A Comparative Study of Transformer-Based Embeddings for Complaint Classification

Lee-Ling Fong1, a), Choo-Yee Ting1, 2, b), Hui-Ngo Goh1, 2, c), Hu Ng1, d) and Theng-Jia Law1, e)

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

*2Center of Bid data and Blockchain technology, CoE for Advanced Cloud, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia*

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

*a) fong.lee.ling1@student.mmu.edu.my*

*c) hngoh@mmu.edu.my*

*d) nghu@mmu.edu.my*

*e) thengjialaw@gmail.com*

**Abstract.** Effective administration of housing-related grievances is vital for guaranteeing tenant fulfillment and public wellbeing. This study pursues three key objectives: (1) establishing a standardized pipeline for generating complaint embeddings using a variety of pretrained transformer-based embedding approaches, (2) conducting comparative analysis of transformer-based embeddings across various machine learning classifiers for complaint categorization, and (3) determining optimal embedding-classifier combinations for optimal classification performance. After a well-crafted computational framework was established to harness integrated text features, twenty distinct transformer models with eleven classifiers were subjected to an exhaustive test. The best performance of combined paraphrase-MiniLM-L3-v2 embeddings with Support Vector Classifier (SVC) resulted in an accuracy of 86.11%. This demonstrates how well even smaller transformer models can penetrate the meaning of complaint texts. This finding will, therefore, help in refining and designing better and efficient housing complaint management systems.

# INTRODUCTION

One of the key components of the maintenance of acceptable living conditions within a residential area is complaint management. The period of time before the clearing of fallen trees or uncollected garbage influences the residents’ satisfaction greatly with safety [1]. Certainly, a mounting number of complaints need faster and more effective handling in the future to justify machine classification [2]. In such situations, residents would stand to benefit from timely and accurate complaint resolution by a transformer-based machine learning model [3]. These models have shown effectiveness in a host of natural language processing (NLP) tasks [4].

Despite these advances, identifying optimal transformer-based models for housing complaint classification remains challenging. The proliferation of transformer architectures with varying domain adaptations complicates model selection for this specific application [5, 6]. Current research lacks comprehensive comparisons of transformer-based embeddings across multiple machine learning classifiers within housing environments [7].

This study addresses these gaps through three key objectives: (1) establishing a standardized pipeline for generating complaint embeddings using a variety of pretrained transformer-based embedding approaches, (2) conducting comparative analysis of transformer-based embeddings across various machine learning classifiers for complaint categorization, and (3) determining optimal embedding-classifier combinations for optimal classification performance. The research scope encompasses housing-related complaints, evaluating multiple transformer models to assess their effectiveness in capturing the nuances of residential maintenance issues.

# LITERATURE REVIEW

Environmental complaint analytics in housing areas has emerged through smart city initiatives [8, 9], providing critical insights for improving living standards [10, 11]. NLP techniques enable effective complaint classification [12, 13], with contextual embedding models capturing semantic meanings [14, 15].

Transformer models like Robustly Optimized BERT Pretraining Approach (RoBERTa), Distilled Bidirectional Encoder Representations from Transformers (DistilBERT), and eXtreme Learning NET (XLNet) excel at processing noisy citizen reports [5, 16, 17], while domain-specialized embeddings show promise for housing environments [18, 19].

DistilBERT’s efficiency makes it well-suited for real-time applications, while A Lite BERT (ALBERT) and RoBERTa further enhance results through optimization techniques [20, 21, 22].

Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa) facilitates multilingual complaint processing [23], while domain-specific models improve specialized classification [24, 25]. Sentence-Bidirectional Encoder Representations from Transformers (Sentence-BERT) creates meaningful embeddings for short texts [14, 26], and large language models demonstrate potential for reasoning-based interpretation [27, 28, 29].

Ensemble techniques and gradient boosting models have shown promise for prioritizing complaints using multimodal data [4, 30, 31], while traditional supervised approaches remain competitive baselines [32, 33].

Current research mainly focuses on customer service complaints, leaving a gap in understanding environmental complaints in housing [34], and few explicitly model location-specific features in complaint patterns.

# METHOD

In this section, a comprehensive computational pipeline is implemented to evaluate the effectiveness of various transformer-based embeddings for environmental complaint classification, as illustrated in Figure 1. The methodology involves a thorough examination of how different embedding techniques capture the nuanced semantic content within environmental complaints. The core objective is to identify the embedding methodologies that demonstrate the highest efficacy in accurately classifying complaints, and to determine the optimal computational approaches for environmental complaint categorization that can potentially streamline regulatory responses. This multifaceted methodological framework enables deeper insights into the complex linguistic characteristics of environmental complaints, thereby facilitating more effective approaches to enhance environmental protection initiatives and regulatory efficiency.

