Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Implementation of IndoNLU Pre-Trained Model for Aspect-Based Sentiment Analysis of Indonesian Stock News

Muhammad Riza Alifi[1], Djoko Cahyo Utomo Lieharyani[2], Bima Putra Sudimulya[3], Mohammad Rizky Maulidhan[4]

[1,2,3,4]Computer Engineering and Informatics Department, Politeknik Negeri Bandung
[1,2,3,4]Jl. Gegerkalong Hilir, Ciwaruga, Kec. Parongpong, Kab. Bandung Barat, Jawa Barat 40559.

Article:
Accepted: July 30, 2023
Revised: May 14, 2023
Issued: October 28, 2023

***Correspondence Address:**
djoko.c.utomo@polban.ac.id

ABSTRACT

Investor numbers in Indonesia are steadily increasing, particularly in mutual funds managed by investment managers. News plays a pivotal role in the decision-making process for investment managers in stock investments. However, the diverse array of news sources and varying writing styles present a challenge in extracting information on each issuer mentioned in the news. This research addresses this challenge by implementing the Aspect-Based Sentiment Analysis (ABSA) method to specifically extract news related to each aspect (issuer) and assess their sentiment. The chosen model is the pre-trained Indonesian Bidirectional Encoder Representations from Transformers (BERT) model, IndoNLU, tailored for Indonesian-language stock news analysis. The primary objective of the study is to enhance the understanding of sentiment analysis within the context of stock news, contributing to the decision-making capabilities of investment managers. The data used to train the model is sourced from online news platforms, including CNBC Indonesia, Bisnis, Indopremier, Investor Daily, and Kontan, with a total of 7685 collected news articles. The results of this research were achieved through a combination of hyperparameters, including batch size of 8, learning rate of 0.00002, and 8 epochs for the IndoNLU model. This combination of hyperparameters demonstrates impressive average evaluation metric value, including precision, recall, f1-score, and accuracy reaching 90%.

Keywords: *ABSA, BERT, indoNLU, stock, news*

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

1. INTRODUCTION

Stock investment is the distribution of financial resources by placing a number of funds to buy securities in the form of shares with the hope of generating certain additional or returns on funds that have been invested in stock trading [1]. In Indonesia, the number of investors in the capital market continues to increase from year to year. Recorded until December 2022, the number of investors in Indonesia has reached around 10.31 million entities [2]. Of this total, mutual funds contributed the most with 9.60 million investors [2].

A fund manager is the person responsible for managing investors' funds in a mutual fund. Investment managers use a different and more comprehensive approach in the investment decision-making process compared to individual investors [3]. News is one of the factors considered to help make decisions, because it contains fundamental information and sentimental information [4]. Furthermore, extracting useful information from online news for financial tasks is more challenging than structured numerical information or traditional news information [5]. This is because online news sources are much more diverse with different writing styles [5].

A possible solution for rapid analysis of a large number of news stories is the use of computational algorithms to automatically analyze text using Natural Language Processing (NLP). One of the tasks in NLP is sentiment analysis which can be used to identify opinions, in terms of positive, negative, or neutral sentiment [6]. Sentiment analysis can be used to identify the sentiment of the text contained in stock news. Investment managers can utilize sentiment analysis of stock news to understand the state, opinion, perception of a particular stock, and how it can affect the stock price.

Previous research [6] conducted sentiment analysis on stock news and provided relevant information for decision making in the stock market. The Bidirectional Encoder Representations from Transformers (BERT) model in the study was carried out to improve the manual labeling process of 582 articles from several news sources as positive, neutral, and negative. This research resulted in an accuracy of 82.5%, precision of 75%, recall of 71.3%, and f1-score of 72.5%. In this research, sentiment analysis cannot be done specifically on one particular issuer, so this research suggests that future research should use methods that can extract stock news specifically for each issuer. One solution to conduct sentiment analysis specifically on one particular issuer can be done by implementing aspect-based sentiment analysis (ABSA). ABSA is an advanced sentiment analysis technique that can define each aspect listed in the text, as well as analyze sentiment towards that aspect [7].

