Journal of Big Data

RESEARCH

Open Access

Argument annotation and analysis using deep learning with attention mechanism in Bahasa Indonesia

Derwin Suhartono[1*], Aryo Pradipta Gema[1], Suhendro Winton[1], Theodorus David[1], Mohamad Ivan Fanany[2] and Aniati Murni Arymurthy[2]

*Correspondence:
dsuhartono@binus.edu
[1] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
Full list of author information is available at the end of the article

Abstract

Argumentation mining is a research field which focuses on sentences in type of argumentation. Argumentative sentences are often used in daily communication and have important role in each decision or conclusion making process. The research objective is to do observation in deep learning utilization combined with attention mechanism for argument annotation and analysis. Argument annotation is argument component classification from certain discourse to several classes. Classes include major claim, claim, premise and non-argumentative. Argument analysis points to argumentation characteristics and validity which are arranged into one topic. One of the analysis is about how to assess whether an established argument is categorized as sufficient or not. Dataset used for argument annotation and analysis is 402 persuasive essays. This data is translated into Bahasa Indonesia (mother tongue of Indonesia) to give overview about how it works with specific language other than English. Several deep learning models such as CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), and GRU (Gated Recurrent Unit) are utilized for argument annotation and analysis while HAN (Hierarchical Attention Network) is utilized only for argument analysis. Attention mechanism is combined with the model as weighted access setter for a better performance. From the whole experiments, combination of deep learning and attention mechanism for argument annotation and analysis arrives in a better result compared with previous research.

Keyword: Argument annotation, Argument analysis, Deep learning, Attention mechanism, Bahasa Indonesia

Introduction

Taking role as one of natural language processing research fields, argumentation mining puts special concern to sentences in type of argumentation. Argument represents certain opinion or point-of-view from one person regarding things that he believed in. An argument must be supported by relevant facts so that it becomes a valid argument and acceptable statement. An argument can be found in an argumentative essay, debate scripts, user comments in a blog/article, scientific articles, and many others. If an article contains opinion which is completed by supporting statements, it can be categorized as an argument.

An argument consists of several components and they show a structure which is based on argumentative relation between components [1]. Formulation of some argumentation scheme in presumptive reasoning was initiated as one of research pioneers in this field [2]. The scheme was utilized by several research in argumentation mining, one of which was essay scoring [3]. Variant of predefined argument schemes drives to further needs with respect to defining features for automatic classification. Certain researchers defined 5 group of features as the characteristics of an argument component [4]. It achieved 77.3% accuracy by using support vector machine (SVM) as the classifier. Figure 1 describes the argument scheme that is used by the research.

The data came from persuasive essays. Argument components consist of 4 type of statements: major claim, claim, premise and non-argumentative. As the continuation of this research, many additional features were defined. The features were grouped into 8 group of features [6]. Structural and contextual features were indicated as the most significant features among others to characterize an argument.

Researchers have observed argumentation mining from various different perspectives. Thus, research in this field reveals in many areas. For example, argument component detection which was well-utilized in legal documents [7]. On the other hands, other researchers used it for public policy formulation [8]. In addition to feature extraction and machine learning, rule-based approach which is commonly used for NLP research, was also utilized as an indicator to classify argument components. Rule-based approach was combined with probabilistic sequence model to automatically detect high-level organizational elements in argumentative discourse [9]. A slightly different approach was done by using ontology-based in detecting argument component. The result could be used in automatic essay scoring [10]. In a more comprehensive level in argumentation, research is not only required to see the argument components, but to see techniques which is capable to measure validity of an argument. Sufficiency measurement of an argument has been done by using support vector machine (SVM) and convolutional neural network (CNN) [11]. In line with that research, estimation of persuasiveness level from an argument in online forum was conducted [12]. Furthermore, other research worked on prediction about convincingness level of an argument [13]. Due to validity or quality that was being assessed, it required more than only 1 statement. Several statements from one certain argumentative discourse were observed to quantify argument validity.

Machine learning evolves from statistical approach to a more semantically aware system; called deep learning. Many researchers implemented deep learning in conventional task which initially used traditional feature engineering with expectation that they can

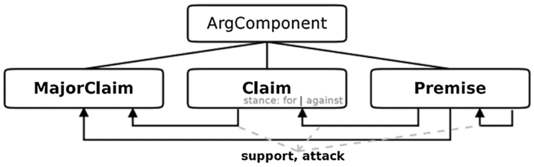

Fig. 1 Relationship between argument components [5]

eliminate tiresome process [14]. They believed by the existence of thousands of not-linear tensor computation, deep learning is able to automatically extract the features. Deep learning itself successfully won a lot of contests in the area of pattern recognition and machine learning. Deep learning can outperform other machine learning algorithms [15]. A lot of research result shows superiority of deep learning compared with regular machine learning. Convolutional neural network was better than machine learning techniques especially for NLP tasks [16]. However, we believed that deep learning is able to achieve better performance in argumentation mining as well as aforementioned NLP tasks.

Argumentation mining

Argumentation mining is a field of study that focuses on argument extraction and analysis from a natural language text [17]. Argumentation mining has 2 phases: argument annotation and argument analysis [18].

A. Argument annotation

Fundamental task in managing arguments is to understand how we can find the location of an argument in documents. For that matter, many supervised machine learning methods are used. The approach is to classify the arguments into argument component or non-argument component.

Data that comes from several sources such as magazine, advertisement, parliamentary notes, judicial summary, etc. were collected to be stored in a database [19]. As a continuation of which, a software named Araucaria was built [20]. This software was used to analyze argumentation and provided a relation among arguments in form of diagram. Initial analysis was conducted from existed corpus [19] and continued by exploration in 2 areas: argumentation surface feature and utilized argumentation scheme [21].

There was different investigation of argument coming from perspective to legal documents based on their rhetoric and visualization [7]. This research was conducted based on feature extraction in which 11 features were utilized. There were 286 words involved as one of the features sets.

Different approach for detecting argument components was done by utilizing combination of rule-based and probabilistic sequence model [9]. High-level organizational element from such argumentative discourse were attempted to identified. Organizational element was also known as shell language. Rule-based was defined by using 25 patterns of handwritten regular expression. Manual annotation without standard guideline was done to 170 essays. The annotation was executed by experts that has been familiar to essay writings. Sequence model was made in accordance to Conditional Random Fields (CRF) by using a number of general features based on lexical frequency. After conducting evaluation, hybrid sequence model was assumed to have best performance in the task.

Argument extraction was applied to support public policy formulation [8]. Result from this research was used to assist policy maker in observing how was the reaction from society in respect to the policy. Tense and mood were the main features as argument indicator.

By using ontology approach, 8 rules were defined to identify arguments from such statements [10]. Rules were defined by research intuition and informal examination to 9 essays. In other research, argumentation scheme was used for essay scoring [3]. It was based on Walton theory [2] involving some adjustments within. This research focused on how annotation protocols intended for argumentative essays were made. Annotation protocol was made for 3 argumentative schemes; they are policy argument scheme, causal argument scheme and argument from a sample scheme.

