

## Feature selection technique on convolutional neural network – multilabel classification task

Regiolina Hayami<sup>1,2</sup>, Nooraini Yusoff<sup>2</sup>, Kauthar Mohd Daud<sup>3</sup>, Harun Mukhtar<sup>1</sup>, Januar Al Amien<sup>1</sup>

<sup>1</sup>Department of Informatics Engineering, Faculty of Computer Science, Universitas Muhammadiyah Riau, Pekanbaru, Indonesia

<sup>2</sup>Faculty of Data Science and Computing, Universiti Malaysia Kelantan, City Campus Kota Bharu, Kelantan, Malaysia

<sup>3</sup>Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

### Article Info

#### Article history:

Received Jan 25, 2024

Revised Apr 15, 2024

Accepted May 12, 2024

#### Keywords:

Chi-square

Deep learning

Feature selection

Job resume analysis

Multilabel classification

### ABSTRACT

Automated text-based recommendation, an artificial intelligence development, finds application in document analysis like job resumes. The classification of job resumes poses challenges due to the ambiguity in categorizing multiple potential jobs in a single application file, termed multi-label classification, deep learning, particularly convolutional neural networks (CNN), offers flexibility in enhancing feature representations. Despite its robust learning capabilities, the black-box design of deep learning lacks interpretability and demands a substantial number of parameters, requiring significant computational resources. The primary challenge in multilabel learning is the ambiguity of labels not fully explained by traditional equivalence relations. To address this, the research employs feature selection techniques, specifically the Chi-square method. The goal is to reduce features in deep learning models while considering label relevance in multi-label text classification, easing computational workload while preserving model performance. Experimental tests, both with and without the Chi-square feature selection technique on the dataset, underscore its substantial impact on the classification model's ability. The conclusion emphasizes the influence of the Chi-square feature selection technique on performance and computational time. In summary, the research underscores the importance of balancing computational efficiency and model interpretability, especially in complex multi-label classification tasks like job applications.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



### Corresponding Author:

Regiolina Hayami

Department of Informatics Engineering, Faculty Computer Science, Universitas Muhammadiyah Riau Pekanbaru, Riau, Indonesia

Email: regiolinahayami@umri.ac.id

## 1. INTRODUCTION

Currently, numerous websites provide job openings as intermediaries between employers and job seekers. These job posting sites act as platforms for companies to disseminate information about job vacancies within their organizations, allowing job seekers to access information on a broader scale during their job searches. With the ease of access available to job seekers and employer companies, a substantial amount of job resume data will probably be received. As a result, a significant amount of time is required to analyze this job resume data manually.

Multilabel text classification (MLTC) is a crucial task in the field of natural language processing (NLP), applicable in various real-world scenarios such as information retrieval [1] and tag recommendation [2], [3]. The objective of the MLTC task is to assign multiple labels to each sample in the dataset. Within the

realm of job classification, MLTC plays a pivotal role in categorizing individuals based on a multitude of skills, thereby facilitating efficient talent acquisition processes. While existing methods [4], [5] already perform remarkably well when extracting skill-related terms from resumes, they are limited by their inability to infer high-level skills not explicitly mentioned in resumes.

In recent years, neural networks have significantly succeeded in many fields, including NLP. One of the developments in the neural network is the introduction of deep learning techniques as a new field in computer science to handle recommended cases [6]. The use of deep learning technique to help determine the results of recommendations has shown an increase in terms of feature extraction [7], [8]. Convolutional neural networks (CNN), initially designed for image processing, have been adapted to text classification tasks with remarkable success [2], [9], [10] shows promising results on the evaluation matrix. CNN's ability to extract salient features from input data has propelled its efficacy in text classification, outperforming traditional machine learning algorithms.

Even though CNN shows promising results in completing text classification tasks, weaknesses are still found in its processing. Some of these are in the case of single-class and multi-class classifications. There are multiple word embedding matrices, making it difficult to extract or display how each embedding matrix is related to each other [11], [12]. In the case of MLTC, CNN can produce maximum performance values on evaluation metrics. However, they need to be analyzed and calculated to determine the contribution of each selected feature to the value predicted by the classifier [13]-[15]. Feature selection has a substantial impact on reducing training time and required storage. The output of feature selection techniques is a lengthy list of features statistically sorted based on their distinctiveness for each class. Features with higher values are chosen as representative features [16]. Therefore, feature selection aimed at filtering relevant features is a crucial step to reduce data dimensionality and enhance learning performance.

