.1109/ICAICTA.2019.8904218

[3] Ridwan, M. (2018). National and official language: The long journey of Indonesian language. Budapest International Research and Critics Institute-Journal (BIRCI-Journal), 1(2): 72-78. https://doi.org/10.33258/birci.v1i2.14

[4] Stankevičius, L., Lukoševičius, M., Kapočiūtė-Dzikienė, J., Briedienė, M., Krilavičius, T. (2022). Correcting diacritics and typos with a ByT5 transformer model. Applied Sciences, 12(5): 2636. https://doi.org/10.3390/app12052636

[5] Rivera-Acosta, M., Ruiz-Varela, J.M., Ortega-Cisneros, S., Rivera, J., Parra-Michel, R., Mejia-Alvarez, P. (2021). Spelling correction real-time american sign language alphabet translation system based on yolo network and LSTM. Electronics, 10(9): 1035. https://doi.org/10.3390/electronics10091035

[6] Hládek, D., Staš, J., Pleva, M. (2020). Survey of automatic spelling correction. Electronics, 9(10): 1670. https://doi.org/10.3390/electronics9101670

[7] Fahda, A., Purwarianti, A. (2017). A statistical and rule-based spelling and grammar checker for Indonesian text. In 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, Indonesia, pp. 1-6. https://doi.org/10.1109/ICODSE.2017.8285846

[8] Sakuntharaj, R., Mahesan, S. (2016). A novel hybrid approach to detect and correct spelling in Tamil text. In 2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), Galle, Sri Lanka, pp. 1-6. https://doi.org/10.1109/ICIAFS.2016.7946522

[9] Liyanapathirana, U., Gunasinghe, K., Dias, G. (2021). Sinspell: A comprehensive spelling checker for sinhala. arXiv preprint arXiv:2107.02983. https://doi.org/10.48550/arXiv.2107.02983

[10] Lubis, A.R., Sitompul, O.S., Nasution, M.K., Zamzami, M. (2022). Feature Extraction of Tweet data Characteristics to Determine Community Habits. In 2022 5th International Conference of Computer and Informatics Engineering (IC2IE), Jakarta, Indonesia, pp. 309-313. https://doi.org/10.1109/IC2IE56416.2022.9970180

[11] Lubis, A.R., Nasution, M.K., Sitompul, O.S., Zamzami, E.M. (2023). A new approach to achieve the users' habitual opportunities on social media. IAES International Journal of Artificial Intelligence, 12(1): 41-47. https://doi.org/10.11591/ijai.v12.i1.pp41-47

[12] Mawardi, V.C., Susanto, N., Naga, D.S. (2018). Spelling correction for text documents in Bahasa Indonesia using finite state automata and Levinshtein distance method. MATEC Web of Conferences, 164: 01047. https://doi.org/10.1051/matecconf/201816401047

[13] Zukarnain, N., Abbas, B.S., Wayan, S., Trisetyarso, A., Kang, C.H. (2019). Spelling checker algorithm methods for many languages. In 2019 International Conference on Information Management and Technology (ICIMTech), Jakarta/Bali, Indonesia, Vol. 1, pp. 198-201. https://doi.org/10.1109/ICIMTech.2019.8843801

[14] Jimale, A.O., Zainon, W.M.N.W., Abdullahi, L.F. (2019). Spell checker for somali language using knuth-morris-pratt string matching algorithm. In: Saeed, F., Gazem, N., Mohammed, F., Busalim, A. (eds) Recent Trends in Data Science and Soft Computing. IRICT 2018. Advances in Intelligent Systems and Computing, vol 843. Springer, Cham. https://doi.org/10.1007/978-3-319-99007-1_24

[15] Gipayana, M. (2017). Spell checker implementation to analyze the narrative essay of sixth-grade elementary school students in Indonesia. Bulletin of Social Informatics Theory and Application, 1(1): 18-25. https://doi.org/10.31763/businta.v1i1.21

[16] Fesseha, A., Xiong, S., Emiru, E.D., Diallo, M., Dahou, A. (2021). Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. Information, 12(2): 52. https://doi.org/10.3390/info12020052

[17] Ayedh, A., Tan, G., Alwesabi, K., Rajeh, H. (2016). The effect of preprocessing on arabic document categorization. Algorithms, 9(2): 27. https://doi.org/10.3390/a9020027

[18] Lubis, A.R., Prayudani, S., Fatmi, Y., Nugroho, O. (2022). Classifying news based on Indonesian news using LightGBM. In 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Surabaya, Indonesia, pp. 162-166. https://doi.org/10.1109/CENIM56801.2022.10037401

[19] Shim, J.G., Ryu, K.H., Lee, S.H., Cho, E.A., Lee, Y.J., Ahn, J.H. (2021). Text mining approaches to analyze public sentiment changes regarding COVID-19 vaccines on social media in Korea. International Journal of Environmental Research and Public Health, 18(12): 6549. https://doi.org/10.3390/ijerph18126549

[20] Lubis, A.R., Prayudani, S., Fatmi, Y., Nugroho, O. (2022). Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP) Models on News Data. In 2022 5th International Conference of Computer and Informatics Engineering (IC2IE), Jakarta, Indonesia, pp. 314-319. https://doi.org/10.1109/IC2IE56416.2022.9970067

