Computer Engineering and Applications Vol. 11, No. 1, February 2022

Identification of Indonesian Authors Using Deep Neural Networks

Firdaus Firdaus[1], Irvan Fahreza[1], Siti Nurmaini[*1], Annisa Darmawahyuni[1], Ade Iriani Sapitri[1], Muhammad Naufal Rachmatullah[1], Suci Dwi Lestari[1], Muhammad Fachrurrozi[2], Mira Afrina[3], Bayu Wijaya Putra[4]

[1]*Intelligent System Research Group, Universitas Sriwijaya, Palembang 30137, Indonesia,*
[2]*Informatics Engineering Department, Universitas Sriwijaya, Palembang 30137, Indonesia,*
[3]*Information System Department, Universitas Sriwijaya, Palembang 30137, Indonesia,*
[4]*Informatics Management Department, Universitas Sriwijaya, Palembang 30137, Indonesia*
**siti_nurmaini@unsri.ac.id*

ABSTRACT

Author Name Disambiguation (AND) is a problem that occurs when a set of publications contains ambiguous names of authors, i.e. the same author may appear with different names (synonyms) in other published papers, or author (authors) who may be different who may have the same name (homonym). In this final project, we will design a model with a Deep Neural Network (DNN) classifier. The dataset used in this final project uses primary data sourced from the Scopus website. This research focuses on integrating data from Indonesian authors. Parameters accuracy, sensitivity and precision are standard benchmarks to determine the performance of the method used to solve AND problems. The best DNN classification model achieves 99.9936% Accuracy, 93.1433% Sensitivity, 94.3733% Precision. Then for the highest performance measurement, the case of Non Synonym-Homonym (SH) has 99.9967% Accuracy, 96.7388% Sensitivity, and 97.5102% Precision.

Keywords: Author Name Disambiguation, Synonym, Homonym, Bibliographic Data, Deep Neural Network.

1. INTRODUCTION

Digital Library (DL) is assumed to provide content and high-quality content to its users, but they fail to provide it [1]. The main mistakes of DL are typography, data conversion, find and replace, copy and paste, meta data, different citation formats, ambiguous author names, and abbreviations of publication titles, etc. [2], [3]. Among these errors, the main problem that occurs is the ambiguous names of the authors of a study.

Author Name Disambiguation (AND) is an issue that occurs when a set of publications contains ambiguous author names. The same author may appear with different names (synonyms) in other published papers, or different authors but have the same name (homonyms). [3].

To overcome this problem, many studies have been carried out using the disambiguation feature, such as the name of the co-authors, the title of the paper or publication, the topic of the article, email/affiliation, etc. AND is a challenging task for scholars mining bibliographic information for scientific knowledge. A constructive approach to resolving name ambiguity is to use a computer algorithm to

identify the authors. Several algorithm-based disambiguation methods have been developed by data scientists [4].

The issue of author name ambiguity is closely related to other research areas such as entity disambiguation[5]–[10], name disambiguation [3], [11], [12], name variant[13], aliases name [14], and global name architecture [15]. Generally, author name ambiguity can be resolved by using different publication attributes such as publication title, affiliation, co-authors, keywords, references, abstract words, place and year of publication. [11], [16]–[18].

Some researchers use the author assignment approach with classification [16], [19], [20]. Even so, the results were unsatisfactory in terms of accuracy [16], [19]. The use of the Neural Network approach has been explored to identify publications. However, its performance gets a poor recall score with good results on accuracy [20]. To improve the performance of the conventional Neural Network algorithm, a DNN with multiple layers is used in this study [4]. Recent studies [21]–[23] have shown strong DNN performance in feature learning across multiple tasks. The internal features studied by DNN are relatively stable for data variants if the training data is sufficiently representative [23]. In addition, the use of Neural Networks also has the advantage of being able to build general models. This model can differentiate author names incrementally as new publications are entered into the data set.

