Enhancing Real Estate Value Estimation with Data Embedding Approach

Ahmad Danial bin Ahmad Fauzi1, a), Choo-Yee Ting1, b), Goh Hui-Ngo1, c), Albert Quek1, d)

*1 Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia*

*b) Corresponding author: cyting@mmu.edu.my*

*a) 1211100824@student.mmu.edu.my*

*c) hngoh@mmu.edu.my*

*d) quek.albert@mmu.edu.my*

**Abstract.** Automated valuation models for real estate increasingly leverage machine learning yet most rely on engineered numeric features alone. In this study, three different pipelines were compared and evaluated with the same seven regression models that are Random Forest, Decision Tree, XGBoost, Gradient Boosting, LightGBM, Lasso Regression and Ridge Regression on a bilingual Malaysian transaction dataset of 191,360 records after data cleaning process. Pipeline one applies feature selection followed by machine learning. Pipeline two combines those selected features with transformer model ‘paraphrase-multilingual-mpnet-base-v2’ to generate embeddings and then feeds the augmented data into the same regressors. Pipeline three applies embeddings solely using the same transformer model after data cleaning process and then inputs into the same regressors. The evaluations metrics carried out by *R*2, RMSE, MAE and MAPE shows that the first pipeline by Random Forest model yields the most accurate price estimates. The findings in this study show that in a mixed English and Malay dataset, out of the box multilingual embeddings could introduce noise when training the Machine Learning models. The study concluded that disciplined feature choice remains essential for property valuation and future work should incorporate fine-tuning language models on local listings and reducing embedding dimensionality.

# INTRODUCTION

Real estate value estimation, or property valuation, is the process of determining out how much a property is worth. This process is vital for homeowners, buyers, sellers and investors. Accurate predictions are critical in this process because it improves decision making and reduce financial risks for all parties involved [1, 2, 3]. Similarly, Budiman et al. [4] further elaborated that accurately valuating residential property values is a major challenge in real estate, because price determination affects everyone from buyers and sellers to developers and policymakers and underpins market stability and trust. Housing demand in Malaysia remains strong as residential transactions grew by 6.1% in the first half of 2024 to 121,964 units according to JPPH Malaysia [5], showing the importance of data driven valuation.

Valuation typically relies on sales comparison, cost, or income approaches [6, 7]. Sales comparison matches a property to nearby recent sales. The cost approach uses replacement cost less depreciation. The income approach values rental properties by their revenue potential. With the common valuation approaches outlined, machine learning use in real estate is still limited in Malaysia [8]. Machine learning algorithms enable pattern recognition, outcome prediction and information classification by learning from existing data to generalizing to new and unseen data [9, 10, 11, 12]. Other than that, emerging tools like Large Language Models (LLMs) are now being explored in this field. LLMs are powerful artifical intelligence systems that understand and produce human language. LLMs are trained on very large datasets that allows them to understand the patterns and structure of language without manual labelling [13, 14]. Chen et al. [15] compare traditional machine learning to LLMs specifically using Chat Generative Pre-Trained Transformer (ChatGPT) in a few-shot forecasting setup. The author’s findings highlight that LLMs have the potential to outperform conventional machine learning models when provided with refined and relevant training data.

# RELATED WORK

Prior to 2025, several Malaysian studies have applied machine learning to residential property price prediction in Malaysia. Rahman et al. [8] evaluated multiple regression, ridge regression (RR), LightGBM (LGBM) and XGBoost (XGB) on a properties dataset focusing on Kuala Lumpur and found that XGB consistently achieved the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) along with the highest adjusted R-squared (*R*2). The author suggested that future work should incorporate house size, amenity proximity and wider location coverage. Lim et al. [16] compared Decision Tree (DT), Linear Regression (LR) and Random Forest (RF) on data from the Malaysian Valuation and Property Services Department and reported that RF delivered the best overall accuracy with the lowest RMSE and MAE. Constrained by limited open datasets, the authors recommended adding geographic area, available facilities, crime rates, property age and resident income in future studies. More recently, Oh et al. [17] applied DT, RF and multilayer perceptron models with principal component analysis (PCA) applied for dimensionality reduction to enhance model performance to terrace-house prices in Kuala Lumpur and demonstrated that RF with PCA achieved the highest *R*2 of 0.74 and the smallest MAE and RMSE.