The challenges in applying job resume data to CNN algorithms and feature selection techniques involve two main aspects. Firstly, as a deep learning algorithm, CNN is commonly utilized in image classification cases, necessitating adjustments for the network architecture to process text-form data. Secondly, feature selection poses a combinatorial optimization problem in discrete space, necessitating a specialized design for coding strategy, crossover operators, and mutation. To address these challenges, this study proposes integrating the Chi-square feature selection technique into the CNN framework for MLTC. Our contributions encompass the construction of a multi-label classification model using CNN on job resume text, feature reduction through Chi-square selection, and the seamless integration of feature selection into CNN architecture.

As the final result of this research, experiments were carried out to see the effect of using feature selection techniques in the case of multi-label text classification using the CNN algorithm. The ensuing sections will delineate our methodology, experimental findings, and the implications thereof on the field of talent acquisition. Through rigorous analysis and experimentation, we aim to showcase the efficacy of our proposed approach in enhancing the accuracy and efficiency of job-resume matching algorithms.

## **2. LITERATURE REVIEW**

### **2.1. MLTC using CNN**

MLTC is an essential task in the field of NLP, which can be applied in many real-world scenarios, such as information retrieval [1] and tag recommendation [2], [3]. CNN was first designed for image processing but has been widely used in text processing. In 2014, Kim proposed an adaptation of CNN for text classification with convolutional filters and max-pooling filters that slide only on one dimension (the y dimension), named 1D-CNN [17]. However, in recent years, it has been proven to analyze natural language and become a model used for sentence classification [18], [19]. Feature extraction on text data using CNN with multi-label classification as an essential part of generating recommendations has achieved maximum accuracy [2].

### **2.2. Chi-square feature selection technique for multi-label classification**

One of the problems that arise in the case of text classification is that textual data contains many words. A large number of words can cause high computational complexity and reduce the accuracy of the classification results [20]. Feature selection can be used to determine dominant features and improve efficiency and performance in text classification. Integration between deep learning and feature selection has been carried out in the case of multi-label text classification (MLTC). Feature selection identifies and filters irrelevant and redundant features, reduces data dimensionality, and determines their contribution to classification [21], [22]. The Chi-square method selects features considered essential for the classification process and can eliminate features that do not affect the target class. Chi-square is a feature selection method that calculates the relationship between existing features and the target class. The use of the Chi-square

feature selection technique also performs well in intrusion detection model classification and sentiment analysis, with accuracy reaching 99% and 100% [23], [24]. Table 1 describes the performance results from research that applies the Chi-square feature selection technique to multi-label text classification [16].

Table 1. Performance of chi-square feature selection on multi-label text classification

Representation	Accuracy	F1	Recall	Precision	Hamming loss
MLTC	81.44	92.00	90.53	93.52	0.022
MLTC-tuning parameter	82.29	92.39	92.40	92.55	0.0215
MLTC-Chi-Square	82.31	92.52	92.44	92.60	0.0214

Based on the discussion above, this study aims to apply the Chi-square feature selection technique to job resume data for multi-label text classification. This feature selection technique is used to rank the features in each class and select the top 50 features in each class to be used in the classification process using the CNN algorithm.

### 3. METHOD

This research was conducted in several stages, from pre-processing to evaluation. The data that has been collected will be pre-processed so that it can be converted into a matrix. The matrix data is then used as input data for the CNN algorithm to create a multi label classification. A feature selection technique will be added to produce reduced features. Furthermore, the CNN algorithm will use these features in the classification process. The Figure 1 shows the process flow of the research conducted.

