International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 1, February 2022, pp. 1018~1029
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i1.pp1018-1029

Automated hierarchical classification of scanned documents using convolutional neural network and regular expression

Rifiana Arief, Achmad Benny Mutiara, Tubagus Maulana Kusuma, Hustinawaty
Faculty of Computer Science and Information Technology, Gunadarma University, Depok, Indonesia

Article Info

Article history:

Received Jan 7, 2021
Revised Jul 5, 2021
Accepted Aug 6, 2021

Keywords:

Classification
Convolutional neural network
hierarchical
Regular expression
Scanned documents

ABSTRACT

This research proposed automated hierarchical classification of scanned documents with characteristics content that have unstructured text and special patterns (specific and short strings) using convolutional neural network (CNN) and regular expression method (REM). The research data using digital correspondence documents with format PDF images from *Pusat Data Teknologi dan Informasi* (Technology and Information Data Center). The document hierarchy covers type of letter, type of manuscript letter, origin of letter and subject of letter. The research method consists of preprocessing, classification, and storage to database. Preprocessing covers extraction using Tesseract optical character recognition (OCR) and formation of word document vector with Word2Vec. Hierarchical classification uses CNN to classify 5 types of letters and regular expression to classify 4 types of manuscript letter, 15 origins of letter and 25 subjects of letter. The classified documents are stored in the Hive database in Hadoop big data architecture. The amount of data used is 5200 documents, consisting of 4000 for training, 1000 for testing and 200 for classification prediction documents. The trial result of 200 new documents is 188 documents correctly classified and 12 documents incorrectly classified. The accuracy of automated hierarchical classification is 94%. Next, the search of classified scanned documents based on content can be developed.

Corresponding Author:

Rifiana Arief
Faculty of Computer Science and Information Technology, Gunadarma University
Margonda Raya 100 Pondok Cina, Depok, 16424, Indonesia
Email: rifiana@staff.gunadarma.ac.id

1. INTRODUCTION

The increase and variety of documents make document classification necessary to direct, summarize and organize documents effectively. Document classification can be defined as the grouping of documents automatically into certain classes based on the similarity of document content [1]. Document classification is one aspect of the fundamental problems experienced in the management of information management and information retrieval tasks. Document classification is the process of grouping documents into predetermined category criteria. In digitalizing the data, document classification needs to perform for effective huge data organization that saves a lot of user time and helps in analyzing customer feedback. In the task of classification, the condition which more than two classes existing is called multi-class classification [2]. The document collections are organized as hierarchical class structure in many application fields: Web taxonomies, email folders and product catalogs, this is called hierarchical classification [3]. The inclining document availability in the organization and the rapid growth of data cause the automated document classification to become an important key method for searching documents quickly and accurately [4].

Machine learning algorithm can be used in designing such new logic that can classify the document and manage it when a new instance of data arrives, without human intervention [5].

To date, most of the text classification methods generally used to assign multiple topics to documents [6], grouping of documents into a fixed number of predefined classes [7], sentiment analysis to determine the viewpoint/polarity of a writer with respect to some topic [8], spam filtering of emails [9], automatic hate speech detection [10]. In the era of big data, the increasing number of complex documents makes traditional machine learning methods difficult to implement because conventional learning processes are not designed for big data and will not work properly with high data volumes. Traditional machine learning algorithms such as Naïve Bayes and others are designed for data that will actually be loaded into memory and cannot be handled should the data refer to large data. Therefore, other algorithms are required to handle it [11]. Newer machine learning method for document classification is taken from deep learning. This becomes increasingly necessary because the performance of conventional methods will decline with the inclining number of documents. Deep learning has been widely used for image processing, but numerous recent studies have implemented deep learning in other domains such as text and data mining [12]. The increase and diversity of data, the validity of uncertain data form, as well as the need for access quickly lead to a classification trend by utilizing deep learning neural network that has more capabilities than conventional methods of machine learning for big data characteristics [13].

Numerous government institution in Indonesia archives various types of correspondence documents digitally through the scanning process. In digital correspondence documents there are various different types of letters and every document has a letter number in which there is a certain meaning, among others, the origin of the letter and the subject of the letter. The existing correspondence documents filing system still requires operator assistance to interpret the information in the digital correspondence documents and classify it according to the criteria in the document hierarchy (letters) manually. The amount of letter from various criteria that continues to increase requires a method that is able to classify each letter automatically into an appropriate hierarchy according to the applicable correspondence procedure in the institution, and subsequently file the classified documents to the database. The characteristics of data possessed by digital correspondence documents are the existence of unstructured text information content and number of letters with short strings with special code specifications. The letter hierarchy consists of several levels and each level consists of several class categories. Analysis of unstructured text information content for main level classification (to get the letter type criteria), and short string with special code specifications for the next level classification/sub-document (to get the criteria for the type of letter script, the origin of the letter and the subject of the letter) is needed.

Problems of image document classification can be done with a text-based approach with the help of optical character recognition (OCR) and machine learning [14]. Various methods are used to solve the problem of classification of scanned documents by adding the preprocessing process to convert the scanned document into a text document first. Tesseract OCR was the best open source available in various languages used to extract and recognize text content from scanned documents in image format [15]. Besides, there are other various OCR applications that we can use, such as free online OCR, online OCR.net, free OCR, i2OCR, Google Vision OCR. Based on small trial, the accuracy performance of Google Vision OCR was the best comparing to other OCR tools [16]. In previous studies, the automatic classification of scanned electronic health record documents done by extracted text using (OCR and multiple text classification machine learning models, including both "bag of words" and deep learning approaches [17], the classifying image spam detection using OCR, machine learning and natural language processing [18] and the classifying promotion images using OCR and Naïve Bayes classifier [19]. From research [17]-[19] show that text-based classification systems can accurately classify scanned documents. The problem of text classification can also be solved by deep learning using the convolutional neural network (CNN) such as for hate speech classification [20], news classification [21] and sentiment analysis [22]. Classification of text documents based on matching input strings efficiently used the regular expression conducted in [23]-[25]. Other research performs hierarchical classification for news article document [26] and multi-level classification for medical datasets [27]. Hierarchical classification by combining CNN and recurrent neural network (RNN) method as deep learning model for learning into each level in the text document hierarchy using WOS-11967, 46985 and 5736 datasets with accuracy of 82.3% was carried out in [28]. The research studies mentioned above, hierarchically classified the documents and its sub-documents, but the method used has not been able to classify different documents based on the extraction of short and unique codes from the contents of the document. Classification tasks in the form of big data sentiment analysis can use big data tools such as Apache Hadoop [29] and Apache Spark [30].

Based on review from several previous studies, the OCR-assisted classification of scanned documents had been carried out. However, there has not been classification of scanned documents based on the document hierarchy that has data characteristics in the form of unstructured text information content and short strings with special code specifications. This research proposes solution of hierarchical

classification/multi-level classification for scanned documents that are automatically supported by OCR with the combination of CNN and regular expressions method. This research aims to automate the digital correspondence documents classification process belonging to Technology and Information Data Center institution in the form of scanned documents in the PDFImages format according to the predetermined rules of official script administration. This proposed method has the advantage of being able to automate the classification process of scanned image-formatted documents with the condition of documents that have unstructured text content and have special patterns (specific and short strings) so that each image-format scanned document will be classified based on the document hierarchy with a depth of 4 levels, namely manuscripts letter -> letter type -> letter origin -> letter subject automatically. With this automation the classification process no longer requires manual human intervention to classify types of letters, types of letter manuscripts, types of letter origins and types of subject matter. By implementing this method, every digital correspondence documents previously will be classified automatically according to the letter hierarchy and after that will be archived into a filing system that has been prepared for big data needs, namely the big data architecture in the Hive database automatically.