2. METHODOLOGY

2.1 DATASET

The Indonesian author's publication dataset was developed using data sourced from Scopus. Scopus is an indexing and abstract database resource with full-text links produced by Elsevier Co[24]. The description of the dataset is shown in table 1. There are six attributes used in this study, namely Author Name, Label Author, Co-Author, Year, Venue, and Title.

TABLE 1.
Dataset Description

Description	Total Data
Name Instances	10212
Distinct authors label	1481
Distinct presented names	1673
Distinct venue	1177
Distinct co-authors names	9774
Year range	1995-2021
Synonym authors/row affected	203/1415
Homonym presented names/row affected	30/319
Non Synonym-Homonym	8422
Synonym-Homonym	56

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

2.2 DATA PREPROCESSING

Pre-processing of data is one of the stages before classification. Some features and labels will be processed in a different process depending on what information is contained in the feature. But generally the concept of feature processing is data normalization, feature extraction, and feature reduction. This procedure aims to make the data easily understood by the machine when performing the computational process and to get better research results. The preprocessing stages in this study follow the steps as shown in Figure 1.

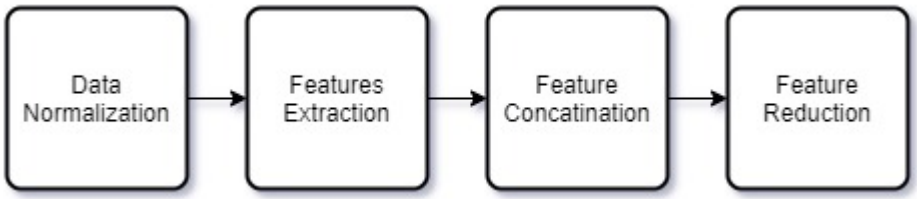


FIGURE 1. Data Preprocessing Steps

Data normalization adjusts values measured at different scales to a common scale. Normalization allows comparison of appropriate values from different data sets. Unnormalized data is difficult to scale because it is very difficult to calculate and compare with other parameters. The normalized features are Author name, venue, year, co-authors, and title.

All features are extracted so that the shape of all features becomes uniform. The Author name and venue features use the label encoder technique, the year feature uses the minmax scaler technique, the co-author feature uses a dictionary technique, and the title feature uses stemming and lemmatizing techniques.

After all the features are processed, then all the features will be combined into a single feature. All data is combined. then the data is split into training data and testing data with a ratio of 4:1, or 80% for training and 20% for testing.

Because all the features have been combined, the number of features becomes very large which can cause the machine to take a long time in processing data information. For this reason, the features are reduced by using Principal Component Analysis (PCA). PCA technique is used to simplify a data or reduce the dimensions of the data without reducing the characteristics of the data. In the data reduction process, PCA will maintain as much as 95% of the data variance.

2.3 EXPERIMENT SETUP

The DNN classifier has several parameters that can be set to produce the best results in classifying data, including the number of epochs, learning rate, number of hidden layers, batch size, number of nodes, and type of optimizer. Epoch is the number of times the process of repeating the model in one training. The learning rate controls how quickly the model can be adapted to the problem. Smaller learning rates require more epochs in training because only small changes are made to each epoch, while larger learning rates require fewer epochs because there are fast changes made to each

epoch. The hidden layer is the layer between the input layers and the output layers, where the artificial neuron takes a set of weighted inputs and produces output through the activation function. Batch size is a parameter that controls the number of training samples that must be carried out in each epoch. Nodes also called neurons or perceptrons are the number of computational units that have one or more input weights. Optimizer is an algorithm or method used to minimize loss function or to maximize production efficiency. The optimizer helps figure out how to change the weight and learning rate of the neural network to reduce losses. The architecture used in this study is shown in Figure 2.

