Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 179 (2021) 624–631

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020

Deep Learning as a Vector Embedding Model for Customer Churn

Tjeng Wawan Cenggoro[a,c,*], Raditya Ayu Wirastari[a], Edy Rudianto[a], Mochamad Ilham Mohadi[a], Dinne Ratj[b], Bens Pardamean[b,c]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480*

[b]*Computer Science Department, BINUS Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta, Indonesia, 11480*

[c]*Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia, 11480*

Abstract

To face the tight competition in the telecommunication industry, it is important to minimize the rate of customers stopping their service subscription, which is known as customer churn. For that goal, an explainable predictive customer churn model is an essential tool to be owned by a telecommunication provider. In this paper, we developed the explainable model by utilizing the concept of vector embedding in Deep Learning. We show that the model can reveal churning customers that can potentially be converted back to use the previous telecommunication service. The generated vectors are also highly discriminative between the churning and loyal customers, which enable the developed models to be highly predictive for determining whether a customer would cease his/her service subscription or not. The best model in our experiment achieved a predictive performance of 81.16%, measured by the F1 Score. Further analysis on the clusters similarity and t-SNE plot also confirmed that the generated vectors are discriminative for churn prediction.

Keywords: customer churn, customer behavior, deep learning, vector embedding, representation learning

1. Introduction

The competition among telecommunication companies in these modern days is ferocious[1]. One of the factors that raise the competition is customer churn, which is defined as the customer behavior to stop their current telecommunication service subscription[2]. The trend of customer churn is currently increasing, which can be seen from the current tendency for telecommunication customers to switch their current service to a different telecommunication provider[3]. If it is not handled thoughtfully, customer churn could severely afflict the telecommunication company[4]. The termination of service subscription definitely decreases the revenue of a telecommunication company. In order to

* Corresponding author.
E-mail address: wcenggoro@binus.edu

1877-0509
10.1016/j.procs.2021.01.048

keep the customers that might stop their subscription, it is essential for the company to have an insight into the behavior of those customers. This goal can be achieved by developing an explainable Artificial Intelligence (AI) model. This is especially true for the telecommunication industry in developing countries, with the increasing interest in AI development[5].

One of the approach that can be taken for developing an explainable model is vector embedding. This approach requires the model to generate an informative vector for each data point it represents. Intended information can be infused to the vector by letting the model to perform a specific task. By inferring the infused information in the vector, hidden knowledge in the data can be revealed. With the explainability via vector embedding, a deep learning model can act as a decision support system for a telecommunication company. This could help the company picks a decision based on the customer behavior, as the recommendation from a decision support system is highly correlated with customer behavior[6]. The model can also be used to forecast the possibility of churning customers. Because a forecasting model can be employed as a decision support system as well[7], the model can also give recommended decisions from a different perspective.

Having observed the urgency, we carried this study to proof the capability of deep learning as an explainable customer churn model. Specifically, we showed that a deep learning model can be used as a vector embedding model to visualize the relationship among churning and loyal customers in a vector space. Therefore, it can be used as a tool to identify types of churning and loyal customers, by observing the cluster formed by the vectors. The contribution of this study are:

- to develop vector embedding deep learning models as explainable customer churn models,
- to show the capability of the developed models to reveal useful information from a customer churn dataset.

2. Literature Review

Deep learning is currently the most preferred AI method for many applications[8,9,10]. It is also currently the preferred method as a churn model for customers in retail industry[11], music streaming service[12], mobile game[13], and also telecommunication industry[14,15,16,17,18,19]. Despite the popularity, the developed models can still be considered as a black box because of the deep learning complicated architecture. Many approaches has been developed in the effort to open the black box nature of deep learning. One of the approach is vector embedding, where the model is allowed to generate a vector that can be analyzed to uncover the learned information inside the deep learning model. Vector embedding has been applied for customer churn in mobile game[13], although the vectors were not analyzed.

