Topic Clustering and Classification on Final Project Reports: a Comparison of Traditional and Modern Approaches

Hendra Bunyamin, Heriyanto, Stevani Novianti, and Lisan Sulistiani

***Abstract*—Text clustering and classification has been studied at large in machine learning literature. For clustering text, topic modeling algorithms are statistical methods to discover unseen structures in archives of documents. Equally important, Convolutional Neural Networks (ConvNets) have been successfully applied for classifying text without knowing information about syntactic and semantic aspects of a language. In this paper, we utilizes both clustering and classification algorithms to organize and classify topics from final project reports. In clustering task, we examine two techniques, that are Latent Dirichlet Allocation (LDA) functioning as a unigram model and LDA supported by a Skip-gram model. Our results show each topical distribution of words found by the techniques are truly representing keywords from every topic; to elaborate, skip-gram model that works hand in hand with LDA are suitable to acquire topical words from the final report topics. For our classification task, we analyze the application of ConvNets, artificial neural nets with ReLU activation functions, and traditional algorithms. Concretely, our findings suggest that selecting parts of a report that contains essential information is very important for ConvNets to learn. Additionally, traditional algorithms is more preferrable than neural nets-based algorithms if the size of dataset is less than 20,000; as a result, our traditional algorithms, specifically Ridge classifier, Passive-Aggressive, and Support Vector Machines outperform neural nets-based algorithms significantly.**

***Index Terms*—convolutional neural networks, deep learning, final project report, latent dirichlet allocation, machine learning, skip-gram model, text classification, topic model**

I. Introduction

MARANATHA Christian University has digitized and stored its students' final project reports; however, it has become increasingly challenging for librarians to assign categories to the reports; moreover, it can be especially time-consuming to assign categories on final project reports manually. Blei [1] suggests project reports can be categorized according to their topics by using topic model algorithms.

Topic models are defined as statistical models that understand patterns of word use and connect documents that exhibit similar patterns from archives of documents founded on probabilistic latent semantic analysis [2], [3], [4], [5], [6], [7], [8]. Since there are no labels in the documents to guide the categorization process, topic models are classified as unsupervised learning algorithms. In contrast, supervised learning algorithms classify documents which have labels.

Reports classification or text categorization (TC) in general is a quintessential problem in natural language processing where one should give predefined labels to unstructured documents. Basically, given a training set $\mathcal{D} = \{X_1, \ldots, X_N\}$, each element of the training set is assigned a label which is taken from a set of k values. Fig. 1 shows a training set is used to train a machine learning model; after being trained, the model is able to assign labels to the test set [9].

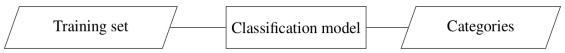

Fig. 1. A classification model that connects training records and categories

Thus far, nearly all algorithms of TC are dependent on words where plain statistics of several ordered combinations of words (such as language models) generally give best results [10], [11], [12]. These techniques are usually called traditional machine learning models. On the contrary, numerous researchers have established convolutional neural networks (ConvNets) as an all-around technique to extract information from images, text, speech, and other raw signals. Specifically, before deep learning gains its popularity, mostly sequential data are modeled by time-delay networks which are basically ConvNets [13], [14], [15].

Applying ConvNets to TC at large has been studied in literature. Specifically, ConvNets have been successfully applied for discrete [16] or distributed [17], [18] embedding of words where any information about syntactic and semantic aspects of a language is unknown. Moreover, these techniques have been competitive to traditional models.

Our study makes two key contributions: first, we provide an analysis of topic patterns in final project reports by utilizing Latent Dirichlet Allocation (LDA) with Skip-gram model and demonstrate that a combination of LDA and Skip-gram has a capability to cluster students' final project abstracts. Second, we provide a systematic comparison of the performance of traditional and modern approaches that model text classification problem for final project reports; specifically, we demonstrate that the performance of word-level traditional machine learning algorithms are better than ConvNets and equally comparable to other modern neural networks in final project reports classification task.

II. Related Work

We first review the concept of Skip-gram model and Latent Dirichlet Allocation (LDA), which is an undecorated

Manuscript received March 16, 2019; revised July 8, 2019. This work was supported by the Office of the Institute for Research and Community Service Maranatha Christian University under Grant 332-15/LPPM/UKM/XI/2017.