From other perspective of data, researcher attempted to see argument aspect from social media [22]. It was started by separating statement from dataset into 2 classes: statements which contains argument and does not contain. It was continued by computation involving Conditional Random Fields (CRF).

Argument extraction from Greek news was experimented [23]. Technique that was used in this research was word embeddings extracted from huge size of not-annotated corpus. From the result, one of interesting conclusions was that word embeddings could positively contribute in extracting argumentative sentence.

Unstructured and various data can be found in a web site. Argument extraction to websites were attempted as well [24]. In their research, a gold standard corpus from user-generated web discourse were built along with direct testing by using several machine learning algorithms.

As the continuation from research that did binary classification, which were argument components classification into 2 classes: argument or not, researchers made a try to formulate specific categories from argumentative statements. Generally, 2 classes were defined: claim and premise. Aside from those classes, there were still other various naming or definitions.

Corpus with claim and evidence as labels was built by extracting argumentative statements from Wikipedia articles [25]. It has been utilized by public to be tested by many approaches. There was an opinion saying that all leaves of tree were arguments [26]. They were premises and conclusions, which were placed together one to another.

A new corpus from persuasive essays was made [5]. It contained argumentative statements. This corpus consisted of 90 essays which was labelled by 3 annotators. This corpus covered 3 components of argumentation: major claim, claim, and premise. Other than that, statements that were not categorized as arguments were classified as non-argumentative. It was the 4th class. In order to see how argument components were related one to another, 2 classes to describe their relationship were defined. They were support class and attack class.

From aforementioned corpus, features formulation was also made such that annotated argumentative components could be recognized automatically [4]. All proposed features were categorized to 5 group of sub-features: structural, lexical, syntactic, indicator and contextual. It achieved an accuracy of 77.3%. Specifically, other researchers took a closer look to discourse marker role which was one feature from argumentative corpus in German language [27]. From several conducted experiments, discourse markers were said to be quite indicative in differentiating claim to premise. One research tried to combine all features that has been proposed before [28]. The results were better yet there was no significant improvement.

Table 1 Current works in argument annotation

No	Authors	Dataset	Methods
1	[3]	Argumentative essays	Annotation protocols
2	[5]	Persuasive essays	5 group of sub-features
3	[7]	Legal documents	11 feature sets
4	[8]	Greek language text	Tense and mood
5	[9]	Argumentative discourse	combination of rule-based and probabilistic sequence model
6	[10]	52 essays written by university students	Ontology: 8 rules
7	[22]	Social media	Conditional Random Fields (CRF)
8	[23]	Greek news	Word embeddings
9	[27]	Argumentative corpus in German language	Discourse markers
10	[28]	Persuasive essays	68 sub-features
11	[29]	Persuasive essays	Argument and domain words; LDA
12	[30, 31]	Political debates	CDCD approach

Caused by phenomenon that big and sparse feature space can result on difficulty of feature selection, a more compact feature was proposed [29]. By utilizing corpus of persuasive essays, n-gram and syntactic rules could be replaced by feature and constraint through extracted argument and domain word. Escalation of argument mining performance can be significantly achieved. After argument components were identified, post processing was conducted by using topic modelling: latent dirichlet allocation (LDA) to extract argument word and domain word.

Analyzing argumentation category was also enriched by contribution in certain fields such as debate technology and assessment of argumentation quality. Given a context, automatic claim detection in one discourse was possible [30]. This technique was then developed further by considering negation detection to each detected claim [31]. Following this current research, evidence detection in unstructured text was also conducted [32]. Specified context of data was used for experiments. After claim and evidence were successfully detected, several approaches to get stance from context-dependent claim was observed [33].

Claim and evidence cannot be separated in forming arguments. If claim does not have evidence, then it will not have meanings. For example, political debates contain many claims followed by evidences as the data to support claims. Given a condition of argumentation summarizer needs, an automatic summarizer for argumentation specifically for political debates was built by some researchers [34]. Not only for political debates, automatic summarizer for online debate forum was also conducted as well [35].

In addition, research on argument mining was also conducted in persuasive online discussion. A computational model that handled micro and macro level of argumentation was proposed [36]. Even further, generating argument using a novel framework named CANDELA was conducted. The argument generation was done with retrieval, planning, and realization [37].

Table 1 summarized all current works in argument annotation which are done so far. For further analysis in completing state-of-the-art of argument annotation research, we concentrate to utilize deep learning methods to handle this argument annotation tasks.

Argument analysis

To assess quality of arguments, not only extrinsic aspects need to be observed, but also intrinsic aspects as well. However, it is different to categorization whose assessment can be done directly by observing the texts (extrinsic aspects). Discourse marker as the main component to differ such argumentative statements is no longer valid to use in scoring quality of arguments. In this case, keywords as discourse marker are not representative as the evaluator.

A good argument is the one that can convince the reader that it is a valid and strong argument. To handle this issue, some researchers started to propose some approaches in measuring argument validity. Persuasiveness level of an argument can be estimated by feature extraction to discussion in the online forum [12]. Posting time and writer reputation were said to be useful to utilize as metadata information. Textual features had worse result compared to argumentation-based features. If the data is an essay, argument quality can be assessed through the essay score. In addition to prompt adherence, coherence and technical quality aspect, argument strength can be involved as well to give grade to essays [38].

Huge number of online communities impacts to the appearance of debates in several issues in blogs or forums. Combination of textual entailment and argumentation theory were attempted to extract argumentation from debates, as well as their acceptability [39].

In other research, convincingness appeared as new terminology in assessing quality of argumentation [13]. Relation between arguments in one whole sequence of statements was assessed. Based on that relation, classification was applied. The output was to find out which argument was more convincing and create a list of arguments sorted by their convincingness level. Furthermore, there was another similar task in assessing argument quality. It was done by observing either the relation was sufficient or not [11]. Long Short Term Memory (LSTM) as one of promising deep learning method for text was modified involving Siamese network to recognize argumentation relation in persuasive essay [40]. Furthermore, Hierarchical Attention Network (HAN) with XGBoost was utilized to similar task and indicated to be a promising method for hierarchical data [41].

Table 2 summarized all current works in argument analysis which are done so far. Slightly different with current works, we concentrate to utilize deep learning methods to handle argument analysis tasks.