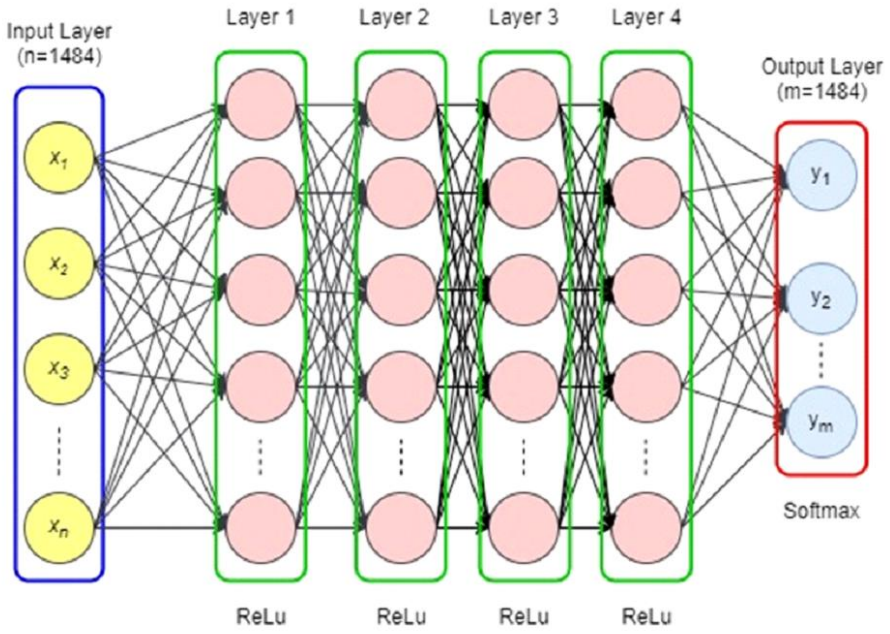


FIGURE 2. Model Architecture

There were a total of 108 trials using the DNN classification to find one best model. In this study, the experiment used the number of neurons of 50, 100 and 150, Learning Rate 0.001 and 0.0001, Batch Size 16, 32 and 64, hidden layer 2, 3 and 4, and Adam and Adamax as Optimizer. For the activation function on the input layer and hidden layer using Rectified Linear Units (ReLu). As for the Output layer activation function using Softmax with a loss function using Categorical Cross Entropy. While the other parameters will be set by default.

2.4 PERFORMANCE METRIC

Performance measurement is used to determine whether something that has been done has succeeded in achieving its goals and meeting the desired targets. The performance measures used in this study are specificity(1), precision(2), sensitivity/recall(3), accuracy(4), error-rate(5), and F1-score(6).

$$Specificity\ Average = \frac{\sum_{i=1}^{I} \frac{tni}{tn+fpi}}{I} \quad (1)$$

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

tn is the number of negative data detected correctly, then fp is the positive data detected as negative data. Specificity measures the proportion of negatives that are identified as true.

$$Precision\ Average = \frac{\sum_{i=1}^{I} \frac{tpi}{tpi+fpi}}{I} \quad (2)$$

tp is the number of positive data correctly classified by the system, then fp is the positive data detected as negative data. Precision is the proportion of positive labeled predictions that are correct to the overall positive prediction.

$$Sensitivity\ Average = \frac{\sum_{i=1}^{I} \frac{tpi}{tpi+fni}}{I} \quad (3)$$

Sensitivity is the proportion of data that is predicted to be positive from data that are indeed positive.

$$Accuracy\ Average = \frac{\sum_{i}^{I} = 1\ \frac{tpi+tni}{tpi+fpi+fni+tni}}{I} \quad (4)$$

Accuracy merupakan tingkat kedekatan antara nilai prediksi dengan nilai aktual.

$$Error\ Rate = \frac{\sum_{i=1}^{I} \frac{fpi+fn1}{tpi+fpi+fni+tni}}{I} \quad (5)$$

Error Rate is the proportion of patterns that have been incorrectly classified.

$$F1\ avg = \frac{(\beta^2+1)Presisi\ average\ \times Recall\ average}{\beta^2 Presisi\ average + Recall\ average} \quad (6)$$

F1-Score summarizes the number of precision and recall by taking the average value of both.

