

Utilizing RoBERTa and XLM-RoBERTa pre-trained model for structured sentiment analysis

Nikita Ananda Putri Masaling¹, Derwin Suhartono²

¹Department of Computer Science, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

²Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Feb 26, 2024

Revised Aug 21, 2024

Accepted Aug 27, 2024

Keywords:

Opinion tuples

RoBERTa

Structured sentiment analysis

XLM-RoBERTa

ABSTRACT

The surge in internet usage has amplified the trend of expressing sentiments across various platforms, particularly in e-commerce. Traditional sentiment analysis methods, such as aspect-based sentiment analysis (ABSA) and targeted sentiment analysis, fall short in identifying the relationships between opinion tuples. Moreover, conventional machine learning approaches often yield inadequate results. To address these limitations, this study introduces an approach that leverages the attention values of pre-trained RoBERTa and XLM-RoBERTa models for structured sentiment analysis. This method aims to predict all opinion tuples and their relationships collectively, providing a more comprehensive sentiment analysis. The proposed model demonstrates significant improvements over existing techniques, with the XLM-RoBERTa model achieving a notable sentiment graph F1 (SF1) score of 64.6% on the OpeNER_{EN} dataset. Additionally, the RoBERTa model showed satisfactory performance on the multi-perspective question answer (MPQA) and DS_{UNIS} datasets, with SF1 scores of 25.3% and 29.9%, respectively, surpassing baseline models. These results underscore the potential of this proposed approach in enhancing sentiment analysis across diverse datasets, making it highly applicable for both academic research and practical applications in various industries.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Nikita Ananda Putri Masaling

Department of Computer Science, BINUS Graduate Program, Master of Computer Science

Bina Nusantara University

Jakarta, 11480, Indonesia

Email: nikita.masaling@binus.ac.id

1. INTRODUCTION

Humans share their thoughts on various subjects, such as products, events, services, and more [1]. These opinions, also known as sentiments, significantly influence human behaviour, decisions, and perceptions of the external world [2]. The widespread use of communication media has further fuelled the expression of opinions. The internet, with 5.158 billion users globally in January 2023 [3], provides numerous platforms for expressing opinions. This ease of access aligns with the abundance of online platforms allowing users to express their views, including e-commerce sites featuring product reviews and ratings. Product reviews play a crucial role in the sustainability of businesses, influencing consumer decisions. Online user reviews have replaced traditional business-to-consumer communication regarding product quality [4], [5]. The quantity and quality of user reviews impact online consumer behaviour [6]. While e-commerce platforms use star ratings, the limitations of this system, such as incomplete reviews or

misleading star ratings, pose challenges for both businesses and consumers. Thus, sentiment analysis is still essential to this day.

Sentiment analysis presents numerous opportunities in the era of big data, but it still has significant challenges [7]. Studies on sentiment analysis examine opinions, sentiments, evaluations, attitudes, and emotions in text and classify them into polarities, such as positive, negative, or neutral [8]-[10]. While early sentiment analysis annotation works annotated for fine-grained sentiment [11], [12], most of research on sentiment analysis either relies on simplified and idealized tasks, such as sentence-level polarity classification [13], or focuses on a variety of sub-tasks that avoid performing the full task, such as targeted [14], [15], aspect-based [16]-[18], or end-to-end sentiment [19]. Aspect based sentiment analysis (ABSA) is one of the fine-grained sentiment analysis methods. ABSA focuses on specific entities and aspects and provides valuable insights for consumers and businesses. Fine-grained sentiment analysis, however, faces challenges, with research arguing against its division into sub-tasks due to a lack of sensitivity to overall sentiment resolution [20].

Research efforts, exemplified by the SemEval-2022 structured sentiment analysis (SSA) shared task [21], continue to explore detailed sentiment analysis. The process of SSA involves identifying and analysing all opinion tuples, denoted as $O=O_1, \dots, O_n$, inside a given text s . Specifically, each opinion O_i is represented by a quadruple (h, t, e, p) , which indicates a holder expressing a polarity that belongs to the set {Positive, Neutral, Negative} towards a target using a sentiment expression. It is important to mention that the variables h , t , and e might be empty in this work. In accordance with a study about SSA, tuples with empty values are considered implicit opinions [22]. Figure 1 shows an example of structured sentiment graph. Even though this study focused on SSA for tuple extraction, the classic polarity classification was also conducted.