[21] Santoso, P.H., Istiyono, E., Haryanto, Hidayatulloh, W. (2022). Thematic analysis of indonesian physics education research literature using machine learning. Data, 7(11): 147. https://doi.org/10.3390/data7110147

[22] Lubis, A.R., Prayudani, S., Lubis, M., Nugroho, O. (2022). Sentiment analysis on online learning during the Covid-19 pandemic based on opinions on Twitter using KNN method. In 2022 1st International Conference on Information System & Information Technology (ICISIT), Yogyakarta, Indonesia, pp. 106-111. https://doi.org/10.1109/ICISIT54091.2022.9872926

[23] Yanfi, Y., Gaol, F.L., Soewito, B., Warnars, H.L.H.S. (2022). Spell checker for the Indonesian language: Extensive review. International Journal of Emerging Technology and Advanced Engineering, 12(5): 1-7.

https://doi.org/10.46338/ijetae0522_01
[24] Lubis, A.R., Nasution, M. K., Sitompul, O.S., Zamzami, E.M. (2022). Spelling checking with deep learning model in analysis of tweet data for word classification process. In 2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Jakarta, Indonesia, pp. 343-348. https://doi.org/10.23919/EECSI56542.2022.9946476
[25] Lubis, A.R., Zarlis, M., Nasution, Z. (2021). Effect of various coordinate points on social media. Journal of Physics: Conference Series, 1830(1): 012004. https://doi.org/10.1088/1742-6596/1830/1/012004
[26] Lubis, A.R., Nasution, M.K., Sitompul, O.S., Zamzami, E.M. (2020). Obtaining value from the constraints in finding user habitual words. In 2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS), Lombok, Indonesia, pp. 1-4. https://doi.org/10.1109/ICADEIS49811.2020.9277443
[27] Rbah, Y., Mahfoudi, M., Balboul, Y., Fattah, M., Mazer, S., Elbekkali, M., Bernoussi, B. (2022). Machine learning and deep learning methods for intrusion detection systems in iomt: A survey. In 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, pp. 1-9. https://doi.org/10.1109/IRASET52964.2022.9738218
[28] Fuentes, A., Yoon, S., Kim, S.C., Park, D.S. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors, 17(9): 2022. https://doi.org/10.3390/s17092022
[29] Tsuneki, M. (2022). Deep learning models in medical image analysis. Journal of Oral Biosciences, 64(3): 312-320. https://doi.org/10.1016/j.job.2022.03.003
[30] Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S.R., Tiede, D., Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote Sensing, 11(2): 196. https://doi.org/10.3390/rs11020196
[31] Rahmat, R.F., Pratama, M.F., Purnamawati, S., Faza, S., Lubis, A.R., Al-Khowarizmi, A.K., Lubis, M. (2022). Astrocytoma, ependymoma, and oligodendroglioma classification with deep convolutional neural network. IAES International Journal of Artificial Intelligence, 11(4): 1306-1313. https://doi.org/10.11591/ijai.v11.i4.pp1306-1313
[32] Bakator, M., Radosav, D. (2018). Deep learning and medical diagnosis: A review of literature. Multimodal Technologies and Interaction, 2(3): 47. https://doi.org/10.3390/mti2030047
[33] Lu, J., Tan, L., Jiang, H. (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. Agriculture, 11(8): 707. https://doi.org/10.3390/agriculture11080707
[34] Ezzat, D., Hassanien, A.E., Ella, H.A. (2021). An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. Applied Soft Computing, 98: 106742. https://doi.org/10.1016/j.asoc.2020.106742
[35] Verawati, I., Hasibuan, I.D.P. (2021). Artificial neural network in classification of human blood cells using faster R-CNN. In 2021 4th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, pp. 86-91. https://doi.org/10.1109/ICOIACT53268.2021.9563974
[36] Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 10(5): 593. https://doi.org/10.3390/electronics10050593
[37] Krizanova, A., Lăzăroiu, G., Gajanova, L., Kliestikova, J., Nadanyiova, M., Moravcikova, D. (2019). The effectiveness of marketing communication and importance of its evaluation in an online environment. Sustainability, 11(24): 7016. https://doi.org/10.3390/su11247016
[38] Maslej-Krešňáková, V., Sarnovský, M., Butka, P., Machová, K. (2020). Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. Applied Sciences, 10(23): 8631. https://doi.org/10.3390/app10238631
[39] Rahutomo, F., Mulyo, A.S., Saputra, P.Y. (2018). Automatic grammar checking system for Indonesian. In 2018 International Conference on Applied Science and Technology (iCAST), Manado, Indonesia, pp. 308-313. https://doi.org/10.1109/iCAST1.2018.8751591
[40] Toleu, A., Tolegen, G., Mussabayev, R., Krassovitskiy, A., Ualiyeva, I. (2022). Data-driven approach for spellchecking and autocorrection. Symmetry, 14(11): 2261. https://doi.org/10.3390/sym14112261
[41] Alharbi, L.M., Qamar, A.M. (2022). Arabic sentiment analysis of eateries' reviews using deep learning. Ingénierie des Systèmes d'Information, 27(3): 503-508. https://doi.org/10.18280/isi.270318
[42] Asqolani, I.A., Setiawan, E.B. (2023). Hybrid deep learning approach and Word2Vec feature expansion for cyberbullying detection on Indonesian twitter. Ingénierie des Systèmes d'Information, 28(4): 887-895. https://doi.org/10.18280/isi.280410