2. RESEARCH METHOD

This research implemented automated hierarchical classification (4 level) of scanned documents with the help of Tesseract OCR, with the character of the document content have unstructured text content and have special patterns with storage on Hadoop architecture database with big data technology. We adapt the research of Kowsari [28] which performs a hierarchical classification (2 level) using the CNN and RNN methods. We use CNN method and change RNN methods with regular expression. CNN method uses to classify letter types and regular expression method use to extract information on letter numbers which contain short and short codes but have special meanings and specifications according to the numbering rules in the established correspondence so that it can classify manuscript of letter, origin of letter and subject of letter accurately quickly. Hierarchical classification is carried out to automatically obtain the criteria for manuscripts of letter, types of letters, origin of letters and subject of letters from every digital correspondence documents as shown in Figure 1.

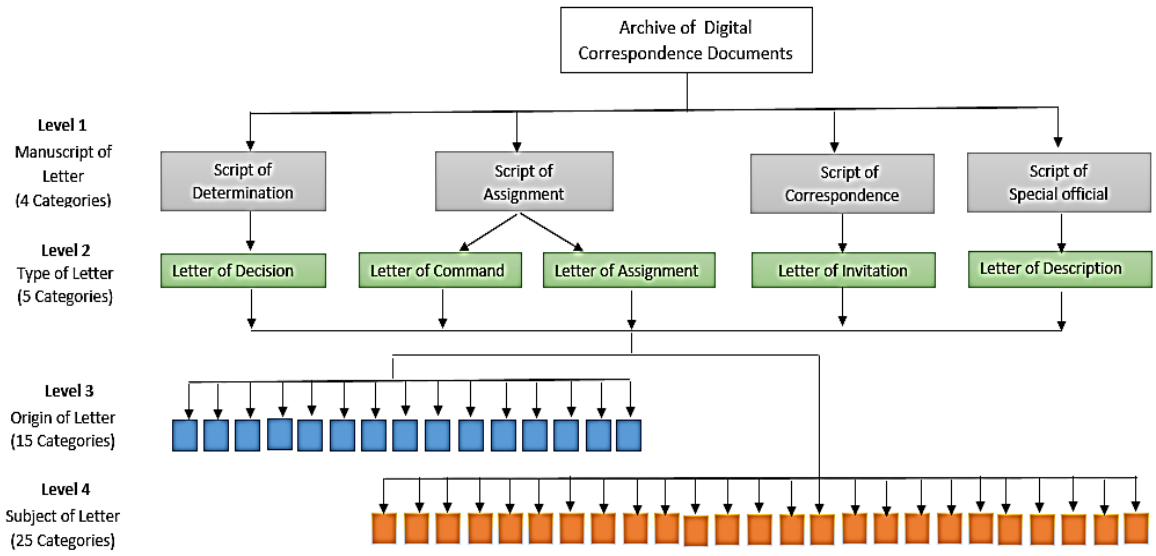


Figure 1. Hierarchical of digital correspondence documents

The object of this research uses digital correspondence documents in PDFImages format as shown in Figure 2 (see in appendix), which is consisted of 5 types of letters, namely letter of decision, letter of statement, letter of command, letter of assignment, letter of invitation. Each document has a unique number with a special code indicating the origin of letter and the subject of letter. By extracting the special code in the letter number on the document it can show the information classification of the letter manuscript (there are 4 categories), the origin of the letter (there are 15 categories) and the subject of the letter (there are 25 categories).

The hierarchical classification of digital correspondence documents consists of 3 stages: preprocessing, classification and storage to database illustrated in Figure 3. The contribution in this research is a classification block that produces a multi-level classification (4 level) model by combining the CNN and regular expression methods so that it can classify scanned documents in the form of digital correspondence documents based on the manuscript of letter, type of letter, origin of the letter and subject of the letter then the classified documents are automatically saved to the database Hive.

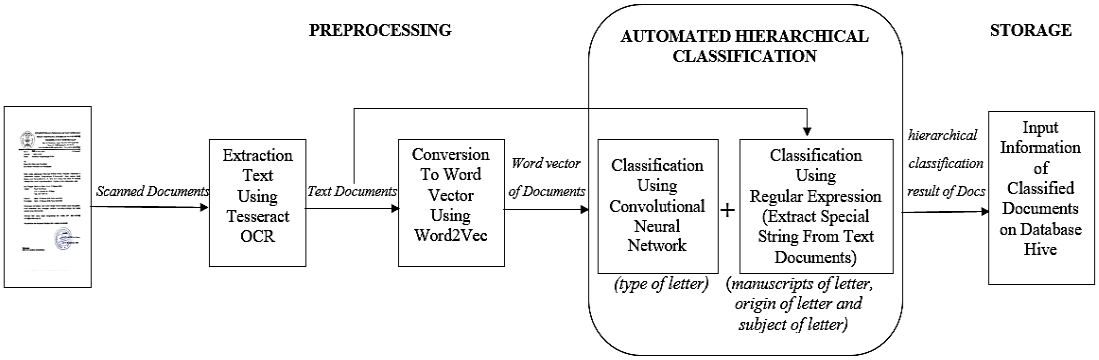


Figure 3. Method of automated hierarchical classification of digital correspondence documents

2.1. Preprocessing

Before the scanned documents are classified, there is preprocessing stage of extraction and stage of word vector formation. The stage of extraction consists of converting PDFImages documents into JPG documents with Apache PDFBox followed by the extraction of text contents from image document using Tesseract OCR [13]. The extraction result of text document is used for classification based on the text approach. The stage of word vector formation is to transform the text of words in a document into the form of word vector. Word vectorization needs to be completed because the document classification process using the convolutional neural network (CNN) is not able to use text. Word vectorization is made based on words from the extraction results of all letter documents using Word2Vec [11]. All the contents of the text documents in the folder of the extraction result are united and stored in one file. Then, the separation of each sentence and tokenization to receive the words from the document. Preprocessing to eliminate all numbers, symbols, and special symbols. The process of a vector model formation of each word with 1 iteration and epoch of 1 time training will form a word vector should a word occurs at least 5 times and the length of each word vector is 100. The result of word vector will be stored in the word vector path for use in the classification stage.

2.2. Hierarchical classification

Hierarchical classification aims to classify each scanned document of digital correspondent documents as a document that is automatically type of letter, type of manuscript letter, origin of letter, and subject of letter. The hierarchical classification model to digital correspondence documents combines CNN and regular expression method as shown in Figure 4. CNN method processes the word vector representing words in text documents to obtain criteria for type of letter. Meanwhile, regular expression method extracts and captures information content in documents with special patterns to obtain the criteria for text, origin, and subject of letter quickly and accurately without having to go through a training process as CNN method. Regular expression will certainly save time and money. The classification criteria consist of 5 types of letters, 4 types of manuscript letters, 15 origins of letters and 25 subjects of letters. Evaluation of classification modeling is made by training and testing as well as predictions of new documents. The results of the classification stage are classified documents to be stored in database.