Table 2 Current works in argument analysis

No	Authors	Tasks	Methods
1	[11]	Argument sufficiency	Feature extraction
2	[12]	Persuasiveness level	Feature extraction
3	[13]	Convincingness level	Relation between arguments in one whole sequence
4	[38]	Argument quality	Textual features
5	[39]	Argument acceptability	Combination of textual entailment and argumentation theory
6	[40]	Argument relation	Siamese network
7	[41]	Argument relation	Hierarchical Attention Network (HAN) with XGBoost

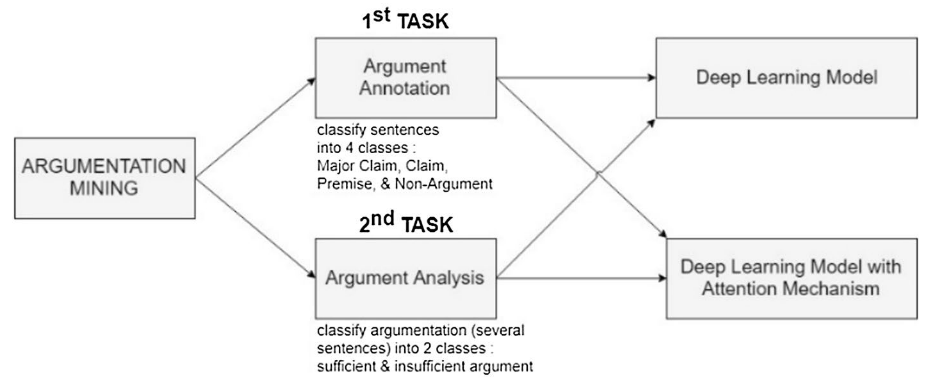
1st TASK
Argument Annotation
classify sentences into 4 classes : Major Claim, Claim, Premise, & Non-Argument
2nd TASK
Argument Analysis
classify argumentation (several sentences) into 2 classes : sufficient & insufficient argument
ARGUMENTATION MINING
Deep Learning Model
Deep Learning Model with Attention Mechanism

Fig. 2 Research framework

Proposed methods

Argumentative statements are the main object for this research. It was initiated by classifying statements into several type of argument components (argument annotation). More than that, categorizing arguments relation into sufficient or not was conducted (argument analysis). Those tasks are described in Fig. 2. Deep learning is used as main methods as well as attention mechanism for a better performance.

Keras [42] was utilized as the main library in all stages from preprocessing (such as tokenizer, vocabulary processor, and indexing) to modeling. Experiments are conducted with a single NVIDIA TITAN X Pascal GPU. Experiment was conducted by involving 402 persuasive essays [6] as dataset which was translated manually into Bahasa Indonesia.

Argument annotation and analysis are included as classification task. Classes that are defined for the classification are:

1. Argument annotation
   This task classifies statements based on their argument type. Statements are classified into 4 classes: Major Claim (MC), Claim (C), Premise (P), and Non-Argumentative (N).
2. Argument analysis
   This task takes a look into relationship between arguments. Relationships are classified into 2 classes: Sufficient (S) and Insufficient (I).

All experiments used dataset (402 persuasive essays) that has been translated to Bahasa Indonesia. FastText was used as word vector representation. Aside from it, we did not use word vector yet utilizing embedding layer (build vector from scratch, without using pre-trained word vector) to compare the performance. Previously, similar works using English dataset was conducted [43] and Glove as word vector representation was used. This research continues to investigate the result from specific language, which is in Bahasa Indonesia.

Figure 3 describes all process from input to output. Each word was saved into dictionary and got its index. Therefore, each statement became sequence of id from all words. Indexing was done to escalate performance or reduce complexity. All words represented by ids were converted to vector representation.

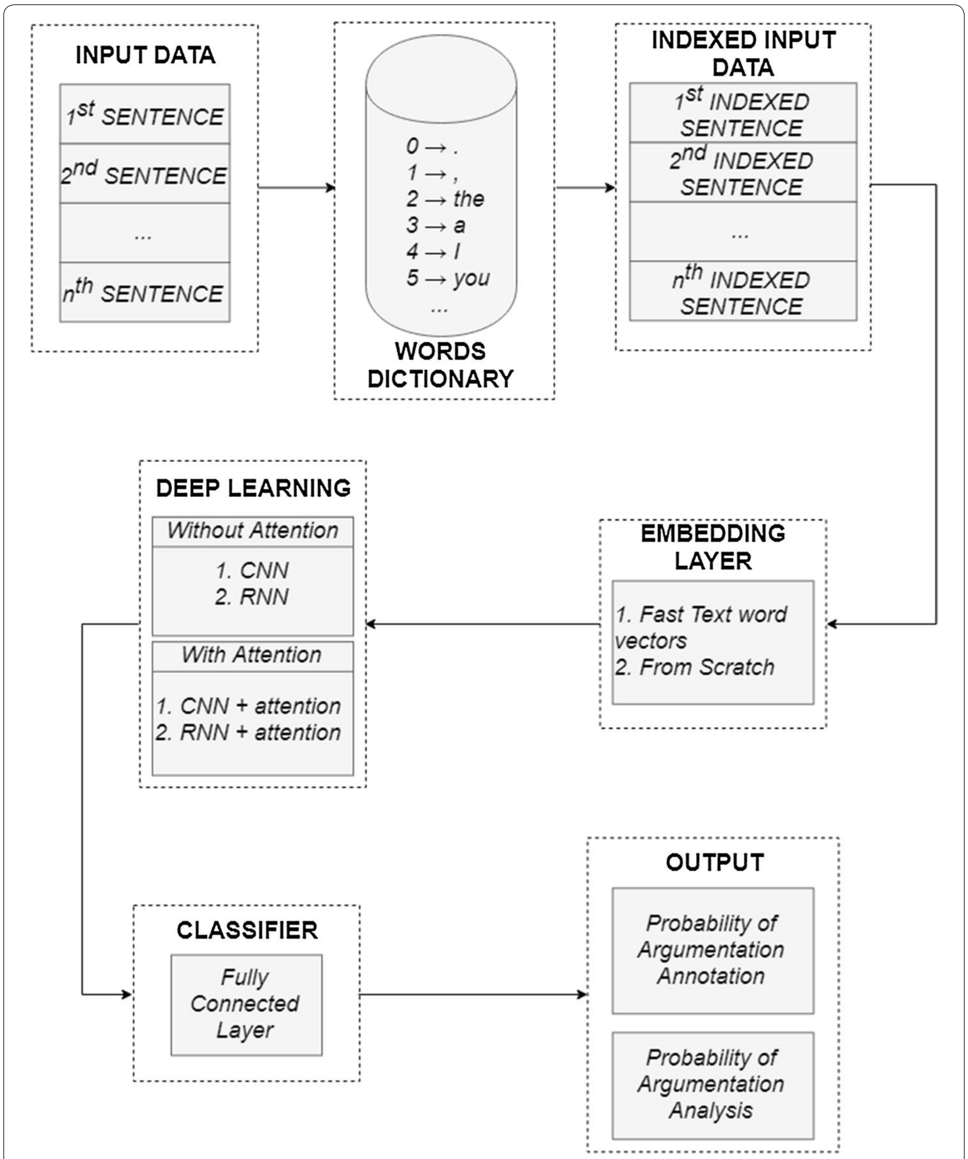

Fig. 3 Workflow of argument annotation and analysis

To compare the result to similar task [4, 6], we did same setting for using cross validation to previous task. For classifying argument component (argument annotation), tenfold cross validation was used while classification layer was using fully connected.

