International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 3
ISSN: 2088-5334

Development of Classification Method for Lecturer Area of Expertise Based on Scientific Publication Using BERT

Didi Rustam [a], Adang Suhendra [a], Suryadi Harmanto [a], Ruddy Suhatril [a], Dwi Fajar Saputra [b,*], Rusdan Tafsili [c], Rizky Prasetya [c]

[a] *Department of Information Technology, Gunadarma University, Depok, West Java, Indonesia*
[b] *Department of Information Science, Universitas Pembangunan Nasional Veteran Jakarta, West Java, Indonesia*
[c] *Postgraduate Learning of Technology Universitas Negeri Malang, East Java, Indonesia*
*Corresponding author: *dwifajar@upnvj.ac.id*

Abstract*— Implementing the Artificial Intelligence concept in higher education can be utilized in the context of Human Resource (HR) talent management. The lecturer portfolio provided by the Integrated Resource Information System (SISTER DIKTI) is expected to give an overview of the profiles of all lecturers and map competencies based on groups of fields of knowledge. However, the map of scientific fields based on SISTER data currently available is still subjective. The data is in the form of a group of lecturers' chosen fields of science, independently selected by each lecturer to recognize their expertise. This study discusses the problem of processing unstructured SISTER data. It looks for mapping solutions and classification methods by developing a strategy for classifying groups of scientific fields from unstructured data input. It is necessary to identify the suitability of the chosen field of science compared to that developed through the tri-dharma through identification based on a mapping of the group of fields that can be extracted from the tri-dharma activity, in this case, research represented by scientific publications recorded on SISTER. Therefore, we need an appropriate model to measure similarity, which can then be classified based on abstract documents and scientific publication titles for the classification of scientific fields using NLP based on classification run on DGX A100. This study aims to develop a classification method from titles and abstracts. Scientific publications contained in SISTER are unstructured data, so a corpus is formed to identify the lecturer's field of science. The results show that the classification method developed in this study can measure the similarity of lecturer publications based on abstracts and titles through a vector formation process based on bidirectional encoders and also produces a deep learning model to classify 24 categories of fields of science with an accuracy of 95,0345 percent on training data and 92.876 percent on the test data.

***Keywords*— Wireless sensor networks; localization; mobile beacon; mobile anchor; RSSI.**

Manuscript received 6 Feb. 2024; revised 14 Mar. 2024; accepted 1 Apr. 2024. Date of publication 30 Jun. 2024.

I. INTRODUCTION

Digital transformation. The application of *Artificial Intelligence* (AI) and *Big Data* in higher education institutions may speed the digitization of institutional tasks such as human resource management. The purpose of digitizing human resource management is to regulate all actions and policies linked to selection, development, training, rewards, and incentives to achieve higher education institutions' goals. Lecturer career development is part of human resource management in educational institutions. Human resource management in educational institutions might concentrate on evaluating teacher performance to make appropriate recommendations and increase the competency of teachers and lecturers [1].

Through the Integrated Resource Information System (SISTER DIKTI), the Ministry of Education, Culture, Research, and Technology's Directorate General of Higher Education facilitates academic documentation and lecturer competence. SISTER overviews all lecturers' profiles and maps competencies based on knowledge domains. To recognize their knowledge, each professor must freely select the data recorded on SISTER.

Conducting research and publishing, part of the higher education tri-dharma can help lecturers increase their abilities and areas of expertise. Lecturers can study themes or problems related to their fields and multidisciplinary topics in interrelated fields. Data must be processed into the models

needed to objectively determine lecturers' abilities to classify transdisciplinary research and publications.

Several document categorization research studies have been conducted, including constructing community detection models. For text classification, further research on Random Multimodal Deep Learning for Classification integrates CNN, DNN, and RNN [2]. The training data distribution can be described using research on text categorization utilizing Naive Bayes on the frequency normalization of word terminology [3].

So far, similarity measures have been based on word vector data, in which each word is encoded independently without regard for the words before and after it. This can lead to misunderstandings about the role of the word in the sentence. The Bidirectional Encoder Representations From Transformers (BERT) approach was developed by Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova to train an immersive two-way representation of unlabeled text [4].

BERT is an artificial neural network-based pre-training technique for Natural Language Processing (NLP), which aims to make it easier for computers to understand language like humans. The BERT pre-training model can be fine-tuned with just one additional output layer for advanced modeling of various tasks, such as language inference queries, without any substantial task-specific architectural modifications. Research related to the BERT model is done by fine-tuning using the top layer of the BERT model for text classification [5].

By establishing a technique for identifying scientific groups from unstructured data input, this research covers the processing of unstructured SISTER data and explores mapping solutions and classification approaches. Documentation of lecturer records at SISTER in the form of a portfolio of tri-dharma activities in higher education represented by scientific publications. As a result, we require an appropriate model to evaluate similarity, which is then classified based on abstract papers and scientific publication titles to classify scientific areas using NLP (Natural Language Processing) based classification, which is performed on the DGX A100.

II. Material and Method

A. Bert

Bidirectional Encoder Representations from Transformers is a learned language representation created in 2018 by Google AI Language researchers. Bert was created using Deep Learning approaches and various methods, including semi-supervised learning, OpenAI Transformers, ULMFiT, ELMo, and Transformers.

BERT is intended to practice in-depth two-way representation of unlabeled text across all levels by co-conditioning on left and right contexts. Without modifying substantial task-specific architecture changes, BERT pre-training models can be fine-tuned with just one additional output layer to generate cutting-edge models for various tasks, such as question answering and language inference.

Many natural language processing tasks have been proven to benefit from language model pre-training [6]–[9]. These include sentence-level tasks like natural language inference [10], [11] and paraphrasing [12], which aim to predict relationships between sentences by analyzing them holistically, as well as token-level tasks like recognizing named entities and answering questions, in which models must produce fine-grained output at the token level [13], [14].

There are two approaches to establishing pre-training language representations for downstream tasks: feature-based and fine-tuning. Feature-based approaches, such as ELMo [9], employ a task-specific architecture with trained representations as an additional feature. Fine-tuning techniques, such as the OpenAI GPT [7], introduce a few task-specific parameters and train on downstream tasks exclusively by fine-tuning all pre-trained parameters. Both approaches use a unidirectional language model to examine shared language representations during pre-training and have the same objective function.

B. Transformers

Transformer is a mechanism for investigating the contextual links between words in text [15]. The Transformer can understand and convert the understanding received through a mechanism known as the self-attention mechanism, which is the Transformer's method of transforming the "understanding" of other similar words into words that the mechanism will process. As with the Transformer, there are two mechanisms: the Encoder reads all text input simultaneously, and the Decoder produces a predicted output sequence. Figure 1 depicts a representation of the Encoder and Decoder.

