Populating Philippine OpenStreetMap Building Type using Large Language Models

Jonathan Christian F. Aceron [1,2], Gilson Andre M. Narciso [1,2], Abdel Jalal D. Sinapilo [1,3], Rose Anne I. Coronado [1], Lorrize Mae L. Guevarra [1,2], Jeromalyn A. Palma [1] , John Harold B. Tabuzo [1]

[1] Department of Science and Technology - Philippine Institute of Volcanology and Seismology (DOST-PHIVOLCS), Philippines - jcaceron.lupa@gmail.com; gmnarciso.lupa@gmail.com; ajsinapilo.lupa@gmail.com; racoronado.lupa@gmail.com; llguevarra.lupa@gmail.com; jeromalyn.palma@phivolcs.dost.gov.ph; harold.tabuzo@phivolcs.dost.gov.ph

[2] Department of Geodetic Engineering, University of the Philippines - Diliman, Philippines
[3] Artificial Intelligence Program, University of the Philippines - Diliman, Philippines

Keywords: street mapping, open-source, BERT, RoBERTa, random forest

Abstract

Accurate building information is essential to a wide range of applications requiring preliminary inputs for humanitarian initiatives, city planning, scientific studies, and navigation systems (Atwal et al., 2022). In the Philippines, which includes the preparation of the Comprehensive Land Use Plan (CLUP), collecting building types can contribute to the development of the demographic maps, exposure maps, and transportation maps, among others. Data collection is an extensive process; thus, community mapping tools such as OpenStreetMap (OSM) present a convenient avenue for collecting descriptive attributes of building types faster. However, the data entry process involves varied tasks beyond basic entry, which results in menial and tedious recording of information that is occasionally inaccurate and lacking. To fill this gap, the study proposes to use Large Language Models (LLMs) trained in Philippine infrastructure contexts. Comparing the LLMS and a classical machine learning model in text classification demonstrates that the classical model performs well on localized features, while LLMs specialize in more complex contextual features. The BERT model, RoBERTa, and the Random Forest model showcased 98.25%, 98.42%, and 92.39% accuracy, respectively, on the testing dataset. This highlights the potential of textual classification using LLMs in urban planning studies.

1. Introduction

The Philippines, although abundant in natural resources, has a rapidly growing population. Proper land use planning is key to ensure that the allocation of resources and urban development is beneficial not just for the current residents but for the future population as well. In land use planning, proper classification of buildings and structures ensures that studies would generate accurate results and corresponding decisions are well-informed. Buildings and structures could be classified in a multitude of themes: socio-economic use, structural, height, etc.

Knowing the building types could lead to various research topics such as inferring population, modelling urban mobility, site suitability analysis, and planning emergency routes. In developing the Comprehensive Land Use Plan (CLUP), knowing the building types could be used in a variety of preliminary inputs, such as demographic maps, exposure maps, and transportation maps. Most of the information gathering about the structures in an area is the responsibility of local government units and their constituents. During the CLUP writing process, a significant amount of time is spent in collecting building information since it is not easily derived from remote sensing images.

Community mapping has the potential to aid in such endeavours and enable the creation of richer information databases since everyone could contribute. An example of a community mapping tool is OpenStreetMap (OSM), an open geospatial database (Ribeiro et al., 2015) with thousands of volunteers around the world. An example is last 2024 where volunteers were able to map buildings affected by the Kanlaon Eruption through OSM (Humanitarian OpenStreetMap Team, 2024). Despite the numerous efforts of both community mappers and LGUs, collecting these types of information are menial and tedious tasks, and information is sometimes inaccurate and lacking. Because of that, the use of artificial intelligence (AI) to speed up the process is welcome.

Large Language Models (LLMs) are text-based AI that can understand and interpret human language. LLMs can be used in a variety of tasks like translation, text generation, summarization, and text classification (Sun et al., 2020). As the name implies, LLMs are trained on a huge amount of data by various corporations. There are various LLMs free for public use, however, most of them have subscription fees during fine-tuning. This study used the Bidirectional Encoder Representations from Transformers (BERT) model developed by Google and the Robustly Optimized BERT Approach (RoBERTa) model developed by Meta AI which builds on BERT. BERT and RoBERTa are free to train and fine-tune to perform text classification. In text classification using LLMs, a specific phrase or prompt is given a class, and the model tries to predict what class the prompt belongs to. This study aims to use BERT and RoBERTa to analyze how buildings are named in the Philippines, using the building type as the class and the building name as the prompt. BERT and RoBERTa will be fine-tuned using OSM datasets and infer the building types from their names.

Moreover, this study also employs the Random Forest model for text classification. Random Forest is a classical machine learning model that uses the ensemble method, which is a method that combines numerous predictions to achieve a grand result. In the case of Random Forest, this is done through multiple decision trees that arrive at a determined classification.

2. Materials and Methods

2.1 Study Area

The area of interest of the study is the Metro Manila region in the Philippines. The region was selected due to the abundance of OSM buildings in the area. It also has the greatest variety in building types and names.