Similar workflow happens for argument analysis as described in Fig. 3. The fundamental difference is in Hierarchical Attention Network (HAN) architecture as hierarchy form of attention mechanism. Attention mechanism process is visualized in Fig. 4. For argument analysis, 20 times fivefold cross validation is chosen as the evaluation scenario.

In identifying sufficiency from an argument, theoretical framework was used [42]. This theory has been used in another research as well [6]. Argument quality measurement happened in various way, such as sufficiency level of categorization [11], persuasiveness

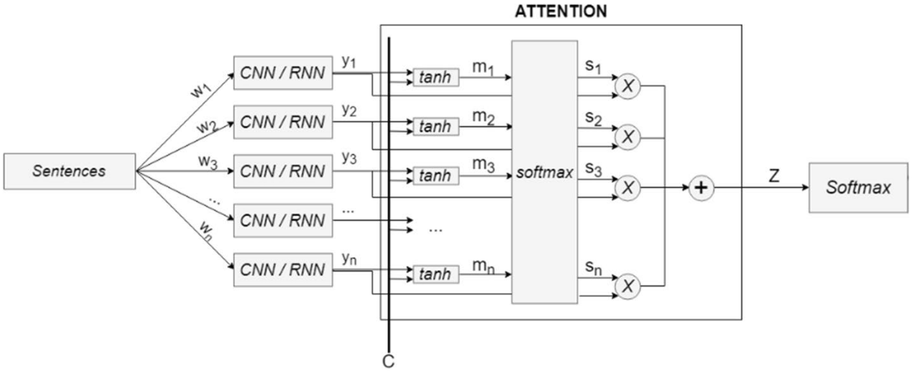

Fig. 4 Attention mechanism attached to deep learning

[12], convincingness [13], and acceptability [39]. In this research, argument analysis focused on sufficiency criteria. This criterion separated which argument was supported sufficiently from others which was not supported sufficiently. The measurement was conducted from contribution given from premise to claim in the argument.

Taking role as main focus to measure impact of attention mechanism to deep learning, layer of attention mechanism was put after deep learning finished in processing the data. Figure 4 explained in detail what happened in "Deep Learning" box in Fig. 3. Output from CNN/RNN was in form of vector that further processed as input for attention layer. 'C' contains information from context of statement for attention layer. Vector of y_1, y_2 till y_n were the output from deep learning model. Tanh was chosen as activation function. All value of m_1, m_2, till m_n were the output after going through activation function which afterwards went into softmax and resulted on vector of s_1, s_2 till s_n. All vectors were combined using vector addition. Final result was 'Z' vector which was vector representation from input statement after going through deep learning model and attention mechanism.

Combining deep learning model with attention mechanism for argument annotation and analysis

Several deep learning models were involved in the experiment, such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). We utilized combination of deep learning with attention mechanism such that the result can justify the impact of attention mechanism in argument annotation and analysis.

Models of deep learning are briefly justified as follows:

1. Convolutional Neural Network (CNN)
   CNN is chosen due to its excellent performance in many different classification tasks such as sentiment classification or question classification [16]. Unlike Recurrent Neural Network architecture, CNN does not rely on the sequential nature of the data per se. Looking into how CNNs process words, it implies that there is a syntactical benefit similar to N-gram windows. Different window size may result into different behavior which may lead into a fairly robust. Through several experiments, a sin-

gle convolutional layer with a window size of 3 and 250 feature maps performs best together with 0.5 dropout rate. Attention mechanism was also added to the architecture in the experiments.

2. Long Short-Term Memory (LSTM)
   LSTM has distinguished characteristics in its effectivity to handle data with sequential nature. LSTM was said to be the best Recurrent Neural Network (RNN) architecture empirically. This happens not only for one directional LSTM, but also bidirectional as well. Based on that background, both LSTMs for one and bidirectional were used for persuasive essays. By observing their parameter through several amount of experiment, 128-unit LSTM, 0.5 for dropout and recurrent dropout rate were used for the experiment. Furthermore, attention mechanism was attached to the architecture.
3. Gated Recurrent Unit (GRU)
   GRU is used due to its performance which is more likely with LSTM and also it has beneficial from the aspect of computation efficiency. Differentiation between LSTM and GRU is the amount of gate in the model [44]. GRU has 2 gates: reset and update while LSTM has 3 gates: input, forget and output. Using the same scenario with LSTM, result comparison was done to GRU and bidirectional GRU. Best parameter for GRU and bidirectional GRU was 128-unit GRU and 0.5 dropout and recurrent dropout rate. Finally, attention mechanism was attached to the architecture.
4. Hierarchical Attention Network (HAN)
   Figure 5 showed HAN architecture using GRU [45]. This architecture worked with 2 level of attention mechanism.

Document was considered as 4-dimensional data consisting of batch size, number of statements, number of words in statement, and vector representation. In the deepest part of the architecture, word-level attention was used by utilizing one bidirectional GRU. This word-level attention was seen as the most influential word representation in one statement.

On the outside of the architecture, other attention was added: sentence-level attention. Similar to word-level attention, this attention mechanism played a role as statement representation which was the most informative one from one document.

At the outermost part of the architecture, softmax layer [46] and negative log likelihood were used. Best setting for HAN was 1-layer bidirectional GRU for word and sentence encoder, along with utilizing 32 unit of GRU. Dropout and recurrent dropout rate were 0.5. Nadam [42] was used as the optimizer, 0.002 learning rate and 32 batch size.

Results and discussion

Corpus was initially created in English [24]. Excellent experts were selected to annotate arguments independently. For this research needs, the dataset was translated into Bahasa Indonesia involving some linguistic experts.

A. Argument annotation

By using translated dataset in form of 402 persuasive essays, result of utilizing several deep learning models was presented in Tables 3 and 4. All experiments used 128 batch size. Classification was made into 4 classes: major claim, claim, premise, dan