H. Bunyamin is with Informatics Engineering, Maranatha Christian University, Bandung, 40164 Indonesia, (e-mail:hendra.bunyamin@it.maranatha.edu)

Heriyanto (e-mail: augheri@yahoo.com) is with Central Library, Maranatha Christian University

S. Novianti (e-mail: stevaninovianti12@gmail.com) is with Faculty of Psychology, Maranatha Christian University

L. Sulistiani (e-mail: lisans1601@gmail.com) is with Informatics Engineering, Maranatha Christian University.

type of topic model. LDA is derived from a statistical latent semantic analysis model [3]; specifically, LDA refines probabilistic latent semantic analysis model by addressing a complete generative process [4]. Moreover, LDA describes a mixture model that employs convex combinations from distributed component sets to model observations. In LDA algorithm, a combination of topics (y) generates one word (t). Particularly, the probability of one word (t) producing a term (w) is

$$P(t = w) = \sum_i P(t = w|y = i)P(y = i), \qquad (1)$$

with $\sum_i P(y = i) = 1$ and every component ($P(t = w|y = k)$) equals to a multinomial distribution for every term which corresponds to an unseen topic $y = i$ from a corpus. Based on equation (1), goals of LDA inference are

1) to discover term distribution $P(w|y = i) = \vec{\varphi}_i$ for every topic i and
2) to discover topic distribution $P(y|d = m) = \vec{\vartheta}_m$ for every document m.

Quality of word vectors which have similarity can be improved by learning from huge data sets of billion of words. On the other hand, those similar word vectors may suffer from a wide range of similarities [19], [20], [21]. With regard to this problem, a word offset technique is used to allow simple algebraic operations to be performed on the vectors [21]. Subsequently, the resulting word vector is literally a result of those algebraic operations. For example, *vector("Emperor") - vector("Man") + vector("Woman")* becomes a vector that represents the word *Empress*. This technique enriches a class of language models which are so-called neural network language models (NNLM). Skip-gram model is a particular type of NNLM; in general, the model is trained by executing two steps as follows: a simple model learns from continuous word vectors. Next, we train an N-gram NNLM on top of the learned model [19], [20].

Equally important, we also review the work of character-level ConvNets for text classification by Zhang et al. [22] and Kim [18] and elaborate traditional machine learning algorithms utilized in this research. We opt for character-level ConvNets as both character-level and word-level ConvNets have a roughly similar performance [23]. The gradients are computed by an optimization algorithm called back-propagation [24].

Specifically, character-level Convolutional Networks consists of several ConvNets modules. In this paper ConvNets modules consist of three key modules such as a temporal convolutional module, a temporal max-pooling module, and a rectifier or thresholding module. Moreover, the ConvNets model accepts a sequence of encoded characters as input. The most important module in CovNets is a temporal convolutional module that calculates a 1-D convolution. Let us have a discrete input function $d(x) \in [1, p] \rightarrow \mathbb{R}$ and a discrete kernel function $k(x) \in [1, q] \rightarrow \mathbb{R}$. The convolution $c(y) \in [1, \lfloor (p-q)/t \rfloor + 1] \rightarrow \mathbb{R}$ between $k(x)$ and $d(x)$ with stride t is defined as

$$c(y) = \sum_{x=1}^{m} k(x) \cdot d(y \cdot t - x + f), \qquad (2)$$

where $f = m - t + 1$ is an offset constant. Similar to ConvNets in computer vision, this module is defined by a set of kernel functions $k_{ij}(x)(i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b)$ that are stated as *weights*, on inputs $d_i(x)$ and outputs $c_j(y)$. Every d_i (or c_j) input (or output) is defined as *features* and a (or b) input (or output) are called feature size. Each output $c_j(y)$ is computed as a sum over i from convolutions between $d_i(x)$ and $k_{ij}(x)$. All other specifications of ConvNets can be read in Zhang et al. [22]. In addition to ConvNets, we also utilize modern neural networks with rectified linear unit (ReLU) defined by an activation function $g(z) = \max\{0, z\}$ [25], [26], [27], [28].