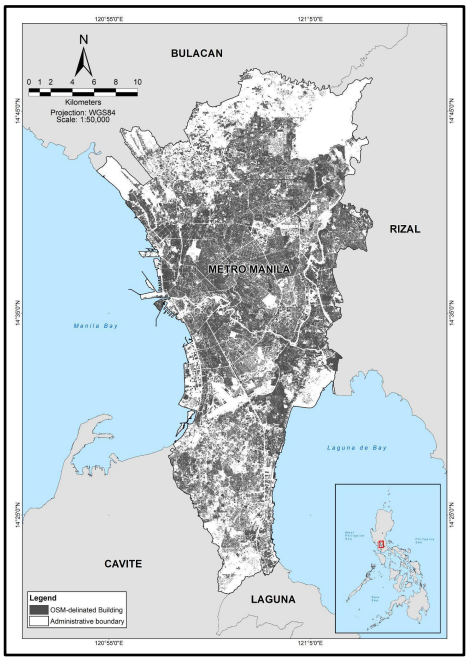

Figure 1. OSM delineated buildings in Metro Manila.

2.2 Methodology

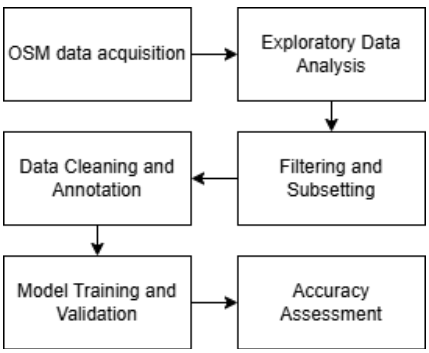

Figure 2. Overall workflow for the research.

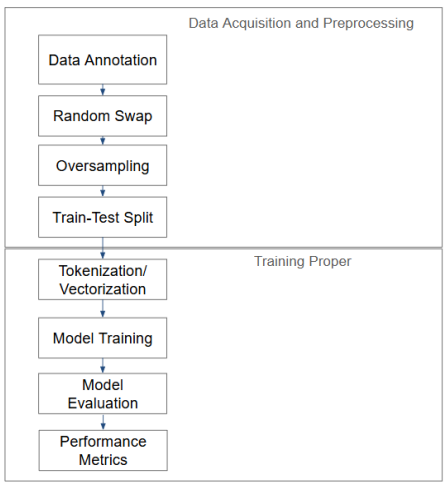

Figure 3. Data preparation and training workflow.

2.2.1 Data Acquisition: Current OSM data as of May 23, 2025, was downloaded for the whole Philippines. Comparing the point dataset with the polygon dataset, the polygons offer a more complete coverage, therefore the building polygons were used. Since it was very heavy for processing, the data for Metro Manila was subsetted. The OSM buildings dataset contains the fields *'osm_id'*, *'code'*, *'fclass'*, '*name*', and '*type*'. The dataset was filtered for areas with a non-empty *'name'* field. It is important for the building to have a name since that is the basis for deriving the building type. For the initial model, the training and validation data are composed of features with both name and type.

The model is implemented through PyTorch, a Python-based, open-source machine learning framework. Using the transformers library, the BERT model is loaded with pretrained weights. Since the geospatial information is not needed in BERT, we transformed the OSM buildings data to csv with only the name and type fields retained. Fine-tuning was then done by feeding the model the filtered dataset. BERT will facilitate the text classification of the *'name'* field into the corresponding class *'type'*.

2.2.2 Exploratory Data Analysis: After exploring the OSM buildings dataset in Metro Manila, it can be seen in Table 1 that only 1.37% of the buildings have filled names. Of these named buildings, only 53.33% have their corresponding building type.

Class	Count
Metro Manila Buildings	1,155,506
Metro Manila Buildings - with *name*	15,802
Metro Manila Buildings - with *name* and *type*	8,427

Table 1. Summary of OSM buildings in Metro Manila.

As seen in Table 2, the filtered OSM dataset contains 87 unique building types. Analyzing each class reveals that the dataset has many noise classes: types that are either redundant or names that have been inputted as types. There are also classes that are too specific and most of them are not thematically aligned, like *'columbarium'*, *'pop-up_store'*, and *'roof'*. The classes are retained during initial training to see if it would be able to be trained properly using those data. After training for 10 epochs, training loss decreased from 1.75 to 1.30 and the accuracy increased from 0.608 to 0.647. The model was able to infer building types from building names with average accuracy. The accuracy could be attributed to the improper annotation of building types in the OSM dataset, and further testing was done to see if cleaning the annotation and trying different building classification schemes will yield better results.

2.2.3 Data Preparation: Given that the original OSM dataset contains numerous repetitive and noise classes, the building types were reduced to 15 based on land use. Data points that were unclassifiable based on name alone were labeled 'unclassifiable' while the rest were annotated loosely based on the classes from CLUP Guidebook Volume 1 by the Housing and Land Use Regulatory Board (2013). Some land use types were subclassified, leading into 15 building types, with counts given in Table 3.