**FIGURE 1.** Flow of the experimental process



## Dataset Preparation and Preprocessing

The study utilizes a real-world public complaint dataset obtained from Majlis Bandaraya Seberang Perai (MBSP), the municipal authority responsible for governing Seberang Perai, a region in Penang, Malaysia. The dataset consists of 26,671 complaint records collected between January and October 2024, encompassing 27 structured attributes that cover administrative information, complainant details, complaint content, handling assignments, categorization, and status tracking. This study uses seven key columns from the MBSP public complaint dataset: "tajuk\_aduan" (com- plaint title), "saluran\_aduan" (complaint channel), "daerah" (district), "dun" (state constituency), "keterangan\_aduan" (complaint description), and "jabatan" (department) as input features for transformer-based text embedding and classification. The "kategori\_utama" (main category) serves as the target variable. Knowing these data up to October 2024 would have facilitated such processes as categorizing complaints through transformer-generated text embeddings, analyzing the performance of various classifiers, recognizing patterns in community grievances, and directing data-informed decision-making by municipalities, in addition to a host of other potential applications for natural language processing and machine learning in this field. All sensitive and personal data have been carefully anonymized in ways that protect the individual’s right to privacy and are consistent with the ethical principles underlying research.

After preprocessing, everything was standardized for case consistency and cleared for some region-specific suffixes. It also involved the identification and removal of missing values and duplicate records. Stratified sampling was applied to split the data into an 80% primary training set and a 20% test set, retaining the same distribution of complaint categories across both sets to ensure fair representation in the subsets and thus a more accurate and realistic evaluation of model performance.

## Textual Feature Integration

This stage is actually formatting and preparation for embedding the data, by merging several features into one text format. Specifically, the fields "tajuk\_aduan", "saluran\_aduan", "daerah", "dun", "keterangan\_aduan", and "jabatan" were combined into a single line of text separated by semicolons for each complaint. As can be seen below, the result is not just that it has become a documentized summary of all environmental complaints, but also that it has all the references to those complaints:

“tajuk\_aduan”; “saluran\_aduan”; “daerah”; “dun”; “keterangan\_aduan”; “jabatan"

This transformation has enabled transformer models to fully exploit the language processing capabilities in processing this data without hindering the original format. This approach makes possible a much more thorough and easier analysis by putting administrative and descriptive data along one another.

## Transformer-Based Embeddings

To generate text embeddings, there are 20 transformer-based models from different fields have been developed in this study, such as:

1. General-purpose models: BERT, RoBERTa, DistilBERT, XLM-RoBERTa, GPT-2
2. Domain-specialized models: SciBERT, BioBERT, Legal-BERT
3. Efficient variants: BERT-tiny, all-MiniLM models
4. Multilingual models: mBERT, XLM-RoBERTa-base
5. Alternative architectures: ALBERT, ELECTRA, DeBERTa, XLNet, BART, T5, MPNet

A variety of critical stages have been gone through by each of the transformer-based models. To begin, tokenization was used to prepare the combined text features, and truncation and padding were applied as necessary. To create embeddings, this step assisted in properly formatting the data. By averaging the token embeddings to generate the sentence-level embeddings, the common meaning of each complaint was then identified. To raise the efficiency and lower the memory usage, the data was then processed in batches. The outputs are then saved for use in the next step.

## Classification Framework

There are 11 classification methods developed and evaluated. The combinations of the different transformer-based embeddings with these classification methods are then evaluated and analyzed to determine the effectiveness of each of them. The 11 classification methods are listed below:

### *Traditional approaches*

* 1. Logistic Regression:A simple method used to estimate the likelihood of various outcomes using a logistic function. Due to the easy understanding and effectiveness in operations, it has become a great option in many classification problems.
  2. K-Nearest Neighbors:A simple technique used to classify a data point based on the distance between itself and the other points in the training data. This technique is effective in identifying local patterns where there is no need to develop a formal model.
  3. Support Vector Machines:It is a useful technique that can identify the ideal boundaries between different classes or groups. It is suitable for use in either simple or complex datasets due to the development of kernel functions to handle the complex decision boundaries.
  4. Gaussian Naive Bayes:A probabilistic technique that is based on Bayes’ theorem, which assumes features are independent and distributed normally. Positive results are frequently obtained using this technique, and it is working well with feature-rich data.
  5. Quadratic Discriminant Analysis:It is a technique that uses a distinct covariance structure for every class. This technique allows it to initiate the curve decision boundaries, which enables it to handle the scenarios in which classes are difficult to divide by a straight line.