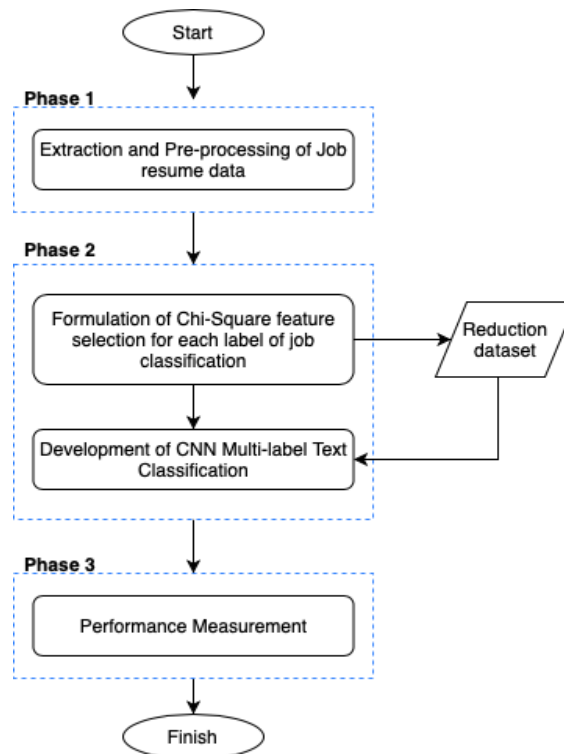


Figure 1. Process flow of the research

#### 3.1. PHASE 1: extraction and pre-processing of dataset

The dataset consists of 28,707 resumes collected from the Indeed.com website and distributed into ten classes [2]. The resume consists of IT classes: Software Developer, Front-End Developer, Network Administrator, Web Developer, Project Manager, Database Administrator, Security Analyst, System Administrator, Python Developer, and Java Developer. Figure 2 illustrates the distribution of the dataset used. In the pre-processing stage, case folding, cleaning, lemmatization, stopword, and tokenizing are carried

out on the dataset to make words more structured and usable for the classification process. The dataset will be modified and described before using CNN in Chi-square feature selection or classification.



Figure 2. Data resume distribution based on skill

### 3.2. PHASE 2: integration of chi-square feature selection and CNN multi-label text classification

At this stage, Chi-square feature selection is formulated on job resume data to generate a reduction dataset. Subsequently, this reduction dataset will be used for developing a CNN model in the case of multi-label text classification of job resume data. The steps undertaken in phase 2 are as follows:

A. Term frequency – inverse document frequency (TF – IDF)

Term frequency (TF) is the number of times a word appears in a document. In contrast, inverse document frequency (IDF) is a word score that is calculated by comparing the number of documents in the corpus with the number of documents containing the word. TF-IDF is used together to reduce the effect of common words appearing in the entire corpus [25]. To calculate the TF-IDF value of the words contained in the data, the steps are [26]:

- Tokenizing
- Calculate the TF value for each word in each sentence with the formula:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

Information:

$n_{i,j}$ : number of occurrences of entry  $w_i$

- Calculate the IDF value for each word with a formula :

$$IDF = \log\left(\frac{|D|}{|\{j: w_i \in x_j\}| + 1}\right) \tag{2}$$

Information:

$|D|$ : total number of files in the corpus

$|\{j: w_i \in x_j\}|$  = number of files containing entry  $w_i$

- Compute the TF-IDF of both sentences:

$$TF - IDF_{w_i} = TF_{i,j} \times IDF_i \tag{3}$$

B. Chi-square feature selection

Chi-square is a feature selection method that, utilizes statistical distribution by measuring the dependency value between features and labels [27]. Chi-square feature selection calculation is performed on at least two groups of labelled documents. The steps are as follows:

- Make a table of the frequency of words in each group of documents
- Calculate the Chi-square value for each word using the following formula [23]:

$$Chi^2 = \frac{N \times ((ad - bc)^2)}{((a+b) \times (c+d) \times (a+c) \times (b+d))} \tag{4}$$

Information:

- a: the number of documents that belong to group 1 and contain that word
- b: the number of documents that belong to group 2 and contain the word
- c: the number of documents that belong to group 1 and do not contain the word
- d: the number of documents that belong to group 2 and do not contain the word
- N: total number of documents

C. CNN multi-label text classification

CNN modeling is carried out to create a multi-label classification model. This model will be used to classify the skills possessed by each resume later. CNN is used to classify resumes according to one class. In other words, CNN classification predicts each class so that the number of initial classifications corresponds to the number of classes in the resume data. The following Figure 3 describes the CNN architecture for performing multi-label classification.