One of the pre-trained language representation models that outperforms many task-specific architectures is BERT. BERT is designed to train deep bidirectional representations of text [8]. A pre-trained BERT model can be returned by simply adding one output layer for a specific task. The use of pre-trained BERT models can avoid the need to train new models from scratch. Many derivatives of BERT models are developed specifically for one particular language called monolingual BERT and for multiple languages called multilingual BERT. This research uses monolingual BERT specifically for Indonesian, because the training data used is stock news in Indonesia using Indonesian. The monolingual BERT model for Indonesian used in this study is IndoNLU. IndoNLU was trained using 4 billion words from the Indo4B dataset collected through 12 sources consisting of social media, blogs, news, and websites [9]. So, this research will implement ABSA on Indonesian stock news using a pre-trained Indonesian BERT model, called IndoNLU.

This research makes a significant contribution by focusing on sentiment analysis in Indonesian stock news, a context that is rich and unique. We adopt the IndoNLU model, a pre-trained Indonesian BERT, and ABSA as the primary tool due to its linguistic suitability for the Indonesian language. Through this approach, the study not only deepens the understanding of sentiment in the context of the Indonesian stock market but also emphasizes the importance of using models optimized for the local language to achieve more accurate and relevant results in the analysis of Indonesian stock news.

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

2. METHODS

2.1. BERT

BERT is a deep learning model that can be used to process text. BERT is designed to help computers understand the meaning of ambiguous language in a text by using surrounding text to build context [8]. BERT is a novel language representation model that generates a pre-trained bidirectional representation of unlabeled text by conditioning on both contexts at all layers. The trained BERT model can be customized by adding only one output layer [8].

BERT uses transformers that apply attention mechanisms to learn contextual relationships between words or subwords in a text [10]. Transformers are sequence-to-sequence architectures based on self-attention mechanisms for encoders and decoders [6]. Self-attention, also known as intra-attention, is an attention mechanism that uses the position of words in a sequence or sentence to determine the representation or relationship of the sequence [11]. Self-attention allows the model to find the relationship between words and other words by calculating their attention weights so that the relationship of words in the sentence and important words in a sentence can be known.

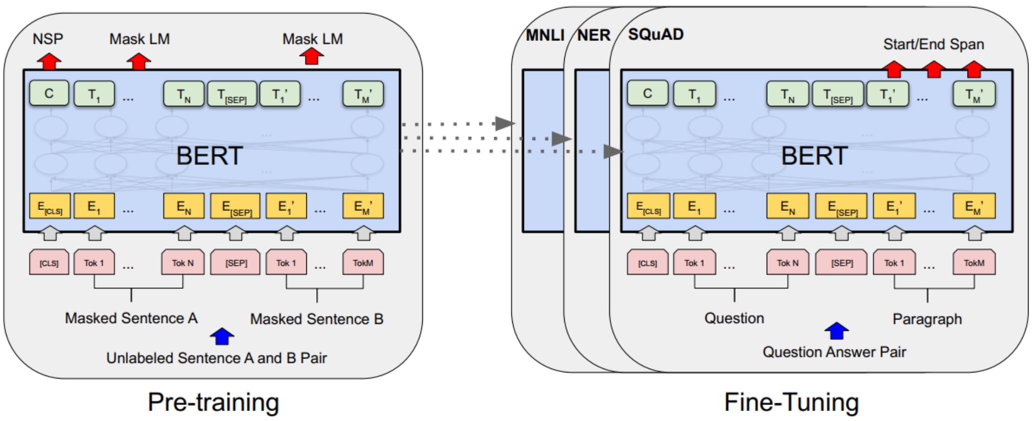

Figure 1. Pre-training and fine-tuning of BERT

Figure 1 shows that BERT has two parts: pre-training and fine-tuning. BERT has a multi-layer bidirectional transformer architecture. The pre-training performed on the BERT model is trained using two tasks that are unsupervised learning, namely masked language modeling (MLM) and next sentence prediction (NSP) [8]. MLM has a function to predict blank or masked sentences in a sentence, while NSP predicts the relationship between two input sentences [8].

Fine-tuning is the process of training a pre-trained model using different datasets by changing some parameters and adding output layers to the model [8]. MLM teaches BERT to understand the relationships between words, while NSP teaches BERT to understand dependencies and relationships across sentences [12].