The concept of vector embedding was popularized by its application in Natural Language Processing (NLP) named as word2vec[20]. It is followed by numerous word embedding models. The popular examples are GLoVE[21], fasttext[22], and ELMo[23]. The vector embedding model in NLP is used to generate a vector for each word in a document, which has a defining characteristic of the word it refers to. For instance, in the word2vec model, the vector of the word "King" is relatively close in Euclidean Space with the vector formed by subtracting the vector of "Woman" from "Queen" and adding the vector of "Man". Despite its starting point in NLP research, vector embedding can be generalized to other types of categorical data such as items[24], books[25], academic papers[26], and authors[27]. The process of vector embedding for categorical data is described in Algorithm 1. Not only for categorical data, but vector embedding can also be generalized to other case such as graph analysis[28], biological data modelling , and drug repositioning[29,30,31].

for *each categorical data* **do**
  convert the data to one hot vector (dimension = $1 \times n$, where n is the number of category)
  multiply the one hot vector with a trainable matrix W (dimension = $n \times d$, where d is the dimension of the embedding vector)
  use the output of the matrix multiplication as the generated embedding vector (dimension = $1 \times d$)
end

Algorithm 1: Categorical embedding algorithm.

3. Dataset

The dataset used in this study is a publicly available dataset for telecommunication customer churn[32]. This comprises the data of 3,333 telecommunication customers which churn attribute is flagged as true or false. To put it in simple terminologies, we refer the churn=true customers as churn customers and churn=false customers as loyal customers. The behavior of each customer is defined with the other 19 attributes in the dataset. Among these attributes, four attributes are categorical data, while the others are numerical data. During the initial inspection of the data, we found that one categorical attribute named "Account Length" has a distribution that is not discriminative towards defining churn customers. Thus, we did not include this attribute in the vector embedding model. In the initial inspection, we also found that the proportion of churn and loyal customers is imbalanced with the ratio of 15:85.

Table 1: Categorical Attributes in the Dataset

Attribute Name	Type	Description
State	Categorical	Location of phone number's state origin.
International plan	Categorical	Whether or not a customer has an international plan.
Voice mail plan	Categorical	Whether or not a customer has a voice mail plan.
Area code	Categorical	Identifier for geographic region of phone number origin.
Number of vmail messages	Numerical	Number of voice mail sent by customers.
Total day minutes	Numerical	Total minutes spent on calls during the day.
Total day calls	Numerical	Total calls made during the day.
Total day charge	Numerical	Total charge for calls during the day.
Total eve minutes	Numerical	Total minutes spent on calls during the evening.
Total eve calls	Numerical	Total calls made during the evening.
Total eve charge	Numerical	Total charge for calls during the evening.
Total night minutes	Numerical	Total minutes spent on calls during the night.
Total night calls	Numerical	Total calls made during the night.
Total night charge	Numerical	Total charge for calls during the night.
Total international minutes	Numerical	Total minutes spent on international calls.
Total international calls	Numerical	Total international calls made.
Total international charge	Numerical	Total charge for international calls.
Customer service calls	Numerical	Number of calls made to customer service.

4. Methodology

This section elaborates the procedure of our experiment and our proposed model. We released the code to replicate our experiment at https://github.com/wawancenggoro/churn-vector-embedding.

4.1. The Proposed Vector Embedding Model

Figure 1 illustrates the architecture of our proposed model for customer churn vector embedding. As depicted in the figure, the numerical and the categorical data are first processed with separate layers before the concatenated intermediate representation is fed to the next layer. For the numerical data, the processing layer is a fully connected layer. Meanwhile, the processing layer of the categorical data is a categorical embedding layer, which is a layer that was used by the word embedding model to generate an embedded vector. This should not be confused with the vector embedding layer generator in our proposed model. The embedded vector in our proposed model is generated using the last fully connected layer, which is colored in orange in Figure 1. In the figure, ReLU means the Rectified Linear Unit activation function, while BN is the Batch Normalization layer[33]. To infuse the generated embedded vector with the discriminative characteristics of churn and loyal customers, we trained the model to classify the churn attribute in a supervised learning framework. The loss function employed is a standard softmax loss.

