Performance of genetic algorithm optimized Doc2Vec-kNN for classifying space science and adjacent fields documents with heterogeneous sampling

Dominic P. Guaña,[1] Arcy Layne L. Sace,[1] Paul Leonard Atchong C. Hilario,*[1]
Efren G. Gumayan,[2] and Gay Jane P. Perez[1,3]

[1] *Philippine Space Agency, Bagumbayan, Quezon City*
[2] *Iloilo Science and Technology University, Iloilo City*
[3] *Institute of Environment Science and Meteorology, University of the Philippines Diliman, Quezon City*
*Corresponding author: atchong.hilario@philsa.gov.ph

Abstract

We assessed the performance of k-nearest neighbor classification on documents consisting of heterogenous class sample. Input data used publicly available titles, abstract, and field of studies of space science and adjacent field articles. Genetic algorithm (GA) was used for hyperparameter tuning, while Doc2Vec was used to transform text into vectors. Results showed that GA optimized Doc2Vec-kNN algorithm performed very well as it can correctly predict the test data >92% on average for all classes. The calculation of the confusion matrix also supported this finding. However, some selected classes performed below 80% due to lower recall and F1-scores.

Keyword: document classification, k-nearest neighbor, doc2vec, machine learning, space science

1 Introduction

Space activities have greatly increased in the recent years, not only in terms of the number of launches, but in terms of number of players as well [1]. These provide opportunities for new space entrants like the Philippines in the fields of space exploration, space observation, and space environment utilization. While new to the field, the Philippines has an endemic pool of expertise, that, although are not directed towards outer space itself, may still be directed towards it, these are called the space adjacent. A review of the scientific works in the recent years will help identify the active space adjacent fields in the country.

Classifying scientific works is a valuable method for reviewing research activities in a certain field. Given the large volume of scientific works, it becomes difficult to manually group them [2, 3] into relevant thematic areas. K-nearest neighbor (kNN) is a simple, yet one of the most popular [4] machine learning classifiers that is used for a wide variety of applications such as disease prediction [5], musical instrument detection [6], and image recognition [7]. It has also been employed in military [8], economics [9], and psychology studies [10]. When paired with a text vectorization, kNN can become a valuable tool for classifying and categorizing text. The use of term frequency-inverse document frequency (TF-IDF) with kNN demonstrated good results irrespective of *k* value, while the use of word2vec with kNN was shown to provide excellent accuracy in topic classification [11,12].

Doc2Vec [13] is another text vectorization algorithm that uses neural networks to transform text. The interest in its use arises from its capability to preserve document semantics when transformed into the desired vector space dimension. However, there is a challenge in determining the appropriate hyperparameters for Doc2Vec-kNN model as this can become time consuming and compute intensive using brute force method arising from the high number of possible permutations. Genetic algorithm [14] is an attractive search algorithm due to its capability to evolve, mutate, or improve individual hyperparameters as search continues. It can achieve optimal hyperparameters faster and efficiently. So far, there is very limited information on the use of genetic algorithm optimized Doc2Vec-kNN model for text classification, thus the merit of this study.

In this paper, we analyzed the performance of document classification using doc2vec and k-nearest neighbor on heterogenous class sample from space science and adjacent fields. Part 1 constitutes the background of the study; Part 2 discusses the dataset; Part 3 presents the algorithm and result; Part 4 is the conclusion, and Part 5 gives the recommendations.

2 Dataset

An initial set of title, abstract, and field of studies (FoS) of 397 scientific articles underwent expert classification validation (Table 1), with each heterogeneous sized-class having an average 33 vetted data. The initial dataset was used as the permanent training subset for all other machine training. We mined another dataset consisting of 10,593 sets of title, abstract, and field of studies of scientific articles. The data miner assigned the classes of the second dataset to the same classes as in Table 1 (unvetted data), with each heterogeneous sized-class having an average of 883 unvetted data. Stop word cleaned-concatenated title, abstract, and FoS of articles served as input data.

