Received: June 1, 2022. Revised: July 15, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Stance Detection on Tweets with Multi-task Aspect-based Sentiment: A Case Study of COVID-19 Vaccination

Cornelius Bagus Purnama Putra[1] **Diana Purwitasari**[1,2]* **Agus Budi Raharjo**[1]

[1]*Informatics Engineering Department, Institut Teknologi Sepuluh Nopember, Indonesia*
[2]*University Center of Excellence on Artificial Intelligence for Healthcare and Society, Indonesia*
* Corresponding author's Email: diana@if.its.ac.id

Abstract: Public opinion analyses on Twitter conducted based on sentiment analysis cannot identify the author's stance regarding agreement or disagreement with a given target. Stance detection determines whether the author of a text is in favor, against, or neutral towards a target. However, stance detection based on text-only is less representative opinion, especially on a tweet, which is a short text with slightly contextual information. Therefore, more information is needed to represent the author's stance better. In previous research, most research on stance detection was carried out using simple sentiment information to measure the support to target. This study addresses multi-task aspect-based sentiment analysis (ABSA) and social features for stance detection based on deep learning models of BiGRU-BERT on tweets. Our contribution combines aspect-based sentiment information with features based on textual and contextual information that does not emerge directly from Twitter texts. ABSA approach can provide more accurate sentiment information at aspect level on tweets, which is possible contains multiple issues discussed. Aspect information on tweets can reflect the issue that influences the author's stance toward a target. Multi-task learning was applied to help improve the generalization performance of ABSA with simultaneous processes. We extracted social attributes and online behavioral features for contextual information. Since same community tends to have the same opinion towards a target, we applied a community detection task and combine with the Twitter social attributes. The proposed method has significantly improved evaluation metrics (>10%) than textual features only for stance detection on tweets.

Keywords: Stance detection, Aspect-based sentiment analysis, Multi-task learning, COVID-19 vaccination.

1. Introduction

Twitter users typically express their opinions publicly on a given topic, product, service, or event. Public opinion analysis on Twitter is often carried out with sentiment analysis to obtain public polarity. However, sentiment analysis cannot identify and conclude the author's stance regarding agreement or disagreement with a target [1]. Stance detection identifies an author's standpoint, including in favor, against, or neutral on a predetermined target. Identifying a stance could be influenced by several aspects, including personality, culture, and social life [2], which makes it challenging for the stance detection task. Previous works on stance detection used political and ideological datasets [1]. Due to the COVID-19 pandemic, most stance detection focus on opinions toward the COVID-19 outbreak [3-5]. The COVID-19 pandemic can affect many sectors and bother public activities [6]. It is important to automatically track positions and public opinions on social media, especially on vaccine topics, as the long-term solution to the COVID-19 pandemic. Furthermore, vaccines have the potential to be refused or delayed acceptance [7]. Hesitancy to vaccines is prevalent, with skeptics in several religious, ethnic groups, and socioeconomic [8]. Therefore, stance detection on vaccination-related opinions is needed as public health surveillance to enhance public acceptance of COVID-19 vaccines.

Previous stance detection studies usually involve machine learning with elaborate feature engineering [9-13] and deep learning with word embedding as text representation [2-5]. Most of them used various features based on raw text only. However, the tweet is a short text with limited characters, making it lack contextual information sufficient for stance detection [12]. Existing stance detection systems include the social actor to represent social and online behavioral attributes, such as network features and Twitter metrics [12]. A few studies use sentiment information to measure the support of a target based on the assumption that sentiment polarity reveals the author's stance [11-13]. Based on the relationship, sentiment has an orthogonal relationship with stance [1].

Conversely, other studies show that sentiment can improve stance detection performance [11-13]. Previous research used lexicon-based sentiment information extracted from sentiment resources, such as MPQA [10], EmoLex [11], Hiu&Liu, and AFINN [12]. Each word in the text will pass to the lexicon dataset as sentiment resources to generate the sentiment attribute. Then, the sum of the numerical values will calculate as the total polarity score. The total polarity score was used as a vector representation of the text. The results showed that sentiment information can improve stance detection in some cases. However, this approach cannot capture semantic features because it does not consider the context of words in the text. Furthermore, this approach cannot handle the *Out of Vocabulary* (OOV) problem, which is a condition of token that does not appear in the vocabulary or sentiment resources. This problem can happen because domain data differences may contain new vocabulary and specific terms that are not recognized. On the other hand, this approach only identified sentiment at document levels, even though a tweet can have multiple issues or subtopics with a different sentiment. Therefore, a complete analysis must be conducted to determine the possible issue/subtopics and identify the sentiment polarity.

Aspect-based sentiment analysis (ABSA) can better mine the fine-grained emotional tendency, thus providing a sentiment polarity more accurately on the aspect level [14]. Unlike sentiment analysis at document and sentence levels, ABSA is concerned with extracting terms to specific aspects and the sentiment in the text toward aspects [15]. This research assumes that aspect information can affect the author’s stance. There are three main sub-tasks in ABSA, including aspect term extraction (ATE), aspect category identification (ACI), and aspect sentiment classification (ASC) [15]. ATE task extracts the linguistic expression that refers to aspect category. Meanwhile, the ACI task aims to identify the aspect category discussed in the text. Lastly, the ASC task aims to classify the sentiment polarity based on the identified aspect category and aspect term. Existing studies have various approaches to solving subtasks of ABSA. Several studies have implemented deep learning methods for ATE and ACI tasks [14-16]. In another way, several rule-based methods have been developed to deal with ATE task [17-21], where Part of Speech (POS) Tagging and dependency parsing was executed to identify terms related to an aspect. Recently, most studies on ABSA only focused on improving the performance of ASC tasks [14-21]. Moreover, most previous ACI and ASC are two separate tasks and could not meet practical application needs [22]. Therefore, an approach that can solve ABSA problems simultaneously is needed.

Multi-task learning is a paradigm to leverage useful information in multiple related tasks to improve the generalization performance of all tasks [22]. With multi-task learning, the ABSA task can be facilitated by sharing information between subtasks [23]. A few works conducted ABSA tasks with a multi-task learning paradigm [22-24]. The results showed that multi-task learning could simultaneously detect aspect terms and sentiment with more discriminative feature representation [24]. The evaluation results show that multi-task learning can obtain competitive performance with reduced complexity [22]. Inspired by this approach, this research carried out multi-task learning for the simply practical application of ABSA.