Considering that the proportion of churn and loyal customers is imbalanced, we hypothesized that special treatment is needed for the optimal performance of our proposed model. It is in fact a common knowledge in deep learning

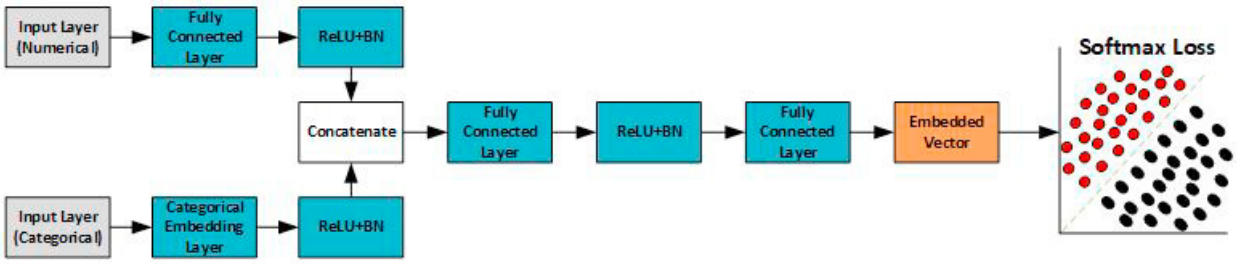


Fig. 1: The proposed architecture for customer churn vector embedding model

research that, without any treatment, imbalance data can degrade the performance of the trained model[34,35,35,36]. Therefore, we experimented on the effect of a weighted softmax loss to provide larger gradient feedback for classifying churn customers. The weight ratio of churn versus loyal customers we explored was 4:6, 3:7, 2:8, and 1:9.

4.2. Evaluation

To detect if the trained models suffer overfitting, we split the dataset into three sets, namely training, validation, and testing. The dataset was split with the ratio of 60:20:20. The training set was used to train the models for 100 epochs. At each epoch, the loss value of the model was plotted together with the training loss value. Based on the plots, we chose the best configuration of the weight ratio of the softmax loss. The best configuration was afterward evaluated using the testing set for comparison with the unweighted model. The evaluation was carried out by using the accuracy and F1 Score. Additionally, we also measure the compactness of the cluster formed by each of the churn and loyal customers embedded vector. We called the compactness measurement as intra-cluster similarity, which is formulated as:

$$\text{intra-cluster similarity} = \sum_{i=0}^{n} \frac{Cos(x_i, \bar{x})}{n} \tag{1}$$

where n is the number of data and $\bar{x}$ is the mean embedded vector of the data. We also measure the similarity between a cluster to the other cluster, which is formulated as:

$$\text{inter-cluster similarity} = \sum_{i=0}^{n} \frac{Cos(x_i, \bar{y})}{n} \tag{2}$$

where $\bar{y}$ is the mean embedded vector of the data from the other cluster. In this case, x can be data from the churn class and y from the loyal class or vice versa. Finally, the ratio between the inter-cluster similarity and intra-cluster similarity is used to evaluate the models performance. By the aforementioned definition, the higher the value of this ratio, the better the model is. To complement the quantitative assessment, we also qualitatively assessed the models by using t-SNE[37] to plot the embedded vectors in two-dimensional space.

5. Results and Discussions

Figure 2 shows the training versus validation loss plot of the model with unweighted softmax loss and the weighted softmax loss with the churn:loyal weight ratio of 4:6, 3:7, 2:8, and 1:9. Only the unweighted and 3:7 weighted model that did not suffer overfitting. Therefore, we picked these models to be compared using the evaluation metrics that are defined in sub-section 4.2. The comparison of the models is summarized in Table 2.