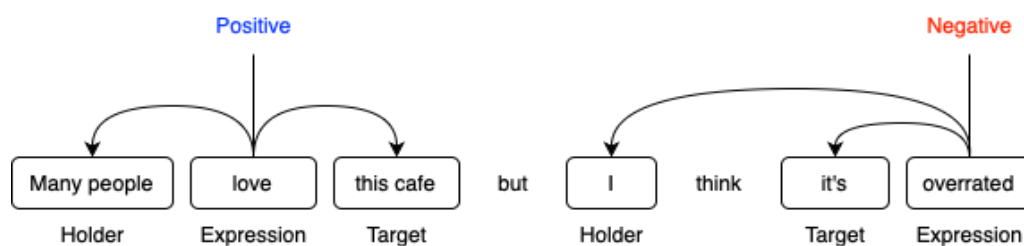


Figure 1. An example of structured sentiment analysis

Sentiment analysis can usually be done using traditional machine-learning methods [23]. In supervised methods, previous studies have employed various supervised machine learning techniques such as support vector machines (SVM), logistic regression, and Naïve Bayes [24]-[26]. Unsupervised methods include various techniques that leverage sentiment lexicons, grammatical analysis, and syntactic patterns [17], [27], [28]. However, the performance of sentiment analysis using traditional machine learning methods still faces challenges [29]. Sentiment analysis research using transformer models gained popularity in 2018 with the release of bidirectional encoder representations from transformers (BERT) model [30]. BERT can capture complex text patterns and enhance natural language processing (NLP) capabilities through transfer learning for various NLP applications, including sentiment analysis. BERT excels in learning vector representations of natural language, adapting to various text domains, utilizing contextual information, and being easy to fine-tune. Several improvements and adjustments have been made for specific NLP tasks. Some developments from BERT include robustly optimized BERT pretraining approach (RoBERTa) and cross-lingual language model of RoBERTa (XLM-RoBERTa or XLM-R). These models have several advantages over the standard BERT model, particularly in the context of sentiment analysis. RoBERTa optimizes the BERT model by training it longer, using larger datasets, and dynamically changing the masking patterns in the training data. These advantages allow RoBERTa to better understand language context, which is crucial in sentiment analysis [31], [32]. Meanwhile XLM-RoBERTa, represents a further development of the RoBERTa model, specifically designed for multilingual tasks. XLM-RoBERTa is trained on 100 different languages, including lesser-used languages, making it highly effective for sentiment analysis across various languages. Consequently, XLM-RoBERTa can deliver superior results in sentiment analysis for non-English texts compared to the standard BERT model [33], [34].

Motivated by this, the study proposes a method of SSA using the RoBERTa and the XLM-RoBERTa model, aiming to predict all opinion tuple elements and their relationships comprehensively. The choice of this topic is motivated by the inadequacies of conventional machine learning methods in sentiment

analysis and the potential of RoBERTa and XLM-RoBERTa to address these challenges. Theoretical and practical reasons underscore the significance of this research within the field and its potential application.

2. RELATED WORKS

The concept of SSA originated from early research about sentiment tuples. Many studies in this field were motivated by the corpus compiled by Wiebe *et al.* [11], which annotated English news documents with sentiment holder, target, expression, intensity, and other variables. Previous research has successfully identified individual components of sentiment in text, but none have been able to analyze the complete structure of an opinion. SSA aims to fill this gap by analyzing the entire opinion structure simultaneously. One study that became the baseline in SemEval 2022 Task 10 as well as this research, used a neural graph parsing model adopted from the neural parser by Dozat and Manning [35]. This parser was trained to score each possible arc and then predict the output structure. The basis of the network structure is a bidirectional LSTM (BiLSTM), POS tag embedding, lemma embedding, and character embedding created by LSTM. In its experiments, the token representation was also added with previously trained contextualized embedding from multilingual BERT. The results achieved were a sentiment graph F1 (SF1) score of 52.1% for the OpeNEREN dataset, 12.5% for the multi-perspective question answer (MPQA) dataset, and 20.4% for the DSUnis dataset [20].

A token graph encoding approach was also proposed for SSA. This research uses several layers in its framework architecture. The first is the encoder layer to produce contextualized word representations from the input sentence using convolution neural networks (CNN) to produce character-level embedding, which is then encoded vectorial token representation into contextualized word representation using BiLSTM. Next is the graph layer where it generates an attention scoring matrix based on the prediction of all labels. Then it goes to the multi-hop reasoning layer and refines the token representation. Finally, the prediction layer is used to consider relationships in essential labels and then decode all opinion tuple components [36]. Another study proposed the CBERT method for performing SSA. The first layer in the system consists of an encoder layer that utilizes the mBERT model. The next layer is a concatenation layer that also uses a word-piece-based BERT vector. After that, the process enters the attention layer and calculates its loss function. The final layer, the decoding layer, will produce opinion pairs or tuples [37]. Sequence labelers are used to train a model to extract sub-elements (holder, target, expression), then tried to classify whether these sub-elements have a relationship or not. Three BiLSTM models were trained separately to extract each holder, target, and expression. Then a relation prediction model was trained, which uses BiLSTM + max pooling to create a contextual representation of the full text, the first element (holder or target), and sentiment expression. The three tuples are then combined and sent to a linear layer, followed by a sigmoid function [38].

Overall, the studies above show various approaches and techniques used to improve SSA in text processing. The results achieved represent advances in understanding and extracting sentiment better from the given text. Studies mentioned also highlight performance of RoBERTa and XLM-RoBERTa model. However, there are still challenges to be overcome to achieve better results in SSA both in terms of accuracy and techniques in implementing the SSA.

3. METHOD

This section outlines the datasets, methodology, and evaluation metrics used in our study, which involves experiments on three datasets: OpeNER_{EN}, MPQA, and DS_{Unis}. The methodology incorporates advanced machine learning techniques for SSA, utilizing the node extractor and edge predictor modules. Performance is measured using the sentiment graph F1 score, assessing the model's ability to accurately detect and classify sentiment structures within diverse texts.

3.1. Datasets

The data used in this study are sourced from three datasets as part of SemEval-2022 Task 10: SSA [21]. These datasets include the OpeNER_{EN} corpus [39], the multi-perspective question answer corpus [11], and the Darmstadt Universities corpus [40]. The OpeNER_{EN} corpus comprises hotel reviews in English. Collected from various booking sites between November 2012 and November 2013, it provides diverse sentiments across languages. The MPQA corpus features English news text. The Darmstadt Universities (DS_{Unis}) corpus originated from the darmstadt service review corpus (DSRC) and contains reviews about online universities and services. All the datasets mentioned are annotated with sentiment tuples like opinion expressions, holders, targets, polarity, and opinion strength.

These datasets provide diverse annotated data for SSA. The study aims to develop and evaluate a SSA system using the pre-trained RoBERTa and XLM-RoBERTa model. Table 1 provides an overview of

the three datasets used in this study: OpeNER_{EN}, MPQA, and DS_{Unis}. The table shows the number of sentences (#), opinion holders (Hold), opinion targets (Targ), and opinion expressions (Exp) in each dataset. OpeNER_{EN} corpus contains 2,492 sentences, with 413 opinion holders, 3,843 opinion targets, and 4,149 opinion expressions. MPQA corpus is larger than OpeNER_{EN}, with 10,048 sentences, 2,265 opinion holders, 2,437 opinion targets, and 2,794 opinion expressions. DS_{Unis} corpus is the smallest of the three, with 2,803 sentences, 94 opinion holders, 1,601 opinion targets, and 1,082 opinion expressions. Overall, the table shows that the MPQA corpus is the largest in terms of sentence count, while the OpeNER_{EN} corpus has the most opinion targets and opinion expressions. The DS_{Unis} corpus is the smallest in all categories.