2.2.1. Classification for type of letter

The classification process with CNN begins with the establishment of CNN model architecture by adding layers and configuring input layer, extraction layer, and output layer specifically for use in the document classification process [18]. After the architecture is configured, the data is loaded to be used for training and testing datasets from OCR extracted text documents in the form of word vectors. The training process or training uses a training dataset on the architecture that has been created and evaluation uses a dataset testing of models that have been trained.

The formation of the CNN architecture for the classification of 5 types of letters will be used to classify documents based on the type of letter. It starts by configuring a basic neural network with batch 100, vector 300, Epochs 10, deduction of word length 256, l2 0.0001, layer of feature maps 100, random data, RELU activation function, updater Adam (0.01). Next, configure the layer on the neural network by adding an input layer, convolution layer cnn3 with kernel size (3, 300), stride (1, 300), convolution layer cnn4 with kernel size (4, 300), Stride (1, 300), convolution layer cnn5 with kernel size (5, 300) , Stride (1, 300), with input from input and output layers to feature layer, pooling layer with maximum type, DropOut (0.5), combined layer (cnn3, cnn4, cnn5), output layer with LossFunction.MCXENT function, SOFTMAX activation function, input (3 * cnnLayerFeatureMaps) and Output of 5 class. The 5 classes for the type of letter (letter of decision, letter of statement, letter of command, letter of assignment, letter of invitation).

The process of loading word vector that will be used for the classification of type of letter. Started by taking the word vectors from the path where the word vectors are formed and taking the training dataset containing the word document vector for the training process and the testing dataset that contains the word document vector for the testing process. The training and testing dataset were previously taken from the folder path of 5 different types of letters that were determined both for the training and testing process. Each document from each folder with a different type will be accommodated in a file and will be mapped to the entire contents of the document in such folder, thus each document will be displayed in the form of a word vector, and each followed by labeling that has been done before. Documents in the form of word vectors will then go through a classification process that is training and testing according to the CNN architecture.

The training process will manage and arrange documents in the form of word vectors prepared for training according to the label given for each document with Epoch 10 to documents labeled with different types of letters that have been set. Next, an evaluation is carried out through the testing process of documents labeled with different types of letters that have been set. The document classification training and testing process will run on the convolutional neural network architecture that has been determined by running the process in the input layer, feature extraction layer and classification layer. Thus, the evaluation results of classification are obtained.

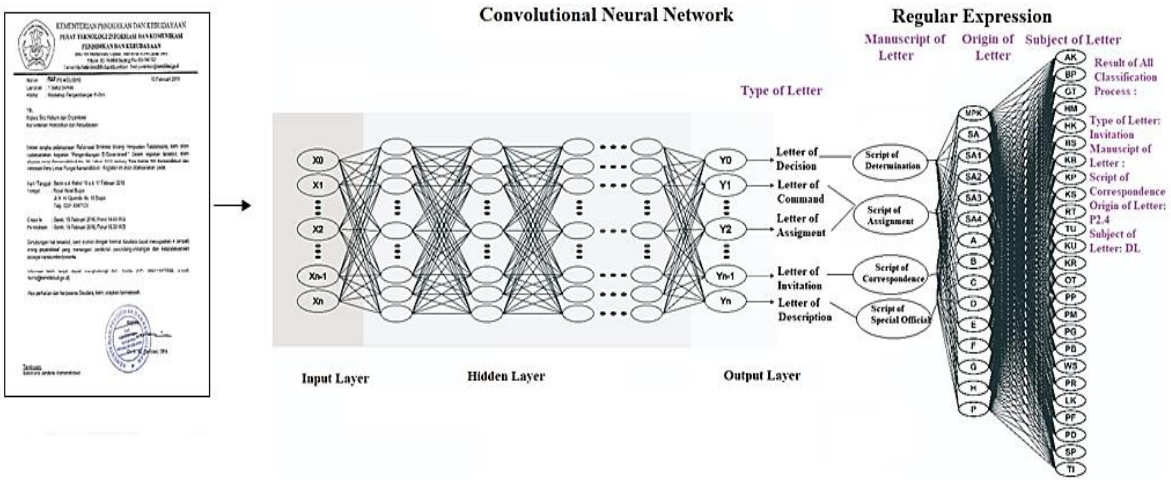


Figure 4. Model of automated hierarchical classification scanned document using combination of convolutional neural network and regular expression

2.2.2. Classification for type manuscript of letter, origin of letter, and subject of letter

More detailed classification for document is completed by regular expression method after classifying the type of letter using the CNN method. Each letter has text content, and inside this, there are certain codes that can describe the origin of the letter and the contents of the letter. The rules in coding the origin of the letter and the subject of the letter are highly simple and short. Therefore, regular expression method is highly effective in being able to find codes with certain patterns in documents quickly and accurately [20].

a) Classification algorithm for manuscript of letter

```
Input: Classification Result base on Type of Letter
Output: Classification for Manuscript of Letter
```

```
1. Retrieve documents that have been classified according to Type of Letter through the
   Convolutional Neural Network method
2. Decide the pattern for Manuscript Letter based on Type of Letter
   If Type of Letter = Letter of Decision Then Manuscript Letter = Script of Determination
   If Type of Letter = Letter of Command and Letter of Assignment
                       Then Manuscript Letter = Script of Assignment
   If Type of Letter = Letter of Invitation Then Manuscript Letter = Script of
   Correspondence
   If Type of Letter = Letter of Description Then Manuscript Letter = Script of Special
   official
   In addition, Manuscript Letter = No Category
3. Match the criteria for Manuscript of Letter that is suitable based on type of Letter
4. Receive the suitable criteria for Manuscript of Letter
5. Save the classification result for Manuscript of Letter from documents
```

b) Classification algorithm for origin letter

```
Input: Content of Classified Document base on certain Type of Letter
Output: Classification for Origin of Letter (15 Categories)
1.  Read Content for Classified Document of Certain Type of Letter
2.  Decide pattern of Regular Expression for the criteria type of origin of letter
3.  Match the content of document with pattern of Regular Expression that has been set
4.  If there is string matching to the criteria for origin of letter in its content then
    value of string is classified as the origin of letter.
5.  If not match Then classification result for the origin of letter = none
6.  Receive value of string found matching with the pattern criteria that has been made
7.  Save the value matching to the pattern found as the classification result for the
    origin of letter
```

Table 1 describes pattern of regular expression for classifying documents based on the origin code of letter. In the origin code of letter, there are various patterns for the origin of letter, from simple pattern, for example, containing only strings/MPK/to numbering details separated by dots and combinations of letters and numbers, for example /SA4.A1/.