3. RESULT AND DISCUSSION

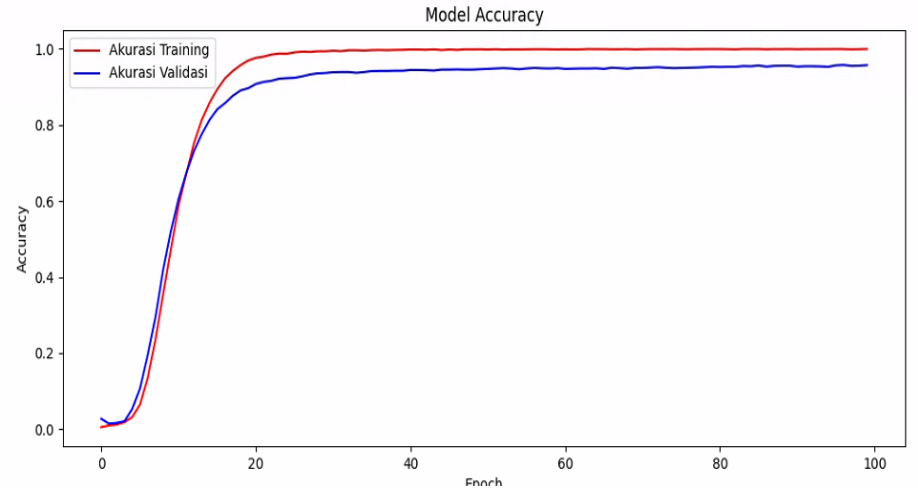
From the results of data pre-processing, a number of features are generated for each attribute. Details of the number of features of each attribute can be seen in table 2.

TABLE 2.
Number of Features Generated from Attributes

Attributes	Features
Author(s) Name	1673
Author(s) (label)	1484
Co-Author(s)	9774
Year	1177
Venue	21
Title	10166
Total Fitur	22811

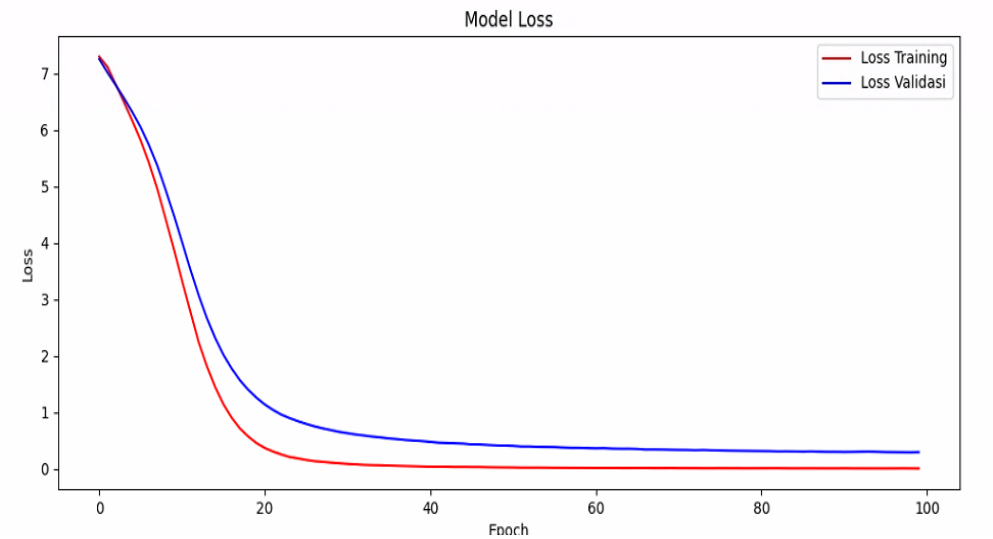
The number of features generated from the pre-processing stage is 22811 features. PCA managed to maintain as much as 95% of the data variance which made PCA managed to reduce approximately 73.38% of the total features to 6071 features. Data splitting with a ratio of 4:1, which is 80% for training data with 8169 lines of data and 20% for testing data with 2043 lines of data.