The traditional machine learning models in this research are those that employing a manually crafted feature extractor and several classifiers. Specifically, the feature extractor is TF-IDF (term-frequency inverse-document-frequency) [29] and the classifiers are listed as follows: Ridge Classifier [30], Perceptron [31], Passive-Aggressive [32], K-Neighbors [33], Random Forest [34], Support Vector Machines [35], Stochastic Gradient Descent [36], Nearest Centroid [37], Multinomial Naïve Bayes, and Bernoulli Naïve Bayes [38], [39]. All classifiers are provided conveniently in Pedregosa et al. [40].

III. Research Methodology

Mainly, our methodology covers two activities that are clustering and classification tasks. We start by preprocessing our corpus. In the clustering task we apply the LDA with Skip-gram model to find similar topics in the data set. After that, we apply character-level ConvNets, modern neural networks, and traditional algorithms to classify documents their topics in the classification task. To create more structure in our analysis, we separate the analysis of the clustering and classification tasks.

A. Preparing Dataset

Fig. 2 shows the monthly average number of visitors from year 2011 to 2016 in Maranatha Christian University Library [41]. Visitors from faculty of psychology has had the most increment since 2013 among all other faculties; hence, psychology students' reports are chosen as dataset for our experiments.

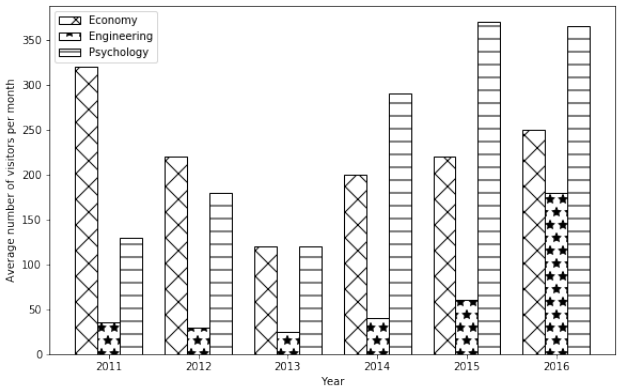

Fig. 2. Monthly number of library visitors in average

The preprocessing applied into the final projects consists of removing empty lines, sentences that encompasses specific words ("ABSTRACT" or "ABSTRAK", "DAFTAR BAGAN DAN SKEMA", "DAFTAR ISI", "DAFTAR TABEL", "DAFTAR LAMPIRAN", non-ASCII characters, and "Universitas Kristen Maranatha"), and page numbers.

More importantly, we also provide labels for our dataset as shown in Table I. In addition, Table II shows distribution of the topics.

TABLE I
SIX TOPICS THAT SERVE AS LABELS IN OUR DATASET

Topic	Name	Topic	Name
0	Educational Psychology	3	Clinical Psychology
1	Industrial Organizational Psychology	4	Developmental Psychology
2	Social Psychology	5	Others

TABLE II
NUMBER OF REPORTS FOR EACH TOPIC

Topic	Name	Percentage
0	Educational Psychology	13.2%
1	Industrial Organizational Psychology	24.9%
2	Social Psychology	11.9%
3	Clinical Psychology	29.1%
4	Developmental Psychology	18.6%
5	Others	2.3%

Our dataset for the clustering task comprises abstracts with the total number of 1,805. Equally important, we use abstracts, chapter 1, chapter 2, and chapter 3 for our classification task.

Our domain expert selects randomly 10 keyphrases from each topic; these keyphrases are used as a comparison for our algorithm outputs. Specifically, for example, keyphrases from Topic 0 (*Educational Psychology*) are kelas (*class*), universitas (*university*), sma (*high school*), and belajar (*study*). Topic 1 (*Industrial Organizational Psychology*) has phrases as follows: efficacy, work, kompetensi (*competency*), and stress. Topic 2 (*Social Psychology*) includes phrases such as remaja (*teenager*), emosional (*emotional*), kuesioner (*questionaire*), and purposive sampling. Topic 3 (*Clinical Psychology*) encompasses keyphrases such as sosial (*social*), engagement, anak (*child*), perilaku (*behavior*). Next, Topic 4 (*Developmental Psychology*) contains phrases such as style, brand image, minat (*interest*), and keputusan membeli (*buying decision*). Lastly, Topic 5 (*Others*) has labeled phrases such as korelasi (*correlation*), berkisar (*range*), dukungan (*support*), and saran (*advice*). Samples of keyphrases annotated by our domain expert are summarized in Table III.