Other than typical data cleaning and encoding steps that precede predictive tasks, recent research in other domains has explored converting structured data into continuous vector representations using a Transformer based on BERT, and then feeding those vectors directly into downstream models such as XGBoost or Random Forest. Embedding begins by converting each table cell into text, concatenating the row into a sentence, and mapping that sentence through a transformer to produce a vector. Do et al. [18] applied this pipeline to corrosion rate prediction by embedding each row of sentence with DistilBERT and feeding the vectors into an XGB model. Law et al. [19] used multilingual BERT to embed each student feature in an employability dataset, then applied Boruta feature selection before applying to multiple machine learning models such as Logistic Regression, Gaussian Naïve Bayes, RF, Adaptive Boosting, XGB, Categorical Boosting, and LGBM.

Bringing these advances back to the real estate field, researchers are now seeing the same advantages with the advancement of LLM. Chen et al. [15] implemented ChatGPT in 0-shot, 1-shot, 5-shot and 10-shot setups on rental price forecasting and found that the 10-shot setup outperformed all traditional regression models on R squared met- rics. The author further highlights LLMs’ promise for more versatile prediction workflows, since LLMs can ingest unstructured text alongside structured inputs. Moreover, Bittencourt et al. [20] leveraged textual descriptions to im- prove house price prediction tasks. The author compared six text extraction methods including binary bag-of-words (Binary), bag-of-words count (Count), TF-IDF, Binary TF-IDF, word embeddings (FastText), and BERT-based mod- els on listing descriptions. The findings showed that models with text features beat a no text baseline on MAE, RMSE and R squared. These results confirm that LLMs can unlock predictive gains beyond standard machine learning.

Despite these innovations, insights from other domains reveal the hurdles of bilingual inputs. Perera et al. [21] enumerate issues such as lack of formal grammar, spelling variations, creative spelling, undetermined mixing rules, noise, and nonstandard abbreviations, all of which inject noise into tokenization and vector-space feature extraction. Simran et al. [22] report that while off-the-shelf multilingual transformer models outperform cross-lingual embed- dings initially, fine-tuning on synthetic code-switched data is required to overcome translation noise and embedding misalignment. Vivek et al. [23] show that standard code-mixing complexity metrics fail when language mixing is inconsistent or unbalanced, risking overfitting to spurious language artifacts.

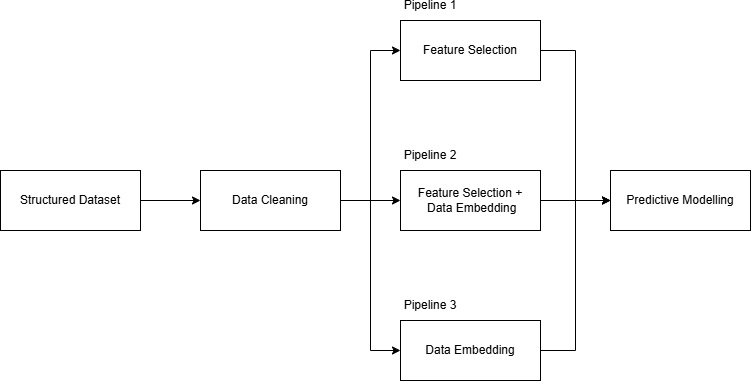
Building on these prior efforts, this study differs from previous work by first enriching this bilingual dataset with point-of-interest count columns that captures how many businesses, landmarks, and transportations that lies within each property’s vicinity [24]. Transformer-based embeddings is then applied both with and without prior feature selection. When these amenity counts are combined with the embeddings, the models are able to capture complex relationships better and may achieve higher accuracy than those relying only on traditional property attributes and simple geographic measures.

# METHOD

This section presents the methods used to develop the regression model. It provides an in-depth overview of the framework including the datasets used, processes and the evaluation metrics employed to gauge the model’s accuracy across three different pipelines. Figure 1 illustrates the overall research framework, showing how the structured dataset was processed through data cleaning and then passed across three different pipelines.