### *Tree-based methods*

1. Decision Trees: A decision tree technique is a classification model that develops a hierarchy of feature- based decisions, which provide simple and comprehensible decision paths. It is good at identifying the complex relationship between the features.
2. Random Forest: A random forest technique is an ensemble learning technique that integrates the output of various decision tree models into a single effective model to reduce overfitting and improve classification performance.
3. AdaBoost: An adaptive boosting technique that iteratively concentrates on misclassified instances, with the potential to improve the model’s performance on complex examples within the embedding space.

### *Gradient boosting frameworks*

1. XGBoost: An optimized gradient boosting algorithm that demonstrates enhanced performance through the application of regularization techniques and parallel processing capabilities, often exhibiting superior results when handling structured data representations.
2. CatBoost: A gradient boosting framework that exhibits proficient management of categorical features and incorporates embedded strategies to minimize overfitting.
3. LightGBM: A computationally efficient gradient boosting framework that utilizes histogram-based techniques and a leaf-wise tree growth strategy, rendering it particularly well-suited for large-scale classification tasks.

Each classifier was trained on the embeddings from the training set and evaluated on the test set. This methodology enabled the isolation of the impact of embedding quality from classifier selection, allowing for a fair comparative analysis of how different embedding spaces represent the semantic content of complaints and support accurate categorization across a range of algorithmic approaches.

## Evaluation Metrics

Performance evaluation was conducted using below key metrics:

* 1. Accuracy: Measures how often the predictions are correct overall.
  2. Macro-averaged Precision: Calculates the average precision for all categories, treating each category equally.
  3. Macro-averaged Recall: Calculate the mean recall of every category while maintaining every category at the same weight.
  4. Macro-averaged F1-score: Determine the harmonic mean of precision and recall on all the categories to add them into a single score.

Some of the categories might be less usual in the dataset, but macro-averaging was still chosen to balance each of the weights. This is to make sure the performance of the minority class has a significant impact on the results.

# FINDINGS

This section discusses the outcome of the comparison between the combination of various transformer-based embeddings used to categorize complaints about housing-related complaints with the different classification methods. Numerous models are evaluated in this study, including efficient, multilingual, domain-specific, general-purpose, and alternative transformers. The models are evaluated by using metrics such as accuracy, precision, recall, and F1-score.

The results indicate the impact of various embedding methods on classification performance and identify the optimal model combinations for enhancing complaint handling.

**TABLE 1.** Top 10 best-performing combinations of Transformer-based embeddings and classifiers for complaint category

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Embedding** | **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Average Score** |
| paraphrase-MiniLM-L3-v2 | SVC | 0.8611 | 0.8451 | 0.7495 | 0.7613 | 0.8042 |
| xlm-roberta-base | SVC | 0.8434 | 0.8170 | 0.7266 | 0.7366 | 0.7809 |
| bert-base-multilingual-cased | SVC | 0.8361 | 0.8371 | 0.7187 | 0.7282 | 0.7800 |
| xlm-roberta-base | Logistic Regression | 0.8411 | 0.7706 | 0.7391 | 0.7484 | 0.7748 |
| gpt2 | SVC | 0.8399 | 0.8063 | 0.7141 | 0.7192 | 0.7699 |
| legal-bert-base-uncased | SVC | 0.8392 | 0.8116 | 0.7149 | 0.7101 | 0.7689 |
| paraphrase-MiniLM-L3-v2 | Logistic Regression | 0.8288 | 0.7551 | 0.7313 | 0.7374 | 0.7631 |
| xlm-roberta-base | CatBoost Classifier | 0.8288 | 0.7964 | 0.7030 | 0.7197 | 0.7620 |
| distilbert-base-uncased | Logistic Regression | 0.8307 | 0.7543 | 0.7259 | 0.7309 | 0.7605 |
| gpt2 | Logistic Regression | 0.8253 | 0.7489 | 0.7288 | 0.7357 | 0.7597 |

Based on the result shown in Table 1, the combination of the sentence-transformers/paraphrase-MiniLM-L3-v2 embeddings with Support Vector Classifier (SVC) generated the optimal outcome with an accuracy of 86.11%, precision of 84.51%, recall of 74.95%, and F1-score of 76.13%, which had the highest score on all metrics. This might be due to the complementarity of the two techniques. A huge volume of sentence pairs has been used to train MiniLM, which benefits its in understand various ways people express similar ideas. Meanwhile, SVC is working great in separating different complaint categories with the detailed vector data from the embeddings. This makes this pair especially useful for handling user-written content, where people often use very different words and phrases to describe the same issues.