Unique building types (MM bldgs w/ name and type)			
church	apartments	hangar	pop-up_store
retail	cathedral	townhall	guardhouse
residential	supermarket	terrace	shed
commercial	chapel	storage	hall
public	roof	religious	manufacture
civic	logistics	multipurpose	Subdivision_ Administ
government	train_station	accomodation	historic
dormitory	condominium	fire_station	clinic
hospital	synagogue	abandoned	bridge
school	garages	medical	sports_hall
college	retain	place of worship	kingdom_hall
university	columbarium	temple	leisure
office	service	institutional	bank
hotel	house	marquee	pharmacy
parking	grandstand	Philippine_ Nurses_As	healthcare
pavilion	garage	Condominium	water station
industrial	monastery	Baicapture HQ	brgy_hall
warehouse	ruins	kindergarten	bakehouse
construction	sports_centre	detached	building
transportation	clubhouse	semidetached_ house	government_ office
Recreational_ Building	Bagumbayan Riverboat	kiosk	wall
security_booth	bakery	museum	

Table 2. Unique building types in Metro Manila.

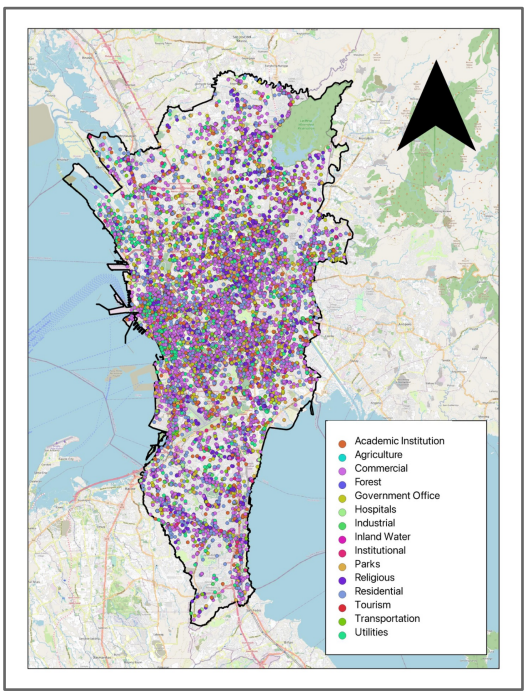

Figure 4. Training Data distribution

Class	Count
Commercial	4,818
Residential	2,096
Government Office	1,139
Religious	1,125
Academic Institution	850
Utilities	611
Parks	599
Tourism	462
Hospitals	461
Institutional	310
Industrial	275
Transportation	216
Inland Water	5
Agriculture	4
Forest	3

Table 3. Final building types and counts of the training data.

.

Since the dataset was reduced due to the removal of 'unclassifiable' points, data augmentation was performed through Random Swap, where randomly selected two words in each data point were swapped, shown in Figure 5. The original and random swap datasets were then concatenated to double the number of data points.

	original name	random swap	category
133	? Barangay Health Center	Barangay ? Health Center	Hospitals
137	(CLOSED) Dela Rosa Carpark 2	Carpark Dela Rosa (CLOSED) 2	Commercial
143	10 Acacia Place	10 Place Acacia	Residential
145	10 West Campus	West 10 Campus	Academic Institution
147	1001 Parkway Residences	Parkway 1001 Residences	Residential
...	...	...	...
15807	Zu Body Foot Spa	Foot Body Zu Spa	Commercial
15808	Zuellig Building	Building Zuellig	Commercial
15809	Zulueta Animal Clinic	Zulueta Clinic Animal	Commercial
15810	Zus Coffee	Zus Coffee	Commercial
15811	Zytek Chapel	Chapel Zytek	Religious

Figure 5. Comparison of original address name and new address name through Random Swap.

The dataset exhibits extreme abnormal distribution and class imbalance. Oversampling was implemented to equalize the value counts of each building type, summarized in Figure 6.

The final oversampled dataset was then split between training and testing set, with 20% validation set.

2.2.4 Model Training: Three distinct machine learning models were employed for text classification: a pre-trained BERT LLM, pre-trained RoBERTa LLM, and a classical Random Forest Classifier. The dataset was consistently preprocessed by extracting text and labels and then split into training, validation, and test sets.

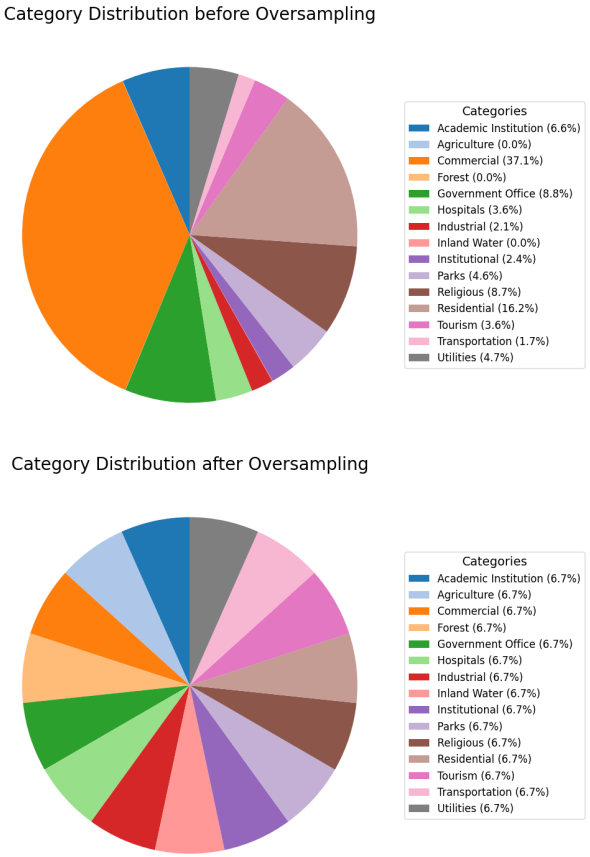

Figure 6. Class distribution before and after Oversampling.