Based on the comparison result, the unweighted model outperforms the weighted model by 0.45% in terms of accuracy. However, it should be noted that the data is imbalanced, hence the accuracy is not a suitable metric to evaluate a model trained using this data. The more suitable metric, the F1 Score, demonstrate the opposite. The F1 Score of the weighted model is improved by 2.21% compared to the weighted model. As for the inter-class:intra-class

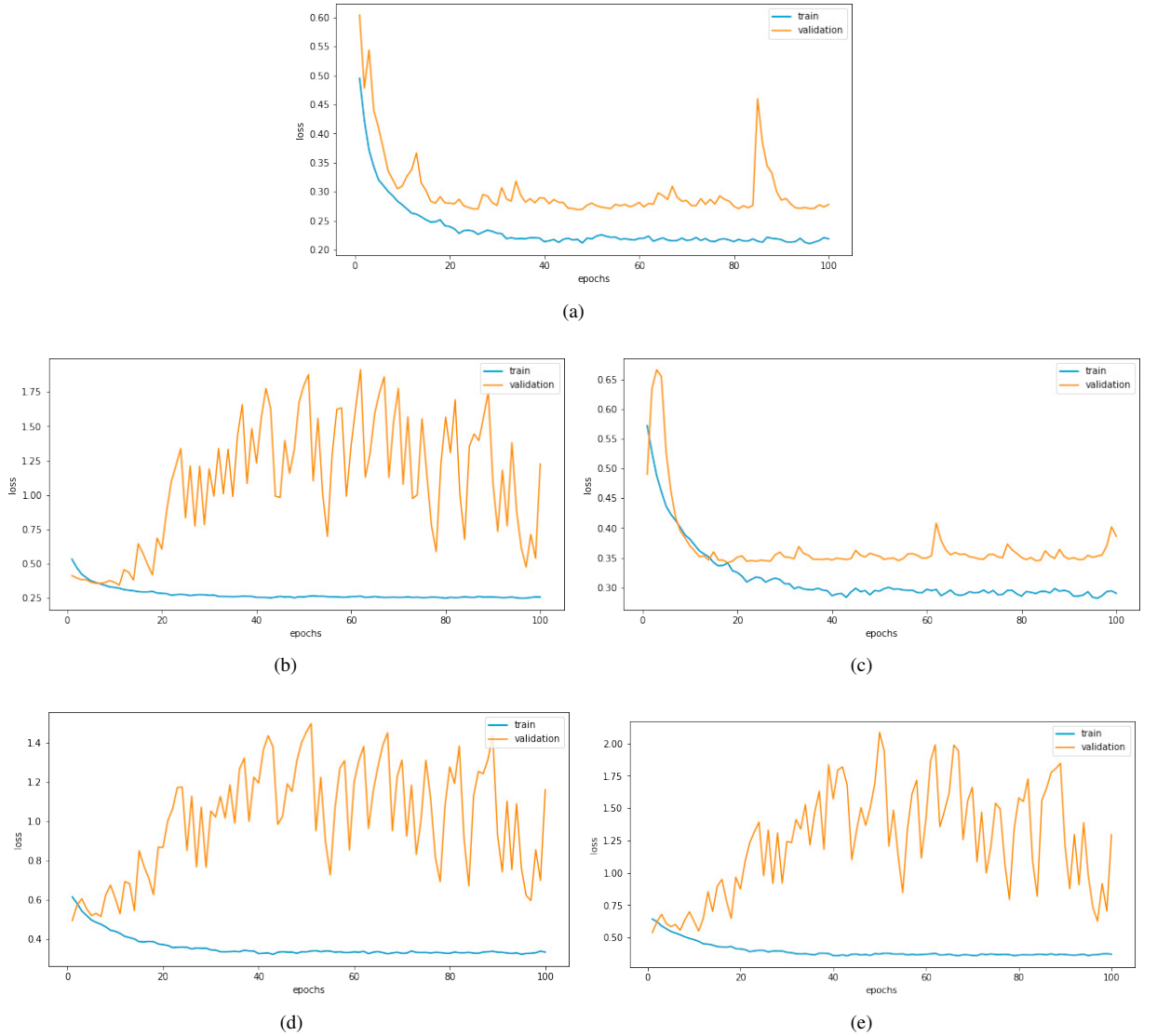


Fig. 2: Training and validation loss plot for: (a) unweighted softmax loss and weighted softmax loss with the churn:loyal weight ratio of (b) 4:6, (c) 3:7, (d) 2:8, and (e) 1:9.

similarity ratio, the unweighted model is better for the loyal customers cluster, while the weighted model is better for the churn customers cluster. This result is expected, as the gradient penalty for churn class is heavier for the weighted model, leading to a better churn class modeling, but might decrease the performance to model the loyal customers behavior.