TABLE III
SAMPLES OF KEYPHRASES IN INDONESIAN WHICH ANNOTATED BY A DOMAIN EXPERT. THE ENGLISH TRANSLATED KEYPHRASES ARE EXPLAINED IN SUBSECTION III-A

Topic 0	Topic 1	Topic 2
studi	efficacy	remaja
universitas	work	emosional
sma	kompetensi	kuesioner
belajar	stress	purposive sampling
Topic 3	**Topic 4**	**Topic 5**
sosial	remaja	korelasi
engagement	brand image	berkisar
anak	minat	dukungan
perilaku	keputusan membeli	saran

B. Clustering: LDA + Skip-gram models

The following are two settings in our clustering experiment. Firstly, we run LDA algorithm with skip-gram setting and without skip-gram setting. Particularly, we apply online learning to cluster final project topics [42] in our first setting. In next experiment, a bigram language model is constructed by employing skip-gram model algorithm. An online learning is activated based on the constructed language model to learn hidden topics [21]. Eventually, a domain expert evaluates the results by comparing the findings with annotated keyphrases whose samples are depicted in subsection III-A.

C. Classification: Artificial Neural Networks & Traditional Algorithms

In this task, we divide the dataset into train set, validation set, and test set with 70%, 20%, dan 10% proportions respectively. The train set is used for training the models, validation set is for tuning parameters, and test set is for measuring the models' performance.

Our modern text classfication models consist of ConvNets and artificial neural networks with ReLU activation functions (ANNwR). To make the ConvNets learn, we have tried several settings for the number of characters, hence we find one thousand characters are the most optimized setting in our model. Furthermore, ANNwR model are also fine-tuned by experimenting various parameters in the maximum number of words to keep, based on word frequency (`num_words`), number of hidden layers (`#hidden_layers`), and number of nodes in each layer (`#nodes`). After running 4-fold cross-validation, ANNwR model with `#hidden_layers`=3, `#nodes`=3, and `num_words`=15,000 gains best performance on validation set.

We prepare 12 text classification models as our traditional text classification models. Specifically, we tune the parameters of these traditional classifiers by running 4-fold cross-validation and choose classifiers with the best performance on validation set. Table IV shows traditional classifiers with their best settings.

TABLE IV
THE BEST SETTINGS AFTER 4-FOLD CROSS-VALIDATION FOR TRADITIONAL CLASSIFIERS

Classifier	Parameter-tuning settings
Ridge Classifier (RC)	$\alpha = 1.0$
Perceptron (P)	$\alpha = 0.0001$
Passive-Aggressive (PA)	loss = hinge
k-Nearest Neighbors (kNN)	n_neighbors = 10
Random Forest	criterion = information gain
Linear SVC (SVC)	penalty = L2
SGD classifier (SGDL2)	$\alpha = 1$ penalty = L2
SGD classifier (SGDL1)	$\alpha = 0.0003$ penalty = L1
SGD classifier (SGDE)	$\alpha = 0.0003$ penalty = elasticnet
Nearest Centroid (NC)	metric = euclidean
Multinomial NB (MNB)	$\alpha = 1$
Bernoulli NB (BNB)	$\alpha = 0.01$

IV. RESULTS: CLUSTERING TASK

We run two experiments in the first algorithm setting. In our first experiment, we run word removals for words whose frequencies $= 1$ and frequencies ≥ 2000. We also find that there are some words, for example, "validitas" (validity), which are considered keywords in several topics as shown in Table V. Therefore, words in this setting are not fully able to separate themselves as keywords.