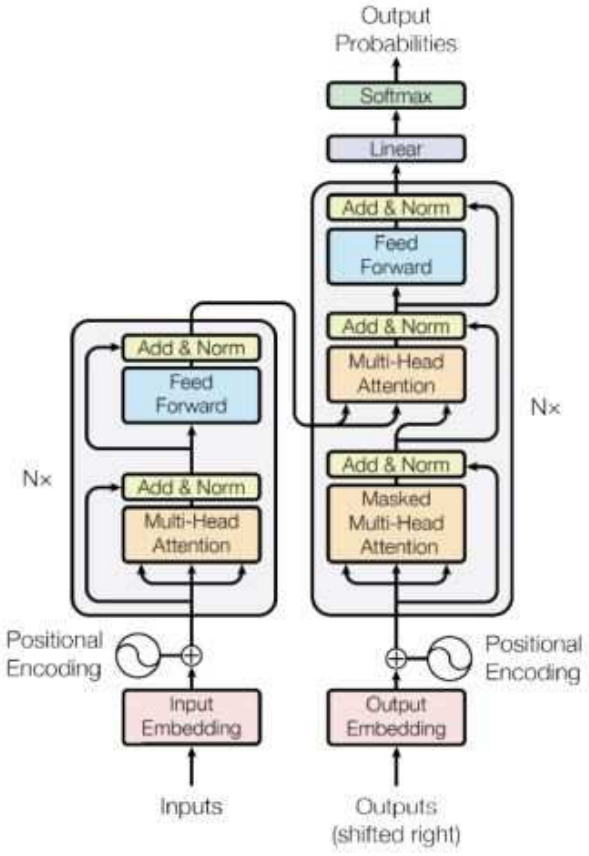


Fig. 1 Encoder and Decoder

The BERT model's architecture is a multi-layer bidirectional Transformer, similar to the original Transformer

implementation, although it only employs a process up to the encoder.

C. Unsupervised Featured Based Approach

For decades, this approach to generally applicable word representation has been a focus of research, including non-neural approaches [16]–[20]. Trained word embedding is an essential component of current NLP systems, improving learning embedding from scratch [21]. A left-to-right language modeling objective [22] was utilized for pre-training word insertion vectors and to identify proper words from erroneous words in left and right contexts [19]. This method has been extended to coarser details such as sentence embedding [23], [24] or paragraph embedding [25]. Previous research has used the objective of ranking next-sentence candidates to train sentence representation [24], [26], left-to-right production of the following phrase's words based on a representation of the preceding sentence [23] or omit the auto-encoder derivation objective [27].

D. Unsupervised Fine-Tuning Approach

This method employs trained word insertion parameters derived from unlabeled text [28]. The benefit of this technique is that some parameters must be learned from the ground up. Because of these benefits, OpenAI GPT [7] has already obtained cutting-edge performance on various sentence-level problems from the GLUE benchmark [29]. For pre-training the model, left-to-right language modeling and auto-encoder objectives were utilized [6]–[8]. The BERT framework consists of two steps: pre-training and fine-tuning.

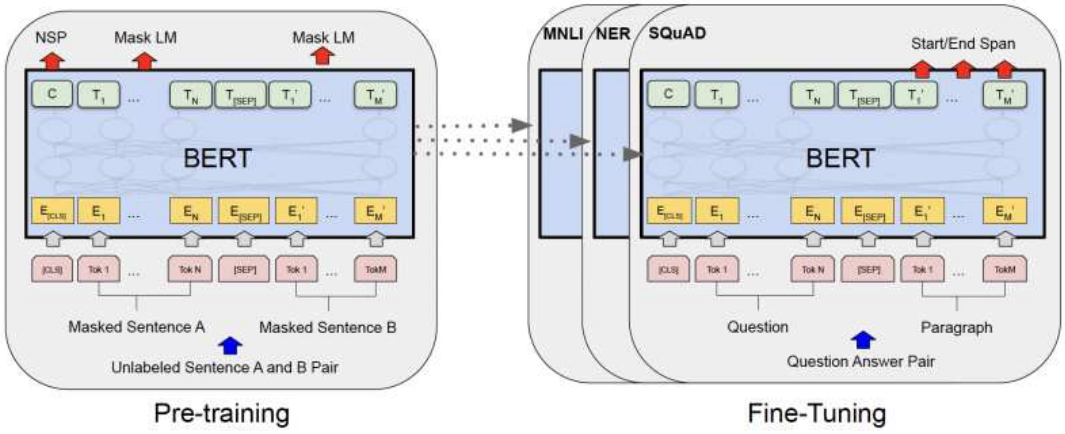


Fig. 2 Pre-Training and Fine-Tuning Procedures on Bert

The hallmark of BERT is its unified architecture across a wide range of tasks. There are minimal differences between pre-trained architecture and final downstream architecture. The BERT architectural model is a multi-layer bidirectional Transformer encoder based on the original implementation described in [15]. Representation of Input/Output To enable BERT to perform various downstream tasks, our input representation may unambiguously represent a single phrase and a pair of sentences (e.g., Question, Answer) in a single token sequence. A "sentence" in this paper can be an arbitrary range of contiguous texts rather than an actual language sentence. A "sequence" is a set of input tokens to a BERT that can be one sentence or two sentences combined.

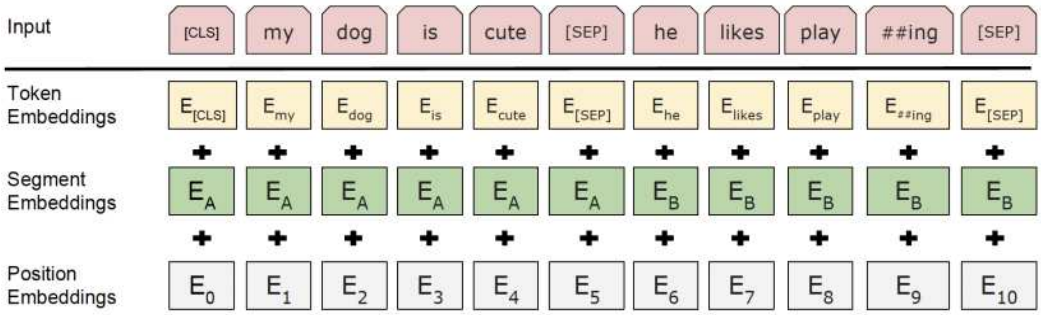


Fig. 3 Representation of Bert's Input.