For the BERT and RoBERTa LLMs, the pre-trained BERT and RoBERTa models from the Hugging Face Transformers library were used. The data was tokenized using the AutoTokenizer function. To reduce computational cost, only the classifier, embeddings, pooler, dense, and LayerNorm layers of BERT, and the embeddings, intermediate, self, and classifier layers of RoBERTa were unfrozen for fine-tuning. A hyperparameter greedy search was performed on both LLMs using the Optuna library to maximize the performance metrics on the validation set, yielding optimal parameters with a learning rate of 2.27884e-05, batch size of 32, 8 epochs, weight decay of 0.0010, and warm-up ratio of 0.1333 for BERT, then a learning rate of 9.1068e-06, batch size of 8, 6 epochs, weight decay of 0.0919, and warm-up ratio of 0.1731 for RoBERTa.

For the Random Forest model, the RandomForestClassifier from the Scikit-learn library was utilized. The text data was first converted into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency). This vectorized data was then used to train the classifier with 100 trees and a random state of 42.

3. Results and Discussion

The BERT training and validation losses and accuracies for 6 epochs are given in Figure 7, with its test set evaluation given in Figure 8, showing a 98.25% prediction accuracy.

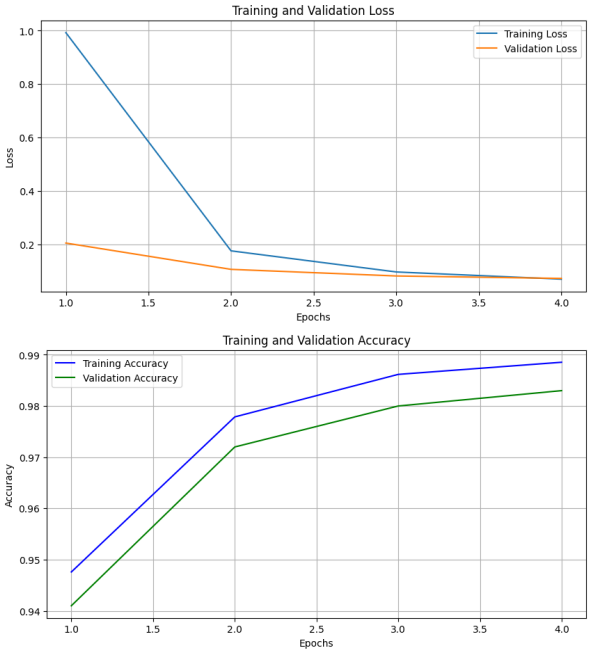

Figure 7. BERT training and validation losses and accuracies.

Comparison of True vs. Predicted Building Types (Test Set):

	Text	True_Building_Type	Predicted_Building_Type
0	Sports Elorde Complex	Parks	Parks
1	SCS Agro	Agriculture	Agriculture
2	25 ML Balara Reservoir	Inland Water	Inland Water
3	Onse Covered Court	Parks	Parks
4	Land Bank of the Philippines	Government Office	Government Office
...	...	...	...
28903	General Malvar Hospital	Hospitals	Hospitals
28904	Triple V (Kamayan, Dad's, Saisaki)	Commercial	Commercial
28905	DOST-NCR CAMANAVA CASTO	Government Office	Government Office
28906	Senior Office Citizens	Government Office	Government Office
28907	Caltex	Utilities	Utilities

28908 rows × 3 columns

Percentage of correct predictions:98.25307873253078%

Figure 8. BERT test set evaluation.

The fine-tuned BERT model exhibited a strong performance, indicated by the evaluation metrics, confusion matrix, and ROC-AUC curve in Table 4 and Figures 9 and 10, respectively.

The evaluation metrics have relatively comparable values for all categories, with Inland Water, Agriculture, and Forest classes achieving a perfect score on all metrics, followed by Transportation, Hospitals, Utilities, and Government Office with F1 scores a few decimal values lower, supported by the confusion matrix with high values in the main diagonal and small values in the off-diagonal, showing minimal misclassification. The ROC-AUC curve with consistently high areas under the curve further validated the strong discriminative power of the fine-tuned model to classify across the categories.