Table 1. Pattern of regular expression for origin of letter (15 categories)

No	Pattern of Regular Expression	Code	Categories
1	"/\/\bMPK\//"	/MPK/	*Kementerian Pendidikan dan Kebudayaan* (Minister of Education and Culture)
2	"/\/\bMPK\.[A-P]\//"	/MPK.A/ until /MPK.P/	Variation code detail of categories Minister of Education
3	"/\/\bMPK\.[A-P]\d{1}\//"	/MPK.A1/ until /MPK.P9/	
4	"/\/\bSA\//"	/SA/	*Staff Ahli Menteri* (Minister's Expert Staff)
5	"/\/\bSA\.[A-P]\d{1}\//"	/SA.A1/ until /SA.P9/	Variation code detail of categories Minister's Expert Staff
6	"/\/\bSA[1-4]//"	/SA1/ until /SA4/	SA1 *Staff Ahli Bidang Inovasi dan Daya Saing* (Expert Staff for Innovation and Competitiveness) SA2 *Staff Ahli Bidang Hubungan Pusat dan Daerah* (Expert Staff for Central and Regional Relations) SA3 *Staff Ahli Bidang Pembangunan Karakter* (Expert Staff for Character Development) SA4 *Staff Ahli Bidang Regulasi Pendidikan dan Kebudayaan* (Expert Staff for Education and Culture Regulation)
7	"/\/\bSA[1-4]\.[A-P]\d{1}\//"	/SA1.A1/ until/SA4.P9/	Variation code detail of categories Expert Staff
8	"/\/[A-P]\//"	/A/ until /P/	A *Sekretariat Jenderal* (General Secretariat) B *Direktorat Jenderal Guru dan Tenaga Kependidikan* (Directorate General of Teachers and Education Personnel) C *Direktorat Jenderal Pendidikan Anak Usia Dini dan Pendidikan Masyarakat* (Directorate General of Early Childhood Education and Community Education) D *Direktorat Jenderal Pendidikan Dasar dan Menengah* (Directorate General of Primary and Secondary Education) E *Direktorat Jenderal Kebudayaan* (Directorate General of Culture) F *Inspektorat Jenderal* (Inspectorate General) G *Badan Pengembangan dan Pembinaan Bahasa* (Ministry of Education and Culture's National Agency for Language Development and Books) H *Badan Penelitian dan Pengembangan* (Research and Development Agency) P *Pusat* (Center)
9	"/\/[A-P]\.[A-P]\d{1}\//"	/A.A1/ until /P.P9/	
10	"/\/[A-P]{1}\d{1}/"	/A1 / until /P9/	Variation code detail of categories from Number. 8
11	"/\/[A-P]{1}\d{1}.\d{1}\//"	/A1.1/ until /P9.9/	

c) Classification algorithm for subject of letter

Classification algorithm for the subject of letter is as same as algorithm for the origin of letter but with a simpler pattern (not varied for 25 categories). Table 2 describes pattern of regular expression for classifying documents based on code for subject of letter. For example, the regular expression pattern "/\/[A][K]\//" is to search for criteria in regard to the subject of letter with the code /AK/. Thus, this will search for text document of OCR result containing a string /AK/ for instance in the document with letter number 0568/A.A1/AK/2016, thus we will receive /AK/ as a criterion for subject of letter indicating subject of letter on Accreditation.

Table 2. Pattern of regular expression for subject of letter (25 categories)

No	Pattern of Regular Expression	Code	Categories	No	Pattern of Regular Expression	Code	Categories
1	"/\/[A][K]\//"	AK	*Akreditasi* (Accreditation)	14	"/\/[H][M]\//"	HM	*Hubungan Masyarakat* (Public Relations)
2	"/\/[B][P]\//"	BP	*Bantuan Pendidikan* (Education Assistance)	15	"/\/[P][P]\//"	PP	*Pendidikan dan Pelatihan* (Education and Training)
3	"/\/[G][T]\//"	GT	*Guru & Tenaga Kependidikan* (Teacher & Education Personel)	16	"/\/[P][M]\//"	PM	*Pendidikan Masyarakat* (Community Education)
4	"/\/[O][T]\//"	OT	*Organisasi dan Tata Laksana* (Organization Administration)	17	"/\/[P][G]\//"	PG	*Penelitian dan Pengembangan* (Research and development)
5	"/\/[H][K]\//"	HK	*Hukum* (Law)	18	"/\/[P][B]\//"	PB	*Perbukuan* (Bookkeeping)
6	"/\/[B][S]\//"	BS	*Kebahasaan* (Language)	19	"/\/[W][S]\//"	WS	*Pengawasan* (Supervision)
7	"/\/[S][P]\//"	SP	*Sarana Prasarana Pendidikan* (Facilities Infrastructure)	20	"/\/[P][R]\//"	PR	*Perencanaan dan Penganggaran* (Planning and Budgeting)
8	"/\/[K][P]\//"	KP	*Kepegawaian* (Staffing)	21	"/\/[L][K]\//"	LK	*Perlengkapan* (Equipment)
9	"/\/[K][S]\//"	KS	*Kerjasama* (Cooperation)	22	"/\/[P][F]\//"	PF	*Perfilman* (Movies)
10	"/\/[R][T]\//"	RT	*Kerumahtanggaan*(household)	23	"/\/[P][D]\//"	PD	*Peserta Didik* (Learners)
11	"/\/[T][U]\//"	TU	*Ketatausahaan* (Administration)		"/\/[K][B]\//"	KB	*Kebudayaan* (Culture)
12	"/\/[K][U]\//"	KU	*Keuangan* (Finance)	25	"/\/[T][I]\//"	TI	*Tek. Inf & Komunikasi* (ICT)
13	"/\/[K][R]\//"	KR	*Kurikulum* (Curriculum)				

2.3. Storage

The storage stage of classified documents to the database is the process after the hierarchical classification of digital correspondence documents.

Algorithm for storage of classified document on hive database (framework Hadoop)

```
Input: Result classification from Document
Output: Information of classified document on Hive Database
1. Read the classification results (the latest index information, the document name to [i]
   in the directory, the content of document, the origin of letter, subject of letter, the
   text of letter, type of letter, classification value for the type of letter.
2. Input the information into the table in the Hive database in the form of document id,
   document name, content, the origin of letter, the subject of letter, the text of letter,
   the type of letter, classification value.
```

3. RESULTS AND DISCUSSION

The total data used is 5200 digital correspondent documents (scanned documents in PDFImages format). Training data of 4000 documents (each type of letter totaling 800 labeled documents) and testing data of 1000 documents (each type of letter totaling 200 labeled documents). For prediction data, 200 new scanned documents, not training data, not testing data, and they are not labeled.

Table 3 shows the summary of classification testing result in the form of accuracy, precision, recall and F1Score from 10 Epochs. High score is epoch 5. Started from epoch 6, the score tends to decrease. In the end of Epoch 10, it reaches the accuracy of 94%. Table 4 shows the results for the confusion matrix of 5 types of text documents classification (decision, statement, command, assignment, and invitation) for epoch values 10. The average accuracy 94%. Table 5 shows the trial results for the classification of 200 documents. There are 188 documents accurately classified (accurately) for all level. However, 12 documents are inaccurately classified that its origin or subject of letters are no match. Accuracy of classification by dividing the number of documents classified accurately with the total number of documents tested is multiplied by 100%, then 188/200x100% is 94%.