E. Feature-based Approach with Bert

All previous BERT outcomes used a fine-tuning approach in which a simple classification layer is added to the pre-trained model, and all parameters are jointly modified for the downstream task. However, the feature-based technique, which extracts fixed features from a pre-trained model, has some advantages. Not all jobs are easily represented by Transformer's encoder design, necessitating extra task-specific model architectures. Second, there is a significant computational benefit to generating expensive training data representations once and then performing several tests with less costly models after these representations.

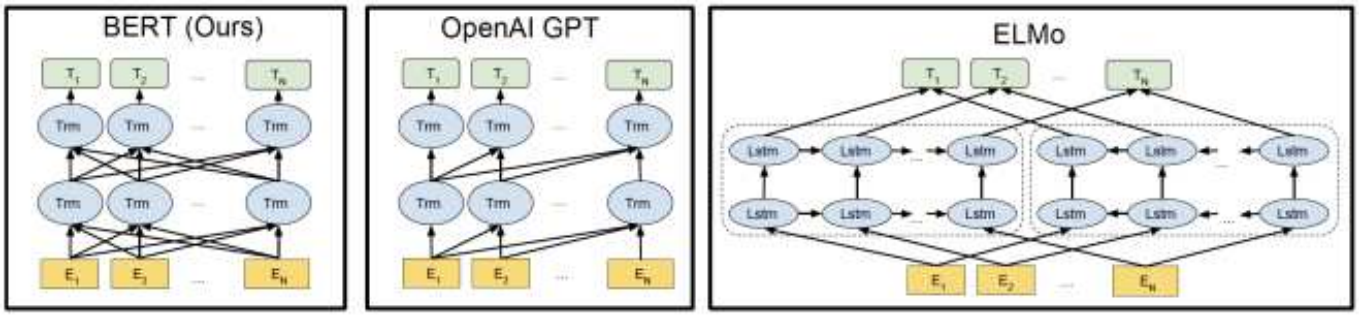


Fig. 4 Differences in Pre-training Model Architecture.

BERT uses a bidirectional transformer, and OpenAI GPT uses left-to-right transformers. ELMo generates features for downstream tasks by using independently trained left-to-right and right-to-left LSTM sets. Only the BERT representations are co-conditioned across all layers in the left and right contexts. Aside from architectural distinctions, BERT and OpenAI GPT are fine-tuning approaches, whereas ELMo is feature-based.

The Multi-Genre Natural Language Inference (MNLI) task is a large-scale crowdsourcing entailment classification problem [11]. The purpose is to predict whether the second phrase is an entailment, contradiction, or neutral concerning the first. QQP Quora Question Pairs is a binary classification problem in which Chen et al. seek to evaluate whether two Quora questions are semantically identical (2017). The QNLI Question Natural Language Inference is a binary classification task based on a variant of the Stanford Question Answering Dataset [14]. A positive example is a pair of (questions or sentences) with a correct answer, while a negative example is a pair of (questions or phrases) from the same paragraph that does not contain an answer.

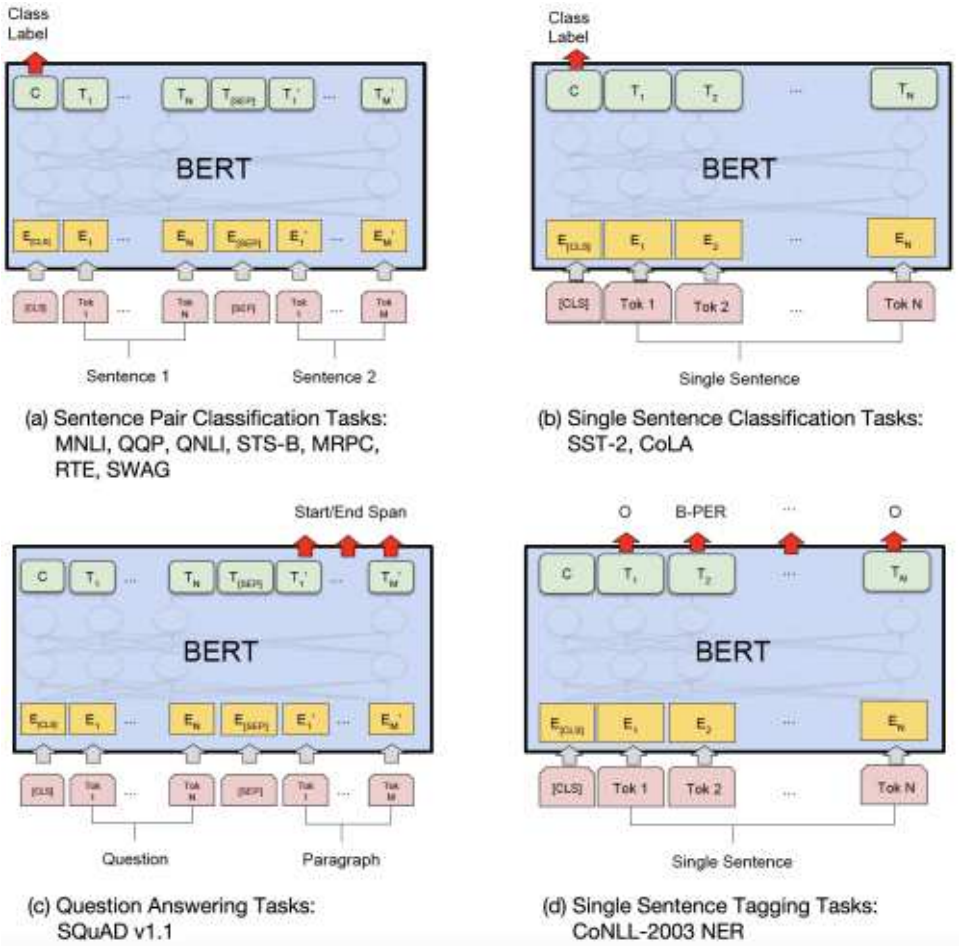


Fig. 5 Fine-Tuning Bert illustrations on a variety of different tasks

Stanford Sentiment SST-2 Treebank is a binary single-sentence classification challenge consisting of sentences derived from film reviews with human comments on their moods [31]. CoLA The Corpus of Linguistic Acceptability is a binary single sentence classification problem aimed at predicting whether or not an English sentence is linguistically "accepted" STS-B The Semantic Textual Similarity Benchmark is a set of sentence pairs derived from news

headlines and other sources [32]. They are scored from 1 to 5, indicating how close the two sentences are in terms of semantic significance. The Microsoft Research Partial Corpus (MRPC) comprises pairs of sentences retrieved mechanically from online news sources, with human annotations indicating whether the sentences in pairs are semantically identical [12]. RTE Recognizing Textual Entailment is a binary task similar to MNLI but requires much less training material. Winograd NLI (Winograd Natural Language Inference) is a tiny natural language inference dataset. According to the GLUE website, there are issues with the data set's design, and any trained system submitted to GLUE performs lower than the 65.1 baseline accuracy in predicting the majority class.