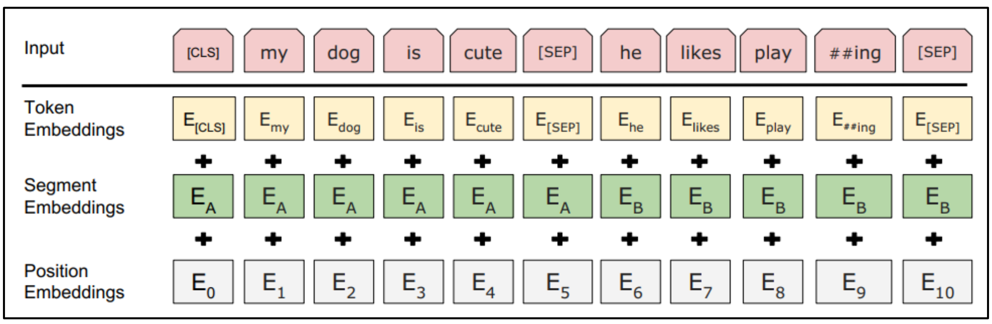

Figure 2. Example embedding process of BERT

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

In the process of word tokenization BERT uses WordPiece tokenizer which is a sub-word tokenization based on vocabulary data where every first token in the sequence is always marked with [CLS] which is a special classification token in BERT and separated or terminated by the token [SEP] [8]. Figure 2 shows the input part illustrates the tokenizer result of BERT which is a sub-word with an example of the word "playing" which becomes two tokens namely "play" and "##ing".

2.2. IndoNLU

IndoNLU represents an Indonesian BERT monolingual model renowned for its robustness and effectiveness in natural language understanding tasks. The model was meticulously trained utilizing the extensive Indo4B dataset, a vast corpus encompassing a staggering 4 billion words distributed across 250 million sentences [9]. Through this extensive training on a diverse array of linguistic contexts, IndoNLU has acquired a comprehensive understanding of the Indonesian language, enabling it to excel in various NLP applications, including sentiment analysis, text classification, and language generation tasks among others. Its proficiency stems from its ability to capture nuanced linguistic nuances and intricate structures within the Indonesian language, making it a pivotal asset in the realm of natural language processing within the Indonesian linguistic landscape.

Dataset	# Words	# Sentences	Size	Style	Source
OSCAR (Ortiz Suárez et al., 2019)	2,279,761,186	148,698,472	14.9 GB	mixed	OSCAR
CoNLLu Common Crawl (Ginter et al., 2017)	905,920,488	77,715,412	6.1 GB	mixed	LINDAT/CLARIAH-CZ
OpenSubtitles (Lison and Tiedemann, 2016)	105,061,204	25,255,662	664.8 MB	mixed	OPUS OpenSubtitles
Twitter Crawl[2]	115,205,737	11,605,310	597.5 MB	colloquial	Twitter
Wikipedia Dump[1]	76,263,857	4,768,444	528.1 MB	formal	Wikipedia
Wikipedia CoNLLu (Ginter et al., 2017)	62,373,352	4,461,162	423.2 MB	formal	LINDAT/CLARIAH-CZ
Twitter UI[2] (Saputri et al., 2018)	16,637,641	1,423,212	88 MB	colloquial	Twitter
OPUS JW300 (Agić and Vulić, 2019)	8,002,490	586,911	52 MB	formal	OPUS
Tempo[3]	5,899,252	391,591	40.8 MB	formal	ILSP
Kompas[3]	3,671,715	220,555	25.5 MB	formal	ILSP
TED	1,483,786	111,759	9.9 MB	mixed	TED
BPPT	500,032	25,943	3.5 MB	formal	BPPT
Parallel Corpus	510,396	35,174	3.4 MB	formal	PAN Localization
TALPCo (Nomoto et al., 2018)	8,795	1,392	56.1 KB	formal	Tokyo University
Frog Storytelling (Moeljadi, 2012)	1,545	177	10.1 KB	mixed	Tokyo University
TOTAL	3,581,301,476	275,301,176	23.43 GB		

Figure 3. Indo4B dataset

Figure 3 shows Indo4B dataset also includes sentences of formal Indonesian and informal everyday language collected from 12 sources. Two sources cover non-formal language, eight sources cover formal Indonesian, and the rest have a mixed style between formal and non-formal language [9]. Indo4B dataset is large in size and diverse in language style.

2.3. Confusion Matrix

Confusion matrix is a performance measurement tool used for classification problems in ML, be it binary or multiple classification [13]. Confusion matrix is a two-dimensional matrix, which contains the dimension of the ground truth class or actual value (actual value) and the dimension of the predicted value class.