## **Data Preprocessing**

The dataset used in this project is provided by one of the companies doing AVM. It consists of 221,273 entries and 35 attributes ranging from year 2021 to 2023. Table 1 indicates the types and description of each variable in the original dataset. The dataset is then enriched with counts of local amenities that is point-of-interest data for each area, providing additional location-based features such as the number of healthcare, retail, businesses, and transport facilities.



**FIGURE 1.** Research framework

## **Data Cleaning**

Data cleaning steps firstly include handling missing values for *mukim* column. The records of data after dropping empty values of *mukim* column were reduced from 221,273 entries to 199,443 entries. Then, quantile-based filtering or quantile-based trimming was applied which trims data points that fall below 1%th percentile and above 99% percentile to remove the extreme values of *updated\_landarea*, *updated\_build\_up* and *consider\_price*. 8054 entries of extreme outliers were removed. After that, there were still 28 values of missing data for *updated\_locality* and it was dropped. The records of data is now 191,360 entries. Lastly, in the dataset, there were updated versions of certain column. If a column in the dataset has an updated version, only the updated version will be included.

## **Feature Selection**

The categorical attributes were formatted to ensure only alphanumeric characters are left. For instance, special characters were replaced to underscore if it is present. The categorical column were encoded with label encoding because there were some categorical column that has high amount of unique values. Then, the Label encoded categor- ical columns underwent ANOVA F-Score, Mutual Information Score and Kendall Correlation to see if any features have an impact or predictive power to the target variable that is *consider\_price*. Correlation matrix was used for the numerical variable to see the correlation between the numerical columns and the target variable *consider\_price*. Then, *land\_to\_build\_ratio* column was made by dividing the column of *updated\_landarea* and *updated\_build\_up*. Afterwards, the models were trained, using the three best performing models, the models were retrained to get rank features by importance and the least important features were dropped.

## **Data Embedding**

Each row is first serialized into a single natural-language string by formatting numeric area with units columns as “feature (sqm) is value”, point-of-interest data that enriched the dataset which is the amenity columns as “feature count is value”, and all other columns as “feature is value”. These formatted phrases are then joined together using a delimiter. These row-sentences are then fed to the SentenceTransformer model “paraphrase-multilingual-mpnet- base-v2” which maps each sentence into a 768-dimensional vector. In this way we obtain a continuous embedding for every original record, ready to serve as input for downstream predictive models. This model was selected because the dataset contains mixed-language content, including both English and Bahasa Melayu, and it is explicitly designed to support over fifty languages. Table 2 illustrates the data embedding process used in this study.

**TABLE 1.** Dataset description

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| id | Int64 | Unique identifier for each record. |
| state | Object | The state in which the property is located. |
| district | Object | The district within the state where the property is found. |
| mukim | Object | A local administrative area within the district. |
| locality | Object | The locality or neighborhood of the property. |
| updated\_locality | Object | Updated information on the locality. |
| address | Object | The specific address of the property. |
| updated\_address | Object | Updated version of the property address. |
| lot\_type | Object | Type of lot classification. |
| updated\_lot\_type | Object | Updated lot classification. |
| lot\_no | Object | Lot number of the property. |
| status | Object | Status of the property transaction. |
| transaction\_date | Object | Date when the transaction was recorded. |
| quarter\_period | Int64 | Quarter of the year when the transaction took place. |
| transaction\_date\_year | Int64 | Year of the transaction. |
| transaction\_date\_month | Int64 | Month of the transaction. |
| transaction\_date\_day | Int64 | Day of the transaction. |
| landarea | Float64 | Area of the land in the specified unit. |
| updated\_landarea | Float64 | Updated area of the land. |
| landarea\_unit | Object | Unit of land area measurement. |
| title\_type | Object | Type of property title. |
| updated\_title\_type | Object | Updated property title type. |
| title\_no | Object | Title number. |
| sector | Object | Sector classification of the property. |
| tenure | Object | Tenure of the property. |
| updated\_tenure | Object | Updated tenure information. |
| leasehold\_period | Float64 | Duration of the leasehold. |
| updated\_leasehold\_period | Int64 | Updated duration of the leasehold. |
| property\_type | Object | Type of property. |
| updated\_property\_type | Object | Updated type of property. |
| floor\_above | Float64 | Number of floors above ground. |
| floor\_below | Float64 | Number of floors below ground. |
| updated\_build\_up | Float64 | Updated build-up area of the property in square meters. |
| consider\_price | Float64 | Consideration price of the property. |
| consider\_price\_group | Object | Classification of the price. |
| Automotive | Int64 | Count of nearby Automotive of the property. |
| Businesses and Services | Int64 | Count of nearby Businesses and Services of the property. |
| Community and Government | Int64 | Count of nearby Community and Government of the property. |
| Healthcare | Int64 | Count of nearby Healthcare of the property. |
| Landmarks | Int64 | Count of nearby Landmarks of the property. |
| Retail | Int64 | Count of nearby Retail of the property. |
| Social | Int64 | Count of nearby Social of the property. |
| Sports and Recreation | Int64 | Count of nearby Sports and Recreation of the property. |
| Transportation | Int64 | Count of nearby Transportation of the property. |
| Travel | Int64 | Count of nearby Travel-related amenities of the property. |