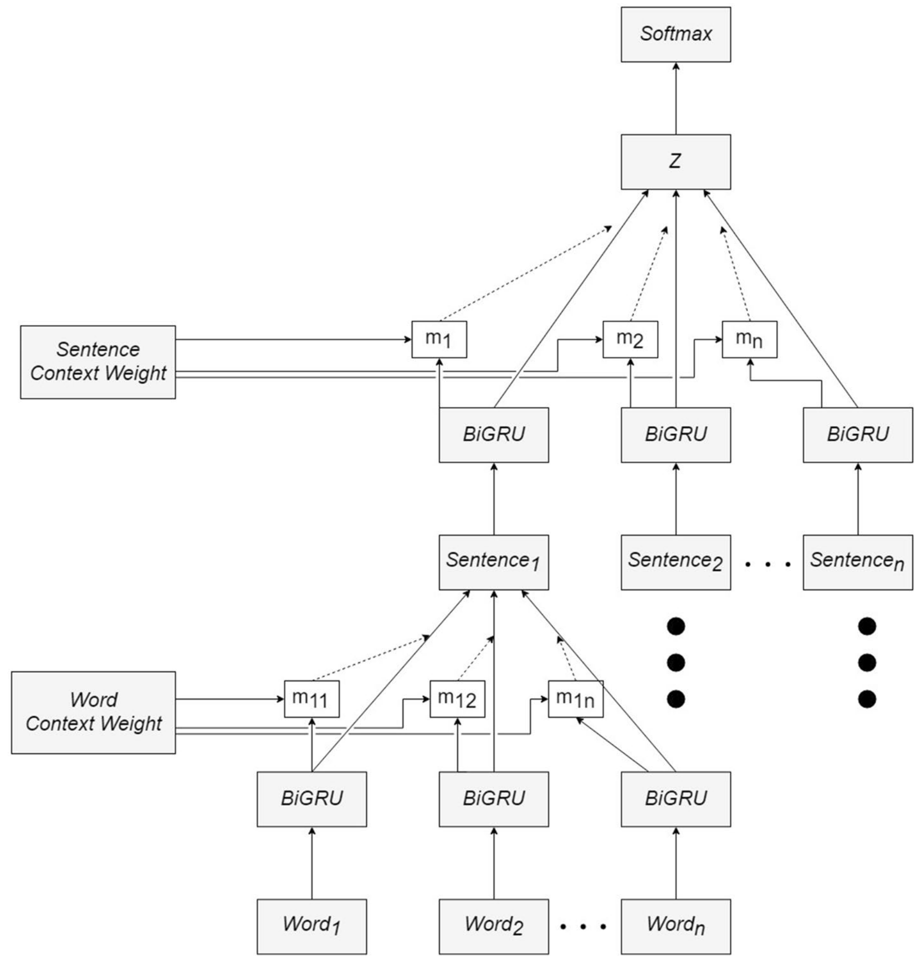

Fig. 5 Hierarchical Attention Network (HAN) for argument analysis

Table 3 Result of argument annotation using deep learning model with attention mechanism (Word Embedding from Scratch)

No	Model name	Accuracy (%)	Recall (%)	Precision (%)	F1 macro (%)
1	CNN	75.43 ± 0.69	61.38 ± 1.38	64.07 ± 1.04	62.21 ± 1.28
2	CNN + Att	76.56 ± 0.66	57.58 ± 1.44	64.87 ± 1.22	57.88 ± 2.04
3	LSTM	75.78 ± 0.55	58.00 ± 1.34	62.68 ± 1.19	58.59 ± 1.56
4	LSTM + Att	75.26 ± 1.23	61.76 ± 1.81	63.74 ± 1.30	62.35 ± 1.70
5	GRU	75.60 ± 1.17	59.22 ± 0.53	64.19 ± 1.13	60.12 ± 0.70
6	GRU + Att	76.37 ± 0.88	59.30 ± 2.36	65.46 ± 1.66	59.92 ± 2.86
7	BiLSTM	75.28 ± 0.55	58.76 ± 2.41	61.79 ± 1.76	58.94 ± 3.16
8	BiLSTM + Att	75.40 ± 1.43	57.89 ± 1.20	60.32 ± 8.53	56.66 ± 3.02
9	BiGRU	75.24 ± 1.29	59.83 ± 0.99	61.45 ± 1.56	59.79 ± 1.41
10	BiGRU + Att	76.36 ± 0.37	60.27 ± 2.60	67.02 ± 3.54	60.52 ± 3.00

Table 4 Result of argument annotation using deep learning model with attention mechanism (FastText Word Embedding)

No	Model name	Accuracy (%)	Recall (%)	Precision (%)	F1 macro (%)
1	CNN	76.50 ± 1.21	61.45 ± 1.92	65.68 ± 1.71	62.10 ± 1.44
2	CNN + Att	74.44 ± 2.11	52.31 ± 1.56	57.50 ± 12.65	49.37 ± 3.30
3	LSTM	75.76 ± 0.82	54.91 ± 3.07	65.19 ± 4.07	53.56 ± 5.03
4	LSTM + Att	75.87 ± 0.51	52.02 ± 1.75	62.03 ± 11.35	48.78 ± 3.62
5	GRU	75.74 ± 0.85	56.95 ± 4.17	64.00 ± 8.51	56.21 ± 5.60
6	GRU + Att	76.17 ± 0.56	56.32 ± 2.14	66.28 ± 1.42	56.18 ± 3.10
7	BiLSTM	75.94 ± 0.24	52.44 ± 2.85	50.35 ± 8.25	48.25 ± 4.25
8	BiLSTM + Att	75.88 ± 0.31	52.79 ± 2.98	63.88 ± 13.77	49.42 ± 5.27
9	BiGRU	76.12 ± 0.30	51.74 ± 1.16	64.30 ± 8.52	48.14 ± 2.27
10	BiGRU + Att	76.41 ± 0.63	55.49 ± 4.28	58.67 ± 8.98	53.38 ± 6.40

non-argumentative. Result of using word embedding from scratch was presented in Table 3, while Table 4 presented result of using FastText [47] as the word embedding.

Generally, result presented in Tables 3 and 4 showed that F1 score did not have significant performance indicating the success of argument annotation. However, this experiment arrived in some conclusions.

Learning mechanism which used word embedding from scratch gave relatively better result compared to FastText as the word embedding. This was caused by a condition where words combination in FastText was a result of crowdsourcing. It did not involve any language experts. Therefore, it was indicated some misuse of words because no quality assurance was dedicated to validating the data.

Other than that, formed word combination using FastText tend to be descriptive rather than argumentative. In the learning process of forming word vector, context of statements was observed such that the way the words be arranged one to another was realized. By the utilization of Wikipedia in Bahasa Indonesia as the ingredients in learning process, word combination that frequently appeared was the descriptive one. Nature of descriptive statements was quite different to argumentative. For example, utilization of word "because" was very rarely used in descriptive statements so given weight to the word "because" would be much different to argumentative statements. In argumentative statements, "because" are very often to be used.

Based on that condition, learning mechanism from scratch is indicated as a better option rather than FastText.

Attention mechanism can refine the performance of almost all deep learning model, such as LSTM (from scratch), BiLSTM (FastText), and BiGRU (from scratch dan FastText). All of them are variants from RNN. This is related with the fact that RNN was claimed as the most suitable deep learning model for text. While for other models, the results were worse compared to deep learning model without attention mechanism. One of them was Convolutional Neural Network (CNN). CNN needs additional spatial information rather than seeing to the context of statements. We arrived in a conclusion that attention mechanism did not play significant role for all deep learning models experimented in this research. This happened because the number of class which was 4 while the total data was only 402 essays. In such case, deep learning did not have enough data to be trained.