Table 3. Evaluation result of classification

No	Epoch	Accuracy	Precision	Recall	F1 Score
1	Epoch 1	0.953	0.956	0.953	0.953
2	Epoch 2	0.955	0.958	0.955	0.955
3	Epoch 3	0.955	0.960	0.955	0.955
4	Epoch 4	0.955	0.958	0.955	0.955
5	Epoch 5	0.955	0.958	0.955	0.955
6	Epoch 6	0.939	0.943	0.939	0.938
7	Epoch 7	0.941	0.944	0.941	0.940
8	Epoch 8	0.939	0.942	0.939	0.938
9	Epoch 9	0.939	0.942	0.939	0.938
10	Epoch 10	0.939	0.942	0.939	0.938

Table 4. Confusion matrix

No	Type of Letter	Decision	Statement	Command	Assignment	Invitation	Accuracy
1	Decision	160	24	0	0	16	80%
2	Statement	2	191	2	0	5	96%
3	Command	0	0	200	0	0	100%
4	Assignment	0	0	2	194	4	97%
5	Invitation	2	0	2	2	194	97%
				Average of accuracy			94%

Table 5. Hierarchical classification accuracy (convolutional neural network & regular expression)

No	Description	Total
1	Total of Document Accurately Classified	188 Documents
2	Total of Document Inaccurately Classified	12 Documents
3	Total of Document	200 Documents
4	Accuracy Rate	188/200*100 %=94%

Table 6 shows the types of errors that occurred while testing the automatic classification of 200 documents. The CNN method is able to classify the type of letter correctly but errors often occur in the regular expression method when classifying the origin of the letter and the subject of the letter. The cause of the error can be in the form of the characters in the letter number are illegible (text characters from OCR were not recognized correctly), data does not match the provided regular expression pattern and unpredictable.

Table 6. Types of errors that occur when classifying documents automatically

No	Document Name	Error Type	Cause
1.	*36612016 Koordinasi dan Evaluasi pelaksanaan*	Origin of the Letter I2.1 and Subject of Letter KP failed to be classified	The characters in the letter number are illegible
2.	*00452016 Penunjang PTP a.n Hardianto*	Origin of the Letter P2.3 and Subject of Letter KP failed to be classified	The characters in the letter number are illegible
3.	*Ceramah Ilmiah dan seminar Nasional*	Origin of the Letter G2 and Subject of Letter TU failed to be classified	The characters in the letter number are illegible
4.	*13303 G1 TU 2017 Kegiatan penyampaian pagu definitif tahun 2018*	Origin of the Letter G1 and Subject of Letter TU failed to be classified	The characters in the letter number are illegible
5.	*4340 G1 SOSIALISASI PROGRAM KPR 19 OKT 2017*	Subject of Letter TU failed to be classified	The characters in the letter number are illegible
6.	*Peringatan Maulid Nabi Muhammad SAW*	Origin of the Letter G and Subject of Letter TU failed to be classified	The characters in the letter number are illegible
7.	*Permohonan pendaftaran BPJS*	Origin of the Letter G1 and Subject of Letter TU failed to be classified	The characters in the letter number are illegible
8.	*Rapat standar kompetensi 2 okt 2017*	Origin of the Letter G1 and Subject of Letter KP failed to be classified	The characters in the letter number are illegible
9.	*Pameran dan Publikasi Kegiatan*	Origin of the Letter G1 and Subject of Letter TU failed to be classified	The characters in the letter number are illegible
10.	*07122016 Ijin Buka Blokir Gedung BPMP Semarang_PENGANTAR*	Subject of Letter LL failed to be classified	Data does not match the provided Regular Expression pattern
11.	*Pameran dan Publikasi dalam Kegiatan Wonderful Sabang and Marine Expo 2017*	Origin of Letter & Subject of Letter failed to be classified. The letter number 008 / MSV-UND / IV / 2017 does not the criteria	Data does not match the provided Regular Expression pattern
12.	*Bimbingan Teknis Kehumasan 2017*	Subject of the Letter TU misclassified to KP	Unpredictable.

This study adapts, modifies and combines the methods in previous studies (scanned document classification with OCR-assisted text approach [17]-[19], hierarchical classification [28], CNN [20]-[22], regular expression [23]-[25] and framework Hadoop [29] which in the end this proposed method is able to overcome the problem of classifying scanned documents (using a text-based approach with the help of OCR) at a depth of 4 levels automatically in a hierarchical manner that is able to classify different document types with document conditions that have unstructured text content using CNN and have special patterns (specific and short strings) using regular expression and implementation of big data technology using Hadoop framework for store and analysis of large-scale data. This method is powerful and effective to overcome the multilevel classification problem in the case of this electronic mail document. The inaccuracy of the scanned document extraction results from Tesseract OCR causes the strings in the text content to be illegal and Errors may occur due to the absence of a regular expression pattern.

4. CONCLUSION

The combination of CNN and regular expression method has successfully solved the problem of hierarchical classification of scanned documents with the characteristics of documents containing unstructured text content and having special codes in the form of short strings to 4 levels according to the automated document hierarchy with an accuracy of 94%. The inaccuracy of the scanned document extraction results from Tesseract OCR causes the strings in the text content to be illegible and errors may occur due to the absence of a regular expression pattern. Some errors are caused by the unavailability of appropriate regular expression patterns, unclear writing, and unpredictable errors. This automatic hierarchical classification method is very necessary and useful to replace inefficient manual classification and store classified documents on hive databases (Hadoop architecture) to anticipate the increasing and varied growth of scanned documents is the right strategies. The future work is to classify documents for the different institution and search for the classified scanned documents based on content.

APPENDIX

KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN
PUSAT TEKNOLOGI INFORMASI DAN KOMUNIKASI PENDIDIKAN
Jalan RE Martadinata, Ciputat. Tromol Pos 7/CPA Ciputat 15411
Telepon: 021-7418808 (hunting), Fax: 021-7401727
Email : pustekkom@kemdikbud.go.id Laman: http://setjen.kemdikbud.go.id/pustekkom

KEPUTUSAN
KEPALA PUSAT TEKNOLOGI INFORMASI DAN KOMUNIKASI PENDIDIKAN
KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN
NOMOR : 0015 /P2/KP/2016

TENTANG

PENGANGKATAN TENAGA PETUGAS KEBERSIHAN DI LINGKUNGAN PUSTEKKOM KEMENDIKBUD

KEPALA PUSAT TEKNOLOGI INFORMASI DAN KOMUNIKASI PENDIDIKAN

Menimbang : bahwa untuk kelancaran pelaksanaan tugas di lingkungan Pustekkom Kemendikbud, perlu menetapkan Keputusan Kepala Pusat Teknologi Informasi dan Komunikasi Pendidikan tentang Pengangkatan Tenaga Petugas Kebersihan di Lingkungan Pustekkom Kemendikbud;

Mengingat : 1. Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional;
2. Peraturan Menteri Pendidikan dan Kebudayaan Nomor 1 Tahun 2012 tentang Organisasi dan Tata Kerja Kemendikbud sebagaimana telah diubah dengan Permendikbud Nomor 69 Tahun 2012;
3. Keputusan Menteri Pendidikan dan Kebudayaan Nomor 73/MPK.A4/KP/2013 tentang Pengangkatan Kepala Pustekkom Kemendikbud;