Confusion Matrix		Prediction		
		Positive	Negative	Neutral
Actual	Positive	True Positive (TP)	False Negative 1 (FNg1)	False Neutral 1 (FNt1)
	Negative	False Positive 1 (FP1)	True Negative (TNg)	False Neutral 2 (FNt2)
	Neutral	False Positive 2 (FP2)	False Negative 2 (FNg2)	True Neutral (TNt)

Figure 4. Confusion matrix

This research uses three classes in sentiment analysis, namely positive, neutral, and negative classes. Figure 4 shows the confusion matrix with nine different combinations of predicted and actual values.

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

2.4. Classification Report

Classification report is an evaluation metrics which is a method used to evaluate the accuracy of classification algorithms [14]. Classification report used in this study consists of accuracy, precision, recall, and f1-score. Classification report aims to provide an understanding of the model performance results.

a. Accuracy

Accuracy is a value that shows the level of closeness between the system prediction value and the human prediction value [15]. In simple terms, accuracy shows how accurate the model is in classifying correctly. The following is the equation for calculating the accuracy of the model based on Figure 4.

$$Accuracy = \frac{TP + TNg + TNt}{Total\ Data} \times 100\%$$

b. Precision

Precision value is the sensitivity value or the value of the system's accuracy between the information provided by the system to correctly indicate negative class data or positive class data [15]. Precision is the ratio of true values in a class compared to all predicted values in that class. The greater the precision value, the better the algorithm's performance. The following is the equation for calculating the precision each class of the model based on Figure 4.

$$Precision\ Positive = \frac{TP}{TP + FP1 + FP2}$$

$$Precision\ Negative = \frac{TNg}{TNg + FNg1 + FNg2}$$

$$Precision\ Neutral = \frac{TNt}{TNt + FNt1 + FNt2}$$

c. Recall

Recall shows the level of success or specificity to retrieve information correctly about data that is negative class or positive text content [15]. Recall is the ratio of the true value in a class compared to the actual value in that class. The Following is the equation for calculating the recall each class of the model based on Figure 4.

$$Recall\ Positive = \frac{TP}{TP + FNg1 + FNt1}$$

$$Recall\ Negative = \frac{TNg}{TNg + FP1 + FNt2}$$

$$Recall\ Neutral = \frac{TNt}{TNt + FP2 + FNg2}$$

d. F1-score

F1-score is a combination of precision and recall function values with the aim of getting a balance value of precision and recall [16]. A good F1-score indicates that the classification model has good precision and recall, and vice versa. The minimum value of f1-score is 0 and the maximum is 1, the higher the f1-score value, the better the quality of the model. The following is the equation for calculating the f1-score of the model based on Figure 4.

$$F1score = 2 \times \frac{recall \ \times precision}{recall + precision}$$

2.5. Research Phases

This section describes the research steps or experimental methods equipped with measurable results and indicators of success, data analysis, experiments conducted and conclusions of the research results.

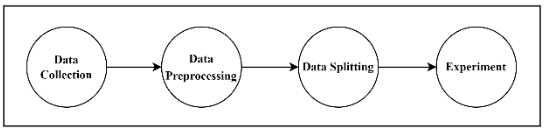

Figure 5. Research phases

Figure 5 shows all the phases of research conducted in this study. All these phases will be carried out to implement the ABSA technique using the IndoNLU pre-trained BERT model.

2.6. Data Collection

During this phase, an extensive compilation of Indonesian stock news data was meticulously gathered to serve as the foundational dataset for both training and testing purposes. This comprehensive information amalgamates findings sourced from an array of distinguished online news platforms, each specializing in finance-related reports. The diverse selection of online news sources includes reputable platforms such as CNBC Indonesia, Bisnis, Indopremier, Investor Daily, and Kontan. This amalgamation ensures a rich and varied dataset, encompassing multifaceted perspectives and insights from the financial domain, facilitating a robust and holistic analysis of stock-related trends and market dynamics within the Indonesian economy.

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

Table 1. Sources of stock news data

Sources	Total
CNBC Indonesia	827
Bisnis	629
Indopremier	3,306
Investor Daily	401
Kontan	2,522
Grand Total	7,685

Table 1 shows the news data sources and amounts used in this study. The total news data collected from the five sources is 7685.

2.7. Data Preprocessing

Data preprocessing is the part of the process that processes and manipulates data so that the data format meets the needs of training and testing. The purpose of this stage is to make the data clean and ready for use. The information processed by the model is expected to be consistent so that the accuracy of the model can be optimized.