Table 1: Distribution of documents for each class.

Classification	**Number of Documents**	
	Expert vetted data	Unvetted data
Astronomy & astrophysics	20	825
Computation, modelling, simulations, & artificial intelligence	65	863
Energy and power generation	42	920
Instrumentations & robotics	30	1,423
Manufacturing, fabrication & rapid prototyping	39	913
Materials, mining & metallurgy	25	762
Medicine & life support	36	870
Optics & photonics	41	883
Physics, biology, chemistry & mathematics	30	652
Security & protection/space situational awareness	11	894
Space colonization & habitat	20	733
Transportation and communications	38	855
Total	397	10,593

3 Performance of the machine learning algorithm

3.1 Machine learning algorithm

Figure 1 shows the flow chart of the machine learning algorithm. In this work we used the Python implementation of the Doc2Vec from the Gensim library [15] for text vectorization. We also used the scikit-learn to implement kNN with Euclidean weight distance [16]. We performed a genetic algorithm (GA) hyperparameterization using DEAP [17] to obtain the optimized parameter for machine learning with expert vetted data as inputs. After the parameter optimization, we then performed machine classification of the unvetted data. Only those that were classified by the kNN and the data miner under the same class were used as machine vetted documents. The machine vetted documents were split into test (10%) and train data (90%). We then merged the train data and the expert vetted data to generate the new kNN model. We then verified the new model with the test data.

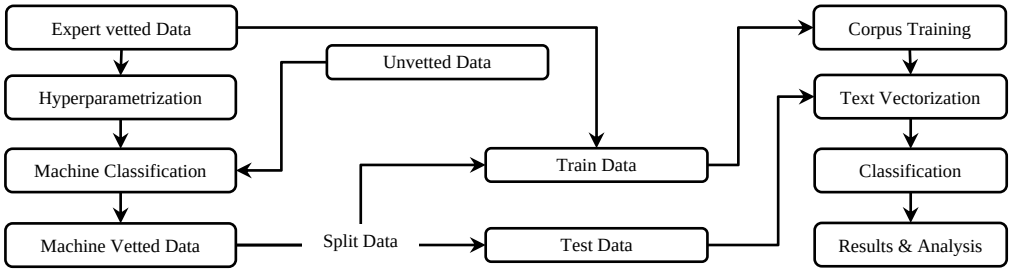

Figure 1: Schematic diagram of the algorithm showing the process of obtaining the machine vetted data and the iterative k-nearest neighbor algorithm.

3.2 Optimization

A representative population fitness for each generation is shown in Figure 2. The fitness of a given individual was obtained by calculating the average accuracy of the Doc2Vec-kNN model for each random selected test data on 12 trials. We observed that the mean accuracy of the population generally increases at early steps, then flattens near the end generation. A similar behavior is also observed for the maximum fitness of the population. The trend of the plot suggests that the genetic algorithm provides optimal improvement on the accuracy of the Doc2Vec-kNN classification of scientific papers across 12 categories.

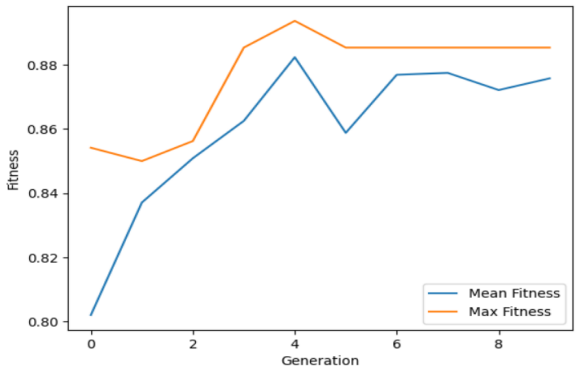

Figure 2: Representative fitness of population over generation. The primary fitness criteria is the accuracy of Doc2Vec-kNN model.