The multilingual models also performed well. For example, xlm-roberta-base with SVC reached 84.34% accuracy, and bert-base-multilingual-cased with SVC got 83.61%. This shows that models designed for multiple languages can handle complaints that come in many different styles and levels of detail. This is especially important in Malaysia, where complaints to MBSP often include both Bahasa Malaysia and English mixed together. Since these multilingual models were trained on over 100 languages, they can handle this kind of language mixing and different word choices without losing accuracy. This flexibility means they stay accurate even when people switch languages or use various ways to describe the same problem.

In addition, the combination of xlm-roberta-base with Logistic Regression achieved competitive results, with an accuracy of 84.11% and balanced precision and recall scores, indicating that traditional classifiers, when paired with rich embeddings, remain highly effective for text classification tasks. Similarly, openai-community/gpt2 embeddings combined with SVC and Logistic Regression also showed relatively strong performance, maintaining accuracies around 83-84%, and demonstrated the potential of generative pre-trained models even when adapted for discriminative tasks like classification. Specialized domain models such as legal-bert-base-uncased achieved respectable results, particularly excelling in precision (81.16%), reflecting their robustness in handling structured and formal complaint narratives commonly found in environmental issues reported by residents.

In contrast, specialized models like legal-bert-base-uncased showed decent results but didn’t emerge as the top performers. This indicates that even though complaint language can sound formal at times, it doesn’t match up closely enough with legal terminology and structure to give these domain-specific models a real advantage over those trained for understanding similar meanings (like MiniLM) or handling multiple languages (like XLM-R). That said, the legal model’s precision of 81.16% stands out as quite impressive. This high precision means the model is usually quite sure and accurate when it decides on a classification. This might be due to its training that allows it to recognize the patterns.

Furthermore, this study also found that those effective transformer-based models, such as MiniLM and DistilBERT perform well when combined with the SVC or Logistic Regression. There is an advantage to these models where there is a limitation on computing power, which allows them to become a good option. In a nutshell, the results show that the efficient and multilingual transformer-based models perform well in handling the housing-related complaints. Especially when combined with suitable classifiers, smaller models perform better than larger models, but general models are more reliable. The purpose of developing this comparison is to help in understanding how the embedding models can perform well with different classification models in order to provide fruitful recommendations for creating systems that are both efficient and expandable.

# CONCLUSION

The performance of the combination of various transformer-based embeddings with several classifiers in order to classify housing-related complaints was investigated in this study. A standardized procedure that includes data cleaning, feature transformation, embedding generation, and model development is developed to measure the performance of models that combine different transformer-based embeddings with various classifiers. Based on the experiments developed in this study, a significant gap is filled due to the first comprehensive comparison that focused on classifying the housing-related complaints.

The best performance outcome is obtained by the combination of Paraphrase-MiniLM-L3-v2 embeddings with the Support Vector Classifier (SVC) across all the combinations of 20 transformer-based embeddings and 11 classification methods, which shows that this combination is a feasible solution for developing large-scale systems to handle the housing-related complaints. Besides, multilingual models such as XLM-RoBERTa-base and bert-base-multilingual-cased also perform well in this study, which proves that their ability to handle multiple languages is effective. Those smaller but effective models, such as MiniLM and DistilBERT, also returned quite good results, which shows that these models are suitable for those cases requiring small processing power. In a nutshell, both multilingual and small transformer models are effective in performing this task. This provides businesses with a range of efficient choices for implementing automated complaint sorting systems that can be customized to meet their language needs and processing capacity. Because of this flexibility, institutions can find the best solution to meet their needs while maintaining a high level of classification accuracy.