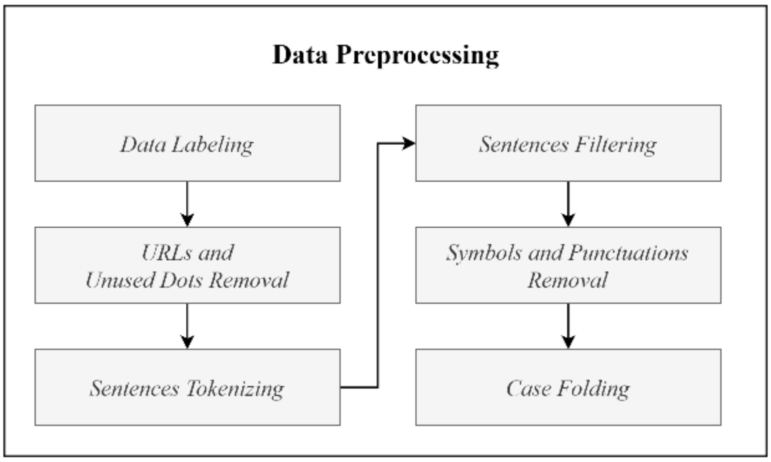

Figure 6.Data preprocessing stages

Figure 6 shows some of the steps taken in the data preprocessing stage. This stage includes data labeling, URLs and unused dots removal, sentences tokenizing, sentences filtering, symbols and punctuations removal, and case folding.

a. Data labeling

Data labeling was done using the ABSA Dataset Prepare Tools provided by the PyABSA library. Before annotation, the data is first divided into smaller pieces, as annotation tools have limitations and cannot read large amounts of data. The data that has been labeled changes its structure to Aspect Polarity Classification (APC). APC consists of three columns: text, aspect, and aspect polarity. The aspect in the news text or article will be marked with "T", in this research the aspect in question is the stock issuer.

Table 2. Total data labeled

Class	Sentences
Negative	5,692
Neutral	4,096
Positive	11,712

Table 2 shows the number of sentences that have been labeled from the three classes of negative, neutral, and positive.

b. URLs and unused dots removal

This stage is done to remove URLs and dots that have no meaning in news articles. This process is done using regular expression (RegEx) and string function techniques. Removal of dots is only done on the words "PT.", "Tbk." and numbering, other than that the dot will not be removed including the dot between numbers or rupiah.

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

c. Sentences tokenizing

Subsequently, within this stage, the articles undergo segmentation into sets of characters structured as sentences. This tokenization process becomes imperative due to the constraints inherent in the BERT model framework, which is characterized by a token length limit of 512 characters. By segmenting the textual content into manageable units, this approach ensures compatibility with the model's token limitations, allowing for comprehensive analysis while circumventing the constraints posed by the model's input length limitations. This segmentation strategy enables the effective utilization of the BERT model's capabilities, ensuring optimal processing of the data without compromising its comprehensiveness.

d. Sentence filtering

Within this stage, an integral facet of the data cleaning process involves filtering the sentence tokenization results. The removal process specifically targets sentences devoid of an aspect tag, denoted by the marker "T". By eliminating sentences lacking aspect tags, the primary aim is to eradicate instances that lack substantial meaning or relevance within the context of the dataset. This meticulous curation ensures the retention of sentences imbued with pertinent aspects, enhancing the dataset's quality by discarding instances that may not contribute meaningfully to the overarching analysis or training of the model.

e. Symbols and punctuations removal

During this elimination process, symbols and punctuation marks devoid of inherent meaning within the sentence structures are systematically removed. These elements, while often present in textual content, may not contribute semantically to the context and can potentially impede the quality of the dataset during training and testing phases. Furthermore, the inclusion of punctuation marks and symbols as distinct tokens could adversely impact the data's quality by introducing unnecessary noise or disrupting the coherent flow of information. Thus, their removal ensures a more refined dataset, conducive to effective training and testing processes, by focusing solely on the meaningful textual content essential for accurate model comprehension and prediction.

f. Case folding

Case folding is the process of uniforming the word form in the sentence by transforming all letters in the sentence into lowercase letters. This is done in order to avoid words that are not issuers but are detected as issuers. For example, "JAKARTA" contains the word "ARTA" which can be detected as an issuer, even though the word is not actually an issuer.