Among 10,593 unvetted scientific papers, only 5,150 documents, or about 49%, were found to belong to the classification identified by both the data miner and kNN. One class, namely Security & protection/space situational awareness, lacks the appropriate number of documents and was not considered for accuracy assessment.

3.3 Doc2Vec-kNN accuracy

Twelve trials were performed to assess the performance of the Doc2Vec-kNN model (Figure 3). The model achieves an accuracy of more than 92% for both seen and unseen corpus. The model also has less tendency to misclassify documents as evident in the confusion matrix. Inspection of the performance of each class (Table 2) supported the results observed in Figure 3. However, we note that two classes namely Materials, Mining & Metallurgy, and Physics, Biology, Chemistry & Mathematics performed below 80% due to either lower recall or F1-scores. A possible source of this could be their fewer test samples. But another fewer sample test from Space Colonization & Habitat suggests otherwise. Investigation of this behavior could be of interest but is not covered in the present work.

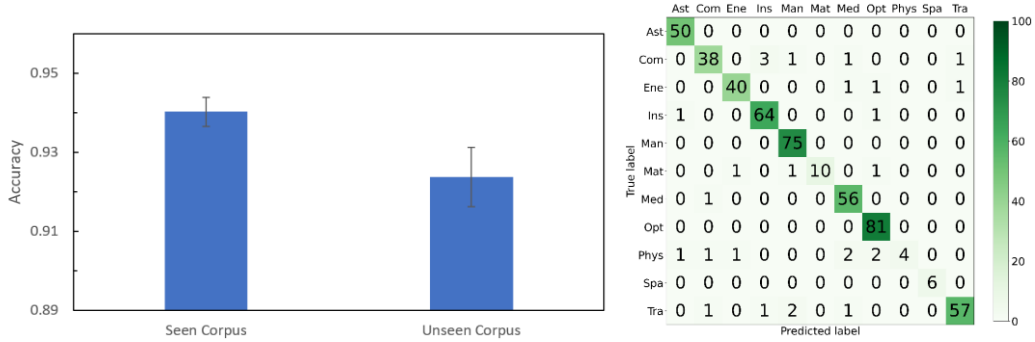

Figure 3: Left image is the plot of model accuracy for seen and unseen corpus. The right image is the plot of confusion matrix of unseen corpus.

Table 2: Precision, recall and F1-score of each class for unseen corpus. Value closest to 1.00 is better.

Classification	Precision	Recall	F1-Score
Astronomy & astrophysics	0.94	0.99	0.97
Computation, modelling, simulations, & artificial intelligence	0.91	0.84	0.87
Energy and power generation	0.93	0.93	0.93
Instrumentations & robotics	0.93	0.96	0.95
Manufacturing, fabrication & rapid prototyping	0.94	0.96	0.95
Materials, mining & metallurgy	0.93	0.70	0.80
Medicine & life support	0.90	0.98	0.94
Optics & photonics	0.93	0.99	0.96
Physics, biology, chemistry & mathematics	0.86	0.39	0.50
Space colonization & habitat	1.00	0.80	0.87
Transportation and communications	0.95	0.92	0.94

4 Conclusions and Recommendations

This work assessed the performance of GA optimized Doc2Vec-kNN classification on documents consisting of heterogeneous class samples of space science and adjacent fields. Results showed that the algorithm performed very well as it can correctly predict the test data >92% on average for all categories. The hyperparameter optimization,

genetic algorithm and machine vetting provides an optimal improvement on the accuracy of the machine classification. The scope of the effect of heterogeneous vs homogeneous data is undetermined.

We suggest increasing high quality datasets for the discarded and lower performing classes to better assess the performance of the model. We also suggest studies for utilizing more than twelve (12) homogenous class size with Doc2Vec-kNN. Investigating the quantitative and qualitative effects of reducing the number of categories may also offer value. Comparison of results with other machine classification methods is also recommended.