**TABLE 2.** Example of data embedding

|  |  |  |  |
| --- | --- | --- | --- |
| Numeric Area with Units Column | Amenity Column | Standard Column | Vector |
| updated\_landarea (sq m) is 208.1 | Transportation count is 12 | state is johor | -2.34583337e-02 -1.96969733e-01... |
| updated\_build\_up (sq m) is 161.0 | Healthcare count is 7 | district is johor bahru | 4.43836898e-02 -3.50780077e-02... |

## **Predictive Modeling**

Predictive modeling without embedding begins by cyclically encoding transaction time features to capture temporal patterns, and then normalizing all numerical attributes using StandardScaler. Next, for both the embedding and non-embedding approaches, the dataset is split into two groups, training and testing sample. In the training sample partition, 80% of the data is used while 20% of the dataset is used in the testing sample. The machine learning techniques used in this study are RF, DT, XGB, GB, Lasso Regression (LSR), RR, and LGBM for the regression task. Model performance is evaluated using *R*2, RMSE, MAE, and MAPE. By comparing these metrics across all models on the same data, we can identify the algorithm whose assumptions and capacity best match our problem, and thereby achieve the highest predictive accuracy.

# FINDINGS

This section reports the evaluation of the the seven regression model that are RF, DT, XGB, GB, LSR, RR, and LGBM under three different pipelines. Model performance is compared using RMSE, MAE, *R*2 and MAPE to show how feature selection and embedding strategies influence the accuracy of property price estimation.

Table 3 shows the result across the three pipelines. In the first pipeline, with hand-picked features and classical machine learning, ensemble methods deliver the strongest results. RF achieves an *R*2 of 0.88 with an RMSE of 136474, an MAE of 72686 and a MAPE of 17.16%. XGB follows closely with an *R*2 of 0.87, an RMSE of 144600, an MAE of 86345 and a MAPE of 21.00%. GB and DT also outperform the linear regressions, which the linear regressions reach *R*2 of 0.51 and MAPE of 48.60%, showing that non-linear ensembles extract more signal from this curated set of predictors.

**TABLE 3.** Comparison of regression model performance across three pipelines: (1) feature selection + machine learning, (2) fea- ture selection + embedding + machine learning, (3) embedding + machine learning. FS = Feature Selection, EMB = Embedding, ML = Machine Learning