In this study, we propose multi-task learning ABSA for stance detection using deep learning models. The multi-task learning was applied to help improve the generalization performance of ABSA. Social attributes and online behavioral features, including linguistic structure, user profile, Twitter metrics, and user community knowledge, were used to reflect opinion better on tweets. We consider to contributed the challenges of non-English short-text classification tasks. The proposed method effectively improved stance detection performance on tweets compared to the baseline classifier.

The organization of the paper is as follows: Section 1 provides the background of our research. The explanation of the proposed method is given in Section 2. The experiments and results are discussed in Section 3. Finally, in Section 4, we conclude the study and provide plans for future work.

2. Proposed method

The proposed method is described in Fig. 1. There are four-phase, including data collection and preprocessing, ABSA, social features extraction, and

Received: June 1, 2022. Revised: July 15, 2022.

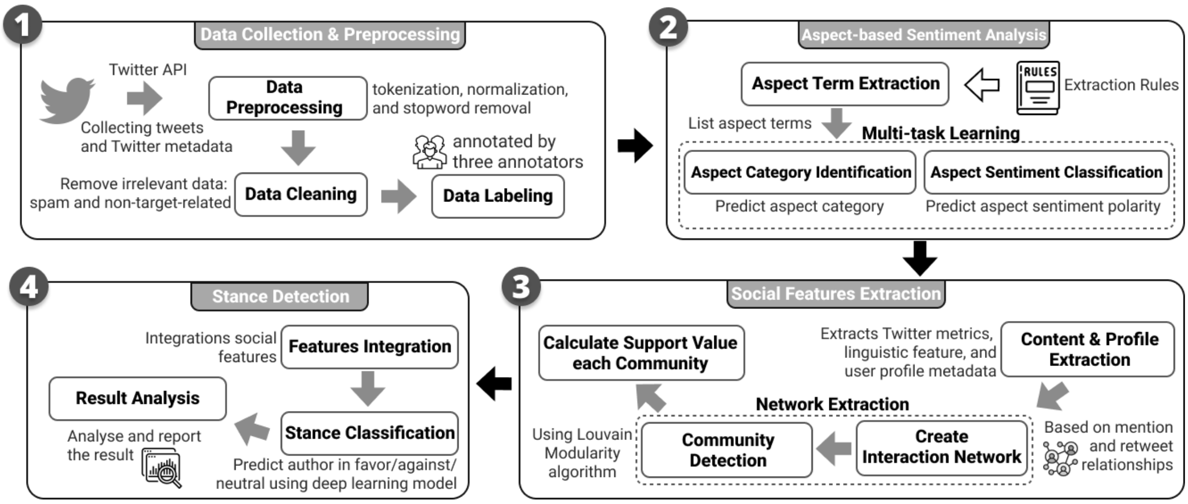

Figure. 1 Proposed method of multi-task aspect-based sentiment for stance detection on tweets

Bahasa: ***Pendaftaran vaksin***$_1$ *buat* ***lanjut usia***$_2$ *di Sleman nih aneh banget emang, keliatan banget kayak gak siap atau terkoordinasi dengan baik, mulai dari* ***link pendaftaran***$_3$ *yang banyak banget, terus* ***sistemnya***$_4$ *juga bikin bingung*

(NN ~ NN) (ADJ ~NN) (NN ~ NN) (NN)

English: **Vaccination registration**$_1$ for the **elderly**$_2$ in Sleman is strange. It looks like it's not ready or well-coordinated, including many **registration links**$_3$, and the **system**$_4$ is also confusing.

Figure. 2 The example tweet of our dataset

stance detection. The following subsections will describe the proposed method in more detail.

2.1 Data collection & preprocessing

The proposed method starts with data collection and preprocessing public opinion from Twitter. The crawling process was conducted using Twitter API services with specific keywords and parameters. A tweet usually contains informal language. Then, data preprocessing was applied, including tokenization, normalization, and stopword removal. Moreover, to remove irrelevant data (i.e., spam and non-target-related), data cleaning was carried out. First, non-target-related tweets were removed based on the collected keywords that were unrelated (i.e., non-English and non-COVID-19 vaccination-related). The duplicated tweets will be classified as spam and removed. In this work, labeled data is required for the supervised learning process. Therefore, we conduct manually annotated sample data by three independent annotators. The majority vote strategy was applied with at least two annotator agreements for the final class label on sample data. Thus, if the notation of each annotator is different, then the tweet will be categorized as an "*Invalid*" class.

2.2 Aspect-based sentiment analysis

2.2.1. Aspect term extraction

Aspect term extraction is the first step of ABSA that extracts single and multiple words that refer to an aspect category. These aspect terms extracted will be used as input for ABSA to generate the word weights. Fig. 2 shows an example tweet from our dataset and demonstrates that phrases "vaccination registration", "elderly", "registration links", and "system" are terms that express the implementation of vaccination programs. Based on Fig. 2, the extraction rules that adopt POS Tagging rules were applied, assuming that aspect terms are nouns and noun phrases [17]. Dependency Parsing is also applied to capture the grammatical relations of words [20]. The dependency relation of adjectival modifier (*amod*) and nominal subject relation (*nsubj*) was employed to extract the aspect-opinion pairs [18]. The *Stanford Parser* was applied to determine POS Tag and dependency relations.

Table 1 shows the extraction rules of the proposed method. Where, NN is a noun, ADJ is an adjective, V is a verb, and ADV is an adverb of word k in tweet j of user i. The result is a set of aspect term $AT_{ij} = \{at_1^{ij}, \dots at_m^{ij}\} | \, at_m^{ij} \in \{TW_{ij}\}$, where m is an index of aspect term and at is an aspect term that element of set tweet j of user i (TW_{ij}).

Table 1. The extraction rules for aspect term extraction

No.	Description	Rule
1	Nouns	Unigram, bigram, and trigram of NN
2	Nouns and Adjectives	Bigram of $(NN_{jk}, ADJ_{ij(k+1)})$
3	Nouns and Verbs	Bigram of $(NN_{jk}, V_{j(k+1)})$
4	Nouns and Adverbs	Bigram of $(NN_{jk}, ADV_{j(k+1)})$
5	Adjectival modifier	If $amod(NN_{jk}, ADJ_{j(k+1)})$, then extract $(NN_{jk}, ADJ_{j(k+1)})$
6	Nominal subject relation	If $nsubj(ADJ_{jk}, NN_{j(k+1)})$, then extract $(NN_{jk}, ADJ_{j(k+1)})$

2.2.2. Multi-task aspect-based sentiment analysis

The present study addresses multi-task ABSA consisting of aspect category identification (ACI) and aspect sentiment classification (ASC) tasks. We develop deep neural networks for multi-layer processing techniques. For all of them, the input will feed with a word embedding. Let text sequence TW_{ij} from tweet j of user i with k words, denoted by $TW_{ij} = \{tw_1^{ij}, \dots tw_k^{ij}\}$, and a set of aspect term AT_{ij}. Each word in tweet tw_k^{ij} and aspect term at_m^{ij} is through to the embedding layer for vector representation. Then, sequence words and aspect term vectors through to neural network model to obtain the output of hidden states.