Acknowledgement

Bing chat was used to generate code snippets.

References

[1] F. Del Canto Viterale, Transitioning to a new space age in the 21st Century: A systemic-level approach, *Systems* **11**, 232 (2023). URL: https://doi.org/10.3390/systems11050232

[2] K. Okamura, Interdisciplinarity revisited: evidence for research impact and dynamism, *Palgrave Commun.* **5**, 141 (2019). URL: https://doi.org/10.1057/s41599-019-0352-4

[3] L. Codignola, K. Schrogl, A. Lukaszczyk, and N. Peter, *Humans in Outer Space — Interdisciplinary Odysseys* (Springer, Vienna, 2009).

[4] N. Ali, D. Neagu, and P. Trundle, Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets, *SN Appl. Sci.* **1**, 1559 (2019). URL: https://doi.org/10.1007/s42452-019-1356-9

[5] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction, *Sci. Rep.* **12**, 6256 (2022). URL: https://doi.org/10.1038/s41598-022-10358-x

[6] F. B. Selerio and P. L. Hilario, Frequency components analysis of piano and violin 4th octave notes for musical instrument detection," in *Proceedings of the 33rd Samahang Pisika ng Pilipinas Physics Congress* (Vigan, 2015), SPP-2015-PA-25. URL: https://proceedings.spp-online.org/article/view/1181

[7] T. K. Hazra, D. P. Singh, and N. Daga, Optical character recognition using KNN on custom image dataset, in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (Bangkok, 2017), 110−114. URL: https://doi.org/10.1109/IEMECON.2017.8079572

[8] E. Fix and J. L. Hodges, Discriminatory analysis−Nonparametric discrimination: Consistency properties, *Int. Stat. Rev.* **57**, 238 (1989). URL: https://doi.org/10.2307/1403797

[9] B. Priambodo, et al., Predicting GDP of Indonesia using k-nearest neighbour regression, *J. Phys. Conf. Ser.* **1339**, 012040 (2019). URL: https://doi.org/10.1088/1742-6596/1339/1/012040

[10] Z. Taha, R. M. Musa, A. P. P. Abdul Majeed, M. R. Abdullah, M. M. Alim, and A. F. A. Nasir, The application of k-Nearest Neighbour in the identification of high potential archers based on relative psychological coping skills variables, *IOP Conf. Ser. Mater. Sci. Eng.* **342**, 012019 (2018). URL: https://doi.org/10.1088/1757-899X/342/1/012019

[11] B. Trstenjak, S. Mikac, and D. Donko, KNN with TF-IDF based framework for text categorization, *Procedia Eng.* **69**, 1356 (2014). URL: https://doi.org/10.1016/j.proeng.2014.03.129

[12] N. G. Ramadhan, Indonesian online news topics classification using Word2Vec and K-Nearest Neighbor, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* **5**, 1083 (2021). URL: https://doi.org/10.29207/resti.v5i6.3547

[13] J. H. Lau and T. Baldwin, An empirical evaluation of doc2vec with practical insights into document embedding generation, in *Proceedings of the 1st Workshop on Representation Learning for NLP* (ACL, Berlin, 2016), 78−86. URL: https://doi.org/10.18653/v1/W16-1609

[14] S. Katoch, S. S. Chauhan, and V. Kumar, A review on genetic algorithm: past, present, and future, *Multimed. Tools and Appl.* **80**, 8091 (2021). URL: https://doi.org/10.1007/s11042-020-10139-6

[15] R. Řehůřek and P. Sojka, Software framework for topic modelling with large corpora, in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (Malta, 2019), 45−50. URL: http://is.muni.cz/publication/884893/en

[16] F. Pedregosa, et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**, 2825 (2011).

[17] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, DEAP: Evolutionary algorithms made easy, *J. Mach. Learn. Res.* **13**, 2171 (2012).