Table 1. Dataset statistics

	Sentiment	Holder	Target	Expression
	#	#	#	#
OpeNER _{EN}	2,492	413	3,843	4,149
MPQA	10,048	2,265	2,437	2,794
DS _{Unis}	2,803	94	1,601	1,082

3.2. Proposed method

The model comprises two base pre-trained model, RoBERTa and XLM-RoBERTa, specifically designed to extract comprehensive contextualised features. The system has two distinctive modules: the node extractor, responsible for extracting expressions, and the edge predictor, which predicts edges. RoBERTa and XLM-RoBERTa are used for all datasets such as OpeNER_{EN}, DS_{Unis}, and MPQA. Figure 2 shows how this system works. First, the system will receive text as input to the model. The node extractor then processes the input and produces tuples for each token in the text. Edge predictor then processes these tokens and determines whether there is a relationship between each token or not. The result is a set of tuples containing holder, target, and expression expressions which are predictions from this system. Next, a polarity classification task will be carried out based on the tuple expression that has been obtained from the node extractor. At this stage, it will be determined whether the sentiment is positive, negative, or neutral. In the end, the results of the tuple extraction and polarity classification tasks will be measured using the sentiment graph F1 score evaluation metric.

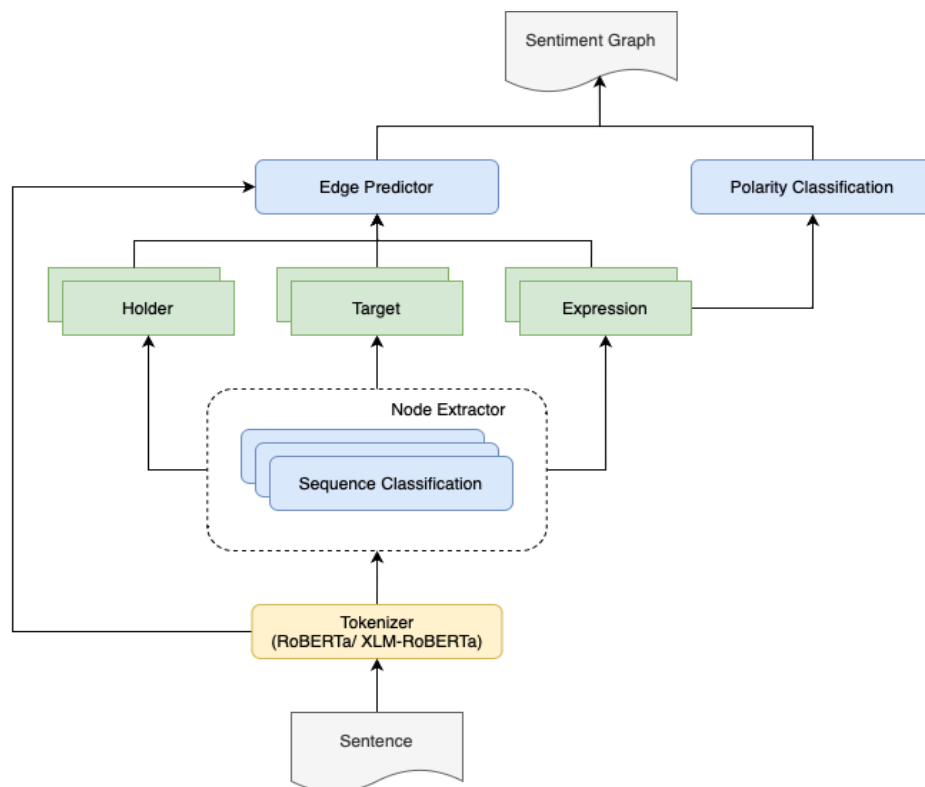


Figure 2. System architecture of proposed method

Node extractor module comprises three feedforward neural networks designed to predict the label of each token in the BIO scheme. Each network is dedicated to a certain entity type. The three networks are used to separately label sequence tokens for expression, target, and holder of sentiment. Rectified linear unit (ReLU) serves as the activation function in these networks, and each network has less than 50,000 parameters. The module uses cross entropy as its loss function. We include the prediction of emotion expression polarity into a sequence labelling task by duplicating BIO labels for negative, neutral, and positive polarities. The node extractor and the simultaneous prediction of sentiment expression and polarity are based on the strategy employed in one of our baselines, namely the one that employs a sequence labelling technique. The objective of the edge predictor was to maximise the use of the pre-trained model's information. To do this, we utilised the attention value calculated by the model to predict edges. In this approach, we calculate the total of attention values for two given phrases in a specified layer and head. We then use the sigmoid function to forecast whether there is an edge between them.

Let us consider two nodes, denoted as 'a' and 'b', which span the intervals (a_{begin} , a_{end}) and (b_{begin} , b_{end}) within a sentence, respectively. The probability of an edge existing between nodes 'a' and 'b', represented as $P_{e_{ab}}$, is computed as equation 1. A in this equation represents the attention value obtained from the pre-trained model. The variables l and h are used to specify a fixed layer and head, respectively. The sigmoid function, symbolized by σ , is used in this probability calculation. This equation suggests that the likelihood of a connection between two nodes depends on the level of attention they give to each other.

$$P_{e_{ab}} = \sigma(\sum_{i=a_{begin}}^{a_{end}} \sum_{j=b_{begin}}^{b_{end}} A_{l \times h \times ij} + A_{l \times h \times ji}) \quad (1)$$

Technically, the tuple extraction process involves several steps. First, text tokens are converted into embedding vectors through the embedding layer, which includes word embeddings, position embeddings, and token type embeddings, each with a dimension size of 768. The encoder layer comprises 12 layers, each with its attention, intermediate, and output mechanisms. Additionally, the model includes a base pooler for pooling and an edge predictor for predicting edges in the sentiment graph, with target and holder edge predictors using sigmoid activation functions. The node extractor module, which includes expression, target, and holder labelers, extracts nodes in the sentiment graph. These labelers are sequence labelers with four layers each, including two linear layers and one ReLU activation layer. Overall, the model consists of multiple layers and components working together to perform structured sentiment predictions, using activation functions like gaussian error linear unit (GELU) and ReLU, and dropout techniques for regularization.