The following chart clearly illustrates the method used in this research. Figure 6 shows titles and abstract writings as components employed in scientific publications. These two components describe the scientific publication's substance. Document classification comprises a vocabulary selection step based on the abstract title of a journal document that may appear in the document and become a representation of the document itself. The BERT stage will undertake pre-training on the vocabulary padding results.

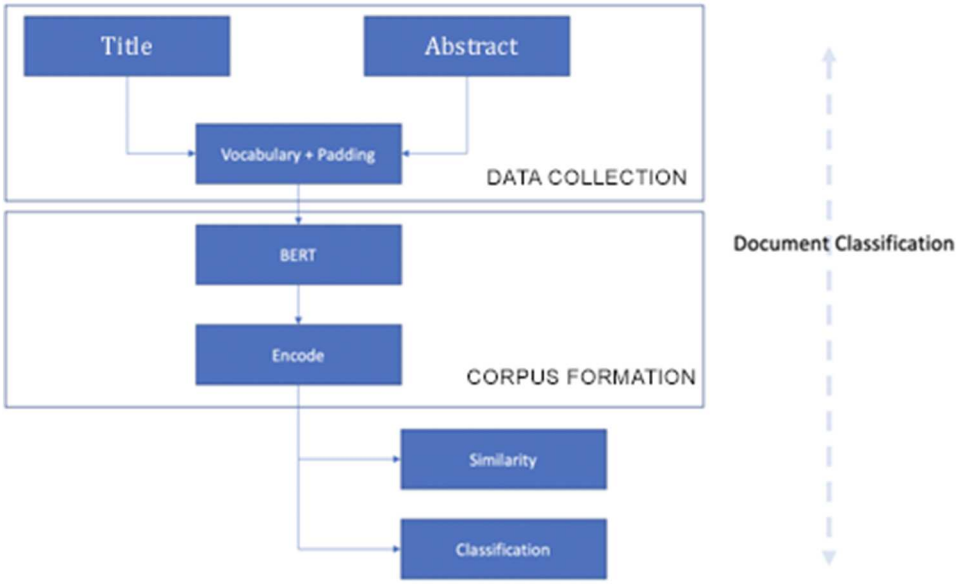


Fig. 6 Process Block.

Similarity and classification coding are steps performed in BERT, especially by transforming a text from the title and abstract into a vector. BERT is a Minimum Viable Product (MVP) NLP; BERT's ability to embed the meaning of words into dense vectors is called a dense vector because each value in the vector has a value from the title and abstract and has contextual values to be in the word from the title and abstract.

F. Data Collection

The SISTER database, linked to the SINTA application (sinta.kemdikbud.go.id) and various other applications within the Directorate General of Higher Education, is used to gather or acquire data for scientific publications. Publication data is supplied in scientific journals, conferences, book chapters, research reports, and other similar publications. In this study, the data was gathered solely from scholarly works in journals, with each piece of information accompanied by a title and abstract. The first step in the acquisition procedure is to return data and clean it using the technique depicted in Figure 7. Stop word removal is the process of deleting non-essential words from text processing, such as punctuation marks, hyphens, and other symbols [33].

Require: *Data Repository*
WHILE *read repository* **do**
id ← *get Unique ID Scientific Publications*
title ← *get Judul Title Scientific Publications*
abstract ← *get Abstract Scientific Publications*
category ← *get Category Field of Science scientific publications*
stopwordremoval *(title)*
stopwordremoval*(abstract)*
savedata *(id, title. abstract, category)*
END WHILE

Fig. 7 Data Acquisition and Data Cleanup Algorithm

G. Making a Corpus

The corpus is created by gathering scientific abstract data from SINTA and the Garuda Portal, where the data collected includes the title of the article, the author's name, the field of research stated by the author, and the abstract. Creating a dictionary will create a BoW, which will be used as a feature in the categorization process later. The corpus is built based on similarity, with the corpus's findings already having similar unified dictionaries. A diagram for creating a corpus algorithm and one for making a corpus using the BERT model are shown below.

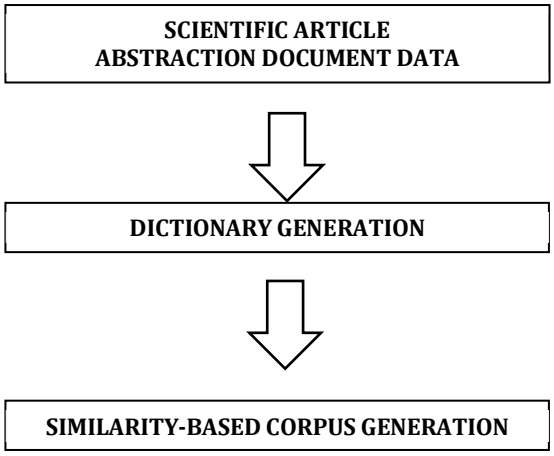


Fig. 8 Diagram of Corpus Creation

Require: *Extracted Data*
  WHILE getdata *(title, abstract)* **do**
    corpus ← *generatecorpussimilarity(title, abstract)*
    savedata(*corpus*)
  END WHILE