2.8. Data Splitting

In this study, the data is divided into two types, namely data from the training and testing process. The general ratio used is 80% for the training process and 20% for the testing process [17], so the ratio is 80:20. After that, the data balancing process is carried out to balance the amount of training data for each class.

The total data before the balancing process is unbalanced with the amount of data in the positive class of 11,707, the neutral class with 4,096, and the negative class with 5,629. After that, the data balancing process is carried out to balance the amount of training data for each class. The data balancing technique used is undersampling of the neutral class, which reduces the amount of data in the positive and negative classes so that the amount is close to the neutral class.

Table 3. Total splitted data

Class	**Train**	**Test**	**Total**
Negative	3,262	806	4,068
Neutral	3,260	808	4,068
Positive	3,269	799	4,068

Table 3 shows the distribution of data for the training and testing process for each class, the data distribution is balanced.

2.9. Experiment

The experimental model necessitates specific hyperparameter values for conducting experiments as a reference. It is essential to note that this study entails a singular simulation, and the selected hyperparameter combination, namely batch size of 8, learning rate of 0.00002, and 8 epochs, is determined as the optimal configuration according to the insights gained from the referenced research [17]. The experimental process was conducted using a server with the specifications shown in Table 4.

.

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (151-160)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

Table 4. Server spesification

Component	Detail
OS	CentOS Linux
CPU Cores	64
RAM	1 TB
GPU	Tesla T4

In this phase the accuracy of the IndoNLU model was analyzed. The analysis examines the deeper performance of the model used in the ABSA case for stock news in Indonesia. Then followed by a discussion or narrative explanation of the model analysis results obtained.

3. RESULTS AND DISCUSSION

The previous research [6] conducted sentiment analysis of stock news and provided relevant information for decision making in the stock market. This research refines the model by manually labeling 582 articles from several news sources as positive, neutral and negative. After the BERT model was refined, the resulting accuracy was 82.5%, precision was 75%, recall was 71.3%, and f1-score was 72.5%. The accuracy of the BERT model generated from this research can still be improved for the better. In addition, this research has not been able to conduct sentiment analysis specifically on one particular issuer.

The difference between the previous research [6] and this research lies in the type of pre-trained model used and the downstream task to be carried out. In the previous research, the model used is a BERT-base that has not been specialized for Indonesian, while in this study using monolingual BERT models specifically for Indonesian, namely IndoNLU. Then the downstream task carried out in previous research [6] is a comprehensive sentiment analysis of stock news, while in this study further sentiment analysis techniques are carried out, namely ABSA which can conduct specific sentiment assessments of each aspect which is stock issuers in stock news and then classified each stock issuers in the news as positive, neutral or negative.

After all the experiments have been carried out, one ABSA model is produced. The model has an evaluation metrics value as a benchmark to assess how well the model predicts the results in each class and overall.

Table 5. Experiment results

Class	Evaluation Metrics			
	Precision	Recall	F1-Score	Accuracy
Negative	0.94	0.95	0.95	0.90
Neutral	0.89	0.89	0.89	
Positive	0.88	0.87	0.87	

Table 6. Evaluation metrics average

Precision	Recall	F1-Score	Accuracy
0.90	0.90	0.90	0.90

Table 5 shows the evaluation metrics value of the experimental model. The evaluation metrics used are precision, recall, and f1-score and accuracy. The results show that the model can predict sentiment in each class well.

Table 6 shows the average evaluation metrics value of the experimental model. The results show that the model has an average value of precision, recall, f1-score, and accuracy of 90%. Overall, the model performed well in performing the sentiment analysis task.

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

CONCLUSION

ABSA is a technique that can be implemented to improve the quality of sentiment analysis on stock news and to extract news specifically on the aspects (issuers) in the news. IndoNLU is a pre-trained Indonesian BERT model that can be implemented for ABSA tasks on Indonesian stock news. The results showed that a combination of hyperparameters consisting of batch size (8), learning rate (0.00002), and epoch (8) for the IndoNLU model can produce an average evaluation metrics value consisting of precision, recall, f1-score, and accuracy of 90%.

In conclusion, this study underscores the pivotal role of utilizing pre-trained BERT models tailored to the language of the data in enhancing model quality. As exemplified in this research, the employment of the language-specific pre-trained BERT model, IndoNLU, for sentiment analysis of Indonesian stock news demonstrates notable accuracy in its results. This underscores the significance of linguistic precision in achieving more accurate and relevant outcomes in sentiment analysis within the specific context of the Indonesian language. The findings affirm that the linguistic alignment of the model with the language of the data is a critical factor contributing to the success of sentiment analysis models.