# REFERENCES

1. W. Yan, L. Xu, J. Tong, and W. Huang, “Automated complaint management for improving customer satisfaction in the service industry,” Journal of Business Research **134**, 325–336 (2021).
2. S. Kraus, S. Feuerriegel, and A. Oztekin, “Deep learning in business analytics: Applications and insights,” Business & Information Systems Engineering **62**, 285–297 (2020).
3. J. Liu, K. H. Lim, K. L. Wood, and M. Li, “Optimizing group utility in itinerary planning: A strategic and crowd-aware approach,” *arXiv preprint* arXiv:2304.08495 (2023).
4. K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, and Z. Liu, “A survey on vision transformer,” IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).
5. Z. Yang, P. Qi, X. Zhang, and S. Bowman, “Transformer-based models for complaint classification: An experimental comparison,” in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2020), pp. 202–215.
6. Y. Sun, Y. Zhang, and Q. Wu, “A study of transformer models for customer feedback classification,” Journal of Machine Learning Research **22**, 1–19 (2021).
7. S. Gunarsson, S. Vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, “A comparative study of machine learning models for text classification in complaint management,” Computers in Industry **127**, 103408 (2021).
8. Y. Zhao, Y. Zhou, and X. Li, “Smart city complaint management using machine learning techniques: A review,” Sustainable Cities and Society **74**, 103167 (2021).
9. F. Ahmad, I. Hussain, and A. Mahmood, “Urban issues mining from citizens’ complaints: A survey,” Cities **120**, 103451 (2022).
10. H. Chen, L. Xu, and L. Wang, “Intelligent complaint management in urban service systems,” Information Processing & Management **59**, 102889 (2022).
11. S. Khalid, M. Ahmed, and M. Nazir, “Smart complaint systems in smart cities: A comprehensive review,” Journal of Urban Technology **30**, 37–58 (2023).
12. J. Li, A. Sun, and J. Han, “Text classification: Advances and challenges,” ACM Computing Surveys **54**, 1–41 (2021).
13. M. Yu, X. Li, and J. Feng, “NLP techniques for social complaint analysis: Trends and future directions,” Information Sciences **622**, 589–605 (2023).
14. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020), pp. 3982–3992.
15. S. Wang, Y. Fang, and W. Yu, “Deep transfer learning for complaint text classification,” Knowledge-Based Systems **203**, 106065 (2020).
16. Y. Liu, M. Ott, N. Goyal, et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692 (2020).
17. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT: A distilled version of BERT,” arXiv:1910.01108 (2020).
18. K. Song, X. Tan, T. Qin, et al., “MPNet: Masked and permuted pre-training for language understanding,” in Advances in Neural Information Processing Systems (NeurIPS), Vol. 33 (2020), pp. 16857–16867.
19. I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2020), pp. 3615–3620.
20. Z. Lan, M. Chen, S. Goodman, et al., “ALBERT: A lite BERT for self-supervised learning of language representations,” in International Conference on Learning Representations (ICLR) (2020).
21. M. Jin and N. Aletras, “Identifying and categorizing complaints on social media,” in Proceedings of the International Conference on Computational Linguistics (COLING) (2020).
22. M. R. R. Rana, et al., “CNN-BiLSTM with RoBERTa embedding for complaint detection,” Cybernetics and Information Technologies (2024).
23. A. Conneau, K. Khandelwal, N. Goyal, et al., “Unsupervised cross-lingual representation learning at scale,” in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2020).
24. I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, “Legal-BERT: The Muppets straight out of law school,” in Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020).
25. J. Lee, W. Yoon, S. Kim, et al., “BioBERT: Biomedical language representation model,” Bioinformatics (2020).
26. P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-enhanced BERT with disentangled attention,” in International Conference on Learning Representations (ICLR) (2021).
27. J. Wei, X. Wang, D. Schuurmans, et al., “Chain-of-thought prompting elicits reasoning,” in Advances in Neural Information Processing Systems (NeurIPS) (2022).
28. L. Prokhorenkova, G. Gusev, A. Vorobev, et al., “CatBoost: Unbiased boosting with categorical features,” in Advances in Neural Information Processing Systems (NeurIPS), Vol. 31 (2020).
29. G. Ke, Q. Meng, T. Finley, et al., “LightGBM: A highly efficient gradient boosting decision tree,” in Advances in Neural Information Processing Systems (NeurIPS), Vol. 30 (2020).
30. L. Pang, X. Zhang, and Y. Zhang, “Environmental complaint prediction with random forests and SVMs,” Environmental Modelling & Software **142**, 105065 (2021).
31. Y. Zhou, Z. Ma, and J. Liu, “Comparative analysis of classifiers for smart complaint handling,” Applied Sciences **12**, 3831 (2022).
32. R. Tang, Y. Liang, and Z. Liu, “Cross-domain complaint detection using deep learning,” Information Processing & Management **58**, 10249 (2021).
33. S. Park, J. Shin, and S. Han, “Smart governance and complaint resolution in smart cities: A machine learning approach,” Government Information Quarterly **39**, 101630 (2022).
34. W.-E. Kong, T.-E. Tai, P. Naveen, and H.A. Santoso, “Performance Evaluation on E-Commerce Recommender System based on KNN, SVD, CoClustering and Ensemble Approaches,” Journal of Informatics and Web Engineering **3**(3), 63–76 (2024)