Fig. 9 Algorithm of Corpus Creation

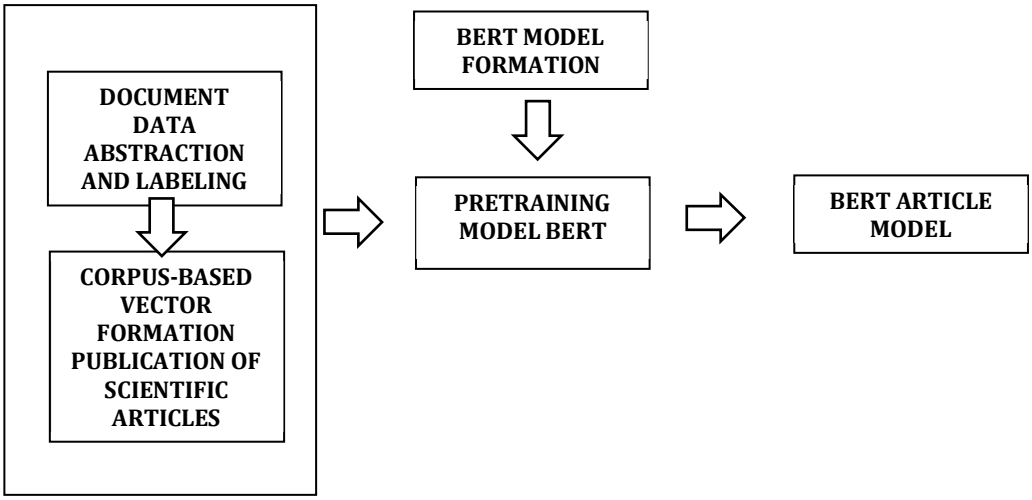


Fig. 10 Diagram of BERT Mode Corpus Creation

H. Classification

The classification process is a Machine Learning function that seeks to develop a predictive model for a class or category by including a set of qualities (also known as features) used to make predictions. The classification method is used to classify scientific fields. The result of the annotation process quantified in the annotation code, is employed on the document representation side (Feature Descriptor). This is true for all documents used throughout the training period. The goal of the training phase is to create a model that can distinguish between shapes or supersets that are included in particular object classes/categories and those that are not. In this phase, the training set prepared is a document that has been represented in vector form which contains the annotation code and the class label it should be.

As a classification method, the proposed algorithm is applicable. The suggested approach uses the semantic annotation features created during the mapping process. This stage begins with developing a classification algorithm that begins with codifying the semantic structure established during the mapping procedure. Build a generic structure for each category based on the training data that supersets all training data in that category. In addition to the structure's creation, the training procedure determines the structure's weight and width limits. The suggested algorithm's training data diagram is illustrated in the following figure.

Fig. 11 Training Diagrams

The testing phase is when the classifier created during training is tested against a new document. The extraction technique, document representation, and other aspects are identical to those in the training phase data preparation, in which the document is cleaned by preprocessing before being extracted in the same manner as in the training phase.

A. Evaluation or Testing

The testing process for measuring classification outcomes using the BERT approach involves testing the accuracy and success rate of text classification using the BERT model established in the previous process. The diagram is shown in Figure 12 below.

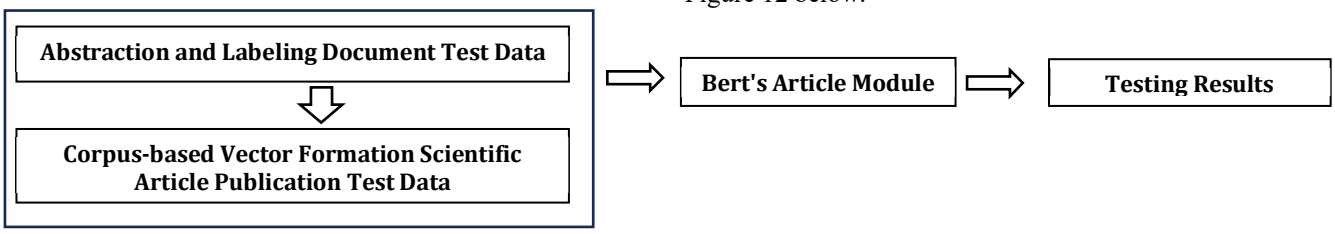


Fig. 12 Training Diagrams

III. Result and Discussion

This part will detail the outcomes of the preceding section's process. Each stage will be detailed, including data collection, vocabulary building, the training procedure, and classification outcomes.

A. Data Collection

The data for this research came from the SISTER database. The publishing statistics used in this study are based on the articles covering the most expertise domains, as shown in Table 1 below. The total number of publications received is 23,444, with 3,865 authors in 35 branches of research. The selection of these 24 disciplines is based on the distribution of the number of articles, which is more than 200 per branch of study, with the aim that the proportions will be sufficient to train each class. The table contains detailed information. Table 2 and Figure 13 show the distribution of articles by category.

TABLE I
EXAMPLE OF REPOSITORY DATA

Id	Title	Knowledge Field
757134	Development of Low Power Carbon Monoxide (CO) Monitoring System Using State Machine Method	Computer Science & IT
757135	Lovebird Feed Composition Optimization Using Particle Swarm Optimization (PSO) Algorithm	Computer Science & IT
757137	Implementation of Complementary Filter in the Design of Feeding Aid for Parkinson's Patients	Computer Science & IT
757139	Development of Medical Record Web Application Case Study of RSIA. Prof. Dr. H. M. Farid Makassar	Computer Science & IT
757142	Implementation of a Web-based Electricity Power Monitoring System and WebSocket Communication Protocol	Computer Science & IT
757143	Optimization of Traveling Salesman Problem on School Transportation Using Genetic Algorithm (Case Study: MI Salafiyah Kasim School Blitar)	Computer Science & IT
757144	Comparative Analysis of the Performance of Multi-Copy and Single-Copy Routing Protocols Based on Node Mobility on Delay Tolerant Networks	Computer Science & IT
757146	Implementation of Dynamic Difficulty Adjustment in Racing Game Using Behavior Tree Method	Computer Science & IT

TABLE II
NUMBER OF ARTICLES DISTRIBUTION

Knowledge Field Category	Total
Economics Econometrics & Finance	382
Education	1325
Social Sciences	335
Computer Science & IT	1522
Agriculture Biological Sciences & Forestry	935
Mathematics	1042
Earth & Planetary Sciences	702
Law Crime Criminology & Criminal Justice	93
Chemical Engineering Chemistry & Bioengineering	622

Knowledge Field Category	Total
Public Health	986
Medicine & Pharmacology	522
Neuroscience	107
Arts	189
Civil Engineering Building Construction & Architecture	430
Nursing	184
Electrical & Electronics Engineering	480
Immunology & Microbiology	227
Health Professions	625
Decision Sciences Operations Research & Management	389
Language Linguistic Communication & Media	262
Biochemistry Genetics & Molecular Biology	224
Sports Science	144
Library & Information Science	199
Religion	144