This study, while primarily focused on the application of RoBERTa and XLM-RoBERTa for SSA, will also encompass the task of polarity classification utilizing the same models. This dual approach aims to achieve a comprehensive evaluation of RoBERTa and XLM-RoBERTa's sentiment analysis capabilities. In the polarity classification stage, the input text first enters the embedding layer, where it is converted into embedding vectors with a dimension of 768. Position embeddings and token type embeddings, each with a dimension of 768, are also applied. The embeddings are then normalized using LayerNorm and regularized with dropout. Next, the embedding vectors pass through an encoder with 12 layers. Each layer includes an attention mechanism with query, key, and value linear layers, along with dropout for regularization. The output from the attention mechanism goes through a hidden layer with a dense layer and GELU activation. This output is combined in the output layer and then passed to the classification head. The classification head consists of two linear layers and dropout for regularization, producing the final class prediction for the input text. While SSA remains the primary focus, polarity classification broadens the investigative scope and enriches the potential insights. This dual approach promises a more comprehensive understanding of RoBERTa and XLM-RoBERTa's capabilities and a deeper analysis of the complex dynamics of sentiment expression within text.

The primary measure for the task is the SF1 score, which aims to assess the extent to which a model accurately captures the whole sentiment graph. Each sentiment graph in SF1 consists of a tuple including the holder, target, expression, and polarity. A true positive is defined as a precise match at the level of the graph, considering the amount of overlap between the predicted and actual spans for each element, and then averaging this over all three h, t, and e spans. The prediction and gold answer of the problem consist of a list of quadruples $q=(q_e, q_t, q_h, q_{pol})$, where the first three entities represent sets of tokens for expression, target, and holder. Polarity refers to the sentiment or emotional tone of a text, and it may be categorised into three values: negative, neutral, and positive. The match score between two given source and target quadruples, denoted as $score(src, tgt)$. This equation is calculating a weighted match over the amount of overlap between entities, averaged across three entity types (e, t, h). The numerator consists of the sum of overlaps between source and target entities, multiplied by an indicator function that equals 1 if the polarities of source and

target are equal. The denominator is the count of source entities, but it's replaced with 1 in case it's empty. If both source and target are empty, their overlap is also set to 1.

$$Score(src, tgt) = \frac{\sum_{s \in \{e, t, h\}} \frac{|src_s \cap tgt_s|}{|src_s|}}{3} \times 1\{src_{pol} = tgt_{pol}\}$$

After calculating the match score of two given source and target quadruples, precision and recall need to be calculated to obtain the main metric of SF1 score. To calculate precision for N input sentences, use $pred_n$ as the list of projected quadruples for the n -th sentence and $gold_n$ as its gold equivalent.

$$Precision = \frac{\sum_{n=1}^N \sum_{p \in pred_n} \max_{q \in gold_n} score(p, q)}{\sum_{n=1}^N |pred_n|}$$

This metric measures how many of the predicted relationships between words (edges) are correct, considering all sentences (N) in the input. While recall is a measure of how many of the correct pieces of sentiment information were predicted by the system. Equation below shows the calculation of recall.

$$Recall = \frac{\sum_{n=1}^N \sum_{q \in gold_n} \max_{p \in pred_n} score(q, p)}{\sum_{n=1}^N |gold_n|}$$

Equation below is used to calculate the SF1 score. The F1 score is a single number that balances precision and recall. It's calculated as the harmonic mean of precision and recall.

$$SF1 = \frac{2 \times precision \times recall}{precision + recall}$$

This SF1 metric provides an in-depth understanding of the model's ability to describe the overall structure of sentiment, considering all the key elements in the sentiment graph. Using these metrics, this research attempts to measure and compare the performance of the proposed model with previous models in various subtasks of sentiment analysis.

4. RESULTS AND DISCUSSION

This section describes the experimental setup of the system, which utilizes RoBERTa and XLM-RoBERTa pre-trained models for SSA. The experiments were conducted on three different datasets: OpeNER_{EN}, MPQA, and DS_{Unis}. The results of these experiments are presented and discussed in charts and tables.

4.1. Experimental setup

The results of the experiments on three datasets: OpeNER_{EN}, MPQA, and DS_{Unis} are reported. RoBERTa, a variant of BERT, has the same architecture as BERT, with 12 layers, 768 hidden units, and 12 attention heads. It was trained on a larger dataset, used dynamic masking during training, removed the next sentence prediction objective, and was trained with larger batches and longer sequences [31]. The XLM-RoBERTa also has 12 layers, 768 hidden units, and 12 attention heads. The model was pre-trained on 100 languages using the Common Crawl corpus [41]. Both models are fine-tuned on each dataset separately, using the AdamW optimizer with a weight decay of 0.01. A learning rate of 1e-4 was used, and a linear warm-up was performed in the first epoch. A step LR scheduler with a gamma of 0.1 and a step size of 9 was applied. The training framework was PyTorch [42] and Pytorch-Ignite [43]. The Experiment Tracking tool was WandB [44]. The batch size is set to 32 and the maximum number of epochs to 30 for each dataset. A fixed random seed is used for reproducibility. The model is evaluated on the test set of each dataset using the sentiment graph F1 score as the main metric. The hyperparameter set up for this work is shown in Table 2.