For aspect sentiment classification, we adopt hidden states operations [16] to enhance interaction between a context and aspect terms. The row-wise maximum and average operations were applied on tweet and aspect terms vector representation. The operation is calculated as follows:

$$\beta = max(\gamma \odot 0.75) \tag{1}$$

$$\delta_1 = (avg(\theta)) - \beta \tag{2}$$

$$\delta_2 = \beta - (max(\theta)) \tag{3}$$

$$\sigma = [\delta_1 \, ; \, \delta_2] \tag{4}$$

where $\odot$ is an element-wise multiplication operator, max and avg are row-wise maximum and average operator, γ is a vector representation of aspect term, and θ is a vector representation of tweet. Finally, result of subtraction δ_1 and δ_2 are concatenated as a vector σ. Then, vector σ through to *sigmoid* and *softmax* layer for predict aspect sentiment P_{ij}, where P_{ij} comes from set $\{POS, NEG\}$ with POS is positive and NEG is negative class.

Meanwhile, for aspect category identification, *GlobalMaxPooling* layer was applied to reduce the dimensionality of tweet representation. Then, the output of *GlobalMaxPooling* layer and vector σ are concatenated. Finally, the vector passed into the *linear* and *softmax* layer to predict aspect category AC_{ij}, where $AC_{ij} \in \{ac_1, ac_2, \dots, ac_t\}$ with t is index of predetermined aspect category. The overall architecture of our model is shown in Fig. 3.

2.3 Social features extraction

Social features extraction will apply to extracting social attributes and behavioral features to take advantage of contextual information by content, profile, and network features. We exploited several linguistic structures of tweets for content features and denoted by $CF_{ij} = [cf_{ij}^1, \dots, cf_{ij}^l]$, where l is index of content features [12]. Meanwhile, profile features extracted from tweet and user metadata, such as account age, count of followers, following, tweet, retweet, reply, and verified status. Profile features will be denoted by $PF_{ij} = [pf_i^1, \dots, pf_i^e]$, where e is index of profile features.

For network feature, we exploited interaction network of users based on mention and retweet relationships. The interaction network can provide a good result for stance detection [1]. Mention relationships represent the user's communication, while retweet relationships represent the agreement with others' opinions [12]. Mention network is extracted by the prefix "@" in a tweet. Meanwhile, the retweet network is extracted by the term "*RT*" at the beginning of a tweet. Louvain Modularity

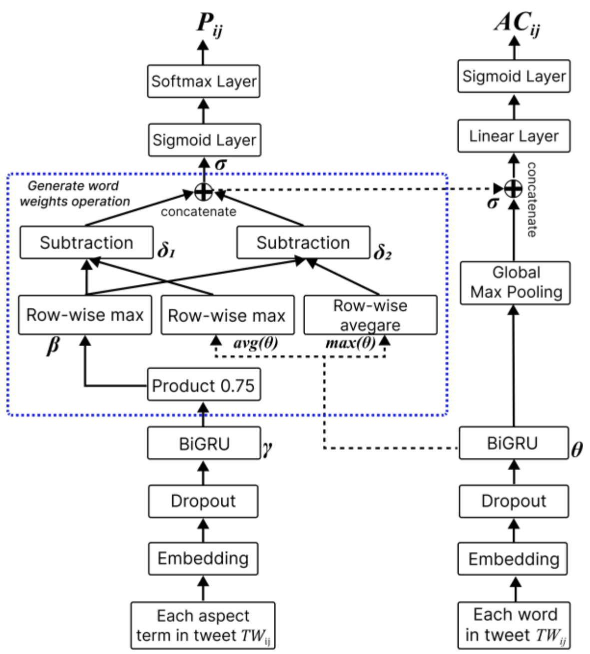

Figure. 3 Architecture multi-task ABSA of our model

Received: June 1, 2022. Revised: July 15, 2022.

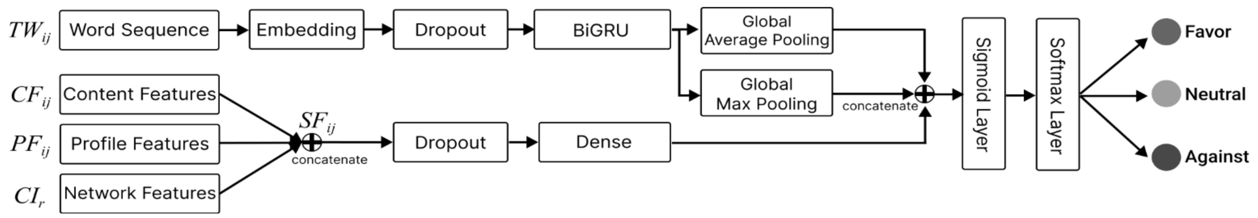

Figure. 4 Architecture of our deep learning model for stance detection

algorithm was applied to extract the community. We implemented Louvain Modularity using *networkX* with default parameters.

We assume that the same community tends to have the same attitude towards a target. Therefore, we generate community information CI_r for each class label to provide information about community by support value. Support value is percentage combinations between community with each label, including stance, aspect, and sentiment class. For instance, given user i is member of community r, and each community calculate support value of each class stance, aspect category, and aspect sentiment as explanied in the Eqs. (5) to (7). Finally, the result will be merged as a set CI_r that represent network features as explained in Eq. (8).

$$S_{stance}\left(ci_r^h\right) = \frac{count(tw_r^h)}{count(tw_r)}, |h \in \{FAV, AGA, NEU\} \quad (5)$$

$$S_{aspect}\left(ci_r^h\right) = \frac{count(tw_r^h)}{count(tw_r)}, |h \in \{ac_1, ac_2, \dots, ac_t\} \quad (6)$$

$$S_{sentiment}\left(ci_r^h\right) = \frac{count(tw_r^h)}{count(tw_r)}, |h \in \{POS, NEG\} \quad (7)$$

$$CI_r = \left[ci_r^{fav}, ci_r^{aga}, ci_r^{neut}, \left[ci_r^{ac_t}\right], ci_r^{pos}, ci_r^{neg}\right] \quad (8)$$

where, h is a class label of user i, r is community's index, ac_t is predetermined aspect category, and set $\{FAV, AGA, NEU\}$ represents of stance class label, including favor, againts, and neutral. Finally, there are 12 numerical attributes represent the support values of each class label.