Fig. 13 Publication Data taken from SISTER

TABLE III
FIELD OF SCIENCE CATEGORY

Label	Knowledge Field Category
0	Economics Econometrics & Finance
1	Education
2	Social Sciences
3	Computer Science & IT
4	Agriculture Biological Sciences & Forestry
5	Mathematics
6	Earth & Planetary Sciences
8	Law Crime Criminology & Criminal Justice
9	Chemical Engineering Chemistry & Bioengineering
10	Public Health
11	Medicine & Pharmacology
12	Neuroscience
13	Arts
14	Civil Engineering Building Construction & Architecture
15	Nursing
16	Electrical & Electronics Engineering
17	Immunology & microbiology
18	Health Professions
21	Decision Sciences Operations Research & Management
22	Language Linguistic Communication & Media
23	Biochemistry Genetics & Molecular Biology
26	Sports Science
28	Library & Information Science
34	Religion

After data collection, the data cleaning technique resulted in a vocabulary of 65,273 words. The obtained words are converted into tokens representing all the words in the data. Several features are not employed because of the omission of stop words and terms that are not even directly related to the publication's title. Table 4 shows the generated vocabulary.

aerizusa	branch	cangang	diekspresikan	equality
aerob	branchionus	cangar	diekstrak	equalization
aerobi	brand	canggih	diekstraksi	equalizer
aerobik	brandan	canggihnya	diekstrapolasi	equation
aerodinamika	branded	canggu	dielaborasi	equationmodeling
aerodinamis	branding	canggung	dielakkan	equations
aerofoil	brandingdepok	cangkang	dielektrik	equestion
aerofood	brandingtaiwan	cangkangnya	dielektriknya	equestrian
aerogenes	branggahan	cangkok	dielektrodeposisi	equevalent
aeroginosa	brangkal	cangkringan	dieliminasi	equilateral
aeromonas	brangkas	cangkul	diemban	equilibrium
aeronautques	brangkasan	cangkupannya	diembannya	equipment
aerosol	brantas	canilaculata	diembargo	equitions
aertembaga	brary	caninum	diemisikan	equity
aeruginosa	brasil	canis	diemulsikan	equityholders

Fig. 14 Examples of Vocabulary Words

B. Making a Corpus

The BERT approach creates a corpus using published data collected. The constructed corpus, known as the BERT article, forms a token that will be employed in the categorization process. BERT's token will generate a vocabulary derivative that counts using the exact substring. Table 4 shows an example of a token generated. The tokens generated have considered the partial use of frequently used terms, reducing the cost of categorization computations. The token can be observed in hashtag tokens.

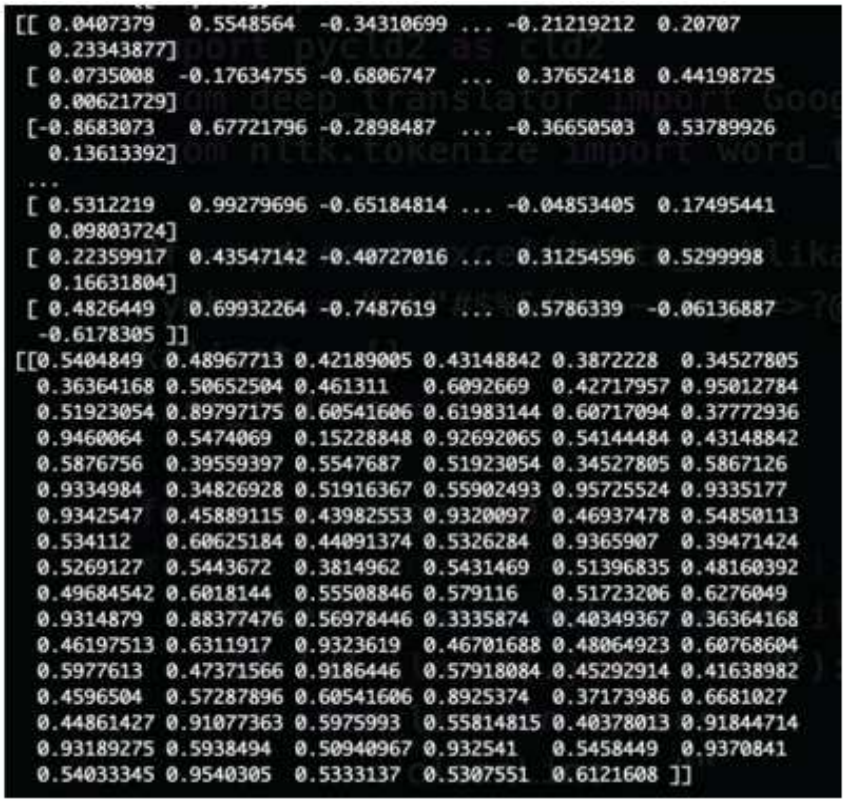
TABLE IV
EXAMPLE OF VOCABULARY TOKEN

[PAD]	##tas	##ura	disi	##yakan	mit	##yyah	##yn
[EOS]	##le	##tri	diba	##yak	mili	##yam	##xi
[UNK]	##ian	##ten	dial	##wati	menca	##wer	##wn
[CLS]	##ti	##ster	dia	##wal	media	##wah	##wl
[SEP]	##et	##son	df	##vasi	medi	##uta	##wit
[MASK]	##ga	##rin	db	##uku	manu	##usia	##was
##s	##aan	##ness	dalam	##uksi	mal	##uni	##ware
##an	##la	##las	cu	##trak	mad	##uat	##uskan
##i	##ce	##he	ci	##tivitas	lea	##tul	##uma
##nya	##lah	##ft	cha	##tip	lap	##tuh	##ulkan
##a	##tor	##fa	cb	##tia	lam	##truksi	##uang
##n	##sis	##esi	bu	##tama	kun	##tim	##tro
##kan	##ve	##ep	bo	##sma	kar	##tam	##tn
##t	##un	##ee	bin	##sip	kala	##sya	##tasnya
##e	##ai	##dang	berse	##sida	iv	##sum	##tannya
##k	##am	##bah	at	##sana	ini	##stasi	##su

The next stage is the feature descriptor formation process. In Figure 15, the results of the feature descriptors obtained from the collection of publications are described. It can be seen below that the formation of a dictionary produces a vector consisting of a collection of words in a scientific article with the frequency of the word.