Table 2. Hyperparameter setup

Hyperparameter	Value
Max. epoch	30
Batch size	32
Optimizer	AdamW
Learning rate	1e-5
Step size	9
Gamma	0.1

4.2. Results

The performance of the RoBERTa and XLM-RoBERTa model are compared against graph parsing method as the baseline model [20]. On the OpeNER_{EN} dataset, XLM-RoBERTa outperformed RoBERTa model and baseline model with a holder F1 score of 94.2%, target F1 score of 70.4%, expression F1 score of 76.4%, and SF1 score of 64.6%. Even though scoring lower than XLM-RoBERTa, the RoBERTa model achieved a commendable SF1 score at 59.9%, exceeding the baseline. For the MPQA dataset, although XLM-RoBERTa's performance is not commendable, RoBERTa model was managed to surpass the baseline with a holder F1 score of 51.5%, target F1 score of 46.3%, expression F1 score of 43.5%, and SF1 score of 25.3%. RoBERTa also stood out in the DS_{Unis} task, exceeding the baseline with F1 scores of 86.2% (holder), 57.9% (target), 36.3% (expression), and 29.9% (SF1). These results, showed in Table 3, underscore the efficacy of RoBERTa and XLM-RoBERTa in SSA tasks especially when applied to diverse datasets like OpeNER_{EN} while also highlighting areas for improvement in future. Table 3 shows the SF1 scores of each model for every dataset.

Figure 3 shows the graph of SF1 result for experiment using RoBERTa in Figure 3(a) and XLM-RoBERTa in Figure 3(b) model. It shows the sentiment graph F1 validation graph for both proposed models. The result of OpeNER_{EN} dataset is much higher than the other 2 datasets using the models. This can be expected because the OpeNER_{EN} dataset has a more balanced comparison of the number of sentiment sentences and tuple expressions compared to the other two datasets. So, it is more possible for the model to be trained to get more optimal results compared to other datasets. When testing the performance of a model using the sentiment graph F1 metric, the validation test shows no signs of overfitting or underfitting and the model reaches a stable point, so the process continued to performance testing using the testing dataset.

Table 3. Experiment result on main metric (model marked with * is the baseline)

Dataset	Model	SF1
OpeNER _{EN}	Graph parsing*	52.0%
	RoBERTa	59.9%
	XLM-RoBERTa	64.6%
MPQA	Graph parsing*	12.5%
	RoBERTa	25.3%
	XLM-RoBERTa	4.3%
DS _{Unis}	Graph parsing*	20.4%
	RoBERTa	29.9%
	XLM-RoBERTa	8.2%

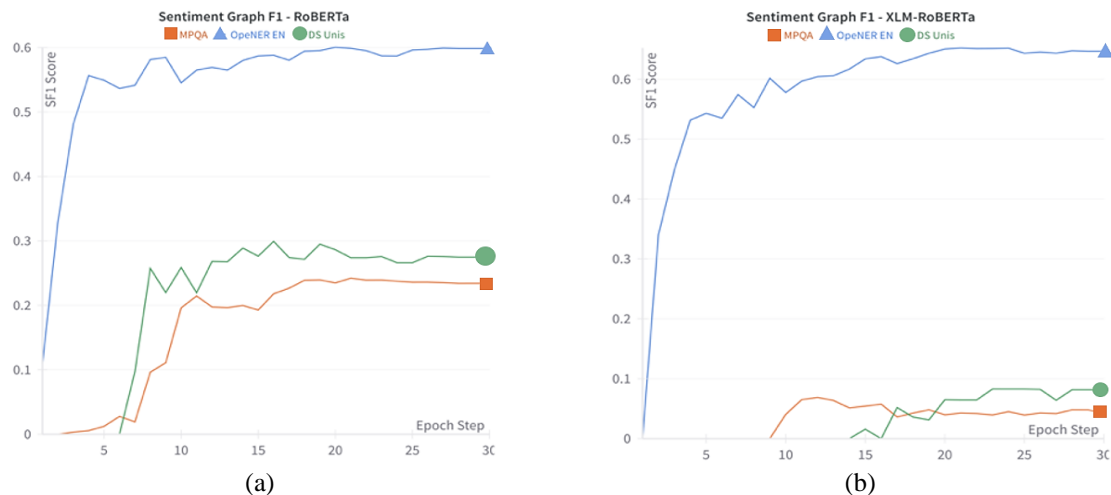


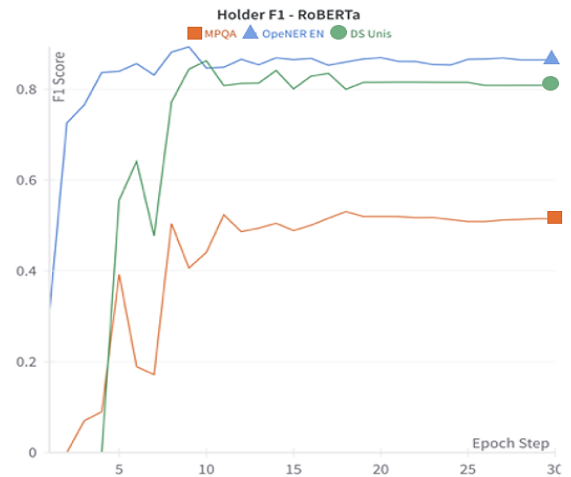
Figure 3. Sentiment graph F1 validation graph using (a) RoBERTa and (b) XLM-RoBERTa model

Table 4 presents the results of the opinion tuple extraction from sentiment sentences, evaluated across three different datasets: OpeNER_{EN}, MPQA, and DS_{Unis}. For each dataset, the table displays the F1 scores of the tuple extraction process for three models: RoBERTa, XLM-RoBERTa, and graph parsing, with baselines marked with an asterisk (*). This unified presentation allows for a clear comparison of the performance of each model on the respective datasets, highlighting the effectiveness of each approach in accurately extracting opinion tuples.