Table 5 Result of argument analysis using deep learning model with attention mechanism (Word Embedding from Scratch)

No	Model name	Accuracy (%)	Recall (%)	Precision (%)	F1 macro (%)	ROC-AUC
1	CNN	81.44 ± 2.54	77.33 ± 4.03	80.41 ± 2.73	78.18 ± 3.59	88.81 ± 0.02
2	CNN + Att	72.59 ± 6.41	65.74 ± 6.39	71.35 ± 7.65	65.89 ± 7.57	79.63 ± 0.05
3	LSTM	66.76 ± 1.88	51.14 ± 2.66	45.36 ± 19.95	42.47 ± 4.89	61.32 ± 0.03
4	LSTM + Att	72.71 ± 5.50	61.58 ± 9.63	60.05 ± 22.23	58.29 ± 15.21	79.03 ± 0.09
5	GRU	65.89 ± 1.87	54.21 ± 3.31	54.83 ± 11.50	51.35 ± 6.45	59.38 ± 0.04
6	GRU + Att	77.17 ± 6.30	71.37 ± 12.01	79.84 ± 3.43	69.34 ± 14.66	83.36 ± 0.10
7	BiLSTM	70.06 ± 5.66	62.56 ± 7.69	67.75 ± 6.46	61.14 ± 9.82	72.45 ± 0.06
8	BiLSTM + Att	75.32 ± 4.46	68.23 ± 8.05	74.10 ± 5.11	68.14 ± 8.94	79.65 ± 0.04
9	BiGRU	65.31 ± 2.56	57.45 ± 6.27	59.27 ± 6.03	55.61 ± 6.88	64.29 ± 0.05
10	BiGRU + Att	75.81 ± 5.78	69.52 ± 10.63	79.07 ± 3.94	67.55 ± 13.95	82.14 ± 0.10

Table 6 Result of argument analysis using deep learning model with attention mechanism (Word Embedding from Scratch)

No	Model name	Accuracy (%)	Recall (%)	Precision (%)	F1 Macro (%)	ROC-AUC
1	CNN	74.35 ± 6.12	66.49 ± 11.45	74.21 ± 9.36	63.84 ± 14.81	80.86 ± 0.05
2	CNN + Att	70.84 ± 3.33	63.02 ± 6.58	71.80 ± 5.80	61.17 ± 9.00	74.06 ± 0.07
3	LSTM	65.79 ± 0.58	51.10 ± 2.04	43.80 ± 13.29	43.54 ± 5.49	56.43 ± 0.06
4	LSTM + Att	69.98 ± 5.46	62.22 ± 8.53	69.14 ± 10.36	61.14 ± 9.51	72.40 ± 0.10
5	GRU	67.15 ± 3.69	53.46 ± 6.54	51.25 ± 14.22	47.48 ± 10.26	57.54 ± 0.09
6	GRU + Att	68.71 ± 4.73	58.89 ± 10.85	53.55 ± 17.62	52.39 ± 15.12	67.69 ± 0.10
7	BiLSTM	67.64 ± 1.56	56.67 ± 3.85	64.01 ± 8.70	54.27 ± 6.46	65.72 ± 0.07
8	BiLSTM + Att	67.54 ± 4.21	57.50 ± 7.11	62.60 ± 16.99	52.86 ± 10.94	73.35 ± 0.09
9	BiGRU	66.96 ± 1.01	54.92 ± 4.01	61.59 ± 4.09	51.28 ± 6.29	61.83 ± 0.05
10	BiGRU + Att	69.10 ± 2.01	61.06 ± 9.05	74.14 ± 6.76	56.28 ± 11.17	71.47 ± 0.10

The best model for argument annotation using Bahasa Indonesia is LSTM with attention mechanism.

B. Argument analysis

Using smaller amount of class, which was 2, argument analysis is categorized as binary classification. ROC (Receiver Operating Characteristics)–AUC (Area Under the Curve) was used as one of evaluation methods. Same dataset was used for argument analysis, yet labelling was only categorized into 2 classes: sufficient and insufficient.

Table 5 presented the result using word embedding from scratch while Table 6 contains result using FastText. Batch size was 128. Different attention mechanism architecture namely Hierarchical Attention Network (HAN) was used. Tables 7 and 8 presented result of HAN.

Tables 5 and 6 described that attention mechanism significantly improved performance of RNN models. ROC-AUC for all RNN models went up after attention mechanism was attached. It clarified discussion from the result of argument annotation

Table 7 Result of argument analysis using hierarchical attention network (Word Embedding from Scratch)

No	Batch size	Accuracy (%)	Recall (%)	Precision (%)	F1 macro (%)	ROC-AUC
1	16	72.69 ± 3.44	61.36 ± 6.10	67.51 ± 17.20	59.69 ± 10.29	81.23 ± 4.13
2	32	74.84 ± 3.01	64.98 ± 5.13	77.19 ± 3.47	65.17 ± 6.36	84.66 ± 1.69
3	64	77.46 ± 3.54	69.28 ± 5.76	78.65 ± 3.74	70.29 ± 6.76	86.16 ± 2.90
4	100	69.79 ± 5.19	58.98 ± 11.14	49.04 ± 19.74	52.87 ± 16.15	73.73 ± 7.86
5	128	69.76 ± 4.77	56.05 ± 8.10	50.91 ± 21.83	50.06 ± 13.19	75.68 ± 7.33

Table 8 Result of argument analysis using hierarchical attention network (FastText Word Embedding)

No	Batch size	Accuracy (%)	Recall (%)	Precision (%)	F1 macro (%)	ROC-AUC
1	16	70.65 ± 3.40	62.97 ± 7.92	61.91 ± 14.69	60.75 ± 11.64	68.00 ± 16.00
2	32	71.33 ± 3.99	64.27 ± 7.86	62.90 ± 15.50	62.41 ± 11.73	70.72 ± 13.47
3	64	73.47 ± 0.88	64.20 ± 3.80	76.37 ± 3.73	63.89 ± 3.78	76.60 ± 2.95
4	100	72.11 ± 1.66	65.35 ± 5.81	71.52 ± 3.48	64.57 ± 6.73	75.83 ± 2.71
5	128	72.51 ± 3.96	65.58 ± 5.13	75.10 ± 5.02	64.57 ± 6.05	77.73 ± 2.87

clearly. Smaller amount of class assisted to better result utilizing 402 persuasive essays. If dataset is enlarged, we hypothesize that argument annotation task will have comparable result with argument analysis.

CNN performed consistently to experiment in argument annotation. It had worse result when attention mechanism was added. Utilization of max pooling layers in CNN for image recognition enables the information to be denser. This information is very useful for recognition task because high level feature extraction will have a denser representation. However, problem in using this layer is loss of spatial information. After condensation has finished, location of certain word is no longer identified whereas location is very important in statements. When the attention mechanism is not used, the fully connected layer that acts as a classifier is assisted in seeing more dense representation patterns. However, changing attention no longer has effect because spatial information from the data has been lost.

Based on all experiments in argument analysis, word embedding from scratch has better performance than FastText. This is relevant with previous discussion in argument annotation.

Best model in argument analysis is HAN with word embedding from scratch with 64 as batch size. This result is in line with experiment using English dataset [43]. HAN has a good performance in dataset with hierarchical characteristics.