In this study, the hyperparameter combination of batch size, learning rate, and epochs used to train the IndoNLU model was determined based on previous research [17]. However, for future research, it is recommended to employ hyperparameter optimization techniques capable of identifying the precise hyperparameter combination to yield improved evaluation metric values. There exist diverse hyperparameter optimization techniques applicable for enhancing machine learning models. Implementing these techniques can aid in discovering more optimal hyperparameter settings, allowing the model to better adapt to data and significantly enhance its performance. Therefore, future research could leverage those techniques to explore a variety of hyperparameter combinations to enhance accuracy, stability, and generalization of the developed model.

ACKNOWLEDGEMENTS

We extend our heartfelt appreciation to Politeknik Negeri Bandung for their invaluable financial support, which has been instrumental in enabling the progress, authorship, and publication of this research. Their generous contribution played a pivotal role in the successful culmination of our study, and their unwavering support remains deeply appreciated.

REFERENCES

[1] I. M. Adnyana, *MANAJEMEN INVESTASI DAN PORTOFOLIO*. Jakarta Selatan: Lembaga Penerbitan Universitas Nasional (LPU-UNAS), 2020.

[2] PT Kustodian Sentral Efek Indonesia, “KSEIdoscope: 25 Tahun Kiparah KSEI Membangun Kemajuan Pasae Modal,” 2022.

[3] H. B. Jaiyeoba, A. A. Adewale, R. Haron, and C. M. H. Che Ismail, “Investment decision behaviour of the Malaysian retail investors and fund managers: A qualitative inquiry,” *Qualitative Research in Financial Markets*, vol. 10, no. 2, pp. 134–151, 2018, doi: 10.1108/QRFM-07-2017-0062.

[4] Y. Nishi, A. Suge, and H. Takahashi, “Construction of news article evaluation system using language generation model,” in *Smart Innovation, Systems and Technologies*, Springer, 2020, pp. 313–320. doi: 10.1007/978-981-15-5764-4_29.

[5] G. Ang and E.-P. Lim, “Investment and Risk Management with Online News and Heterogeneous Networks,” *ACM Transactions on the Web*, Jan. 2023, doi: 10.1145/3532858.

[6] M. G. Sousa, K. Sakiyama, L. D. S. Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, “BERT for stock market sentiment analysis,” in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, IEEE Computer Society, Nov. 2019, pp. 1597–1601. doi: 10.1109/ICTAI.2019.00231.

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.33791

[7] M. M. Abdelgwad, T. H. A. Soliman, A. I. Taloba, and M. F. Farghalya, "Arabic aspect based sentiment classification using BERT," Jul. 2021, doi: 10.48550/arXiv.2107.13290.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[9] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[10] V. Atliha and D. Šešok, "Text augmentation using BERT for image captioning," *Applied Sciences (Switzerland)*, vol. 10, no. 17, Sep. 2020, doi: 10.3390/app10175978.

[11] A. Vaswani *et al.*, "Attention Is All You Need," 2017.

[12] Y. Zhang, F. Zhang, and N. Chen, "Migratable urban street scene sensing method based on vision language pre-trained model," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113. Elsevier B.V., Sep. 01, 2022. doi: 10.1016/j.jag.2022.102989.

[13] A. Rezki Purnajaya, R. Jelita, E. Tesvara, M. Nestelrody, and J. Irwansyah, "Penerapan Metode Radial Basis Function (RBF) dalam Mengklasifikasikan Penyakit Demam Berdarah," 2023.

[14] M. H. Mori Hovipah, E. Hearani, J. Jasril, and F. Syafria, "Klasifikasi Clickbait Menggunakan Transformers," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 1, pp. 172–181, Apr. 2023, doi: 10.37859/coscitech.v4i1.4713.

[15] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 640, Apr. 2021, doi: 10.30865/mib.v5i2.2937.

[16] K. M. Rani, "Diabetes Prediction Using Machine Learning," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 294–305, Nov. 2020, doi: 10.32628/CSEIT206463.

[17] S. M. Isa, G. Nico, and M. Permana, "INDOBERT FOR INDONESIAN FAKE NEWS DETECTION," *ICIC Express Letters*, vol. 16, no. 3, pp. 289–297, Mar. 2022, doi: 10.24507/icicel.16.03.289.