Table 4. Opinion tuples extraction f1 scores (model marked with * is the baseline)

Dataset	Model	Holder F1	Target F1	Expression F1
OpeNER _{EN}	Graph parsing*	-	-	-
	RoBERTa	86.4%	68.5%	77.6%
	XLM-RoBERTa	94.2%	70.4%	76.4%
MPQA	Graph parsing*	43.8%	51.0%	48.1%
	RoBERTa	51.5%	46.3%	43.5%
	XLM-RoBERTa	11.7%	14.1%	9.3%
DS _{Unis}	Graph parsing*	28.0%	39.9%	40.3%
	RoBERTa	86.2%	57.9%	36.3%
	XLM-RoBERTa	23.4%	14.2%	10.8%

From the results in Table 4, the performance of the RoBERTa model tends to be more stable on each dataset. Even though the OpenNER_{EN} dataset with XLM-RoBERTa has higher F1 score results, on the other two datasets, the RoBERTa model still performs stably well, exceeding the baseline. It should be noted that opinion tuples extraction F1 scores for the OpenNER_{EN} dataset with the baseline method are not available. Figure 4 shows the F1 score for tuple holder extraction in the RoBERTa in Figure 4(a) and XLM-RoBERTa in Figure 4(b) models using data validation, illustrating how the performance of both models changes as the number of epochs increases. The graph also highlights that with the XLM-RoBERTa model, tuple extraction performed best on the OpenNER_{EN} dataset, achieving a score of 94.2%, and outperformed the other two methods.



(a)

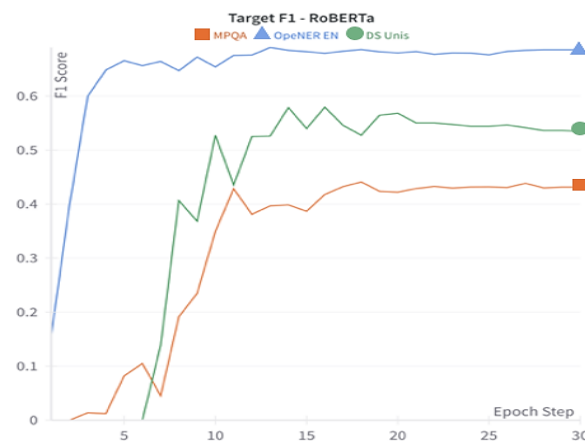


(b)

Figure 4. Opinion holder extraction validation graph using (a) RoBERTa and (b) XLM-RoBERTa

Figure 5 displays the performance of both RoBERTa in Figure 5(a) and XLM-RoBERTa in Figure 5(b) models in the target tuple extraction task using validation data. The graph shows that the RoBERTa model exhibits better performance in the early stages of training compared to XLM-RoBERTa; however, both models reach a plateau after about 15 epochs. Notably, the XLM-RoBERTa model performs best on the OpeNER_{EN} dataset, achieving a score of 70.4% and surpassing the other two methods in target extraction. Figure 6 illustrates the performance validation graph of the RoBERTa in Figure 6(a) and XLM-RoBERTa in Figure 6(b) models in the tuple expression extraction task using validation data. In the initial stages of training, the RoBERTa model outperforms XLM-RoBERTa; however, after approximately 15 epochs, the performance of both models stabilizes. Significantly, the XLM-RoBERTa model achieves the highest score on the OpeNER_{EN} dataset with a score of 77.6%, surpassing the other methods in the expression extraction task.

The polarity classification task was carried out on both proposed models, RoBERTa and XLM-RoBERTa, and their performance was compared with the baseline, as shown in Table 5. Both models consistently outperformed the baseline in sentiment polarity classification across various datasets. Although the baseline's polarity classification result for the OpeNER_{EN} dataset is not available, both proposed models performed exceptionally well on this dataset, with RoBERTa achieving an accuracy of 90.0% and XLM-RoBERTa slightly surpassing it with an accuracy of 91.0%. In the MPQA dataset, both models exhibited similar performance, with RoBERTa and XLM-RoBERTa attaining accuracies of 71.0% and 71.4%, respectively, highlighting their robustness. For the DS_{Unis} dataset, RoBERTa outperformed XLM-RoBERTa, achieving an accuracy of 83.7% compared to XLM-RoBERTa's 74.5%. This suggests that while both models are effective for polarity classification, RoBERTa may be more suited for certain datasets. In summary, both models demonstrate high accuracy in polarity classification, but the optimal model choice may depend on the specific characteristics of the dataset being analyzed.



(a)



(b)

Figure 5. Opinion target extraction validation graph using (a) RoBERTa and (b) XLM-RoBERTa

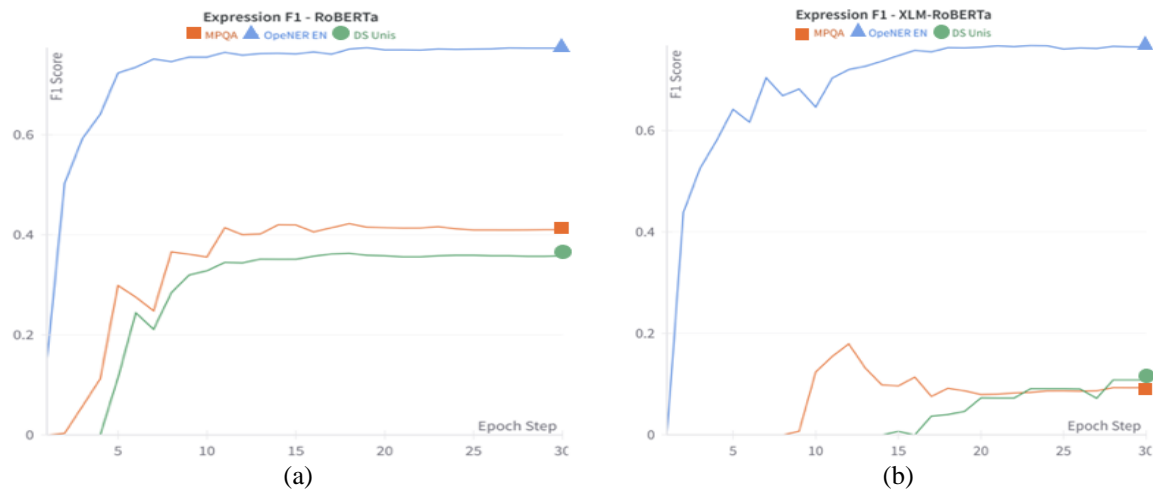


Figure 6. Opinion expression extraction validation graph using (a) RoBERTa and (b) XLM-RoBERTa