Meanwhile, the qualitative assessment of the t-SNE plots reveals that the generated embedded vectors of both models are discriminative towards churn versus loyal customers. As depicted in Figure 3, the churn embedded vectors have a tendency to flock on the top part of the plot, where the loyal cluster is less dense. From these plots, the quantitative results can also be confirmed. It can be seen that the churn cluster of the weighted model embedded vectors is denser than the weighted model churn cluster on the top part of the plots. Because of the denser cluster in this part, more churn class embedded vectors can be separated from the loyal customers cluster, which causing the better F1 Score for the weighted model.

Table 2: Performance Comparison

Metrics	Unweighted	Churn:Loyal Weight Ratio = 3:7
Accuracy	**89.82%**	89.37%
F1 Score	78.95%	**81.16%**
Loyal Customers inter-class similarity	**0.5440**	0.5327
Loyal Customers intra-class similarity	0.5904	**0.5546**
Loyal Customers intra-class:inter-class similarity ratio	**1.0854**	1.0411
Churn Customers inter-class similarity	**0.5547**	0.5345
Churn Customers intra-class similarity	0.6486	**0.6352**
Churn Customers intra-class:inter-class similarity ratio	1.1693	**1.1883**

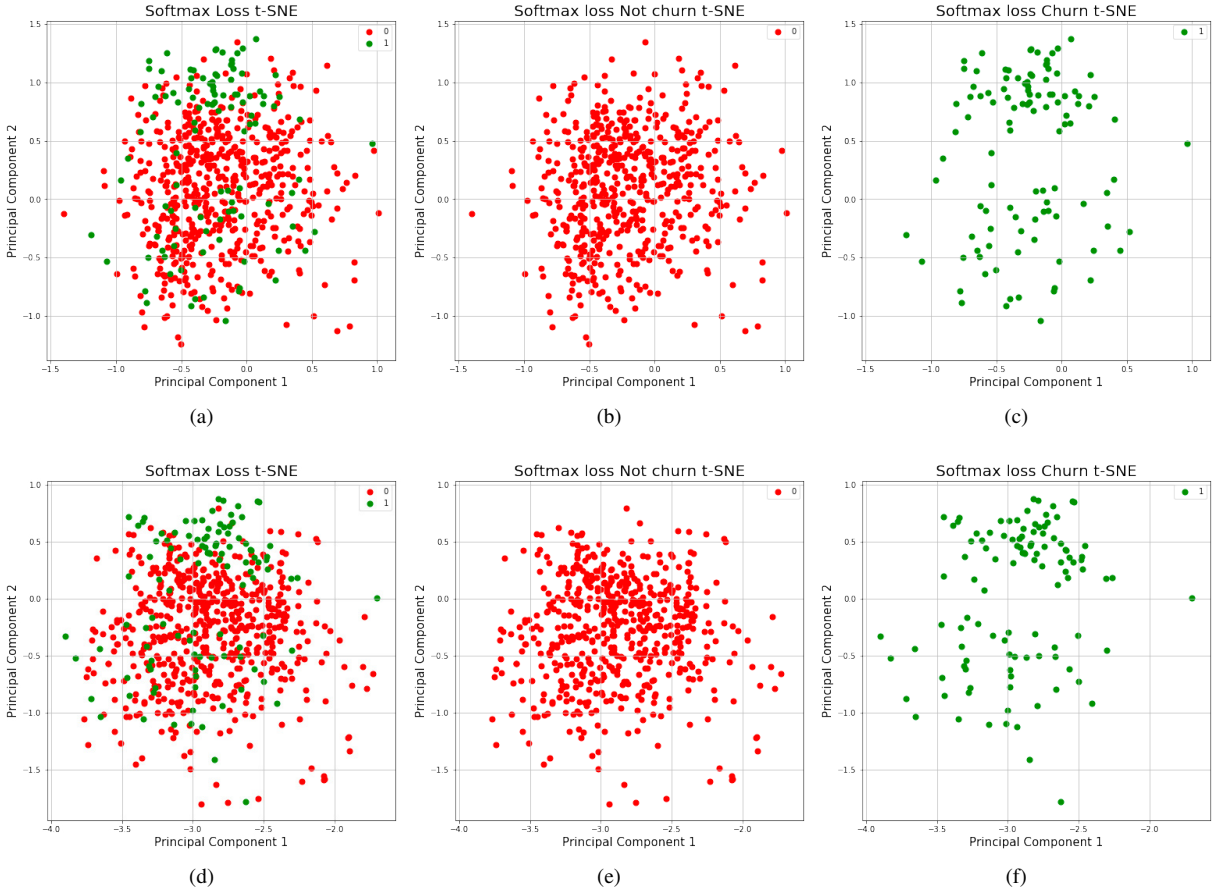
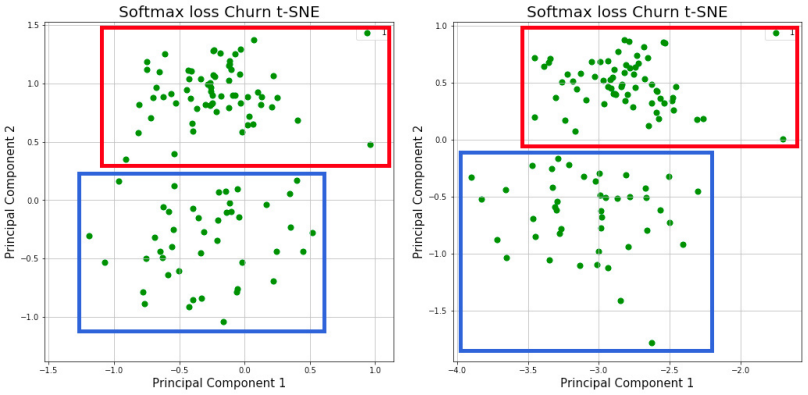


Fig. 3: The t-SNE plot of the embedded vectors generated by: (a) unweighted softmax loss model and (d) weighted softmax loss model with churn:loyal weight ratio of 3:7. The red dot represents the vector of a loyal customer and green dot represents the vector of a churning customer. For better visualization, we also depict the plot for the vector of loyal customers only (b,c) and churning customers only (e,f). (b) and (c) are the vectors from the unweighted softmax loss model. (e) and (f) are the vectors from the 3:7 weighted softmax loss model.

With the highly discriminative vectors generated by both models, we can guarantee that the conclusion drawn by observing the t-SNE plots are valid. Having observed the t-SNE plots, especially the plots for only the churning customers, we can conclude that there are two clusters of churning customer. The two clusters are marked with red box and blue box in Figure 4. On the one hand, the customer vectors in the blue box cluster are mapped to the area

where most of the loyal customers reside. This means that the behavior of customers in this cluster is similar to the loyal customers. Therefore, we could argue that, although the customers are churning customer, they can potentially be converted back as loyal customer. On the other hand, the customers in the red box are visibly separated with the cluster of loyal customers. Thus, the customers in this cluster is not a potential target to be converted back as a loyal customer. This observation can be utilized by a telecommunication company to focus their marketing effort to the customers in blue box.



(a) (b)

Fig. 4: The t-SNE plot of churning customer vectors from (a) unweighted softmax loss model and (b) weighted softmax loss model reveals two clusters. The cluster in the blue box are customers that can potentially be converted back to use the previous telecommunication service, while the customers in the red box cluster are not likely to be converted back.