TABLE V
RESULTS OF THE FIRST EXPERIMENT: TOPICS WITH THEIR HIGHEST PROBABILITY WORDS

Topic 0	Topic 1	Topic 2
reliabilitas	derajat	derajat
teknik	kuesioner	kuesioner
deskriptif	reliabilitas	validitas
kota	hubungan	rendah
teori	universitas	efficacy
hubungan	psikologi	teknik
saran	teknik	work
aspek	fakultas	responden
validitas	kerja	saran
responden	saran	reliabilitas
Topic 3	**Topic 4**	**Topic 5**
teori	karyawan	kerja
dimensi	derajat	validitas
saran	kerja	aspek
kuesioner	kota	reliabilitas
rendah	deskriptif	saran
reliabilitas	dimensi	teori
efficacy	teori	uji
responden	kuesioner	kuesioner
validitas	hubungan	hubungan
derajat	reliabilitas	responden

In order to purify words in each topic, we remove all words in the intersection between every two topics in the second experiment. Table VI exhibits that LDA has discovered words that distinguish themselves as keywords. We see that keywords from topic 1, for example, universitas (*university*), belajar (*study*), and kelas (*class*) are in fact keywords of "Educational Psychology" topic. In general, words in each topic represents the topic respectively.

TABLE VI
RESULTS OF THE SECOND EXPERIMENT: TOPICS WITH THEIR HIGHEST PROBABILITY WORDS

Topic 0	Topic 1	Topic 2
rancangan	efficacy	value
motivasi	work	remaja
perusahaan	perawat	motivasi
belajar	kompetensi	universitas
program	sumber	pengolahan
sma	sampling	emosional
pengolahan	universitas	stres
universitas	stress	studi
studi	pengolahan	sampling
kelas	emosional	rumah
Topic 3	**Topic 4**	**Topic 5**
of	anak	universitas
sosial	bidang	remaja
rancangan	sma	korelasi
berkisar	remaja	anak
engagement	style	pt
anak	pengolahan	rancangan
sampling	kelas	pengolahan
perilaku	sampling	sampling
pengolahan	universitas	berkisar
korelasi	of	dukungan

For our third setting we employ two algorithms altogether, a combination of online LDA [42] and skip-gram model [21]. When our results are compared with the dataset annotated by the domain expert; it shows that this combination of algorithms has an intuitive capability to capture phrases that represent each topic. For example, our expert is in accord about words in topic 1, such as *individuated*, profil (*profile*), and kerja (*work*), are keywords from "Industrial Organizational Psychology" topic.

TABLE VII
TEN WORDS WITH THE HIGHEST PROBABILITY FOR EVERY TOPIC IN THE THIRD SETTING

Topic 0	Topic 1	Topic 2
karakteristik	item	sampel
faktor	kuesioner	profil
orang	data	sesuai
telepon_genggam	kemandirian_emosional	populasi
rendah	berdasarkan_pengolahan	peneliti
individuated	teori	menggunakan_metode
profil	holland	kuesioner
derajat	rank_spearman	metode
kerja	validitas	purposive_sampling
aspek	koefisien_korelasi	berusia_tahun
Topic 3	**Topic 4**	**Topic 5**
maranatha_bandung	rancangan	nokia
rendah	brand_image	orangtuanya
universitas_kristen	mahasiswa	orangtua
mahasiswa_fakultas	minat	mahasiswa
psikologi	keputusan_membeli	saran
mahasiswa	tipe	untuk_mengetahui
untuk_mengetahui	psikologi	derajat
derajat	kesimpulan	telepon_genggam
dimensi	faktor	anak
universitas_x	aspek	subyek

V. RESULTS: CLASSIFICATION TASK

Table VIII displays all the accuracies of classifiers in our experiments. Surprisingly, our linear models such as Ridge classifier, Passive-Aggressive algorithm, and Support Vector Machines outperform non-linear models that are Random Forest and k-Nearest Neighbours although machine learning literature suggests that non-linear models, specifically Random Forest is the best classifier [43]. The ANNwR model has some potential to outperform traditional learning models; however, the ANNwR model lacks of training data instances. In order to learn holistically, number of training instances for the model should be more than 20,000.

TABLE VIII
ACCURACIES OF TRADITIONAL AND MODERN TEXT CLASSIFICATION CLASSIFIERS ON THE TEST SET

Name of classifier	Accuracy
Ridge Classifier	**76.67%**
Perceptron	75.56%
Passive-aggressive	**77.22%**
K-neighbors	55.00%
Random Forest	70.00%
Linear SVC	**76.67%**
SGD classifier	**76.67%**
Nearest centroid	71.11%
MultinomialNB	70.56%
BernoulliNB	70.00%
ANNwR	72.78%

A. Error analysis: ConvNets model

Table IX shows the performance of ConvNets. Although the performance of ConvNets is poor, interestingly, we examine misclassifications by ConvNets on several train instances. Table X contains several instances whose topics ConvNets misclassify as *clinical psychology*. The first row in Table X shows that the true value of the instance is *industrial psychology*. The second until the fourth rows show that the true topics are *social psychology*, *educational psychology*, and *social psychology* respectively. Our domain expert finds the comments such as *purposive sampling*, *descriptive statistics*, *correlation coefficient*, and *descriptive analysis* are indeed keywords from *clinical psychology*. Since our ConvNets

learn roughly one thousand characters and there are other words that can be signals for each topic, the ConvNets should learn from various parts of a report.