Memperhatikan : DIPA Pustekkom Kemendikbud Tahun 2015;

MEMUTUSKAN :

Menetapkan : KEPUTUSAN KEPALA PUSAT TEKNOLOGI INFORMASI DAN KOMUNIKASI PENDIDIKAN KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN TENTANG PENGANGKATAN PETUGAS KEBERSIHAN DI LINGKUNGAN PUSTEKKOM KEMENDIKBUD

Pertama : Mengangkat mereka yang namanya tercantum dalam Lampiran Keputusan Kepala Pustekkom ini dalam jabatan/tugas tercantum dalam Lampiran Keputusan Kepala Pustekkom ini dan kepadanya diberikan upah/honorarium setiap bulan;

Kedua : Keputusan ini bukan sebagai dasar pengangkatan mereka yang namanya dimaksud diktum pertama sebagai Calon Pegawai Negeri Sipil;

Ketiga : Apabila karena suatu keadaan tertentu, Kepala Pustekkom dapat memberhentikan mereka yang namanya dimaksud diktum pertama dari jabatan/tugasnya sebelum masa berkalu Keputusan ini berakhir dengan ketentuan tidak dapat diganggu gugat;

Keempat : Biaya yang timbul sebagai akibat pelaksanaan Keputusan ini dibebankan pada anggaran DIPA Pustekkom Kemendikbud;

Kelima : Keputusan ini berlaku sejak tanggal 4 Januari 2016 s.d. 31 Desember 2016 dengan ketentuan apabila terdapat kekeliruan akan diperbaiki.

Ditetapkan di Jakarta
Pada tanggal 4 Januari 2016

Kepala Pustekkom Kemendikbud,

Dr. Ir. Ari Santoso, DEA.
NIP. 196602181991021001

KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN
PUSAT TEKNOLOGI INFORMASI DAN KOMUNIKASI PENDIDIKAN DAN KEBUDAYAAN
Jalan RE Martadinata, Ciputat. Tromol Pos 7/CPA Ciputat 15411
Telepon: 021-7418808 (hunting), Fax: 021-7401727
Laman: http://setjen.kemdikbud.go.id/pustekkom Surel: pustekkom@kemdikbud.go.id

SURAT KETERANGAN
Nomor 1701 /I2.1/HM/2016

Kepala Bagian Tata Usaha atas nama Kepala Pusat Teknologi Informasi dan Komunikasi Pendidikan dan Kebudayaan Kementerian Pendidikan dan Kebudayaan menerangkan bahwa siswa SMK Cyber Media Jakarta sebagai berikut:

No.	Nama	NIS	Tingkat/Program Studi
1.	Husin	141696	XI Multimedia
2.	Muhammad Jaka Baruna	141700	XI Multimedia
3.	Teuku Daffa Gamal	141715	XI Multimedia

telah melaksanakan Praktik Kerja Industri di Pustekkom Kemendikbud tanggal 1 Maret s.d. 29 April 2016 pada:
1. Bidang Pengembangan Teknologi Pembelajaran Berbasis Multimedia dan Web; dan
2. Bidang Pengembangan Teknologi Pembelajaran Berbasis Radio, Televisi, dan Film.

Surat keterangan ini dibuat untuk dipergunakan sebagaimana mestinya.

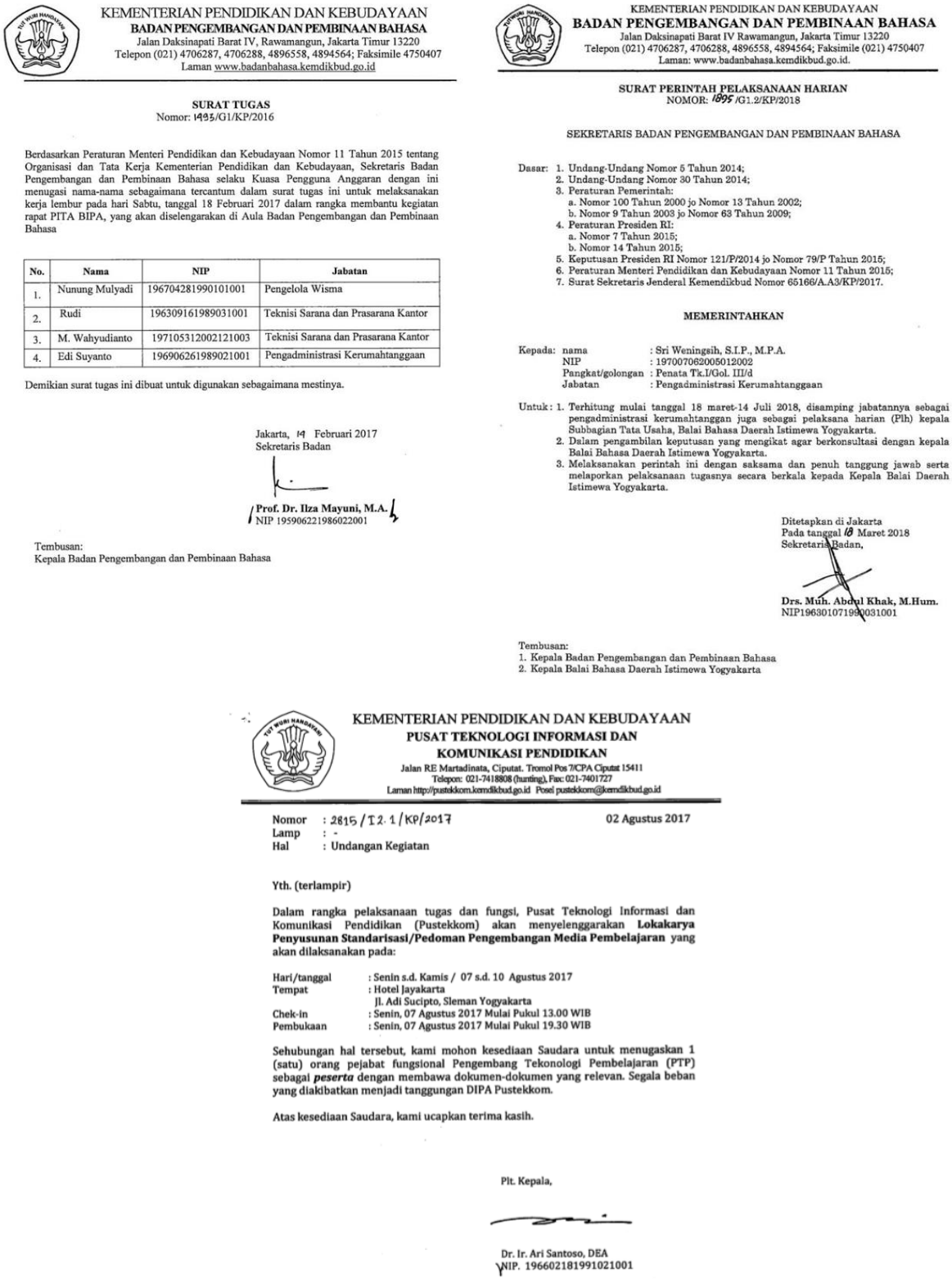
02 Mei 2016

a.n. Kepala
Kepala Bagian Tata Usaha,

Drs. Kusdianto, M.Si.
NIP 196009201983031006

Tembusan:
Kepala Pustekkom

Figure 2. Five letter types of digital correspondence documents (PDFImages format)



KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN
BADAN PENGEMBANGAN DAN PEMBINAAN BAHASA
Jalan Daksinapati Barat IV, Rawamangun, Jakarta Timur 13220
Telepon (021) 4706287, 4706288, 4896558, 4894564; Faksimile 4750407
Laman www.badanbahasa.kemdikbud.go.id

SURAT TUGAS
Nomor: 1493/G1/KP/2016

Berdasarkan Peraturan Menteri Pendidikan dan Kebudayaan Nomor 11 Tahun 2015 tentang Organisasi dan Tata Kerja Kementerian Pendidikan dan Kebudayaan, Sekretaris Badan Pengembangan dan Pembinaan Bahasa selaku Kuasa Pengguna Anggaran dengan ini menugasi nama-nama sebagaimana tercantum dalam surat tugas ini untuk melaksanakan kerja lembur pada hari Sabtu, tanggal 18 Februari 2017 dalam rangka membantu kegiatan rapat PITA BIPA, yang akan diselengarakan di Aula Badan Pengembangan dan Pembinaan Bahasa

No.	Nama	NIP	Jabatan
1.	Nunung Mulyadi	196704281990101001	Pengelola Wisma
2.	Rudi	196309161989031001	Teknisi Sarana dan Prasarana Kantor
3.	M. Wahyudianto	197105312002121003	Teknisi Sarana dan Prasarana Kantor
4.	Edi Suyanto	196906261989021001	Pengadministrasi Kerumahtanggaan

Demikian surat tugas ini dibuat untuk digunakan sebagaimana mestinya.

Jakarta, 14 Februari 2017
Sekretaris Badan

Prof. Dr. Ilza Mayuni, M.A.
NIP 195906221986022001

Tembusan:
Kepala Badan Pengembangan dan Pembinaan Bahasa

KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN
BADAN PENGEMBANGAN DAN PEMBINAAN BAHASA
Jalan Daksinapati Barat IV Rawamangun, Jakarta Timur 13220
Telepon (021) 4706287, 4706288, 4896558, 4894564; Faksimile (021) 4750407
Laman: www.badanbahasa.kemdikbud.go.id.

SURAT PERINTAH PELAKSANAAN HARIAN
NOMOR: 1895 /G1.2/KP/2018

SEKRETARIS BADAN PENGEMBANGAN DAN PEMBINAAN BAHASA

Dasar: 1. Undang-Undang Nomor 5 Tahun 2014;
2. Undang-Undang Nomor 30 Tahun 2014;
3. Peraturan Pemerintah:
a. Nomor 100 Tahun 2000 jo Nomor 13 Tahun 2002;
b. Nomor 9 Tahun 2003 jo Nomor 63 Tahun 2009;
4. Peraturan Presiden RI:
a. Nomor 7 Tahun 2015;
b. Nomor 14 Tahun 2015;
5. Keputusan Presiden RI Nomor 121/P/2014 jo Nomor 79/P Tahun 2015;
6. Peraturan Menteri Pendidikan dan Kebudayaan Nomor 11 Tahun 2015;
7. Surat Sekretaris Jenderal Kemendikbud Nomor 65166/A.A3/KP/2017.

MEMERINTAHKAN

Kepada: nama : Sri Weningsih, S.I.P., M.P.A.
NIP : 197007062005012002
Pangkat/golongan : Penata Tk.I/Gol. III/d
Jabatan : Pengadministrasi Kerumahtanggaan

Untuk: 1. Terhitung mulai tanggal 18 maret-14 Juli 2018, disamping jabatannya sebagai pengadministrasi kerumahtanggan juga sebagai pelaksana harian (Plh) kepala Subbagian Tata Usaha, Balai Bahasa Daerah Istimewa Yogyakarta.
2. Dalam pengambilan keputusan yang mengikat agar berkonsultasi dengan kepala Balai Bahasa Daerah Istimewa Yogyakarta.
3. Melaksanakan perintah ini dengan saksama dan penuh tanggung jawab serta melaporkan pelaksanaan tugasnya secara berkala kepada Kepala Balai Daerah Istimewa Yogyakarta.

Ditetapkan di Jakarta
Pada tanggal 18 Maret 2018
Sekretaris Badan,

Drs. Muh. Abdul Khak, M.Hum.
NIP196301071990031001

Tembusan:
1. Kepala Badan Pengembangan dan Pembinaan Bahasa
2. Kepala Balai Bahasa Daerah Istimewa Yogyakarta

KEMENTERIAN PENDIDIKAN DAN KEBUDAYAAN
PUSAT TEKNOLOGI INFORMASI DAN KOMUNIKASI PENDIDIKAN
Jalan RE Martadinata, Ciputat. Tromol Pos 7/CPA Ciputat 15411
Telepon: 021-7418808 (hunting), Fax: 021-7401727
Laman http://pustekkom.kemdikbud.go.id Posel pustekkom@kemdikbud.go.id

Nomor : 2815/I2.1/KP/2017 02 Agustus 2017
Lamp : -
Hal : Undangan Kegiatan

Yth. (terlampir)

Dalam rangka pelaksanaan tugas dan fungsi, Pusat Teknologi Informasi dan Komunikasi Pendidikan (Pustekkom) akan menyelenggarakan **Lokakarya Penyusunan Standarisasi/Pedoman Pengembangan Media Pembelajaran** yang akan dilaksanakan pada:

Hari/tanggal : Senin s.d. Kamis / 07 s.d. 10 Agustus 2017
Tempat : Hotel Jayakarta
Jl. Adi Sucipto, Sleman Yogyakarta
Chek-in : Senin, 07 Agustus 2017 Mulai Pukul 13.00 WIB
Pembukaan : Senin, 07 Agustus 2017 Mulai Pukul 19.30 WIB

Sehubungan hal tersebut, kami mohon kesediaan Saudara untuk menugaskan 1 (satu) orang pejabat fungsional Pengembang Tekonologi Pembelajaran (PTP) sebagai ***peserta*** dengan membawa dokumen-dokumen yang relevan. Segala beban yang diakibatkan menjadi tanggungan DIPA Pustekkom.

Atas kesediaan Saudara, kami ucapkan terima kasih.

Plt. Kepala,

Dr. Ir. Ari Santoso, DEA
NIP. 196602181991021001

Figure 2. Five letter types of digital correspondence documents (PDFImages format) (*Continue*)

ACKNOWLEDGEMENTS

We would like to thank to Technology and Information Data Center, Ministry of Education and Culture Republic Indonesia for allowing the use of non-confidential e-mail, and Faisal Arkan for his contribution to the trial test of the method in this research.

REFERENCES

[1] C. S. Yoganand, N. Praveen, N. Saranya, and V. G. Karthikeyan, "Survey on document classification based on keyword and key phrase extraction using various algorithms," *International Journal of Engineering Research and Technology (IJERT)*, vol. 3, no. 2, pp. 1804-1808, 2014.