Fig. 15 Dictionary

The next procedure is the formation of the corpus, which is obtained from word frequency. In this procedure, a corpus is formed on the basis of similarity, where words with similar meanings have the same value. Figure 16 depicts the outcomes of the corpus creation. The generated corpus is referred to as the BERT article.



```
[[ 0.0407379   0.5548564  -0.34310699 ... -0.21219212  0.20707
   0.23343877]
 [ 0.0735008  -0.17634755 -0.6806747  ...  0.37652418  0.44198725
   0.00621729]
 [-0.8683073   0.67721796 -0.2898487  ... -0.36650503  0.53789926
   0.13613392]
 ...
 [ 0.5312219   0.99279696 -0.65184814 ... -0.04853405  0.17495441
   0.09803724]
 [ 0.22359917  0.43547142 -0.40727016 ...  0.31254596  0.5299998
   0.16631804]
 [ 0.4826449   0.69932264 -0.7487619  ...  0.5786339  -0.06136887
  -0.6178305 ]]
[[0.5404849  0.48967713 0.42189005 0.43148842 0.3872228  0.34527805
  0.36364168 0.50652504 0.461311   0.6092669  0.42717957 0.95012784
  0.51923054 0.89797175 0.60541606 0.61983144 0.60717094 0.37772936
  0.9460064  0.5474069  0.15228848 0.92692065 0.54144484 0.43148842
  0.5876756  0.39559397 0.5547687  0.51923054 0.34527805 0.5867126
  0.9334984  0.34826928 0.51916367 0.55902493 0.95725524 0.9335177
  0.9342547  0.45889115 0.43982553 0.9320097  0.46937478 0.54850113
  0.534112   0.60625184 0.44091374 0.5326284  0.9365907  0.39471424
  0.5269127  0.5443672  0.3814962  0.5431469  0.51396835 0.48160392
  0.49684542 0.6018144  0.55508846 0.579116   0.51723206 0.6276049
  0.9314879  0.88377476 0.56978446 0.3335874  0.40349367 0.36364168
  0.46197513 0.6311917  0.9323619  0.46701688 0.48064923 0.60768604
  0.5977613  0.47371566 0.9186446  0.57918084 0.45292914 0.41638982
  0.4596504  0.57287896 0.60541606 0.8925374  0.37173986 0.6681027
  0.44861427 0.91077363 0.5975993  0.55814815 0.40378013 0.91844714
  0.93189275 0.5938494  0.50940967 0.932541   0.5458449  0.9370841
  0.5403334S 0.9540305  0.5333137  0.5307551  0.6121608 ]]
```

Fig. 16 Bert Article Generation

C. *Classification Result*

The following is the classification results utilizing several methods, including SVM, Naive Bayes, and Deep Learning, with diverse outcomes.

accuracy: 36.42%

	true Co...	true Ilm...	true Ke...	true Ma...	true Ma...	true Ma...	true Pe...	true Pe...	true Per...	true Per...	true Te...	class pr...
pred. C...	3	0	0	0	0	0	0	0	0	0	0	100.00%
pred. Il...	0	61	2	0	0	0	0	0	0	4	0	91.04%
pred. K...	0	1	75	0	0	0	0	0	0	27	0	72.82%
pred. M...	0	0	0	42	1	1	0	0	0	0	0	95.45%
pred. M...	2	7	7	25	111	7	7	3	4	16	4	57.51%
pred. M...	0	0	0	0	1	1	0	1	1	0	0	25.00%
pred. P...	0	0	0	0	0	1	31	1	0	0	18	60.78%
pred. P...	1	0	0	0	0	3	0	8	7	0	0	42.11%
pred. P...	72	78	62	57	51	90	75	108	137	58	62	16.12%
pred. P...	0	0	3	0	0	0	1	0	1	17	0	77.27%
pred. T...	1	0	0	0	0	1	1	0	0	0	14	82.35%
class re...	3.80%	41.50%	50.34%	33.87%	67.68%	0.96%	26.96%	6.61%	91.33%	13.93%	14.29%	

Fig. 17 Test Results using the SVM Algorithm

accuracy: 56.45%

	true Co...	true Ilm...	true Ke...	true Ma...	true Ma...	true Ma...	true Pe...	true Pe...	true Per...	true Per...	true Te...	class pr...
pred. C...	26	1	2	0	2	10	3	18	24	0	5	28.57%
pred. Il...	0	115	5	0	1	0	0	0	0	5	0	91.27%
pred. K...	0	5	117	2	0	0	1	0	1	33	1	73.12%
pred. M...	0	1	0	96	21	5	0	1	3	2	1	73.85%
pred. M...	4	4	1	12	107	10	2	5	6	3	0	69.48%
pred. M...	14	2	0	6	15	39	2	25	35	2	2	27.46%
pred. P...	1	5	0	1	1	1	73	1	2	2	17	70.19%
pred. P...	16	1	0	2	4	17	1	35	43	2	3	28.23%
pred. P...	10	0	1	3	7	18	3	28	32	1	3	30.19%
pred. P...	1	11	21	2	5	1	5	0	0	71	2	59.66%
pred. T...	7	2	2	0	1	3	25	8	4	1	64	54.70%
class re...	32.91%	78.23%	78.52%	77.42%	65.24%	37.50%	63.48%	28.93%	21.33%	58.20%	65.31%	

true Computer Software

Fig. 18 Test Results using the Naive Bayes Algorithm

accuracy: 52.59%

	true Co...	true Ilm...	true Ke...	true Ma...	true Ma...	true Ma...	true Pe...	true Pe...	true Per...	true Per...	true Te...	class pr...
pred. C...	47	6	3	7	8	34	10	62	70	4	17	17.54%
pred. Il...	0	107	4	0	1	0	3	2	1	5	1	86.29%
pred. K...	0	4	116	0	0	1	0	0	0	2	51	2	65.91%
pred. M...	2	6	0	93	8	1	0	1	0	4	0	80.87%
pred. M...	2	6	0	14	132	9	1	2	3	2	1	76.74%
pred. M...	4	4	3	3	2	28	0	7	13	0	2	42.42%
pred. P...	7	4	5	5	7	5	98	5	6	11	54	47.34%
pred. P...	12	3	0	1	2	19	0	29	41	1	3	26.13%
pred. P...	3	0	0	0	0	5	0	11	11	0	0	36.67%
pred. P...	0	4	17	1	2	2	0	0	0	43	0	62.32%
pred. T...	2	3	1	0	1	1	3	2	3	1	18	51.43%
class re...	59.49%	72.79%	77.85%	75.00%	80.49%	26.92%	85.22%	23.97%	7.33%	35.25%	18.37%	

Fig. 19 Test results using the Deep Learning Algorithm and Rectifier activities

accuracy: 53.61%

	true Co...	true Ilm...	true Ke...	true Ma...	true Ma...	true Ma...	true Pe...	true Pe...	true Per...	true Per...	true Te...	class pr...
pred. C...	54	25	7	5	29	21	58	69	7	32	17.31%	
pred. Il...	0	96	4	0	0	0	0	0	0	8	0	88.89%
pred. K...	1	9	130	3	5	1	1	0	1	71	1	58.30%
pred. M...	0	2	0	95	11	2	1	3	2	4	0	79.17%
pred. M...	3	8	2	14	137	11	0	4	5	3	2	72.49%
pred. M...	3	1	0	0	2	27	0	7	8	0	1	55.10%
pred. P...	0	1	1	1	1	1	82	1	1	6	28	66.67%
pred. P...	2	0	0	0	0	4	1	8	9	0	0	33.33%
pred. P...	14	2	1	4	3	29	1	39	54	1	3	35.76%
pred. P...	0	3	6	0	0	0	0	0	0	22	0	70.97%
pred. T...	2	0	0	0	0	0	8	1	1	0	31	72.09%
class re...	68.35%	65.31%	87.25%	76.61%	83.54%	25.96%	71.30%	6.61%	36.00%	18.03%	31.63%	

true Ilmu Kefarmasian

Fig. 20 Trial Results using Maxout activities.