Table 5. Polarity classification accuracy

Dataset	Model	Accuracy
OpeNER _{EN}	Graph parsing*	-
	RoBERTa	90.0%
	XLM-RoBERTa	91.0%
MPQA	Graph parsing*	38.5%
	RoBERTa	71.0%
	XLM-RoBERTa	71.4%
DS _{Unis}	Graph parsing*	44.5%
	RoBERTa	83.7%
	XLM-RoBERTa	74.5%

5. CONCLUSION

The research on SSA with RoBERTa and XLM-RoBERTa models successfully conducted SSA by extracting tuples consisting of holder, target, and expression, along with sentiment polarity classification. Both RoBERTa and XLM-RoBERTa models outperformed baselines across all datasets, with XLM-RoBERTa achieving an SF1 score of 64.6% on the OpeNER_{EN} dataset, while RoBERTa scored 25.3% and 29.9% on MPQA and DS_{Unis} datasets, respectively. However, challenging datasets like MPQA and DS_{Unis} indicate that further research is needed to enhance performance due to uneven tuple distribution. Nonetheless, the proposed method showcased promising results, surpassing baselines in both tuple extraction and polarity classification tasks, suggesting its value in sentiment analysis applications. These findings suggest that RoBERTa and XLM-RoBERTa models are valuable tools for sentiment analysis applications, although further enhancements are necessary to improve performance on more complex datasets. Future work should focus on addressing the limitations identified in this study to further advance the capabilities of SSA.

REFERENCES





- [1] M. Venkataramaiah and N. Achar, "Twitter sentiment analysis using aspect-based bidirectional gated recurrent unit with self-attention mechanism," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 97–110, Oct. 2020, doi: 10.22266/ijies2020.1031.10.
- [2] B. Liu, *Sentiment Analysis and Opinion Mining*. Cham: Springer International Publishing, 2012.
- [3] S. Kemp, "Digital 2023: global overview report," 2023. [Online]. Available: <https://datareportal.com/reports/digital-2023-global-overview-report>.
- [4] R. Y. Kim, "When does online review matter to consumers? The effect of product quality information cues," *Electronic Commerce Research*, vol. 21, no. 4, pp. 1011–1030, Dec. 2021, doi: 10.1007/s10660-020-09398-0.
- [5] C. Changchit, T. Klaus, and R. Lonkani, "Online reviews: what drives consumers to use them," *Journal of Computer Information Systems*, vol. 62, no. 2, pp. 227–236, Mar. 2022, doi: 10.1080/08874417.2020.1779149.
- [6] M. R. Shihab and A. P. Putri, "Negative online reviews of popular products: understanding the effects of review proportion and quality on consumers' attitude and intention to buy," *Electronic Commerce Research*, vol. 19, no. 1, pp. 159–187, Mar. 2019, doi: 10.1007/s10660-018-9294-y.
- [7] R. Guha and T. Sutikno, "Natural language understanding challenges for sentiment analysis tasks and deep learning solutions," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 11, no. 3, pp. 247–256, Dec. 2022, doi: 10.11591/ijict.v11i3.pp247-256.
- [8] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1495–1545, Oct. 2019, doi: 10.1007/s10462-017-9599-6.