FS + ML FS + EMB + ML EMB + ML

Model *R*2 RMSE MAE MAPE (%) *R*2 RMSE MAE MAPE (%) *R*2 RMSE MAE MAPE (%)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RF | **0.88** | **136474** | **72686** | **17.16** | 0.81 | 174291 | 93232 | 23.16 | 0.71 | 219345 | 127078 | 33.81 |
| DT | 0.81 | 176131 | 89451 | 20.11 | 0.60 | 257443 | 123548 | 28.37 | 0.29 | 342837 | 183016 | 43.53 |
| XGB | 0.87 | 144600 | 86345 | 21.00 | 0.75 | 201345 | 122739 | 31.48 | 0.64 | 244315 | 153907 | 40.06 |
| GB | 0.76 | 199536 | 124494 | 30.22 | 0.49 | 291274 | 187617 | 50.31 | 0.40 | 315203 | 208711 | 59.07 |
| LSR | 0.51 | 282995 | 185166 | 48.60 | 0.59 | 259534 | 171540 | 46.53 | 0.46 | 299282 | 199759 | 53.15 |
| RR | 0.51 | 282995 | 185166 | 48.60 | 0.53 | 278018 | 181631 | 48.77 | 0.41 | 312476 | 207020 | 55.17 |
| LGBM | 0.83 | 163412 | 101278 | 25.27 | 0.67 | 231859 | 146876 | 38.92 | 0.57 | 266149 | 172741 | 47.37 |

When we augment that same feature set with paraphrase-multilingual-mpnet-base-v2 to generate embeddings, every top performing model’s accuracy falls. RF drops to an *R*2 of 0.81, an RMSE of 174291, an MAE of 93232 and a MAPE of 23.16%. XGB reaches an *R*2 of 0.75, an RMSE of 201345, an MAE of 122739 and a MAPE of 31.48%. Interestingly, the linear models such as LSR and RR improve under this setting, with *R*2 increasing from 0.51 to 0.59 and 0.53 respectively, while their errors decrease relative to the first pipeline. However, their performance still falls well behind the top performing models. This suggests that the raw embeddings introduce noise or redundant signals that outweigh any additional information.

Relying solely on embeddings without any feature selection produces the weakest outcomes. RF under this pipeline fell further to *R*2 of 0.71, an RMSE of 219345, an MAE of 127078 and a MAPE of 33.81%. XGB falls to an *R*2 of 0.64 with an RMSE of 244315, an MAE of 153907 and a MAPE of 40.06%. All models underperform their counterparts in the previous pipelines, indicating that high dimensional embeddings alone fail to generalize when left unchecked.

Overall, these findings confirm that careful feature selection remains essential for high accuracy property price estimation. The embeddings generated by paraphrase-multilingual-mpnet-base-v2 may struggle to capture the nuanced relationships in a bilingual dataset or may simply introduce too much redundant information. Future work could investigate fine-tuning that embedding model on domain specific text or applying dimensionality reduction to the embeddings before modeling.

# CONCLUSION

This study presents a comprehensive evaluation of three modeling pipelines to predict property prices with seven regression algorithms. The first pipeline applied feature selection before model training. The second pipeline combined selected features with transformer-based embeddings before model training. The third pipeline relied solely on transformer-based embeddings before model training. On this dataset of 191,360 Malaysian residential transactions, the first pipeline produced the most accurate price estimates. In particular, the RF regressor trained on carefully chosen structural and locational attributes achieved the best *R*2 and the lowest RMSE, MAE, and MAPE values. This confirms that for bilingual property data rigorous selection of numeric and categorical predictors is more effective than directly using generic multilingual embeddings. Future research should explore fine-tuning transformer models on domain specific English and Malay listings and dimensionality reduction for embeddings that could remove redundancy before regression. By combining feature selection with targeted use of text embeddings next generation automated valuation models can reach even higher accuracy in property price estimation.