6. Conclusion

In this paper, we explored the use of deep learning models for generating embedded vectors that exhibit discriminative characteristics for modeling customer churn. We showed that the discriminative vectors can be visualized in a 2D space to reveal two groups of churning customers: one group is highly probable to leave and another group can potentially be retained. This information can provide a decision support to telecommunication companies to effectively target their marketing effort to retain as much customers as possible. Despite the satisfying result of this study, the vector embedding model can potentially be improved by applying more advanced techniques. One of the promising technique to be employed in future works is to use cosine-distance-based loss functions, which have been proved for face recognition to improve the discrimination among embedding vectors in different classes.

References

1. Yan, L., Fassino, M., Baldasare, P.. Predicting customer behavior via calling links. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*; vol. 4. IEEE; 2005, p. 2555–2560.
2. Mattison, R.. *The telco churn management handbook*. Lulu. com; 2006.
3. Khan, A.A., Jamwal, S., Sepehri, M.M.. Applying data mining to customer churn prediction in an internet service provider. *International Journal of Computer Applications* 2010;**9**(7):8–14.
4. Kirui, C., Hong, L., Cheruiyot, W., Kirui, H.. Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. *International Journal of Computer Science Issues (IJCSI)* 2013;**10**(2 Part 1):165.
5. Muchtar, K., Rahman, F., Cenggoro, T.W., Budiarto, A., Pardamean, B.. An improved version of texture-based foreground segmentation: block-based adaptive segmenter. *Procedia Computer Science* 2018;**135**:579–586.
6. Atmojo, R.N.P., Pardamean, B., Abbas, B.S., Cahyani, A.D., Manulang, I.D., et al. Fuzzy simple additive weighting based, decision support system application for alternative confusion reduction strategy in smartphone purchases. *American Journal of Applied Sciences* 2014; **11**(4):666.
7. Caraka, R.E., Bakar, S.A., Pardamean, B., Budiarto, A.. Hybrid support vector regression in electric load during national holiday season. In: *2017 International Conference on Innovative and Creative Information Technology (ICITech)*. IEEE; 2017, p. 1–6.