[2] M. M. Saad, N. Jamil, and R. Hamzah, "Evaluation of support vector machine and decision tree for emotion recognition of Malay folklores," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 7, no. 3, pp. 479-486, 2018, doi: 10.11591/eei.v7i3.1279.

[3] C. Ying and D. R. Ying, "Novel top-down methods for Hierarchical Text Classification," *Procedia Engineering*, vol. 24, pp. 329-334, 2011, doi: 10.1016/j.proeng.2011.11.2651.

[4] N. Khan, M. S. Husain, and M. R. Beg, "Big data classification using evolutionary techniques: a survey," *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, pp. 243-247, 2015.

[5] M. Basha, K. Bagyalaksmi, C. Ramesh, R. Rahim, R. Manikandan, and A. Kumar, "Comparative study on performance of document classification using supervised machine learning algorithms: KNIME," *Australian Journal of Emerging Technologies and Society, vol.* 10, no. 1, pp 148-153, 2019.

[6] M. Nuser and E. Al-Horani, "Medical documents classification using topic modeling," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 17, no. 3, pp. 1524-1530, 2020, doi: 10.11591/ijeecs.v17.i3.pp1524-1530.

[7] T. Winarti, H. Indriyawati, V. Vydia, and F. W. Christanto, "Performance comparison between naive bayes and k-nearest neighbor algorithm for the classification of Indonesian language articles," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 2, pp. 452-457, 2021, doi: 10.11591/ijai.v10.i2.pp452-457.

[8] N. Seman and N. A. Razmi, "Machine learning-based technique for big data sentiments extraction," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 3, pp. 473-479, 2020, doi: 10.11591/ijai.v9.i3.pp473-479.

[9] J. A. Jupin, T. Sutikno, M. A. Ismail, M. S. Mohamad, S. Kasim, and D. Stiawan, "Review of the machine learning methods in the classification of phishing attack," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 8, no. 4, pp. 1545–1555, 2019, doi: 10.11591/eei.v8i4.1344.

[10] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, "Automatic hate speech detection using machine learning: a comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 484-491, 2020, doi: 10.14569/IJACSA.2020.0110861.

[11] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "Erratum to: a survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 85, 2016, doi: 10.1186/s13634-016-0382-7.

[12] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0032-1.

[13] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-014-0007-7.

[14] A. Kölsch, M. Z. Afzal, M. Ebbecke and M. Liwicki, "Real-time document image classification using deep CNN and extreme learning machines," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1318-1323, doi: 10.1109/ICDAR.2017.217.

[15] P. Chakraborty and A. Mallik, "An open-source tesseract-based tool for extracting text from images with application in braille translation for the visually impaired," *International Journal of Computer Applications*, vol. 68, no. 16, pp. 26–32, 2013, doi: 10.5120/11664-7254.

[16] R. Arief, A. Benny, T. Maulana, and Hustinawaty, "Automated extraction of large scale scanned document images using Google Vision OCR in apache hadoop environment," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, pp. 112-116, 2018, doi: 10.14569/IJACSA.2018.091117.

[17] H. Goodrum, K. Roberts, and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," *International Journal of Medical Informatics*, vol. 144, 2020, doi: 10.1016/j.ijmedinf.2020.104302.

[18] Y. K. Yaseen, A. K. Abbas, and A. M. Sana, "Image spam detection using machine learning and natural language processing," *Journal of Southwest Jiaotong University*, vol. 55, no. 2, 2020, doi: 10.35741/issn.0258-2724.55.2.41.

[19] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying promotion images using optical character recognition and naïve bayes classifier," *Procedia Computer Science*, vol. 179, pp. 498-506, 2021, doi: 10.1016/j.procs.2021.01.033.

[20] D. A. N. Taradhita and I. K. G. D. Putra, "Hate speech classification in Indonesian language tweets by using convolutional neural network," *Journal of ICT Research and Applications*, vol. 14, no. 3, pp. 225–239, 2021, doi: 10.5614/itbj.ict.res.appl.2021.14.3.2.

[21] M. A. Ramdhani, D. S. A. Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 19, no. 2, pp. 1000-1009, 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.

[22] S. A. Aljuhani and N. S. Alghamdi, "A comparison of sentiment analysis methods on Amazon reviews of mobile phones," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 6, pp. 608-617, 2019, doi: 10.14569/IJACSA.2019.0100678.

[23] M. Mowbray, W. Horne, P. Rao, "Efficient classification of strings using regular expressions," *Hewlett Packard Labs*, 2017.

[24] P. Prasse, C. Sawade, N. Landwehr and T. Scheffer, "Learning to identify concise regular expressions that describe email campaigns," *The Journal of Machine Learning Research*, vol. 16, no 1, pp. 3687–3720, 2015.

[25] C. A. Flores, R. L. Figueroa, J. E. Pezoa and Q. Zeng-Treitler, "CREGEX: a biomedical text classifier based on automatically generated regular expressions," in *IEEE Access*, vol. 8, pp. 29270-29280, 2020, doi: 10.1109/ACCESS.2020.2972205.

[26] I. C. Irsan and M. L. Khodra, "Hierarchical multi-label news article classification with distributed semantic model based features," *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 1, pp. 40-47, 2019, doi: 10.26555/ijain.v5i1.168.

[27] L. He, Y. Jia, Z. Ding, and W. Han, "Hierarchical classification with a topic taxonomy via LDA," *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 4, pp. 491-497, 2013, doi: 10.1007/s13042-013-0203-3.

[28] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "HDLTex: hierarchical deep learning for text classification," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[29] J. Mehta, J. Patil, R. Patil, M. Somani, and S. Varma, "Sentiment analysis on product reviews using hadoop," *International Journal of Computer Applications*, vol. 142, no. 11, pp. 38-41, 2016, doi: 10.5120/ijca2016909892.

[30] N. D. Zaki, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for Iraqi tweets using spark streaming," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 9, no. 4, pp. 1411–1419, 2020, doi: 10.11591/eei.v9i4.1897.

BIOGRAPHIES OF AUTHORS

Rifiana Arief Lecture at the Faculty of Industrial Technology, Gunadarma University. Deputy Head of Computer Network Development Laboratory, her research interest includes artificial intelligence, big data and computer science. She can be contacted at email: rifiana@staff.gunadarma.ac.id.

Achmad Benny Mutiara Professor, Dean Faculty of Computer Science and Information Technology, Gunadarma University. His research interest includes computer modeling and simulation (esp. molecular dynamics simulation and Monte Carlo simulation in physics), parallel computing (PC-clustering) and computational science. He can be contacted at email: amutiara@staff.gunadarma.ac.id.

Tubagus Maulana Kusuma Associate Professor, Director of Master's Program in Technology and Engineering, Gunadarma University. His research interests include multimedia communications, image/video processing and analysis, quality of experience (QoE), digital broadcast engineering, embedded system, robotics, and artificial intelligence. He can be contacted at email: mkusuma@staff.gunadarma.ac.id.

Hustinawaty Associate Professor, Head of Programme Magister Information Management, Gunadarma University. Her research interests are image processing. She can be contacted at email: hustina@staff.gunadarma.ac.id.