TABLE IX
RESULT OF THE CONVNETS EXPERIMENT WITH SEQUENCE LENGTH 1014, BATCH SIZE 1,605, NUMBER OF EPOCHS 10, DROP PROBABILITY 0.5, AND NUMBER OF CLASSES 6. THE TENTH AND ELEVENTH ROWS SHOW THE FINAL TRAIN ACCURACY AND THE DEV ACCURACY RESPECTIVELY.

Data	Epoch	Loss	Accuracy
Train	1	4.22	17.00%
Train	2	12.50	24.74%
Train	3	8.07	23.30%
Train	4	5.16	15.00%
Train	5	2.96	18.26%
Train	6	2.06	26.54%
Train	7	1.86	27.54%
Train	8	1.75	23.18%
Train	9	1.68	26.17%
Train	10	**1.72**	**25.17%**
Dev	-	**1.64**	**12.50%**

TABLE X
SOME RESULTS WHERE CONVNETS MISCLASSIFY INSTANCES AS A CLINICAL PSYCHOLOGY TOPIC (TOPIC **0**). HEADERS ARE EXPLAINED AS FOLLOWS: **TRUE** = TRUE CATEGORY OF AN INSTANCE, **0** = CLINICAL PSYCHOLOGY, **1** = EDUCATIONAL PSYCHOLOGY, **2** = DEVELOPMENTAL PSYCHOLOGY, **3** = INDUSTRIAL PSYCHOLOGY, **4** = SOCIAL PSYCHOLOGY, **5** = OTHERS, AND **COMMENTS** = COMMENTS ABOUT THE REPORT BEING PREDICTED.

True	0	1	2	3	4	5	Comments
3	✓						purposive sampling
4	✓						descriptive statistics
1	✓						correlation coefficient
4	✓						descriptive analysis

TABLE XI
SOME RESULTS WHERE CONVNETS MISCLASSIFY INSTANCES AS A EDUCATIONAL PSYCHOLOGY TOPIC (TOPIC **1**).

True	0	1	2	3	4	5	Comments
3		✓					population, sample
4		✓					qualitative, interview, variables
3		✓					sample, questionaire
3		✓					observations, respondents
4		✓					cross tabulation

TABLE XII
SOME RESULTS WHERE CONVNETS MISCLASSIFY INSTANCES AS A INDUSTRIAL PSYCHOLOGY TOPIC (TOPIC **3**).

True	0	1	2	3	4	5	Comments
0				✓			foundation,interpersonal,emotional
4				✓			consument,level of satisfaction
1				✓			validity,confidence,reliable
1				✓			self-esteem,cohesion,correlational
0				✓			self-compassion,work,family

We also find a number of train instances that are akin to the analysis in Table X, Table XI, and Table XII. Table XI contains train instances that are misclassified as *educational psychology* and Table XII consists of misclassifications as *industrial psychology*. Moreover, the comments in both Table XI and Table XII are indeed keywords for educational psychology and industrial psychology, respectively.

VI. CONCLUSION

Firstly, this research presents an exploration of LDA and skip-gram model for a clustering task on an archives of final project abstract documents in an unsupervised manner. Moreover, our experiments present that the collaboration of LDA and skip-gram model are able to cluster documents based on their topics. Clusters of documents are determined by the similarity of keywords and keyphrases from the topics of documents. Furthermore, the question of quantitative evaluations need to be addressed in future work.