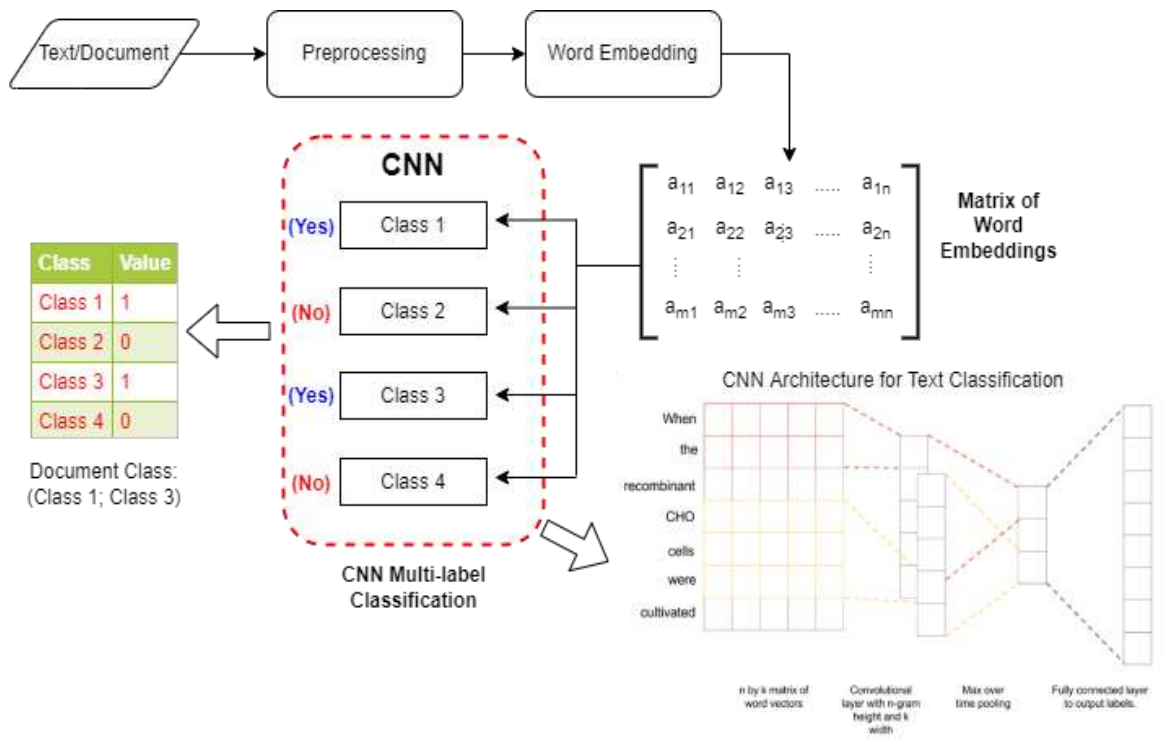


Figure 3. Architecture of CNN for multi-label text classification

To form a training set from each basic classification result, the original multilabel data set is divided into  $n$  single-label data sets (where  $n$  is the total number of classes in the original data set). Each sub-dataset generated corresponds to a binary classification problem focusing on each class. In CNN sentence classification, the sentence must be converted into a vector of real numbers. This vector will be used as input for text processing using CNN. The process of creating the CNN model will be built using a Python library.

**3.3. PHASE 3: evaluation of integration chi-square feature selection and CNN for multi-label text classification**

The evaluation stage is conducted using several criteria, such as accuracy and model performance. The evaluation is conducted by comparing the performance of the model between CNN and CNN integrated with Chi-square feature selection. Three commonly used measurements in the literature for evaluating multi-label classification are accuracy, precision, and recall. Performance measurement on the multi-label model is carried out by calculating the average values of accuracy, precision, and recall for all instances [28].

**4. RESULTS AND DISCUSSION**

This study used 28,707 data on job resumes in the IT field, the Table 2 illustrates sample of the dataset used.

Table 2. Sample dataset resume along with job categories

Resume	Web_developer	Project_manager	Database_Administrator	Front_End_Developer
Database administrator database Administrator. My experience includes SQL Server 2005, 2008 and 2012...	0	0	1	0
Consultant - supply chain management Consultant - supply chain management Supply chain consultant germantown	0	1	1	0

Before the dataset is used for the classification process, the pre-processing stage is first carried out to prepare the data to be used optimally in the classification process. After the pre-processing stage was carried out, a Chi-square feature selection technique was added in the second experiment, which resulted in a reduction of the dataset to 50 features with the highest ranking in each class contained in the resume. The following Table 3 is a sample of the top 10 features in each class based on calculations using Chi-square feature selection.

Table 3 presents the ranking results of the top 10 features for each job resume class based on the highest Chi-square scores in each class. For instance, to become a Database Administrator, the dominant features representing its job skills are database, oracle, dba, and so forth. Another example, the feature "network" appears as a top-10 feature in both the Network Administrator and Software Developer classes. However, in both classes, the "network" feature has different ranking orders. The variation in feature rankings across classes indicates the order of importance of features within each class. Analyzing the importance level of features derived from Chi-square feature selection for job skills will strengthen the contribution of terms used as features in the final decision-making process of multi-label classification on job resumes.