2.4 Stance detection

In the present study, we address supervised learning for stance detection. There are two inputs for stance detection: word sequence of tweet (TW_{ij}) and social features, which is result from previous section (Section 2.3). Firstly, three set of social features (CF_{ij}, PF_{ij}, and CI_r) will be combined, and denoted by $SF_{ij} = \left[sf_1^{ij}, \dots, sf_z^{ij}\right]$, where z is index of element in set social features. For instance, the input is tweet j of user i with k words, denoted by $TW_{ij} = \left\{tw_1^{ij}, \dots\ tw_k^{ij}\right\}$ and social features SF_{ij}.

Then, word sequence TW_{ij} will through to embedding layer for vector transformation. Next, *GlobalAveragePooling* dan *GlobalMaxPolling* was implemented for reduce variance and overcome overfitting problem. While, the social features SF_{ij} were direct to *dense* layer, and then to *concatenate* layer for concate with tweet vectors representation. A dropout layer was also applied to overcome overfitting on social features. Finally, *dense* layer with *sigmoid* and *softmax* activation function was performs to predict stance ST_{ij}, where ST_{ij} comes from set $\{FAV, AGA, NEU\}$. Fig. 4 shows the architecture of our model for stance detection.

3. Experimental result

3.1 Data collection & preprocessing

This study used a case study of COVID-19 vaccine-related tweets in Indonesia. Indonesia is a middle-income country with relatively high vaccine hesitation and lower vaccine coverage [25].

Table 2. Overview of label distribution in our final dataset using majority vote strategy

Stance	Sentiment	Aspect Category							Total
		Implementation	Services	Costs	Participants	Apps	Vaccine	General	
Favor	Positive	579	156	129	350	181	622	1009	3026
	Negative	276	84	36	83	88	78	278	923
Against	Positive	36	2	14	31	29	61	259	432
	Negative	387	44	70	178	179	328	370	1556
Neutral	Positive	437	65	62	291	183	254	825	2117
	Negative	386	53	30	136	154	188	327	1274
Total		2101	404	341	1069	814	1531	3068	9328

Moreover, various cultures and religions in Indonesia make it attractive to investigate. The dataset was collected using Twitter API services for "*vaksinasi*" (vaccination) and "*vaksin*" (vaccine) keywords that were posted for ten months between January and October 2021. Finally, Indonesian COVID-19 vaccine-related tweets of 2,468,970 were collected. Then, data cleaning was conducted to remove irrelevant data, including non-Bahasa (Indonesia Language) and non-COVID-19 vaccination-related. We obtained the cleaned dataset of 248,604 (representing 10% of the initial dataset). It is challenging to annotate over 200k tweets. Therefore, tweets of 9,030 were selected and annotated by three annotators as sample data for modeling processes. We engaged three annotators: two researchers in *Natural Language Processing* (one MSc-level and one BSc-levels) and one communication science expert (BSc-level).

For the stance detection task, each tweet has been annotated into three classes: in *favor*, *against*, or *neutral*. The tweet is in favor or against if the author supports or opposes the target. The tweet is *neutral* class if a tweet is not in *favor* or *against* class. Meanwhile, for the ABSA task, each tweet has been annotated into predetermined aspects to the target. In this study, six predetermined aspects were used that can represent challenges and issues of Indonesian COVID-19 vaccination [25], including "*Services*", "*Implementation*", "*Apps*", "*Costs*", "*Participants*", and "*Vaccine-products*". However, there is a possibility that the author expresses an opinion without stating any aspects above. Therefore, this opinion will be categorized into the "*General*" aspect. Each aspect category will have two possible sentiment values: *positive* and *negative*. Cohen's kappa coefficient was applied to evaluate the agreement of inner-annotators.

We obtained inter-annotator Cohen's kappa coefficients of 0.6517 for stance and 0.6135 for sentiment. It means that the inter-annotator has a moderate agreement. A tweet with multi-aspects is a possibility that has a conflicting sentiment between aspects. In this study, a tweet with conflict sentiment is imbalanced (representing 1.3% of the labeled dataset). Therefore we only use tweets with same sentiment between positive or negative. Moreover, the invalid class from the majority vote strategy also will not be used. Finally, our final dataset contains 8,442 tweets, of which 3,558 are favored, 1,811 are against, and 3,073 are neutral. The distribution of labeled dataset is summarized in Table 2.

3.2 Experiment setup

We develop our proposed method using the Python Keras library. Four deep neural networks were applied, including Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), Bidirectional GRU (Bi-GRU), and Bidirectional LSTM (Bi-LSTM). The batch size of 32 is used for training models with a learning rate of 0.001 and epochs of 5. The dropout rate of 0.5 is selected to prevent the model from overfitting. Adam optimizer is used for optimizing the network with categorical cross-entropy as a loss function for stance detection and binary cross-entropy for ABSA. Three models of pre-trained word embeddings were applied: Word2Vec, FastText, and BERT. For Word2Vec and FastText, we trained on 467K documents of Indonesian Wikipedia for 300-dimensional of word representation. Meanwhile, IndoBERT model was used with a 1024-dimension word representation.

We performed *Stratified K-Fold* with 5-fold cross-validation to use 80% for learning and 20% for testing the model. Each experiment will

Table 3. Top-3 most frequently of single and multi aspect term in Bahasa and English translation

Aspect Category	List of aspect terms	
	Bahasa	English
Implementation	*anak, masyarakat, pemerintah, pelaksanaan vaksinasi, program vaksinasi, petugas vaksin*	child, society, government, implementaion of vaccination, vaccination program, vaccinator
Services	*petugas, panitia, pelayanan, petugas vaksin, panitia vaksin, setifikat vaksin*	officer, committee, service, vaccinator, vaccine committee, vaccine certificate
Costs	*biaya, pemerintah, perusahan, vaksin gratis, biaya vaksin, vaksin mandiri*	cost, government, company, free vaccine, vaccine fee, independent vaccine
Participants	*anak, usia, sekolah, orang tua, vaksin anak, pelayanan publik*	child, age, school, parents, child vaccine, public service
Aplications	*aplikasi, data, sertifikat, sertifikat vaksin, aplikasi peduli, kartu vaksin*	application, data, certificate, vaccine certificate, peduli apps, vaccine card
Vaccine-product	*anak, antibodi, efek, vaksin nusantara, vaksin haram, varian delta*	child, antibody, effect, *nusantara* vaccine, haram vaccine, delta variant
General	*anak, pemerintah, sekolah, orang tua, tolak vaksin, petugas vaksin*	child, government, school, parents, decline vaccine, vaccinator