Category	Sensitivity	Specificity	Precision	F1
Hospitals	1.000	0.999	0.991	0.996
Commercial	0.829	0.998	0.967	0.893
Residential	0.947	0.996	0.938	0.943
Academic Institution	0.990	0.999	0.986	0.988
Utilities	1.000	0.999	0.988	0.994
Government Office	0.985	0.999	0.983	0.994
Tourism	1.000	0.998	0.976	0.988
Inland Water	1.000	1.000	1.000	1.000
Transportation	1.000	0.999	0.992	0.996
Industrial	1.000	0.998	0.973	0.986
Parks	0.997	0.997	0.962	0.979
Religious	0.989	1.000	0.995	0.992
Institutional	1.000	0.999	0.987	0.993
Agriculture	1.000	1.000	1.000	1.000
Forest	1.000	1.000	1.000	1.000

Table 4. BERT evaluation metrics.

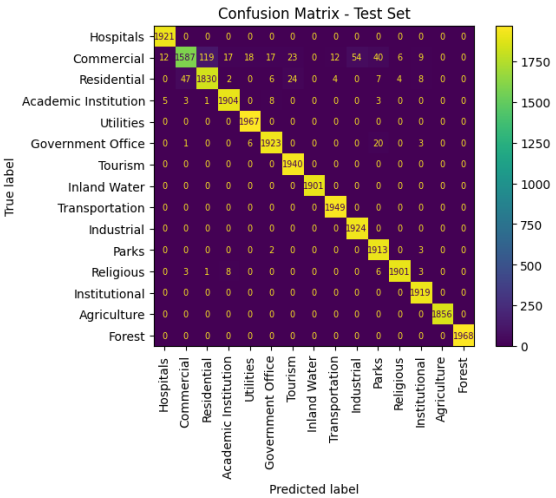

Figure 9. BERT confusion matrix.

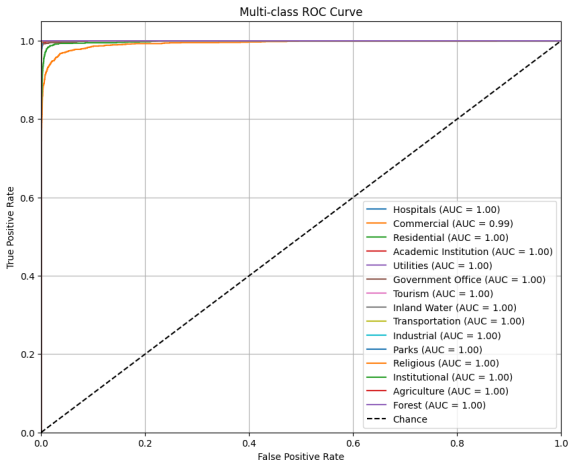

Figure 10. BERT ROC-AUC curve.

Strongly similar to the results of BERT, the RoBERTa training and validation losses and accuracies for 6 epochs are given in Figure 11, with its test set evaluation given in Figure 12, showing a 98.42% prediction accuracy.

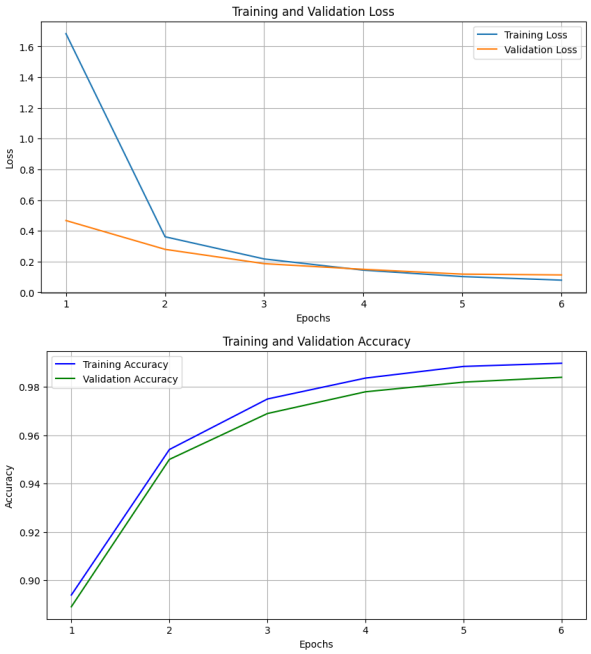

Figure 11. RoBERTa training and validation losses and accuracies.

Comparison of True vs. Predicted Building Types (Test Set):

	Text	True_Building_Type	Predicted_Building_Type
0	PLDT InnoLabs	Utilities	Utilities
1	MDRRMO Staging Area	Government Office	Government Office
2	Lagoon Pavillion	Inland Water	Inland Water
3	ML 19 Balara Reservoir	Inland Water	Inland Water
4	Annex NU Building	Academic Institution	Academic Institution
...	...	...	...
28903	REMAN Hospital Medical and Laboratory Equipment	Hospitals	Hospitals
28904	Creative Arts Center	Academic Institution	Academic Institution
28905	1 Building KAVI	Commercial	Commercial
28906	Foundation BigHands	Institutional	Institutional
28907	Philippine Biodome	Forest	Forest

28908 rows × 3 columns

Percentage of correct predictions:98.41566348415664%

Figure 12. RoBERTa test set evaluation.