Secondly, this paper analyzes the utilization of modern approach that is deep learning and traditional machine learning algorithms on final project reports in order to do automatic topic classification. For the case of having limited computation resources, selecting parts of a report that contains essential information is very important for ConvNets to learn. Our experiment shows that the traditional algorithms outperform neural nets-based algorithms significantly. Notably, our linear models, such as Ridge classifier, Passive-Aggressive, and Support Vector Machines from traditional algorithms are more accurate than the non-linear ones (Random Forest and k-Nearest Neighbours). All in all, traditional algorithms is more preferrable than deep learning algorithms if the size of dataset is less than 20,000.

REFERENCES

[1] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[3] T. Hofmann, "Probabilistic latent semantic indexing," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] W. Buntine and A. Jakulin, "Applying discrete pca in data analysis," *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 59–66.

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[7] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[8] Y. Wang and A. Maeda, "Twitter user's interest detection by using followee information based on lda topic model," *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019, IMECS 2019, 13-15 March, 2019, Hong Kong*, pp. 40–44.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2011.

[10] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," *Icml*, vol. 97, 1997, pp. 412–420.

[11] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *European conference on machine learning*. Springer, 1998, pp. 137–142.

[12] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[13] L. Bottou, F. F. Soulié, P. Blanchet, and J.-S. Lienard, "Experiments with time delay networks and dynamic time warping for speaker independent isolated digits recognition," *First European Conference on Speech Communication and Technology*, 1989.

[14] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Readings in speech recognition*. Elsevier, 1990, pp. 393–404.

[15] N. Shigei, K. Mandai, S. Sugimoto, R. Takaesu, and Y. Ishizuka, "Land-use classification using convolutional neural network with bagging and reduced categories," *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019, IMECS 2019, 13-15 March, 2019, Hong Kong*, pp. 7–11.

[16] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *CoRR*, vol. abs/1412.1058, 2014. [Online]. Available: http://arxiv.org/abs/1412.1058

[17] C. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69–78.

[18] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014. [Online]. Available: http://arxiv.org/abs/1408.5882

[19] T. Mikolov, "Language modeling for speech recognition in czech," Ph.D. dissertation, Masters thesis, Brno University of Technology, 2007.

[20] T. Mikolov, J. Kopecky, L. Burget, O. Glembek *et al.*, "Neural network based language models for highly inflective languages," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4725–4728.

[21] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." *HLT-NAACL*, vol. 13, 2013, pp. 746–751.

[22] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, 2015, pp. 649–657.

[23] S. Zhang, X. Zhang, and J. Chan, "A word-character convolutional neural network for language-agnostic twitter sentiment analysis," *Proceedings of the 22Nd Australasian Document Computing Symposium*, ser. ADCS 2017. New York, NY, USA: ACM, 2017, pp. 12:1–12:4. [Online]. Available: http://doi.acm.org/10.1145/3166072.3166082

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, p. 533, 1986.

[25] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 2146–2153.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104322.3104425

[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[29] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[30] R. M. Rifkin and R. A. Lippert, "Notes on regularized least squares," 2007.

[31] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine Learning*, vol. 37, no. 3, pp. 277–296, Dec 1999. [Online]. Available: https://doi.org/10.1023/A:1007662407062

[32] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551–585, 2006.

[33] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011. [Online]. Available: http://doi.acm.org/10.1145/1961189.1961199

[36] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism in perspective*, pp. 143–155, 1989.

[37] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.

[38] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," *AAAI-98 workshop on learning for text categorization*, vol. 752, 1998, pp. 41–48.

[39] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes?" *CEAS*, vol. 17, 2006, pp. 28–69.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] H. Bunyamin and L. Sulistiani, "Automatic topic clustering using latent dirichlet allocation with skip-gram model on final project abstracts," *The 21st International Computer Science and Engineering Conference, ICSEC*, 2017.

[42] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 856–864. [Online]. Available: http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf

[43] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

Hendra Bunyamin (1976.12-) Hendra Bunyamin was born in Bandung, West Java, Indonesia. He is a lecturer in Faculty of Information Technology at Maranatha Christian University. He is an avid reader and mostly interested in machine learning and natural language processing.

Heriyanto, male, acts as the head of central library at Maranatha Christian University. His interest is information science.

Stevani Novianti, female is an undergraduate student at faculty of psychology Maranatha Christian University. Her interest is educational psychology.

Lisan Sulistiani, female is an undergraduate student at faculty of information technology Maranatha Christian University. Her interest is educational programming.