Table 4. Experimental results for aspect category identification and aspect sentiment classification of multi-task ABSA on the proposed method

Method	Text Representation	Aspect Category Identification				Aspect Sentiment Classification			
		Acc	Macro Pre	Macro Rec	Macro F1	Acc	Macro Pre	Macro Rec	Macro F1
GRU	Word2Vec	56.87	63.40	54.39	57.26	71.57	71.93	72.00	71.04
Bi-GRU		56.13	61.82	57.39	58.75	72.36	72.38	72.69	71.87
LSTM		56.32	62.26	56.40	58.21	72.24	71.52	71.44	71.30
Bi-LSTM		57.41	62.64	56.66	58.82	72.69	72.23	72.49	72.03
GRU	FastText	55.91	61.65	52.61	55.35	72.02	71.11	70.69	70.81
Bi-GRU		54.83	62.07	52.78	55.67	71.58	71.01	70.72	70.50
LSTM		55.91	60.76	52.30	55.36	70.86	70.21	70.43	70.10
Bi-LSTM		55.68	61.54	52.93	55.95	71.80	71.20	71.18	70.92
GRU	BERT	59.78	66.00	57.70	60.41	73.92	73.32	72.96	72.87
Bi-GRU		59.13	65.05	59.98	61.35	**74.56**	**73.94**	**73.82**	**73.69**
LSTM		59.39	65.66	56.68	59.61	73.85	73.39	72.32	72.49
Bi-LSTM		**60.06**	**66.26**	**60.33**	**62.31**	74.03	73.58	72.84	72.34

evaluate by four metrics: accuracy, macro precision, macro recall, and macro F1 score. As baselines, three traditional machine learning were implemented, including Naïve Bayes, K-Nearest Neighbor, and Decision Tree. The models did not carry out any optimization procedures and used their default configuration of *scikit-learn* implementation.

3.3 Scenario: find best model of multi-task ABSA

3.3.1. Aspect term extraction

This study determined seven issues related to COVID-19 vaccination as aspect categories. Rule-based aspect extraction was implemented to identify aspect terms in a tweet. Each tweet has a list of aspect terms that relate to the aspect category. The top-3 most frequently single and multi-terms in each aspect category are shown in Table 3. Table 3 shows that the process was able to extract aspect terms for each aspect category from the dataset successfully and demonstrates that aspect terms may differ but have similar meanings [20] (i.e., "*panitia*" and "*petugas*" terms in *services* aspect is different, but in this context have the same meaning: the person in charge of the vaccination program). Although the General aspect contains terms from other aspects. Nevertheless, this method is sufficient to extract aspect terms in the other six predetermined aspects as a guide to word weighting for multi-task ABSA.

3.3.2. Multi-task learning ABSA

We report the result in Table 4 and highlight the best result in bold. We carried out two tasks of ABSA: aspect category identification (ACI) and aspect sentiment classification (ASC). BiLSTM-BERT obtained the best performance for the ACI task with an accuracy of 60.06%, macro precision of 66.26%, macro recall of 60.33%, and macro F1 score of 62.31%. However, BiLSTM-BERT is slightly different in macro F1 score (< 1%) from the second-best model, BiGRU-BERT, with a macro F1 score of 61.35%. Meanwhile, BiGRU-BERT is the best model for ASC. BiGRU-BERT obtained an accuracy of 74.56%, macro precision of 73.94%, macro recall of 73.82%, and macro averaged F1 score of 73.69%. This result is contrary to the ACI task, in which BiGRU-BERT has the lowest accuracy on the model with BERT (59.13%). Overall, the evaluation results of ASC tend to be better than ACI. We hypothesize that this is because the limited data must be classified into seven labels of aspect category. While the distribution of aspect category is imbalanced, that makes limited training data on minority aspect category. Overall, BiGRU-BERT shows good performance on both tasks.

Table 5. Evaluation of multi-task learning on the proposed method compared with single-task learning

Framework	Method	Task	
		ASC	ACI
Single-task	Naïve Bayes	59.97	28.59
	KNN	65.24	47.02
	Decision Tree	62.26	56.48
	BiLSTM- W2V	71.90	59.75
	BiLSTM- FastText	70.41	57.94
	BiGRU- BERT	72.16	61.35
Multi-task (proposed method)	BiLSTM- W2V	71.86	58.99
	BiLSTM- FastText	70.85	56.91
	BiGRU- BERT	**73.69**	**61.35**

As shown in Table 4, the result shows that BERT outperforms in word embedding models. We hypothesize that large dimensions of BERT (1024-dimensions) can better understand the text's context. Meanwhile, FastText with the OOV mechanism does not perform well. We found that some words were represented too differently by FastText. It

Table 6. Performance evaluation stance detection between the proposed method and text-only feature

Method	Text Representation	Text-only Feature				Proposed Method			
		Acc	Macro Pre	Macro Rec	Macro F1	Acc	Macro Pre	Macro Rec	Macro F1
GRU	Word2Vec	56.09	55.38	53.31	53.74	64.47	64.71	63.02	63.47
Bi-GRU		57.70	57.01	57.22	56.78	64.27	64.60	62.88	63.15
LSTM		56.12	55.36	53.97	54.23	63.80	63.92	62.53	62.58
Bi-LSTM		56.74	56.56	54.83	54.90	64.35	64.37	63.02	63.41
GRU	FastText	53.83	53.63	51.01	51.17	64.93	64.56	64.36	64.26
Bi-GRU		54.58	54.06	52.52	52.58	64.11	64.18	62.62	62.96
LSTM		54.26	53.54	51.83	52.16	64.44	64.36	63.35	63.54
Bi-LSTM		54.63	54.01	53.00	52.85	64.32	64.72	62.36	63.14
GRU	**BERT**	58.64	58.63	56.24	56.45	65.50	65.75	63.72	64.34
Bi-GRU		59.57	59.20	57.32	57.61	**66.17**	**66.41**	**65.05**	**65.26**
LSTM		59.15	58.59	56.79	57.03	65.04	64.79	63.89	64.04
Bi-LSTM		59.11	58.78	57.33	57.35	65.22	65.15	63.84	64.17

Table 7. A comparison of features set of proposed method using BiGRU-BERT

Run	Feature set	Acc	Macro Pre	Macro Rec	Macro F1
0	Baseline (text only)	59.57	59.20	57.32	57.61
1	0 + Linguistic features: emoticons count, question mark count, words count, exclamation marks count, punctuations count, characters count, words in capital letters counts, mentions count, hashtags count, and bag of hashtags	59.67	59.05	58.83	58.46
2	1 + Twitter metrics and user profile metadata: replies count, retweets countand, likes count, followers count, following count, tweets count, verified status, and account ages	59.90	59.45	58.69	58.36
3	2 + Network features: user community knowledge and bag of mention	64.76	65.03	64.81	63.92
4	**3 + Aspect-based sentiment information: bag of aspect term, aspect category, frequency sentiment of emocitons, and aspect sentiment**	**66.17**	**66.41**	**65.05**	**65.26**

indicates that the domain of documents used to train the word embedding model is also important for better representation of words [19]. Moreover, we found no indication that using a bidirectional can improve performance. Based on our experiments, bidirectional usage performs better than the regular model only in some cases.