The fine-tuned RoBERTa model exhibited a strong performance indicated by the evaluation metrics, confusion matrix, and ROC-AUC curve in Table 5 and Figures 13 and 14 respectively.

The evaluation metrics have relatively comparable values for all categories, with Inland Water and Forest classes achieving a perfect score on all metrics, followed by Agriculture, Transportation, and Hospitals with F1 scores a few decimal values lower, supported by the confusion matrix with high values in the main diagonal and small values in the off-diagonal, showing minimal misclassification. The ROC-AUC curve with consistently high areas under the curve further validated the strong discriminative power of the fine-tuned model to classify across the categories.

Category	Sensitivity	Specificity	Precision	F1
Hospitals	1.000	1.000	0.996	0.998

Commercial	0.840	0.999	0.978	0.904
Residential	0.953	0.995	0.927	0.940
Academic Institution	0.996	0.999	0.983	0.989
Utilities	1.000	1.000	0.993	0.997
Government Office	0.991	0.999	0.989	0.990
Tourism	0.998	0.998	0.975	0.987
Inland Water	1.000	1.000	1.000	1.000
Transportation	1.000	1.000	0.996	0.998
Industrial	1.000	0.998	0.978	0.989
Parks	0.997	0.998	0.966	0.982
Religious	0.992	0.999	0.991	0.992
Institutional	1.000	0.999	0.992	0.996
Agriculture	1.000	1.000	0.998	0.999
Forest	1.000	1.000	1.000	1.000

Table 5. RoBERTa evaluation metrics.

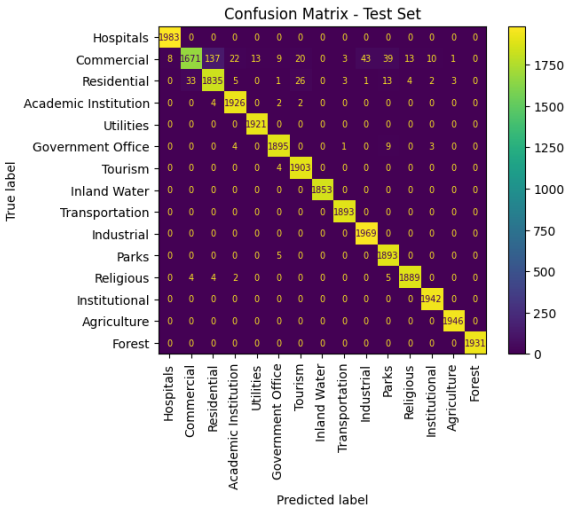

Figure 13. RoBERTa confusion matrix.

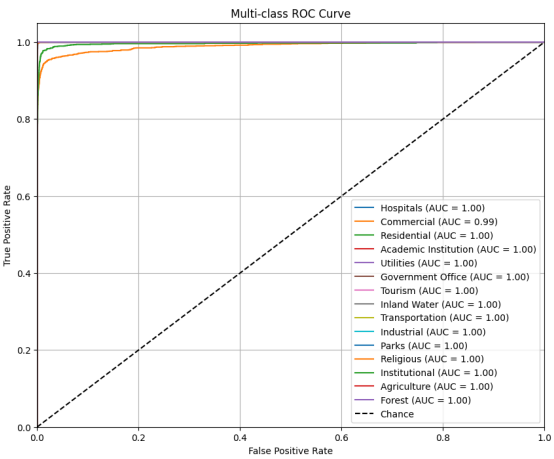

Figure 14. RoBERTa ROC-AUC curve.

The Random Forest model test set evaluation is given in Figure 15, showing a 92.39% prediction accuracy, less than that of the BERT and RoBERTa models.

```
Comparison of True vs. Predicted Categories (Test Set):
                                          Text         True_Category  \
0                        Court Covered Romblon                 Parks
1      Technology of Industrial College Building  Academic Institution
2              El Deposito Underground Reservoir         Inland Water
3                              Avila Polyclinic              Hospitals
4                           Hall Multi-Purpose          Institutional
...                                        ...                   ...
28903               Court Silang Covered Bagong                 Parks
28904                           Westgate Plaza           Residential
28905                       Kambingan ni Tsong            Commercial
28906                                      LBC            Industrial
28907                      Barangay 631 Zone 65    Government Office

         Predicted_Category
0                     Parks
1      Academic Institution
2              Inland Water
3                 Hospitals
4             Institutional
...                     ...
28903                 Parks
28904                 Parks
28905            Commercial
28906            Industrial
28907    Government Office

[28908 rows x 3 columns]
Percentage of correct predictions: 92.39%
```

Figure 15. Random Forest test set evaluation.

Figure 16 shows a plot of Random Forest Feature Importance, which reveals which vectors, or words for text classification, provide the most influence on the prediction in decision trees. In this case, the feature word 'hotel,' which is associated with the Tourism category, has the most relative importance that helps the model definitively decide whether a data point should be classified in the Tourism category or another category.