Table 3. Top-10 features of job skill based on chi-square feature selection technique

Classes	Top 10 features
Database_Administrator	(database: 3171.64), (oracle: 1272.71), (dba: 1209.76), (rman: 1173.07), (tuning: 897.38), (rac: 882.19), (sql: 721.54), (12c: 646.78), (11g: 538.06), (replication: 442.54)
Java_Developer	(spring: 3722.49), (java: 2338.50), (hibernate: 2047.48), (jsp: 1324.89), (j2ee: 1282.95), (junit: 1098.71), (strut: 1072.18), (jdbc: 839.97), (maven: 823.19), (servlets: 812.28)
Front_End_Developer	(front: 2327.39), (end: 1387.33), (css3: 666.18), (ui: 660.86), (react: 598.79), (javascript: 592.11), (jquery: 570.99), (angular: 568.64), (page: 561.40)
Network_Administrator	(network: 3138.84), (cisco: 1259.73), (switch: 732.65), (firewall: 632.50), (router: 512.10), (administrator: 430.57), (lan: 322.67), (vpn: 319.88), (networking: 303.41), (directory: 287.99)
Project_manager	(project: 1928.93), (manager: 1397.74), (budget: 593.55), (management: 312.02), (stakeholder: 272.23), (managed: 241.86), (strategic: 213.29), (led: 209.69), (using: 203.78), (vendor: 202.16)
Python_Developer	(python: 7207.95), (django: 4266.65), (flask: 957.42), (panda: 938.16), (numpy: 750.97), (matplotlib: 539.43), (mysql: 501.72), (using: 366.47), (amazon: 330.62), (postgresql: 306.91)
Security_Analyst	(security: 3727.62), (vulnerability: 1346.27), (nist: 1257.75), (analyst: 954.03), (assessment: 888.29), (risk: 887.72), (cyber: 864.80), (threat: 691.39), (incident: 548.73), (compliance: 509.16)
Software_Developer	(web: 1322.69), (developer: 1305.62), (network: 1164.69), (python: 1001.63), (security: 944.29), (administrator: 916.47), (javascript: 891.29), (java: 738.67), (using: 721.19), (jquery: 667.85)
Systems_Administrator	(administrator: 1043.04), (system: 987.69), (vmware: 674.92), (network: 636.34), (directory: 587.41), (active: 480.12), (server: 397.49), (hardware: 380.32), (window: 376.52), (exchange: 369.96)
Web_Developer	(web: 1368.16), (wordpress: 549.12), (developer: 544.37), (website: 532.60), (javascript: 478.55), (php: 395.01), (jquery: 393.74), (network: 350.61), (cs: 349.40), (page: 313.30)

#### 4.1. CNN multilabel classification

Before data can be classified using the CNN algorithm, the resume that has gone through pre-processing and reduced features in the second experiment is converted into a vector matrix because CNN is a convolution network and requires input in matrix form. CNN modeling is built with a layered architecture starting with the sequential function, with one input and output tensor. The input tensor represents a matrix of input data obtained using the Keras library embedding layers. CNN is used to classify resumes according to one class. In other words, CNN classification predicts each class so that the number of initial classifications corresponds to the number of classes in the resume data. In the case of classification, the dataset is divided into two parts, namely training data and test data. In this study, the test and train data distribution was carried out at 90% for train data and 10% for test data taken from each job label.

In this study, the model built uses the CNN layer to process text, using the ReLu and Sigmoid activation functions on the output layer. In the first layer (embedding layer), the output dimension is defined to be 100. The input layer is set to 500, according to the length of the padded text in the previous step. Then,

compile the model using the Adam optimizer and the loss function Binary\_crossentropy. The metrics calculated during model evaluation are accuracy, precision, and recall. After the parameter settings have been completed, the model can be used in each label's training and testing stages. The predicted output is rounded to 0 or 1 with np\_round, resulting in the appropriate label class. Accuracy, precision, and recall values in the train data are calculated using the functions provided by the scikit-learn library and stored in a previously created list.

**4.2. Evaluation**

To determine the performance of the CNN model in the case of multi-label text classification data resumes, a confusion matrix is used with indicators of accuracy, precision, and recall values for the training and testing processes. The evaluation was carried out for each class label, and the average performance of the models built in experiments with and without Chi-square was calculated. Evaluation was carried out in all classes and obtained the following results on Table 4.