In addition, we also explored the performance of single-task learning, which trains the model for each task separately for comparison. The difference in the network architecture lies in the output layers. There is only one output unit in single-task, while in multi-task, there are two outputs. We evaluate them based on the macro F1 score shown in Table 5. The result shows that multi-task learning obtains competitive performance over single-task. Nevertheless, multi-task ABSA can predicts aspect categories and aspect sentiment simultaneously, which is unreachable for the single-task model. Moreover, BiGRU-BERT outperformed other multi-task ABSA models, and even the single-task models aim for ACI or ASC tasks. Therefore, we further use BiGRU-BERT as multi-task ABSA models for stance detection.

3.4 Scenario: find best model of stance detection to prove the proposed method

We consider aspect-based sentiment information and social features for stance detection. Firstly, our interaction network contains edges of 9,840 and nodes of 10,299, generating a community of 3,648 based on Louvain Modularity. However, 3,431 out of 6,888 users (representing ± 50%) are not part of any community. It is because there is a limited user that has interaction relationships. Table 6 shows that BiGRU-BERT was in top performance with an accuracy of 66.17%, macro precision of 66.41%, macro recall of 65.05%, and macro F1 score of 65.26%. In our proposed methods, word embedding models did not significantly impact with a margin of around 3%. Overall, the proposed method significantly improved all evaluation metrics by over 8% for the stance detection task on tweets. Feature set analysis was conducted to investigate the features set for stance detection, as shown in Table 7. We used BiGRU-BERT as the best model of the proposed method with five types of feature sets. The result shows that the all (run 4) features set has

Received: June 1, 2022. Revised: July 15, 2022.

Table 8. Classification performance of proposed method using BiGRU-BERT

Features	Class	Predict			Evaluation Metrics			
		Favor	Neutral	Against	Pre	Rec	F1	Acc
Textual only	Favor	435	252	25	67.65	61.1	64.21	58.79
	Neutral	124	**476**	15	51.52	**77.40**	61.86	
	Against	84	196	82	**67.21**	22.65	33.88	
Proposed method w/o aspect (sentiment analysis convensional)	Favor	470	168	74	**70.78**	66.01	68.31	64.77
	Neutral	126	433	56	61.51	70.41	65.66	
	Against	68	103	191	59.50	52.76	55.93	
Proposed method	Favor	**536**	112	64	69.97	**75.28**	**72.53**	**68.32**
	Neutral	162	394	59	**68.40**	64.07	**66.16**	
	Against	68	70	**224**	64.55	**61.88**	**63.19**	

Table 9. Comparison the proposed method with other methods

Model	Acc	Macro F1
SVM-ngrams [1]	53.88	52.45
SVM-unigram-NER [1]	55.18	53.56
Two-phase SVM [10]	50.10	54.43
SVM+stylistic, structural, affective and contextual features [12]	61.59	60.37
WKNN+PCA+sentiment [13]	61.36	56.60
BiGRU+Glove+Data Distillation [3]	60.56	59.18
Proposed Methods (BiGRU-BERT)	**68.32**	**67.29**

improved the performance of the model significantly (>8%) than baseline (text only). It indicates that community information (support value of aspect-category, aspect sentiment, and stance class distribution in each community) can represent a user's stance by assuming that users tend to interact with users with the same stance [12].

To better understand the improvement, we compared the proposed method with the one without aspect information (classical sentiment), as shown in Table 8. Firstly, we used the holdout method to split the dataset into 80% for training and 20% for testing. BiGRU-BERT was applied as the best model for our previous experiment. The result shows that sentiment information can improve performance (>5%) of stance detection than textual only. Sentiment information improves the precision score of each class. Interestingly, the highest recall value was obtained only in the favor class with a textual feature. Inversely, the precision value of the favor class becomes the lowest value. However, the proposed method obtained the highest F1 score in each class. Notably, improvement is significantly on against and favor classes. For detailed analysis, the confusion matrix was demonstrated in Table 8.

The confusion matrix in Table 8 shows that the proposed method is useful in correcting neutral and favor instances misclassified as against class. Interestingly, only 22% of the dataset is in the against class, 36% in the neutral class, and 42% in the favor class (see Table 2). This condition makes against is minority class. Reflects that added aspect-based sentiment with combined social features can discern the inclination author to express the opinions. Moreover, 75% of misclassified cases are between neutral and favor classes. This happens because both classes contain similar keywords, which can mislead the model [3]. Finally, the results show that the proposed method using BiGRU-BERT can improve stance detection on tweets.

3.5 Discussion

This research used Indonesian opinion toward COVID-19 vaccination as an example of a topic that might be frequently discussed on Twitter during the crisis of the COVID-19 pandemic. From Table 4 and Table 6 show that BiGRU-BERT was the top-performing classifier for the proposed method. We use multi-task ABSA to identify aspect category and aspect sentiment simultaneously. From Table 5, we saw that multi-task learning has a competitive performance over single-task learning for ABSA. Although it has no significant impact, multi-task ABSA can reduce the complexity of ABSA because it only uses one module [22]. In our case, the aspect sentiment classification module directly depends on the aspect category identification outcome.