By using the DNN classifier, 108 experiments have been conducted with predetermined scenarios to get the best performance. From the experimental results, the best accuracy is generated by a model with a hyperparameter of 2 hidden layers, 32 batch size, 0.0001 learning rate, 100 neuron nodes, and the optimizer using adamax. With validation accuracy of 95.2521%. From the graphs of accuracy and loss (figure 3), the model is best fit for training and testing data.



(a) Accuracy model

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)



(b) Loss model

FIGURE 3. Model Accuracy and Loss

The performance of the model on AND problems shows good results. For homonym and synonym problems, the proposed method produces very good accuracy values. However, this method has a rather poor performance for the combination of synonym-homonym problems (table 3).

TABLE 3.
Proposed Method Performance for AND Problems

Performance Measurement (%)	Synonym	Homonym	Synonym-Homonym	Non Synonym-Homonym
Accuracy	99,8118	99,8231	94,8052	99,9967
Specificity	99,9057	99,9116	97,2727	99,9984
Precision	77,3032	94,3396	78,5714	97,5102
Recall	74,5540	92,1384	78,5714	96,7388
Erorr Rate	0,1882	0,17689	5,19481	0,00329
F1-Score	74,3855	92,9560	76,1905	96,9258

4. CONCLUSION

The proposed method has a very good performance to solve the identification of Indonesian authors. The proposed method can also solve the AND problem well. However, the combination of synonym and homonym problems is still a challenge that must be solved in further research.

REFERENCES

[1] B. E. Coggins and P. Zhou, “Clean,” New Dev. NMR, vol. 2017-Janua, no. 11, pp. 169–219, 2017.
[2] P. Christen, “A comparison of personal name matching: Techniques and

practical issues," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 290–294, 2006.
[3] A. A. Ferreira and M. A. Gonçalves, "A Brief Survey of Automatic Methods for Author Name Disambiguation," vol. 41, no. 2, 2012.
[4] Firdaus et al., "Author identification in bibliographic data using deep neural networks," Telkomnika (Telecommunication Comput. Electron. Control., vol. 19, no. 3, pp. 911–919, 2021.
[5] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, pp. 1–35, 2007.
[6] E. L. Murnane, B. Haslhofer, and C. Lagoze, "RESLVE: Leveraging user interest to improve entity disambigution on short text," WWW 2013 Companion - Proc. 22nd Int. Conf. World Wide Web, pp. 1275–1283, 2013.
[7] A. Chisholm and B. Hachey, "Entity Disambiguation with Web Links," Trans. Assoc. Comput. Linguist., vol. 3, pp. 145–156, 2015.
[8] S. Oramas, L. Espinosa-Anke, M. Sordo, H. Saggion, and X. Serra, "ELMD: An automatically generated Entity Linking gold standard dataset in the Music Domain," Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016, pp. 3312–3317, 2016.
[9] A. Krzywicki, W. Wobcke, M. Bain, J. Calvo Martinez, and P. Compton, "Data mining for building knowledge bases: Techniques, architectures and applications," Knowl. Eng. Rev., vol. 31, no. 2, pp. 97–123, 2016.
[10] L. Zhu, M. Ghasemi-Gol, P. Szekely, A. Galstyan, and C. A. Knoblock, "Unsupervised entity resolution on multi-type graphs," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9981 LNCS, pp. 649–667, 2016.
[11] P. Mitra, J. Kang, D. Lee, and B. On, "Comparative study of name disambiguation problem using a scalable blocking-based framework," in Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05), 2005, pp. 344–353.
[12] D. Shin, T. Kim, J. Choi, and J. Kim, "Author name disambiguation using a graph model with node splitting and merging based on bibliographic information," pp. 15–50, 2014.
[13] G. Louppe, H. T. Al-Natsheh, M. Susik, and E. J. Maguire, "Ethnicity sensitive author disambiguation using semi-supervised learning," in International Conference on Knowledge Engineering and the Semantic Web, 2016, pp. 272–287.
[14] I. Johannes, C. Scholtes, F. Peter, and E. Maes, "System and method for authorship disambiguation and alias resolution in electronic data," vol. 2, no. 12, 2016.
[15] R. L. Pyle, "Towards a global names architecture: The future of indexing scientific names," Zookeys, vol. 2016, no. 550, pp. 261–281, 2016.
[16] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. F. Laender, "Effective self-training author name disambiguation in scholarly digital libraries," in Proceedings of the 10th annual joint conference on Digital libraries, 2010, pp. 39–48.
[17] S. Elliott, "Survey of author name disambiguation: 2004 to 2010," Libr. Philos. Pract., vol. 2010, no. NOVEMBER, 2010.
[18] L. V. B. Esperidião et al., "Reducing Fragmentation in Incremental Author Name Disambiguation," Jidm, vol. 5, no. 3, pp. 293–307, 2014.
[19] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, "Two supervised