IV. CONCLUSION

Methods for identifying scientific fields have been developed through research. The BERT classification method forms the corpus, which was produced from the BERT Model article. This research also successfully identified scientific disciplines based on abstracts and titles from the scholarly publications of 23.499 lecturers. The method used to determine the level of similarity on research topics of lecturers' scientific publications yielded a level of accuracy in the test set of 92.876 from the training test 95.0345time, system throughput, and energy consumption), demonstrating that the RGL scheme outperforms well-known RSSI-based schemes.

REFERENCES

[1] A. M. and S. Salama, "Discovering Performance Evaluation Features of faculty Members using Data Mining Techniques to Support Decision Making," *Int. J. Comput. Appl.*, vol. 178, no. 49, pp. 25–29, Sep. 2019, doi: 10.5120/ijca2019919417.

[2] R. J. Suhatril, B. Mutiara, A. Suhendra, and I. M. Wiryana, “Strategi Klasifikasi Melalui Pembobotansimilaritas Berbasis Semantic Network,” Universitas Gunadarma, 2015.

[3] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, “Some Effective Techniques for Naive Bayes Text Classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 11, pp. 1457–1466, Nov. 2006, doi:10.1109/tkde.2006.180.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *ArXiv*, vol. abs/1810.0, 2019.

[5] A. M. and S. Salama, “Discovering Performance Evaluation Features of faculty Members using Data Mining Techniques to Support Decision Making,” International Journal of Computer Applications, vol. 178, no. 49, pp. 25–29, Sep. 2019, doi: 10.5120/ijca2019919417.

[6] A. M. Dai and Q. V Le, “Semi-supervised Sequence Learning,” In Advances in Neural Information Processing Systems (pp. 3079–3087). Montreal, Canada: Curran Associates, Inc, 2015.

[7] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” Tech. Rep., 2018.

[8] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” 2018, arXiv:1801.06146. [Online]. Available: http://arxiv.org/abs/1801.06146

[9] M. E. Peters *et al.*, “Deep Contextualized Word Representations,”. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT’18): Vol 1, 2018, Jun 1-6, New Orleans, LA, USA. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 2227-2237.

[10] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” 2015, arXiv:1508.05326

[11] A. Williams, N. Nangia, and S. R. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference,” 2017, arXiv:1704.05426. [Online]. Available: http://arxiv.org/abs/1704.05426

[12] Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases.InProc. of the 3rd Int. Workshop on Paraphrasing, pp. 9–16, Jeju island, Korea, 2005

[13] Tjong Kim Sang, E. and Fien De Meulder. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.” Conference on Computational Natural Language Learning, 2003.

[14] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” Association for Computational Linguistics (ACL), 2016.

[15] Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al., “Attention is All you Need,” Proceedings.neurips.cc, [Online]. Available: https://proceedings.neurips. cc/paper/7181-attention-is-all-you-need.2023

[16] P. F. Brown, V. J. Della Pietra, P. V de Souza, J. C. Lai, and R. L. Mercer, “Class-Based n-gram Models of Natural Language,” *Comput. Linguist.*, vol. 18, pp. 467–479, 1992.

[17] R. K. Ando and T. Zhang, “A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data,” *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.

[18] J. Blitzer, R. T. Mcdonald and F. Pereira, "Domain adaptation with structural correspondence learning", Proc. Conf. Empirical Methods Natural Lang. Process., pp. 22-23, Jul. 2007.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean: Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 26, pp.3111--3119, 2013.

[20] J. Pennington, R. Socher, and C. Manning, ‘‘GloVe: Global vectors for word representation,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[21] J. Turian, L. Ratinov and Y. Bengio, Word representations: A simple and general method for semi-supervised learning, 2010.

[22] A. Mnih and G.E. Hinton, "A Scalable Hierarchical Distributed Language Model", Proceedings of Neural Information Processing Systems 21 (NIPS 2008), pp. 1-8, 2008.

[23] Kiros R, Zhu Y K, Salakhutdinov R, Zemel R, Torralba A, Urtasun R, Fidler S. Skip-thought vectors. arXiv: 1506.06726, 2015. https://arxiv.org/abs/1506.06726, June 2017.

[24] L. Logeswaran and H. Lee, “An efficient framework for learning sentence representations,” *ArXiv*, vol. abs/1803.0, 2018.

[25] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents", 1st Workshop on Representation Learning for NLP, 2015.

[26] Y. Jernite, S. R. Bowman, and D. A. Sontag, “Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning,” *ArXiv*, vol. abs/1705.0, 2017.

[27] F. Hill, K. Cho, and A. Korhonen, “Learning Distributed Representations of Sentences from Unlabelled Data,” arXiv preprint arXiv:1602.03483, 2016.

[28] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multi task learning", Proc. of ICML, 2008.

[29] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Glue: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding,” *ArXiv*, vol. abs/1804.0, 2018.

[30] Z. Qu, X. Song, S. Zheng, X. Wang, X. Song, and Z. Li, “Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification,” 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Jan. 2018, doi: 10.1109/bigcomp.2018.00124.

[31] R. Socher, J. Wu, A. Perelygin, C.D. Manning, J. Chuang et al., "Recursive deep models for semantic compositionality over a sentiment treebank", EMNLP, 2013. Association for Computational Linguistics, Seattle, 2013.

[32] B. Walek and V. Fojtik, “A hybrid recommender system for recommending relevant movies using an expert system,” Expert Systems with Applications, vol. 158, p. 113452, Nov. 2020, doi:10.1016/j.eswa.2020.113452.

[33] H. Saif, M. Fernández, Y. He, and H. Alani, “On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter,” The Open University: Milton Keynes, UK, 2014.