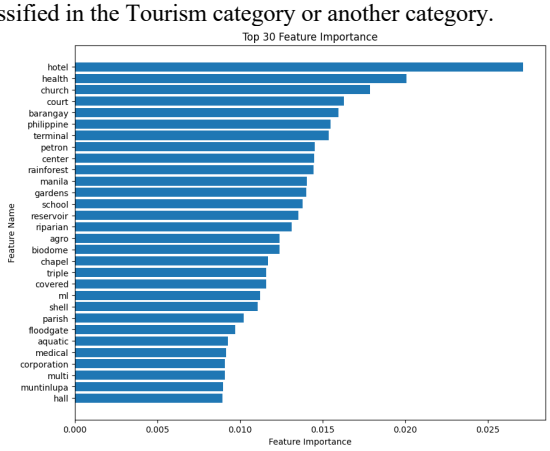

Figure 16. Random Forest point feature importance comparison.

The Random Forest model exhibited a strong but less consistent performance compared to RoBERTa, indicated by its evaluation metrics, confusion matrix, and ROC-AUC curve in Table 6 and Figures 17 and 18, respectively.

The evaluation metrics have relatively comparable values for all categories, with Inland Water, Agriculture, and Forest classes achieving a perfect score on all metrics, also followed by Religious, Transportation, and Hospitals with F1 scores a few decimal values lower, supported by the confusion matrix with high values in the main diagonal. However, the model struggled in classifying with the Commercial category, showing significantly lower Precision and F1 scores of 0.552 and 0.667, respectively. This is further validated by the confusion matrix showing more misclassifications along the rows and columns of

the Commercial category, and by the ROC-AUC curve having Commercial with the least area under the curve.

Category	Sensitivity	Specificity	Precision	F1
Hospitals	0.956	0.998	0.973	0.964
Commercial	0.842	0.951	0.552	0.667
Residential	0.824	0.991	0.869	0.845
Academic Institution	0.917	0.998	0.966	0.941
Utilities	0.911	0.999	0.984	0.946
Government Office	0.923	0.999	0.986	0.953
Tourism	0.912	0.997	0.962	0.936
Inland Water	1.000	1.000	1.000	1.000
Transportation	0.982	0.997	0.958	0.970
Industrial	0.794	0.997	0.942	0.862
Parks	0.897	0.996	0.936	0.916
Religious	0.966	0.999	0.986	0.976
Institutional	0.934	0.997	0.956	0.945
Agriculture	1.000	1.000	1.000	1.000
Forest	1.000	1.000	1.000	1.000

Table 6. Random Forest evaluation metrics.

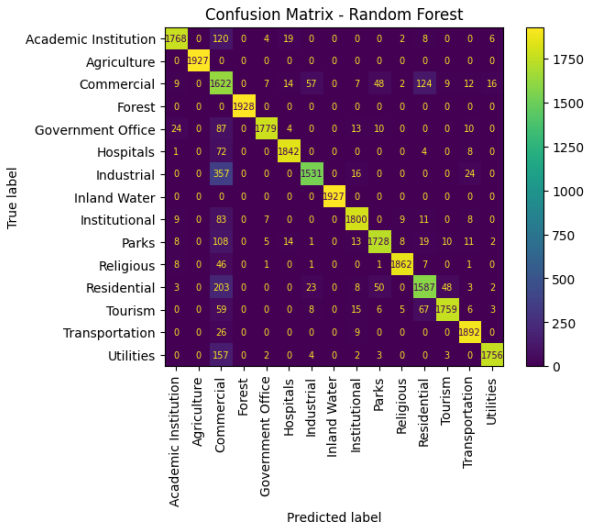

Figure 17. Random Forest confusion matrix.

This significant difference in performance between the LLMs and the classical model lies in their architecture. Random Forest classification relies on the TF-IDF algorithm, which sets weights on specific vectors or words that isolate each word as a local vector, effective on words that are optimally associated to a certain category or class.

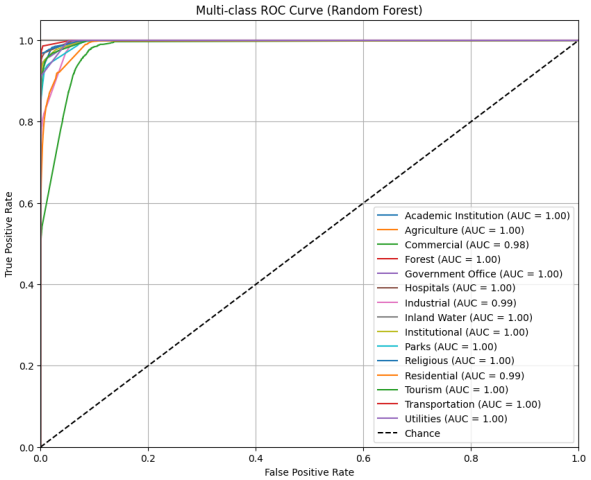

Figure 18. Random Forest ROC-AUC curve.

On the other hand, the pre-trained BERT and RoBERTa LLMs are built on the transformers architecture that employs contextual embeddings, where tokens or words are given weights and values in relation to their neighboring words, that considers the semantic relationship between words in phrases, uncovering more complex patterns.