Based on Table 4, we can see the performance comparison between using traditional CNN for multi-label job resume classification and CNN integrated with Chi-square feature selection. It can be observed that both the accuracy, precision, and recall of each class have improved with the integration of the chi-square feature selection into the CNN architecture. Based on the evaluation of each label above, the average evaluation of the two experiments was calculated. Table 5 describes the comparison of the average performance of the multi-label job resume classification model using CNN with and without the Chi-square feature selection technique.

Table 4. Performance comparison with and without chi-square on each class label

Label	CNN			CNN + Chi-square		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Web developer	86.96	75.69	90.84	94.12	90.59	92.45
Project manager	92.24	79.14	86.41	97.32	93.50	89.27
Database administrator	98.07	89.41	91.13	99.09	94.85	95.95
Network administrator	93.39	81.79	75.41	95.88	88.38	85.46
Front end developer	96.00	90.22	90.71	98.92	97.13	97.74
Python developer	97.84	89.21	95.30	99.49	96.81	99.43
System administrator	92.45	85.19	76.20	95.45	88.73	88.98
Software developer	96.83	98.17	96.91	97.92	98.03	98.77
Security analyst	97.64	86.80	89.01	98.39	90.01	93.72
Java developer	96.41	86.89	91.56	99.15	95.70	99.18

Table 5. Performance comparison of multi-label classification models

Performance indicators	CNN	CNN+Chi-square
Accuracy	94.98	97.58
Precision	86.25	93.37
Recall	88.35	94.09

In the case of multilabel job resume classification using CNN and the chi-square feature selection technique, performance improvement is achieved in performance indicators. Furthermore, regarding training time, the CNN+Chi-square model outperforms the traditional CNN model by approximately 40.74%. The importance of chi-square feature selection is evident in the increased accuracy of the CNN model, particularly in identifying labels in job resumes. Additionally, the reduction in training time can be interpreted as a significant advantage in the efficiency of model utilization. The results of this comparison provide additional insights into the potential performance enhancement by integrating appropriate feature selection techniques in the processing of multilabel resume data.

**5. CONCLUSION**

From this research, it can be concluded that selecting relevant and informative features is crucial in multi-label text classification for better decision-making. Performance evaluation results indicate that using CNN in the case of multi-label text classification yielded an accuracy of 94.98%, precision of 86.25%, and recall of 88.35%. Meanwhile, employing CNN along with Chi-square feature selection resulted in an increased accuracy of 97.58%, precision of 93.37%, and recall of 94.09%. The use of Chi-square in job resume feature selection influenced the process of determining dominant features for each target class. The application of Chi-square in the case of multi-label text classification using CNNs has proven to enhance the performance of CNN models and achieve better computational efficiency than traditional CNNs. In future research, experiments with other text datasets are necessary to explore the impact of the chi-square feature

selection technique on the CNN multi-label classification model further. Other performance metrics can also be utilized to understand the effects of feature selection on various aspects, such as error rates, and so forth. Additionally, we plan to explore the application of alternative feature selection techniques to improve the model's performance in classifying job resumes.

## ACKNOWLEDGEMENTS

The authors wish to express their gratitude for the support and assistance provided by Universitas Muhammadiyah Riau, Indonesia in making this article possible. Additionally, the authors extend their appreciation to fellow researchers who contributed, either formally or informally, to the preparation of this paper.