- [9] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Computing Surveys*, vol. 50, no. 2, pp. 1–33, Mar. 2018, doi: 10.1145/3057270.
- [10] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.
- [11] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2–3, pp. 165–210, May 2005, doi: 10.1007/s10579-005-7880-9.
- [12] Y. Bian, R. Ye, J. Zhang, and X. Yan, "Customer preference identification from hotel online reviews: a neural network based fine-grained sentiment analysis," *Computers & Industrial Engineering*, vol. 172, p. 108648, Oct. 2022, doi: 10.1016/j.cie.2022.108648.
- [13] J. Su, Q. Chen, Y. Wang, L. Zhang, W. Pan, and Z. Li, "Sentence-level sentiment analysis based on supervised gradual machine learning," *Scientific Reports*, vol. 13, no. 1, p. 14500, Sep. 2023, doi: 10.1038/s41598-023-41485-8.
- [14] E. Setiawan, F. Ferry, J. Santoso, S. Sumpeno, K. Fujisawa, and M. Purnomo, "Bidirectional GRU for targeted aspect-based sentiment analysis based on character-enhanced token-embedding and multi-level attention," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 392–407, Oct. 2020, doi: 10.22266/ijies2020.1031.35.
- [15] C. Gan, L. Wang, Z. Zhang, and Z. Wang, "Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis," *Knowledge-Based Systems*, vol. 188, p. 104827, Jan. 2020, doi: 10.1016/j.knsys.2019.06.035.
- [16] A. Firmanto and R. Sarno, "Aspect-based sentiment analysis using grammatical rules, word similarity and SentiCircle," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 5, pp. 190–201, Oct. 2019, doi: 10.22266/ijies2019.1031.19.
- [17] R. Abdullah, S. Suhariyanto, and R. Sarno, "Aspect based sentiment analysis for explicit and implicit aspects in restaurant review using grammatical rules, hybrid approach, and SentiCircle," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 5, pp. 294–305, Oct. 2021, doi: 10.22266/ijies2021.1031.27.
- [18] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah, "Aspect-based sentiment analysis: a survey of deep learning methods," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 6, pp. 1358–1375, Dec. 2020, doi: 10.1109/TCSS.2020.3033302.
- [19] Y. Bie and Y. Yang, "A multitask multiview neural network for end-to-end aspect-based sentiment analysis," *Big Data Mining and Analytics*, vol. 4, no. 3, pp. 195–207, Sep. 2021, doi: 10.26599/BDMA.2021.9020003.
- [20] J. Barnes, R. Kurtz, S. Oepen, L. Øvrelid, and E. Vellidal, "Structured sentiment analysis as dependency graph parsing," in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2021, pp. 3387–3402, doi: 10.18653/v1/2021.acl-long.263.
- [21] J. Barnes *et al.*, "SemEval 2022 Task 10: structured sentiment analysis," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 1280–1295, doi: 10.18653/v1/2022.semeval-1.180.
- [22] H. Zhang *et al.*, "Complete quadruple extraction using a two-stage neural model for aspect-based sentiment analysis," *Neurocomputing*, vol. 492, pp. 452–463, Jul. 2022, doi: 10.1016/j.neucom.2022.04.027.
- [23] D. Suhartono, K. Purwandari, N. H. Jeremy, S. Philip, P. Arisaputra, and I. H. Parmonangan, "Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews," *Procedia Computer Science*, vol. 216, pp. 664–671, 2023, doi: 10.1016/j.procs.2022.12.182.
- [24] N. A. P. Masaling, T. Lubis, A. Amalia, A. R. Lubis, and M. S. Lydia, "Category based sentiment analysis for basic skin-cares using LDA and SVM approach," in *2022 6th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, Nov. 2022, pp. 182–189, doi: 10.1109/ELTICOM57747.2022.10037876.
- [25] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, "Twitter sentiment analysis towards COVID-19 Vaccines in the Philippines using Naïve Bayes," *Information*, vol. 12, no. 5, p. 204, May 2021, doi: 10.3390/info12050204.
- [26] T. H. J. Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Computer Science*, vol. 197, pp. 660–667, 2022, doi: 10.1016/j.procs.2021.12.187.
- [27] M. Kaity and V. Balakrishnan, "Sentiment lexicons and non-English languages: a survey," *Knowledge and Information Systems*, vol. 62, no. 12, pp. 4445–4480, Dec. 2020, doi: 10.1007/s10115-020-01497-6.
- [28] A.-D. Vo, Q.-P. Nguyen, and C.-Y. Ock, "Semantic and syntactic analysis in learning representation based on a sentiment analysis model," *Applied Intelligence*, vol. 50, no. 3, pp. 663–680, Mar. 2020, doi: 10.1007/s10489-019-01540-2.
- [29] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knsys.2021.107134.
- [30] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Naacl-Hlt 2019*, 2018, no. Mlm, pp. 4171–4186.
- [31] Y. Liu *et al.*, "RoBERTa: a robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>.
- [32] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," *Applied Intelligence*, vol. 51, no. 6, pp. 3522–3533, Jun. 2021, doi: 10.1007/s10489-020-01964-1.
- [33] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [34] A. Kumar and V. H. C. Albuquerque, "Sentiment analysis using XLM-R transformer and zero-shot transfer learning on Resource-poor Indian Language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–13, Sep. 2021, doi: 10.1145/3461764.
- [35] T. Dozat and C. D. Manning, "Simpler but more accurate semantic dependency parsing," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 484–490, doi: 10.18653/v1/P18-2077.
- [36] W. Shi, F. Li, J. Li, H. Fei, and D. Ji, "Effective token graph modeling using a novel labeling strategy for structured sentiment analysis," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022, vol. 1, pp. 4232–4241, doi: 10.18653/v1/2022.acl-long.291.
- [37] P. Sarangi, S. Ganesan, P. Arora, and S. Joshi, "AMEX AI Labs at SemEval-2022 Task 10: contextualized fine-tuning of BERT for Structured Sentiment Analysis," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 1296–1304, doi: 10.18653/v1/2022.semeval-1.181.
- [38] K. Anantharaman *et al.*, "SSN MLRG1 at SemEval-2022 Task 10: structured sentiment analysis using 2-layer BiLSTM," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 1324–1328, doi: 10.18653/v1/2022.semeval-1.184.





- [39] R. Agerri, M. Cuadros, S. Gaines, and G. Rigau, "OpeNER: open polarity enhanced named entity recognition," *Procesamiento del Lenguaje Natural*, vol. 51, pp. 215–218, 2013.
- [40] C. Toprak, N. Jakob, and I. Gurevych, "Sentence and expression level annotation of opinions in user-generated discourse," in *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2010, pp. 575–584.
- [41] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
- [42] A. Paszke *et al.*, "PyTorch: an imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [43] V. Fomin, J. Anmol, S. Desroziers, J. Kriss, and A. Tejani, "High-level library to help with training neural networks in PyTorch," *GitHub repository*, 2020, [Online]. Available: <https://github.com/pytorch/ignite>.
- [44] Lukas Biewald, "Experiment tracking with weights and biases," 2020. <https://wandb.ai/Site/> (accessed Jan. 09, 2024).

BIOGRAPHIES OF AUTHORS



Nikita Ananda Putri Masaling     received her first degree from Computer Science Program, Universitas Sumatera Utara, Indonesia in 2020. She is currently a student of Master of Computer Science, Bina Nusantara University, Indonesia. She is also a computer science researcher. Her main research interests focus on natural language processing, data science, and artificial intelligence. She can be contacted at email: nikita.masaling@binus.ac.id.



Derwin Suhartono     received the Ph.D. degree in Computer Science from Universitas Indonesia, in 2018. He is currently a Faculty Member of Bina Nusantara University, Indonesia. His research interest includes natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a National Scientific Association in Indonesia, IndoCEISS, and Aptikom. He has his professional memberships in ACM, INSTICC, and IACT. He also takes role as a reviewer in several international conferences and journals. He can be contacted at email: dsuhartono@binus.edu.