# REFERENCES

1. N. S. Nidamanuru, K. Rajesh, U. Srinivasarao, K. Koppolu, K. Balavani, and M. V. Kumar, “Property price estimation using machine learning and implementing a website using html, css, and javascript,” in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (2024) pp. 1–5.
2. H. Verma, A. Singh, and S. Avasthi, “Linear regression-based prediction model for real estate properties in india,” in *2023 Seventh International Conference on Image Information Processing (ICIIP)* (2023) pp. 158–163.
3. A. Mukherjee, R. Nambiar, and D. Choudhury, “Machine learning based real estate price prediction,” in *2024 8th International Conference on Inventive Systems and Control (ICISC)* (2024) pp. 351–358.
4. B. Budiman, N. Alamsyah, T. P. Yoga, R. Y. R. Alamsyah, and C. Habibi, “Machine learning-based classification of house prices: A comparative study,” in *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)* (2023) pp. 1–6.
5. JPPH, “Jpph laporan harta tanah,” https://napic2.jpph.gov.my/storage/app/media/3-penerbitan/pasaran-harta-tanah/Laporan-Pasaran-Harta- H1-2024.pdf, accessed: 2025-01-20.
6. S. Sakri, Z. Ali, and N. H. A. Ismail, “Assessment of the hybrid deep learning models and hedonic pricing model for house price prediction,” in *2024 Seventh International Women in Data Science Conference at Prince Sultan University (WiDS PSU)* (2024) pp. 127–133.
7. A. Farkas, and B. Porumb, “A Multi-Attribute Sales Comparison Method for Real Estate Valuation,” Period. Polytech. Soc. Man. Sci. **28**(1), 1–11 (2019).
8. S. Abdul-Rahman, N.H. Zulkifley, I. Ibrahim, and S. Mutalib, “Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur,” IJACSA **12**(12), (2021).
9. S. Harshada, Belsare and K.V. Warkar, “A Novel Model for House Price Prediction with Machine Learning Techniques,” IJSRST, 743–754 (2023).
10. M.I. Jordan, and T.M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” Science **349**(6245), 255–260 (2015).
11. R. Bernardes, “Machine learning ‐ Basic principles,” Acta Ophthalmologica **102**(S279), aos.16281 (2024).
12. Shaveta, “A review on machine learning,” International Journal of Science and Research Archive **9**, 281–285 (2023).
13. L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, and E. Chen, “A survey on large language models for recommendation,” World Wide Web **27**(5), 60 (2024).
14. S. Makridakis, F. Petropoulos, and Y. Kang, “Large language models: Their success and impact,” Forecasting **5**(3), 536–549 (2023).
15. T. Chen and S. Si, “Predicting rental price of lane houses in shanghai with machine learning methods and large language models,” (2024), arXiv:2405.17505 [cs.LG].
16. L.W. Yee, N.A.A. Bakar, N.H. Hassan, N.M.M. Zainuddin, R.C.M. Yusoff, and N.Z.A. Rahim, “Using Machine Learning to Forecast Residential Property Prices in Overcoming the Property Overhang Issue,” in *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, (IEEE, Kota Kinabalu, Malaysia, 2021), pp. 1–6.
17. S.Y. Oh, S.P. Hang, and J.T.W. Wang, “Prediction of residential property prices using machine learning algorithms,” ITM Web Conf. **67**, 01042 (2024).
18. T. Do, B.D. Shrestha Gurung, S. Aryal, A. Khanal, S. Chataut, V. Gadhamshetty, C. Lushbough, and E.Z. Gnimpieba, “Utilizing XGBoost for the Prediction of Material Corrosion Rates from Embedded Tabular Data using Large Language Model,” in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (IEEE, Istanbul, Turkiye, 2023), pp. 4497–4499.
19. T.-J. Law, C.-Y. Ting, H. Ng, H.-N. Goh, and Q. Albert, “A Multilingual BERT Embeddings Approach in Identifying Factors Influencing Employability Among Pre-University Students,” in *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, (IEEE, Kota Kinabalu, Malaysia, 2024), pp. 130–135.
20. L. Bittencourt, O. Parraga, D. Ruiz, I. Manssour, S. Musse, and R. Barros, “Leveraging textual descriptions for house price valuation,” (2022).
21. A. Perera, and A. Caldera, “Sentiment Analysis of Code-Mixed Text: A Comprehensive Review,” Jucs **30**(2), 242–261 (2024).
22. S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, and M. Choudhury, “GLUECoS: An Evaluation Benchmark for Code-Switched NLP,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Association for Computational Linguistics, Online, 2020), pp. 3575–3585.
23. V. Srivastava, and M. Singh, “Challenges and Limitations with the Metrics Measuring the Complexity of Code-Mixed Text,” in *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, (Association for Computational Linguistics, Online, 2021), pp. 6–14.
24. Z.Y. Poo, C.Y. Ting, Y.P. Loh, and K.I. Ghauth, “Multi-Label Classification with Deep Learning for Retail Recommendation,” Journal of Informatics and Web Engineering **2**(2), 218–232 (2023).