## REFERENCES




- [1] C. Gan, Q. Feng, and Z. Zhang, "Scalable multi-channel dilated CNN-BiLSTM model with attention mechanism for Chinese textual sentiment analysis," *Future Generation Computer Systems*, vol. 118, pp. 297–309, 2021, doi: 10.1016/j.future.2021.01.024.
- [2] K. Florentin and F. Jiechieu, "Skills prediction based on multi-label resume classification using CNN with model predictions explanation," *Neural Computing and Applications*, vol. 0123456789, 2020, doi: 10.1007/s00521-020-05302-x.
- [3] P. Zhang, W. Gao, J. Hu, and Y. Li, "Multi-label feature selection based on the division of label topics," *Inf Sci (N Y)*, vol. 553, pp. 129–153, 2021, doi: 10.1016/j.ins.2020.12.036.
- [4] L. Sayfullina, E. Malmi, and J. Kannala, "Learning representations for soft skill matching," in *Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018, Revised Selected Papers 7*, Springer, 2018, pp. 141–152, doi: 10.1007/978-3-030-11027-7\_15.
- [5] F. Javed, P. Hoang, T. Mahoney, and M. McNair, "Large-scale occupational skills normalization for online recruitment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 4627–4634, doi: 10.1609/aaai.v31i2.19086.
- [6] M. Naumov *et al.*, "Deep learning recommendation model for personalization and recommendation systems," *arXiv preprint arXiv:1906.00091*, 2019, doi: 10.48550/arXiv.1906.00091.
- [7] Z. Batmaz, A. Yurekli, A. Bilge, and C. Kaleli, "A review on deep learning for recommender systems : challenges and remedies," *Artificial Intelligence Review*, 2018, doi: 10.1007/s10462-018-9654-y.
- [8] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.
- [9] W. Liao, Y. Wang, Y. Yin, X. Zhang, and P. Ma, "Improved sequence generation model for multi-label classification via CNN and initialized fully connection," *Neurocomputing*, vol. 382, pp. 188–195, 2020, doi: 10.1016/j.neucom.2019.11.074.
- [10] B. Guo, C. Zhang, J. Liu, and X. Ma, "Improving text classification with weighted word embeddings via a multi-channel TextCNN model," *Neurocomputing*, vol. 363, pp. 366–374, 2019, doi: 10.1016/j.neucom.2019.07.052.
- [11] Z. Shen and S. Zhang, "A novel deep-learning-based model for medical text classification," in *Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition*, 2020, pp. 267–273, doi: 10.1145/3436369.3436469.
- [12] M. Pota, M. Esposito, G. De Pietro, and H. Fujita, "Best practices of convolutional neural networks for question classification," *Applied Sciences (Switzerland)*, vol. 10, no. 14, 2020, doi: 10.3390/app10144710.
- [13] B. Liu, Y. Zhou, and W. Sun, "Character-level text classification via convolutional neural network and gated recurrent unit," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 8, pp. 1939–1949, 2020, doi: 10.1007/s13042-020-01084-9.
- [14] J. Xu *et al.*, "Incorporating context-relevant concepts into convolutional neural networks for short text classification," *Neurocomputing*, vol. 386, pp. 42–53, Apr. 2020, doi: 10.1016/j.neucom.2019.08.080.
- [15] L. Lenc and P. Král, "Deep neural networks for Czech multi-label document classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 460–471. doi: 10.1007/978-3-319-75487-1\_36.
- [16] A. Omar, T. M. Mahmoud, T. Abd-El-Hafeez, and A. Mahfouz, "Multi-label arabic text classification in online social networks," *Information Systems*, vol. 100, Sep. 2021, doi: 10.1016/j.is.2021.101785.
- [17] Y. Song, Q. V. Hu, and L. He, "P-CNN: Enhancing text matching with positional convolutional neural network," *Knowl Based Syst*, vol. 169, pp. 67–79, 2019, doi: 10.1016/j.knosys.2019.01.028.
- [18] A. Jacovi, O. S. Shalom, and Y. Goldberg, "Understanding convolutional neural networks for text classification," *arXiv preprint arXiv:1809.08037*, 2018.
- [19] R. Hayami, J. Al Amien, and N. T. Utami, "Tweet spam detection using the convolutional neural network (CNN) model," in *AIP Conference Proceedings*, AIP Publishing, Jun. 2023, p. 020031. doi: 10.1063/5.0130468.
- [20] F. S. Nurfikri, M. S. Mubarak, and Adiwijaya, "News topic classification using mutual information and bayesian network," in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, IEEE, May 2018, pp. 162–166. doi: 10.1109/ICoICT.2018.8528806.
- [21] Z. Zhou, J. Wong, K. Yu, G. Li, and S. Chen, "Feature selection on deep learning models: an interactive visualization approach," 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:260140100>
- [22] H. Ebrahimi, K. Majidzadeh, and F. S. Gharehchopogh, "Integration of deep learning model and feature selection for multi-label classification," *International Journal of Nonlinear Analysis and Applications*, vol. 13, pp. 2008–6822, 2022, doi: 10.22075/ijnaa.2021.25379.2998.
- [23] I. S. Thaseen, Ch. A. Kumar, and A. Ahmad, "Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers," *Arabian Journal for Science and Engineering*, vol. 44, no. 4, pp. 3357–3368, Apr. 2019, doi: 10.1007/s13369-018-3507-5.
- [24] M. Hussein and F. Özyurt, "A new technique for sentiment analysis system based on deep learning using chi-square feature selection methods," *Balkan Journal of Electrical and Computer Engineering*, Oct. 2021, doi: 10.17694/bajece.887339.
- [25] S. Boukil, M. Biniz, F. El Adnani, L. Cherrat, and A. E. El Moutaouakkil, "Arabic text classification using deep learning technics," *International Journal of Grid and Distributed Computing*, vol. 11, no. 9, pp. 103–114, 2018, doi: 10.14257/ijgcd.2018.11.9.09.