8. Prabowo, H., Cenggoro, T.W., Budiarto, A., Perbangsa, A.S., Muljo, H.H., Pardamean, B.. Utilizing mobile-based deep learning model for managing video in knowledge management system. *International Journal of Interactive Mobile Technologies (iJIM)* 2018;**12**(6):62–73.
9. Cenggoro, T.W., Tanzil, F., Aslamiah, A.H., Karuppiah, E.K., Pardamean, B.. Crowdsourcing annotation system of object counting dataset for deep learning algorithm. In: *IOP conference series: earth and environmental science*; vol. 195. 2018, p. 012063.
10. Pardamean, B., Cenggoro, T.W., Rahutomo, R., Budiarto, A., Karuppiah, E.K.. Transfer learning from chest x-ray pre-trained convolutional neural network for learning mammogram data. *Procedia Computer Science* 2018;**135**:400–407.
11. Dingli, A., Marmara, V., Fournier, N.S.. Comparison of deep learning algorithms to predict customer churn within a local retail industry. *International journal of machine learning and computing* 2017;**7**(5):128–132.
12. Jie, Z., YAN, J.f., Lu, Y., Meng, W., Peng, X.. Customer churn prediction model based on lstm and cnn in music streaming. *DEStech Transactions on Engineering and Technology Research* 2019;(aemce).
13. Liu, X., Xie, M., Wen, X., Chen, R., Ge, Y., Duffield, N., et al. A semi-supervised and inductive embedding model for churn prediction of large-scale mobile games. In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE; 2018, p. 277–286.
14. Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., Ghatasheh, N.. Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis. *Life Science Journal* 2014;**11**(3):75–81.
15. Mishra, A., Reddy, U.S.. A novel approach for churn prediction using deep learning. In: *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE; 2017, p. 1–4.
16. Agrawal, S., Das, A., Gaikwad, A., Dhage, S.. Customer churn prediction modelling based on behavioural patterns analysis using deep learning. In: *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*. IEEE; 2018, p. 1–6.
17. Karanovic, M., Popovac, M., Sladojevic, S., Arsenovic, M., Stefanovic, D.. Telecommunication services churn prediction-deep learning approach. In: *2018 26th Telecommunications Forum (TELFOR)*. IEEE; 2018, p. 420–425.
18. Cao, S., Liu, W., Chen, Y., Zhu, X.. Deep learning based customer churn analysis. In: *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE; 2019, p. 1–6.
19. Kumar, S., Kumar, M.. Predicting customer churn using artificial neural network. In: *International Conference on Engineering Applications of Neural Networks*. Springer; 2019, p. 299–306.
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. 2013, p. 3111–3119.
21. Pennington, J., Socher, R., Manning, C.D.. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, p. 1532–1543.
22. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 2017;**5**:135–146.
23. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics; 2018, p. 2227–2237. doi:\bibinfo{doi}{10.18653/v1/N18-1202}. URL https://www.aclweb.org/anthology/N18-1202.
24. Barkan, O., Koenigstein, N.. Item2vec: neural item embedding for collaborative filtering. In: *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE; 2016, p. 1–6.
25. Rahutomo, R., Perbangsa, A.S., Soeparno, H., Pardamean, B.. Embedding model design for producing book recommendation. In: *2019 International Conference on Information Management and Technology (ICIMTech)*; vol. 1. IEEE; 2019, p. 537–541.
26. Ganguly, S., Pudi, V.. Paper2vec: Combining graph and text information for scientific paper representation. In: *European Conference on Information Retrieval*. Springer; 2017, p. 383–395.
27. Ganguly, S., Gupta, M., Varma, V., Pudi, V., et al. Author2vec: Learning author representations by combining content and link information. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee; 2016, p. 49–50.
28. Narayanan, A., Chandramohan, M., Rajasekar Venkatesan, L., Chen, Y.L., Jaiswal, S.. graph2vec: Learning distributed representations of graphs. In: *13th International Workshop on Mining and Learning with Graphs*. 2017, .
29. Ngo, D.L., Yamamoto, N., Tran, V.A., Nguyen, N.G., Phan, D., Lumbanraja, F.R., et al. Application of word embedding to drug repositioning. *Journal of Biomedical Science and Engineering* 2016;**9**(1):7–16.
30. Xu, Y., Song, J., Wilson, C., Whisstock, J.C.. Phoscontext2vec: a distributed representation of residue-level sequence contexts and its application to general and kinase-specific phosphorylation site prediction. *Scientific reports* 2018;**8**(1):1–14.
31. Lumbanraja, F.R., Mahesworo, B., Cenggoro, T.W., Budiarto, A., Pardamean, B.. An evaluation of deep neural network performance on limited protein phosphorylation site prediction data. *Procedia Computer Science* 2019;**157**:25–30.
32. Churn in Telecom's dataset. 2020. URL https://www.kaggle.com/becksddf/churn-in-telecoms-dataset.
33. Ioffe, S., Szegedy, C.. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*. 2015, p. 448–456.
34. Johnson, J.M., Khoshgoftaar, T.M.. Survey on deep learning with class imbalance. *Journal of Big Data* 2019;**6**(1):27.
35. Cenggoro, T.W., Isa, S.M., Kusuma, G.P., Pardamean, B.. Classification of imbalanced land-use/land-cover data using variational semi-supervised learning. In: *2017 International Conference on Innovative and Creative Information Technology (ICITech)*. IEEE; 2017, p. 1–6.
36. Pardamean, B., Muljo, H.H., Cenggoro, T.W., Chandra, B.J., Rahutomo, R.. Using transfer learning for smart building management system. *Journal of Big Data* 2019;**6**(1):110.
37. Maaten, L.v.d., Hinton, G.. Visualizing data using t-sne. *Journal of machine learning research* 2008;**9**(Nov):2579–2605.