Going over Table 7, we note that interaction networks as network features provide significant improvements (± 6%). User community knowledge can influential independently by the type of the target of interest [12]. We also investigate the effect of sentiment information in our proposed method. Sentiment and stance have orthogonal relationships because negative opinion is not always against to target, and vice versa [1]. This relationship is proven in our data (Table 2). However, our experiments from Table 7 show that sentiment information provides more significant improvements (± 8%). Sentiment information is highly influences the stance [13]. Moreover, Table 8 shows that aspect-based sentiment information could improve by

Table 10. Error labels examples of the proposed method in Bahasa and English translation

Ex.	Tweet (in Bahasa)	English Translation	Actual	Predicted
(1)	*@detikcom Yuk vaksin lagi yuk... kan nanggung udah 2 kali vaksin. Yuk negara, beli vaksin lagi yuk. Kan yg buatan anak negri ga boleh beredar.*	@detikcom Let us get vaccinated... you have already had second vaccines. Come on, let us buy another vaccine. The ones made by the country's children cannot be circulated.	**Aspect**: Vaccine-products **Sentiment**: Negative **Stance**: Against	**Aspect**: General **Sentiment**: Positive **Stance**: Favor
(2)	*@TofaTofa_id anak siapa dia tu ya.daya imun hukum nya kuat banget. mau juga dong vaksin hukum.*	@TofaTofa_id, whose child is he? His legal immunity is powerful. Do you want legal vaccines?	**Aspect**: General **Sentiment**: Positive **Stance**: Neutral	**Aspect**: General **Sentiment**: Positive **Stance**: Against
(3)	*Diminta ayo vaksin, ayo vaksin Giliran daftar vaksin ribet e setengah mati: masih untuk pedagang, lansia, nakes, bank dll, blm ketambah kudu bawa surat domisili (luarkota)*	Asked, let's get vaccinated, let's get vaccinated. It's the turn of the vaccine list which is complicated: it's still for traders, the elderly, health workers, banks, etc., and you have to bring a domicile letter (out of town)	**Aspect**: Implementation, Participants **Sentiment**: Negative **Stance**: Against	**Aspect**: Implementation **Sentiment**: Positive **Stance**: Favor
(4)	*@patio9eneral Ijin copas jawaban ya Pak. kalau2 suatu saat saya didatangin petugas vaksin juga..🙏😭😭😭*	@patio9eneral ask for permission to use the answer, sir. if one day I am visited by a vaccine officer too..🙏😭😭😭	**Aspect**: General **Sentiment**: Positive **Stance**: Against	**Aspect**: Sevices **Sentiment**: Positive **Stance**: Favor
(5)	*Mau pulang ke Indonesia ku tercinta, tapi mesti karantina walaupun udh vaksin full. Biaya karantina ya aja udh dikit lagi harga pulang pergi pesawat nya😥 gak pulang kangen emak😥*	I want to return to my beloved Indonesia, but I have to quarantine even though I'm fully vaccinated. The quarantine fee is just a little bit more than the price for the flight 😥if I don't come home, I miss my mom😥	**Aspect**: General **Sentiment**: Positive **Stance**: Against	**Aspect**: General **Sentiment**: Positive **Stance**: Favor

around 4% more than sentiment information conventional. Aspect information can provide more accurate sentiment [21] and the details subtopic or issue the author discusses [18].

Table 9 compares our proposed model and other studies in stance detection, all using our labeled dataset. For our system, we report the performance of using the BiGRU-BERT as the best model from our experiments. We compare against current state-of-the-art methods of stance detection on tweets, e.g. [3, 12, 13], etc. The results show that the proposed method outperformed the best result reported in the state-of-the-art method (i.e. [3]) in both accuracy and macro F1 score by 7.76% and 8.16%, respectively. It can happen because the method ([3]) only detects stance based on raw text. Moreover, we also compare against current methods that use sentiment information for stance detection, e.g. [12, 13]. The results show that the proposed method achieved better performance (±7%) than current methods (namely, [12, 13]) that used sentiment information. However, the sentiment information was extracted only based on a limited lexicon dictionary that does not consider the context of words in the text [15].

To better understand, we analyzed several error cases from our proposed method. Table 8 shows that 535 tweets were misclassified, representing 32% of testing data. We summarized five reasons for misclassification. For instance, we demonstrated each reason with the example of a tweet, as shown in Table 10. Firstly, a tweet possibly contains ironic or satirical expressions that can inverse the polarity and mislead the model. In Ex. (1) in Table 10, the actual label is "*against*", but is classified as "*favor*". This tweet contains ironic expressions that represent its support, but in reality, it is satirical against vaccine policy. Moreover, this expression also misleads the model to predict aspect category and aspect sentiment.

Secondly, some tweets express irrelevant targets, as in Ex. (2) in Table 10. This tweet contains related keywords to this study, such as "*vaccine*", but the tweet does not intend toward the vaccination program contextually. Thirdly, there are similar keywords that can mislead the model. As Ex. (3) in Table 10, there is the phrase "*ayo vaksin*" ("let us get vaccinated" in English translation), which means to persuade people to get vaccinated so that it can be classified as "*favor*". However, the real meaning is to criticize the implementation of the vaccination and must classify it as "*against*". Moreover, the aspect sentiment was also misclassified as positive. It means that the phrase "*ayo vaksin*" has a supportive and positive meaning to the target.

Fourthly, lack of contextual information on tweets. We found that a tweet possible is part of a conversation/thread, as Ex. (4) in Table 10. The model cannot understand the whole conversation or thread of tweets because the tweet is a comment from another tweet [12]. Lastly, there is a possibility that the author expresses their opinion implicitly. For example, Ex. (5) in Table 10 shows a tweet that is classified as "*favor*", but if investigated deeply, the tweet should be "*against*". The author expresses his attitude against vaccines indirectly. Overall, the proposed method using BiGRU-BERT obtained higher accuracy and F1 score than using text-only.

4. Conclusion and future works

This paper discusses multi-task aspect-based sentiment for stance detection on tweets by adopting multi-task learning for extracting aspect category and aspect sentiment on tweets simultaneously. Aspect-based sentiment information and social features were used to represent user behavior and support stance detection. COVID-19 vaccination-related tweets were used and there were seven aspects that represent the concerned issues. The result shows that BiGRU-BERT in our proposed method have outperformed for ABSA and stance detection tasks with F1 score 65% for stance detection with 10% improvements.

Despite the theoretical and experimental results achievements, our study has several limitations. First, imbalanced data with moderate inter-annotator agreement could improve. Second, we only used tweets with positive and negative classes. However, there is a possibility that a tweet has a multi-aspect category with a different sentiment (conflict), which may affected the stance. Lastly, the determination of the aspect category that we used was modest. We only determined aspect categories based on sample tweets. However, in real-life scenarios, there are possible any others aspect categories.

In future work, we plan to improve performance stance detection by exploring data augmentation to overcome imbalanced data. Moreover, feature extraction for better contextual information is also needed, such as information on previous tweets, tweets in the same thread, and user connection networks to improve stance detection on tweets.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Cornelius Bagus Purnama Putra and Diana Purwitasari; *Methodology and Software*, Cornelius Bagus Purnama Putra; *Writing—original draft preparation*, Cornelius Bagus Purnama Putra; *Writing—review and editing*, Diana Purwitasari and Agus Budi Raharjo; *Supervision*, Diana Purwitasari and Agus Budi Raharjo; *Project administration*, Diana Purwitasari.