Some points that need to be highlighted from this research are as follows:

1. Word vectors utilization
   Based on the experiments conducted, performance of FastText is worse than word embeddings from scratch. It is in line with previous research using English dataset

[43]. We arrived in a conclusion that pre-trained word vector is not suitable to work on argumentative statements.

2. Number of classes
   More classes will drive to smaller amount of data in each class. The more the number of classes, the more difficult to learn the pattern. Argument analysis results on better performance than argument annotation.
3. Role of attention mechanism
   Most experiments using deep learning with attention mechanism have better results, such as LSTM, GRU, BiLSTM, and BiGRU. Commonly, new features are added to improve performance, yet attention mechanism has its role to strengthen current features involvement. It works by identifying which part of whole sequences contributes in learning process such that the model can perform well.
   Attention mechanism improves result from bidirectional RNN. This is caused of RNN's behavior which involve future context in the process. Hierarchical Attention Network (HAN) performs well in argument analysis, due to HAN's characteristics in form of hierarchy. Attention layer in HAN is divided into 2 layers: word-level and sentence-level. HAN will perform in its best if the data is in form of hierarchy, for example paragraph statement word.
4. Form of language
   Comparing our result with previous similar research utilizing English dataset [43], there is no extreme differences. F1 and ROC-AUC score are relatively close. Fundamental difference is on utilized word embedding. In English, word vector representation such as Glove or Word2vec can be used because they are trained with huge size of data. They can be used as universal feature extractor for several tasks related with text. Research in different language results on many variants of word vector representation, such as FastText for Bahasa Indonesia [47]. FastText is utilized in our research and it has no better result compared with word embeddings from scratch.

Therefore, utilization of other language except English still need to consider how big is the data. We can have better and more representative word embeddings for the features.

Conclusion

Some conclusions related to all experiments conducted in this research are:

1. Pre-trained word vector has no high significance in improving performance argument annotation and analysis
2. Combining attention mechanism with deep learning model results on better performance, especially for Recurrent Neural Network (RNN)
3. Hierarchical Attention Network (HAN) as one variant of attention mechanism works well in hierarchical data, for example: one paragraph contains several statements, and one statement contains several words.
4. Word embedding will play an important role as feature only if it is trained by huge amount of data, otherwise it won't.

Abbreviations
AUC: Area under the curve; CNN: Convolutional neural network; CRF: Conditional random fields; GRU: Gated recurrent unit; HAN: Hierarchical attention network; LSTM: Long short-term memory; LDA: Latent dirichlet allocation; NLP: Natural language processing; RNN: Recurrent Neural Network; ROC: Receiver operating characteristics; SVM: Support vector machine.

Acknowledgements
We would like to thank Universitas Indonesia for grant "Hibah Tugas Akhir Mahasiswa Doktor" year 2018 numbered 1263/UN2.R3.1/HKP.05.00/2018 which support our research. Supports from School of Computer Science Bina Nusantara University and Machine Learning and Computer Vision Laboratory Faculty of Computer Science Universitas Indonesia for supporting all experiments in this research.

Authors' contributions
DS contributed as the research principal in this work. APG, SW and TD take role for technical issues. MIF and AMA advise all process for this work. Regarding the manuscript, DS, APG, SW and TD wrote the manuscript, while MIF and AMA revised the manuscript. All authors read and approved the final manuscript.

Author information
Derwin Suhartono is faculty member of Bina Nusantara University, Indonesia. He got his PhD degree in computer science from Universitas Indonesia in 2018. His research fields are 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. He has his professional memberships in ACM, INSTICC, and IACT. He also takes role as reviewer in several international conferences and journals

Aryo Pradipta Gema received his bachelor's degree from Bina Nusantara University majoring Computer Science with Intelligent Systems specialty. He is also one of many awardees for best students in the university. In his final year of study, he underwent an enrichment program provided by his university as a junior researcher. Main topic that he is interested in is deep learning. He writes and presents widely on argumentation mining research, a subfield of natural language processing field of research as well as several image processing tasks.

Suhendro Winton received his bachelor's degree from Bina Nusantara University majoring Computer Science with Intelligent Systems specialty. In his final year of study, he undergoes an enrichment program provided by his university as a junior researcher. Main topic that he is interested with is deep learning. He writes and presents widely on argumentation mining research, a subfield of natural language processing field of research.

Theodorus David received his bachelor's degree from Bina Nusantara University majoring Computer Science with Intelligent Systems specialty. In his final year of study, he undergoes an enrichment program provided by his university as a junior researcher. Main topic that he is interested with is deep learning. He writes and presents widely on argumentation mining research, a subfield of natural language processing field of research.

Mohamad Ivan Fanany is a researcher and lecturer at Faculty of Computer Science.-Universitas Indonesia. His research interests include machine learning, data science, and combining vision and graphics, remote sensing, climate modeling, biomedical engineering. Before joining the faculty, he worked at Future Project Div. Toyota Motor Corp, Japan, as a member of middleware development and recognition team; NHK ES Inc., as a researcher of IT21 Millennium Project on Advanced High Resolution and Highly Sensible Presence 3D Content Creation funded by NICT Japan; and a JSPS Fellow and Research Assistant at Imaging Science and Engineering, Tokyo Institute of Technology (Tokyo Tech). He served as the Chairman of Titech IEEE student branch 2002-2003. Currently a member of IEEE Consumers Electronics and IEEE Geoscience and Remote Sensing. In January 2015, he was elevated to Senior Member of IEEE.

Aniati Murni Arymurthy is professor in computer science with specialty in computer vision and image processing. She got her MSc from Computer and Information Sciences Department in The Ohio State University (OSU), Columbus, Ohio, USA. She got her PhD from Universitas Indonesia with sandwich program in Pattern Recognition and Image Processing Lab (PRIP Lab), Department of Computer Science, Michigan State University (MSU), East Lansing, Michigan, USA. Currently, she is active as lecturer in Faculty of Computer Science, Universitas Indonesia. Her research interests include pattern recognition, image processing, and spatial data.

Funding
All of this works is fully supported by Universitas Indonesia research Grant numbered 1263/UN2.R3.1/HKP.05.00/2018.

Availability of data and materials
The datasets for this study are available on request to the corresponding author.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia. [2] Machine Learning and Computer Vision Laboratory, Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia.