The Random Forest model performs less accurately when classifying data points in the Commercial category because the words in the address names in that class vary indefinitely, which is where the LLMs outperform Random Forest; whereas in other classes such as Tourism, Hospitals, and Religious, certain words are strongly associated to them such as hotel, health, and church respectively–which are the top 3 important features, where Random Forest performs as well as LLMs.

4. Conclusion

The text classification tasks demonstrated that the pre-trained BERT and RoBERTa Large Language Models performed excellently by distinctively determining classes consistently across all set categories with negligible misclassifications. On the other hand, while the classical Random Forest machine learning still performed satisfactorily, it lagged in categorizing data points in specific classes, as shown in its evaluation metrics, confusion matrix, and ROC-AUC curve compared to the results of the LLMs.

The classical Random Forest still offers passable classification performance on classes with localized features, but modern LLMs such as BERT and RoBERTa uncover more complex patterns and contextual nuances between words, at the expense of computational cost given its deep learning architecture.

This methodology showcases the potential of LLMs in inferencing building types through accurate text-based classification. It enables the rapid filling of data gaps, enabling the deeper analysis of big data such as OSM collected data points for different purposes.

Acknowledgements

This research was done through the Department of Science and Technology (DOST) Grants-in-Aid Project entitled: Spatio-Temporal Land Cover Mapping for Risk Assessments in Land Use Planning through AI (LUPA Project) which is under the

Accelerated Earthquake Multi-hazards Mapping and Risk Assessment Program of the Philippines (ACER Program) monitored by the Philippine Council for Industry, Energy & Emerging Technology (DOST-PCIEERD) and implemented by the Philippine Institute of Volcanology and Seismology (DOST-PHIVOLCS).

References

Atwal, K. S., Anderson, T., Pfoser, D., & Züfle, A., 2022: Predicting building types using OpenStreetMap. *Scientific Reports*, *12*(1), 19976. doi.org/10.1038/s41598-022-24263-w

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J., 2022: A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management, 59(2)*, 102798. doi.org/10.1016/j.ipm.2021.102798

Escouflaire, L., Descampe, A., & Fairon, C., 2024: Automated text classification of opinion vs. news French press articles. A comparison of transformer and feature-based approaches. *Language & Communication, 99*, 129–140. doi.org/10.1016/j.langcom.2024.09.004

Housing and Land Use Regulatory Board, 2013: CLUP Guidebook: A Guide to Comprehensive Land Use Plan Preparation. Volume 1 - The Planning Process.

Kostina, A., Dikaiakos, M., Stefanidis, D., & Pallis, G., 2025: Large Language Models For Text Classification: Case Study And Comprehensive Review. arXiv. doi.org/10.48550/arXiv.2501.08457

Lindholz, M., Burdenski, A., Ruppel, R., Schulze-Weddige, S., Georg, L.B., Schobert, I., Haack, A., Eminovic S., Milnik, A., Hamm, C.A., Frisch, A., & Penzkofer, T., 2025: Comparing large language models and text embedding models for automated classification of textual, semantic, and critical changes in radiology reports. *European Journal of Radiology, 191,* 112316–112316. doi.org/10.1016/j.ejrad.2025.112316

Ribeiro, A., & Fonte, C. C., 2015. A Methodology for Assessing OpenStreetMap Degree of Coverage for Purposes of Land Cover Mapping. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5, 297–303. doi.org/10.5194/isprsannals-II-3-W5-297-2015

Rostam Z.R., & Kertész, G., 2024: Fine-Tuning Large Language Models for Scientific Text Classification: A Comparative Study. *2024 IEEE 6th International Symposium on Logistics and Industrial Informatics (LINDI)* (pp. 000233-000238). IEEE. doi.org/10.1109/LINDI63813.2024.10820432

Sun, C., Qiu, X., Xu, Y., & Huang, X., 2020: How to Fine-Tune BERT for Text Classification? (arXiv:1905.05583). arXiv. doi.org/10.48550/arXiv.1905.05583

Wang, Y., Gong, C., Ji, X., & Yuan, Q., 2025: Text classification for evaluating digital technology adoption maturity based on BERT: An evidence of Industrial AI from China. *Technological Forecasting and Social Change, 211*, 123903. doi.org/10.1016/j.techfore.2024.123903

Wang, Z., Pang, Y., Lin, Y., & Zhu, X., 2024: Adaptable and Reliable Text Classification using Large Language Models. *2024 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 67-74). IEEE. doi.org/10.1109/ICDMW65004.2024.00015

Yan, J., 2024: Large Language Model (LLM) Text Generation Detection: A BERT-BiLSTM Approach with Attention Mechanisms. *2024 4th International Conference on Electronic Information Engineering and Computer (EIECT)* (pp. 404-407). IEEE. doi.org/10.1109/EIECT64462.2024.10866488

Yuan, L., Ouyang, X., Bai, R., Zhu, C., Bai, R., Zhu, X., Zhou, Y., & Zhang, Y., 2024: A Framework for Categorizing Complaint Text via Large Language Model. *2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE)* (pp. 519-523). IEEE. doi.org/10.1109/ICAACE61206.2024.10549750