Acknowledgments

This work was supported by Indonesian Ministry of Research and Technology under Grant No. 084/E5/PG.02.00.PT/2022 with an institutional contract No. 1440/PKS/ITS/2022.

References

[1] A. Aldayel and W. Magdy, "Stance Detection on Social Media: State of The Art and Trends", *Information Processing & Management*, Vol. 58, p. 102597, 2021.

[2] L. Zhu, Y. He, and D. Zhou, "Neural Opinion Dynamics Model for the Prediction of User-Level Stance Dynamics", *Information Processing & Management*, Vol. 57, p. 102031, 2020.

[3] L. Miao, M. Last, and M. Litvak, "Tracking Social Media During the COVID-19 Pandemic: The Case Study of Lockdown in New York State", *Expert Systems with Applications*, Vol. 187, p. 115797, 2022.

[4] L. A. Cotfas, C. Delcea, I. Roxin, C. Ioanas, D. S. Gherai, and F. Tajariol, "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month Following the First Vaccine Announcement", *IEEE Access*, Vol. 9, pp. 33203-33223, 2021.

[5] M. A. Wani, N. Agarwal, and P. Bours, "Impact of Unreliable Content on Social Media Users during COVID-19 and Stance Detection System", *Electronics*, Vol. 10, p. 5, 2020.

[6] Apriantoni, H. A. Thooriqoh, C. Fatichah, and D. Purwitasari, "Topic Detection in Sentiment Analysis of Twitter Texts for Understanding The COVID-19 Effect in Local Economic Activities", In: *Proc. of 13th International Conference on Information & Communication Technology and System*, pp. 354-359, 2021.

[7] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the Public Opinion About the Vaccination Topic from Tweets Analysis", *Expert Systems with Applications*, Vol. 116, pp. 209-226, 2019.

[8] O. J. Wouters, K. C. Shadlen, M. S. Konrad, A. J. Pollard, H. J. Larson, Y. Teerawattananon, and M. Jit, "Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment", *The Lancet*, Vol. 397, pp. 1023-1034, 2021.

[9] U. Can and B. Alatas, "A Novel Approach For Efficient Stance Detection in Online Social Networks with Metaheuristic Optimization", *Technology in Society*, Vol. 64, p. 101501, 2021.

[10] K. Dey, R. Shrivastava, and S. Kaushik, "Twitter Stance Detection - A Subjectivity and Sentiment Polarity Inspired Two-Phase Approach", In: *Proc. of 2017 IEEE International Conference on Data Mining Workshop*, pp. 365-372, 2017.

[11] A. I. A. Ghadir, A. M. Azmi, and A. Hussain, "A Novel Approach to Stance Detection in Social Media Tweets by Fusing Ranked Lists and Sentiment", *Information Fusion*, Vol. 67, pp. 29-40, 2021.

[12] M. Lai, A. T. Cignarella, D. I. H. Farías, C. Bosco, V. Patti, and P. Rosso, "Multilingual Stance Detection in Social Media Political Debates", *Computer Speech and Language*, Vol. 63, p. 101075, 2020.

[13] Q. Sun, Z. Wang, S. Li, Q. Zhu, and G. Zhou, "Stance Detection via Sentiment Information and Neutral Network Model", *Frontiers of Computer Science*, Vol. 13, pp. 127-138, 2019.

[14] J. A. G. Díaz, M. C. García, and R. V. García, "Ontology-driven Aspect-Based Sentiment Analysis Classification: An Infodemiological Case Study Regarding Infectious Diseases in Latin America", *Future Generation Computer Systems*, Vol. 112, pp. 641-657, 2020.

[15] M. M. Abdelgwad, T. H. A. Soliman, A. I. Taloba, and M. F. Farghaly, "Arabic Aspect based Sentiment Analysis using Bidirectional GRU Based Models", *Journal of King Saud University - Computer and Information Sciences*, 2021.

[16] W. Ali, Y. Yang, X. Qiu, Y. Ke, and Y. Wang, "Aspect-Level Sentiment Analysis Based on Bidirectional-GRU in SIoT", *IEEE Access*, Vol. 9, pp. 69938-69950, 2021.

[17] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid Sentiment Classification On Twitter Aspect-Based Sentiment Analysis", *Applied Intelligence*, Vol. 48, pp. 1218-1232, 2018.

[18] Z. Luo, S. Huang, and K. Q. Zhu, "Knowledge Empowered Prominent Aspect Extraction from Product Reviews", *Information Processing & Management*, Vol. 56, pp. 408-423, 2019.

[19] R. Abdullah, Suhariyanto, and R. Sarno, "Aspect Based Sentiment Analysis for Explicit and Implicit Aspects in Restaurant Review using Grammatical Rules, Hybrid Approach, and SentiCircle", *Int. J. Intell. Eng. Syst*, Vol. 14, pp. 294-305, 2021, doi: 10.22266/ijies2021.1031.27.

[20] S. Poria, E. Cambria, and A. Gelbukh, "Aspect Extraction for Opinion Mining with A Deep Convolutional Neural Network", *Knowledge-Based Systems*, Vol. 108, pp. 42-49, 2016.

[21] M. K. A. Venkataramaiah and N. A. N. Achar, "Twitter Sentiment Analysis using Aspect-Based Bidirectional Gated Recurrent Unit with Self-Attention Mechanism", *Int. J. Intell. Eng. Syst*, Vol. 13, pp. 99-110, 2020, doi: 10.22266/ijies2020.1031.10.

[22] M. S. Akhtar, T. Garg, and A. Ekbal, "Multi-task Learning for Aspect Term Extraction and Aspect Sentiment Classification", *Neurocomputing*, Vol. 398, pp. 247-256, 2020.

[23] H. Yang, B. Zeng, J. Yang, Y. Song, and R. Xu, "A Multi-Task Learning Model for Chinese-Oriented Aspect Polarity Classification and Aspect Term Extraction", *Neurocomputing*, Vol. 419, pp. 344-356, 2021.

[24] X. Wang, G. Xu, Z. Zhang, L. Jin, and X. Sun, "End-to-end Aspect-based Sentiment Analysis with Hierarchical Multi-task Learning", *Neurocomputing*, Vol. 455, pp. 178-188, 2021.

[25] G. B. S. Wirawan, P. N. T. Y. Mahardani, M. R. K. Cahyani, N. L. P. S. P. Laksmi, and P. P. Januraga, "Conspiracy Beliefs and Trust as Determinants of COVID-19 Vaccine Acceptance in Bali, Indonesia: Cross-Sectional Study", *Personality and Individual Differences*, Vol. 180, p. 110995, 2021.