- [26] Z. Qi, "The text classification of theft crime based on TF-IDF and XGBoost model," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Jun. 2020, pp. 1241–1246, doi: 10.1109/ICAICA50127.2020.9182555.
- [27] N. Peker and C. Kubat, "Application of Chi-square discretization algorithms to ensemble classification methods," *Expert Systems with Applications*, vol. 185, Dec. 2021, doi: 10.1016/j.eswa.2021.115540.
- [28] J. M. Moyano, E. L. Gibaja, K. J. Cios, and S. Ventura, "Review of ensembles of multi-label classifiers: Models, experimental study and prospects," *Information Fusion*, vol. 44, pp. 33–45, 2018, doi: 10.1016/j.inffus.2017.12.001.

## BIOGRAPHIES OF AUTHORS






**Regiolina Hayami**    graduated with a bachelor's degree at Informatics Engineering Department, State Islamic University of Sultan Syarif Kasim Riau, and master's degree in Master of Information Technology at Putra Indonesia University Padang. Now works as a lecturer at the Faculty of Computer Science, University Muhammadiyah of Riau. With research interests in the field of Machine learning algorithms and AI. She can be contacted at email: [regiolinahayami@umri.ac.id](mailto:regiolinahayami@umri.ac.id).






**Nooraini Yusoff**    is currently an Associate Professor at Faculty of Data Science and Computing at Universiti Malaysia Kelantan. She received her BSc. in Information Technology specialising in Artificial Intelligence, and MSc. in Intelligent System from Universiti Utara Malaysia. In 2012, she obtained her Ph.D. from University of Surrey, United Kingdom, in the field of Computational Neuroscience. Nooraini has more than 20 years, experience in Artificial Intelligence, Data Analytics, Machine Learning and system integration. She has involved in various AI and Data Analytics projects including flood disaster managements, Ministry of Higher Education graduate data analysis project, analysis on effectiveness and relevancy of ICT programs at Malaysia public universities, telecommunication fraud detection and agriculture data analytics projects. Nooraini has published a number of national and international journals and conference proceedings. She is also the reviewer of some reputable journals in the related fields. She can be contacted at email: [nooraini.y@umk.edu.my](mailto:nooraini.y@umk.edu.my).






**Kauthar Mohd Daud**    currently serves as a Senior Lecturer in the Center for Artificial Intelligence Technology, Faculty of Information Science and Technology in Universiti Kebangsaan Malaysia. She received her BSc. in Bioinformatics and MSc in Bioinformatics from Multimedia University and the University of Malaya. In 2019, she received her Ph.D. in computer science from Universiti Teknologi Malaysia. Her expertise includes optimization, metabolic modeling, artificial intelligence, and machine learning. She can be contacted at email: [kauthar.md@ukm.edu.my](mailto:kauthar.md@ukm.edu.my).



**Harun Mukhtar**    obtained his S.Kom degree in Informatics Engineering from STMIK-Amik Riau and his Master's degree in Computer Science from Universitas Putra Indonesia "YPTK" Padang, Indonesia in 2007 and 2010. He is currently pursuing his Ph.D in Data Science at Universiti Malaysia Kelantan, Malaysia. From 2008 to 2010 he joined STIKOM Pelita Indonesia as a lecturer, and from 2010 until now he has been a Lecturer at the Faculty of Computer Science, Universitas Muhammadiyah Riau (Umri), Indonesia. In 2023 he was awarded as Associate Professor at Umri. His main research interests are cryptography, cloud computing, data science, computer networks, and open-source applications. He has published several scientific papers indexed by Scopus and other indexes. He can be contacted at email: [harunmukhtar@umri.ac.id](mailto:harunmukhtar@umri.ac.id).



**Januar Al Amien**    completed education bachelor's degree in the Informatics Engineering Department, STMIK-AMIK Riau. And master's degree in Master of Information Technology at Putra Indonesia University Padang. Now working as a lecturer in the Department of Computer Science, University Muhammadiyah of Riau. With research interests in the field of machine learning algorithms and AI. He can be contacted at email: [januaralamen@umri.ac.id](mailto:januaralamen@umri.ac.id).