Received: 5 August 2020 Accepted: 8 October 2020
Published online: 19 October 2020

References

1. Peldszus A, Stede M. From argument diagrams to argumentation mining in texts: a survey. Int J Cognit Informat Nat Intel. 2013;7(1):1–31.
2. Walton DN. Argumentation schemes for presumptive reasoning. Mahwah: Lawrence Erlbaum; 1996.
3. Song Y, Heilman M, Klebanov BB, Deane P. Applying argumentation schemes for essay scoring. In: Proceedings of the first workshop on argumentation mining. 2014. p. 69–78.
4. Stab C, Gurevych I. Identifying argumentative discourse structures in persuasive essays. In: Proceedings of conference on empirical methods in natural language processing. 2014. p. 46–56.
5. Stab C, Gurevych I. Annotating argument components and relations in persuasive essays. In: Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. 2014. p. 1501–1510.
6. Stab C, Gurevych I. Parsing argumentation structures in persuasive essays. Comput Linguist. 2017;43(3):619–59.
7. Moens MF, Boiy E, Palau RM, Reed C. Automatic detection of arguments in legal texts. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law. 2007. p. 225–230.
8. Florou E, Konstantopoulos S, Kukurikos A, Karampiperis P. Argument extraction for supporting public policy formulation. In: Proceedings of the 7th Workshop on language technology for cultural heritage, social sciences, and humanities. 2013. p. 49–54.
9. Madnani N, Heilman M, Tetreault J, Chodorow M. Identifying high-level organizational elements in argumentative discourse. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012. p. 20–28.
10. Ong N, Litman D, Brusilovsky A. Ontology-based argument mining and automatic essay scoring. In: Proceedings of the First Workshop on Argumentation Mining. 2014. p. 24–28.
11. Stab C, Gurevych I. Recognizing insufficiently supported arguments in argumentative essays. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017. p. 980–990.
12. Wei Z, Liu Y, Li, Y. Is This Post Persuasive? Ranking argumentative comments in the online forum. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. p. 195–200.
13. Habernal I, Gurevych I. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. p. 1589–1599.
14. LeCun Y, Bengio Y, Hinton G. Deep learning. Nat Int J Sci. 2015;521(7553):436.
15. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
16. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014. p. 1746–1751.
17. Boltuzic F, Šnajder J. Fill the Gap! Analyzing implicit premises between claims from online debates. In: Proceedings of the 3rd Workshop on Argument Mining ACL. 2016. p. 124–133.
18. Addawood AA, Bashir MN. "What is Your Evidence?" A study of controversial topics on social media. In: Proceedings of the 3rd Workshop on Argument Mining ACL. 2016. p. 1–14.
19. Katzav J, Reed C, Rowe G. An argument research corpus. Corpora: Practical Appl.s of Ling; 2003.
20. Reed C, Rowe G. Araucaria: software for argument analysis, diagramming and representation. Int J Artificial Intel Tools. 2004;13(04):961–79.
21. Reed C. Preliminary results from an argument corpus. Linguistics in the twenty-first century. 2006. p. 185–196.
22. Goudas T, Louizos C, Petasis G, Karkaletsis V. Argument extraction from news, blogs, and the social web. Int J Artif Intel Tools. 2015;24(05):1540024.
23. Sardianos C, Katakis IM, Petasis G, Karkaletsis V. Argument extraction from news. In: Proceedings of the 2nd Workshop on Argumentation Mining. 2015. p. 56–66.
24. Habernal I, Gurevych I. Argumentation mining in user-generated web discourse. Comput Linguist. 2016;43(1):125–79.
25. Aharoni E, Polnarov A, Lavee T, Hershcovich D, Levy R, Rinott R, Gutfreund D, Slonim N. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In: Proceedings of the First Workshop on Argumentation Mining. 2014. p. 64–68.
26. Palau RM, Moens MF. Argumentation mining: the detection, classification and structure of arguments in text. In: Proceedings of the 12th International Conference on Artificial Intelligence and Law, 2009.
27. Eckle-Kohler J, Kluge R, Gurevych I. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In: Proceedings of the 2nd Workshop on Argumentation Mining. 2015. p. 22–28.
28. Desilia Y, Utami VT, Arta C, Suhartono D. An attempt to combine features in classifying argument components in persuasive essays. In: Proceedings of 17th Workshop on Computational Models of Natural Argument. 2017. p. 71–75.
29. Nguyen HV, Litman DJ. Extracting argument and domain words for identifying argument components in texts. In: Proceedings of the 2nd Workshop on Argumentation Mining. 2015. p. 2–28.
30. Levy R, Bilu Y, Hershcovich D, Aharoni E, Slonim N. Context dependent claim detection. In: Proceedings the 25th International Conference on Computational Linguistics: Technical Papers. 2014. p. 1489–1500.
31. Bilu Y, Hershcovich D, Slonim N. Automatic Claim Negation: Why, How and When. In: Proceedings of the 2nd Workshop on Argumentation Mining. 2015. p. 84–93.
32. Rinott R, Dankin L, Perez CA, Khapra MM, Aharoni E, Slonim N. Show me your evidence–an automatic method for context dependent evidence detection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2015. p. 440–450.
33. Bar-Haim R, Bhattacharya I, Dinuzzo F, Saha A, Slonim N. Stance classification of context-dependent claims. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Long Papers. 2016; 1: 251–261.
34. Egan C, Siddharthan A, Wyner A. Summarising the points made in online political debates. In: Proceedings of the 3rd Workshop on Argument Mining. 2016. p. 134–143.

35. Chowanda AD, Sanyoto AR, Suhartono D, Setiadi CJ. Automatic debate text summarization in online debate forum. Procedia Comput Sci. 2017;116:11–9.
36. Chakrabarty T, Hidey C, Muresan S, Mckeown K, Hwang A. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019. p. 2933–2943.
37. Hua X, Hu Z, Wang L. Argument generation with retrieval, planning, and realization. proceedings of the 57th annual meeting of the association for computational linguistics, 2019. p. 2661–2672.
38. Persing I, Ng V. Modeling argument strength in student essays. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Long Papers; 2015; 1: 543–552.
39. Cabrio I, Villata S. Combining textual entailment and argumentation theory for supporting online debates interactions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers; 2012; 2: 208–212.
40. Gema AP, Winton S, David T, Suhartono D, Shodiq M, Gazali W. It takes two to tango: modification of siamese long short term memory network with attention mechanism in recognizing argumentative relations in persuasive essay. Procedia Comput Sci. 2017;116:449–59.
41. Suhartono D, Gema AP, Winton S, David T, Fanany MI, Arymurthy AM. "Hierarchical Attention Network with XGBoost for Recognizing Insufficiently Supported Argument". International Workshop on Multi-disciplinary Trends in Artificial Intelligence (MIWAI), p. 174–188, Gadong, Brunei Darussalam, 2017.
42. Chollet. Keras. https://github.com/fchollet/keras, 2015.
43. Suhartono D, Gema AP, Winton S, David T, Fanany MI, Arymurthy AM. Attention-based argumentation mining. Int J Comput Vision Robot. 2019;9(5):414–37.
44. Chung J, Gülçehre C, Cho K, Bengio Y. Gated feedback recurrent neural networks. In: Proceedings of International Conference on Machine Learning; 2015. p. 2067–2075.
45. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. p. 1480–1489.
46. Hinton GE, Salakhutdinov RR. Replicated softmax: an undirected topic model. Advances in Neural Information Processing Systems. 2009. p. 1607–1614.
47. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T. Learning Word Vectors for 157. Languages. 2018;. arXiv:1802.06893.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.