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

learning approaches for name disambiguation in author citations," in Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004., 2004, pp. 296–305.
[20] B. Cao, T. Srikanthan, and C. H. Chang, "Authorship recognition and disambiguation of scientific papers using a neural networks approach," vol. 152, no. 20045116, pp. 0–9, 2005.
[21] D. Cires and U. Meier, "Multi-column Deep Neural Networks for Image Classification," pp. 3642–3649, 2012.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," 2012.
[23] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks," pp. 1–9, 2013.
[24] J. F. Burnham, "Scopus database: A review," Biomed. Digit. Libr., vol. 3, pp. 1–8, 2006.
[25] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," Adv. Neural Inf. Process. Syst., vol. 4, pp. 2843–2851, 2012.
[26] C. Szegedy et al., "Going deeper with convolutions," Res. Methods Appl. Settings, pp. 319–338, 2021.
[27] R. Collobert, L. Bottou, J. Weston, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch Ronan," Proc. - 2017 IEEE 3rd Int. Conf. Collab. Internet Comput. CIC 2017, vol. 2017-Janua, pp. 328–338, 2017.
[28] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," Empir. Methods Nat. Lang. Process., no. October, pp. 1631–1642, 2004.
[29] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, 2013.
[30] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3361–3368, 2011.
[31] G. Montavon et al., "Machine learning of molecular electronic properties in chemical compound space," New J. Phys., vol. 15, pp. 0–16, 2013.
[32] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science (80-.)., vol. 313, no. 5786, pp. 504–507, 2006.
[33] N. Hou, H. Dong, Z. Wang, W. Ren, and F. E. Alsaadi, "Non-fragile state estimation for discrete Markovian jumping neural networks," Neurocomputing, vol. 179, pp. 238–245, 2016.
[34] F. Yang, H. Dong, Z. Wang, W. Ren, and F. E. Alsaadi, "A new approach to non-fragile state estimation for continuous neural networks with time-delays," Neurocomputing, vol. 197, pp. 205–211, 2016.
[35] Y. Yu, H. Dong, Z. Wang, W. Ren, and F. E. Alsaadi, "Design of non-fragile state estimators for discrete time-delayed neural networks with parameter uncertainties," Neurocomputing, vol. 182, pp. 18–24, 2016.
[36] Y. Yuan and F. Sun, "Delay-dependent stability criteria for time-varying delay

neural networks in the delta domain," Neurocomputing, vol. 125, pp. 17–21, 2014.

[37] J. Zhang, L. Ma, and Y. Liu, "Passivity analysis for discrete-time neural networks with mixed time-delays and randomly occurring quantization effects," Neurocomputing, vol. 216, pp. 657–665, 2016.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)