Automated Categorization of Complaints in Indonesia's National Complaint System Using Traditional and Transformative NLP Techniques

Summary

In an era of increasing citizen participation, effective complaint handling is critical for ensuring public trust and responsive governance. Indonesia's SP4N-LAPOR! platform, which centralizes public complaints, currently relies on manual categorization—a process prone to delays and human error. This study explores the application of machine learning (Naïve Bayes, Support Vector Machine, Multilayer Perceptron) and deep learning (Bidirectional Encoder Representations from Transformers - BERT) methods to automate this categorization process. Using a real-world dataset of over 73,000 verified complaints categorized into 18 fields, the study compares the performance of these models based on accuracy, precision, recall, and F1-score. Results demonstrate that while traditional models perform reasonably well, the BERT model consistently outperforms others, achieving a 74% accuracy and 0.73 weighted F1-score. The findings suggest significant potential for automating public complaint handling systems, improving efficiency, and paving the way for more advanced AI-driven solutions in e-governance.

1. Introduction:

Citizen complaint systems are a vital link between the public and government institutions, providing a platform for expressing grievances, reporting issues, and influencing policy. Indonesia's SP4N-LAPOR! system serves this purpose at a national scale, allowing citizens to submit complaints across diverse sectors including health, education, infrastructure, and governance.

However, a key operational bottleneck in the platform is the manual categorization of complaints, which introduces inefficiencies and inconsistencies. Manual categorization relies on the complainant's self-selection or administrative staff's classification, leading to variable accuracy, increased processing time, and resource-intensive workflows.

This study investigates whether machine learning, particularly natural language processing (NLP) techniques, can improve the speed and accuracy of complaint categorization. A special focus is placed on comparing traditional machine learning models with modern transformer-based models like BERT.

2. Data and Exploration:

The dataset utilized comprises 73,106 complaints submitted between January and June 2023. Each complaint is labeled into one of 18 categories, such as Public Works and Spatial Planning, Population Administration, Education, and ICT.

Key Observations:

- The **Population Administration** and **Public Works** categories dominate, reflecting citizens' primary concerns.
- Text lengths vary significantly, with some complaints containing just a few words and others extensive narratives.
- Imbalance exists across categories, with rare classes presenting a modeling challenge.

Initial exploratory data analysis (EDA) identified trends in complaint length, category distribution, and regional patterns, providing insight into the complexity and variability of the data.

---

3. Methodology:

3.1 Traditional Machine Learning Models:

- **Naïve Bayes (MultinomialNB)**: A probabilistic model suited for high-dimensional text classification.
- **Support Vector Machine (SVM)**: A robust model capable of handling sparse feature spaces typical in text data.
- **Multilayer Perceptron (MLP)**: A basic neural network architecture to capture non-linear patterns.

TF-IDF Vectorization was applied to transform raw text into numerical features suitable for model input.

3.2 Transformative NLP Model:

- **IndoBERT (Indonesian version of BERT)**: A pre-trained transformer model fine-tuned on the complaint dataset.
    - Input text was tokenized and padded to match BERT input requirements.
    - Fine-tuning involved adjusting the pre-trained layers for specific complaint categories.

3.3 Evaluation Metrics:

- **Accuracy**: Overall correct predictions.
- **Precision, Recall, F1-Score** (macro and weighted averages): Especially critical due to category imbalance.
- **Confusion Matrix**: To visualize classification performance across categories.

---

4. Results:

The entire dataset is partitioned into two distinct subsets: the training data and the test data. The training data, in this case constituting 80% of the total data, serves as the foundation for the classification model's learning process. Through exposure to this dataset, the machine learning algorithm identifies patterns, relationships, and relevant features within the input information. Conversely, the test data, kept separate from the training phase, assumes a critical role in evaluating the model's performance. Withholding this data during training ensures an unbiased assessment of the model's ability to generalise to new, unseen data.

A primary benefit of utilising a test dataset is its capacity to mitigate the risk of overfitting. Overfitting occurs when a model becomes excessively tailored to the training data, inadvertently capturing noise and specific idiosyncrasies not representative of the broader data population. Consequently, such a model struggles to accurately classify data points that deviate from the training set. The test dataset acts as a safeguard against this by revealing how well the model performs on data it has never encountered before.

The training and test datasets allow for fair comparisons between different machine learning models or algorithms. By using the same data for training and testing, we can determine which model performs the best. We will evaluate and compare the models based on overall accuracy, recall, and F1 score. These evaluation parameters will show how well the models perform in each category as well as their average performance overall, both in terms of macro and weighted scores. The macro average gives a rough average across all categories, while the weighted average takes into account the number of units in each category.

4.1. Naïve Bayes

The Naive Bayes (NB) classification algorithm operates by computing the prior probability associated with each class and subsequently determining the likelihood of specific features occurring within the context of a given class. Before finalising the model, a 5-fold cross-validation was employed to

determine the optimal Laplace smoothing factor, known as alpha. This parameter is crucial as it addresses the issue of zero probabilities that can arise when a particular feature does not appear in the training data for a specific class. The cross-validation process identified an alpha value of 0.01 as optimal, resulting in an accuracy of 66.68%. This represents a significant improvement over the default alpha value of 1, indicating that careful tuning of this parameter can substantially enhance model performance.

Table 1. Accuracy of different alpha

Alpha	Accuracy
0.1	66.68%
1.0	58.99%
10.0	51.64%

After selecting the optimal alpha parameter, the model underwent a training process using the entire training dataset, followed by a testing process on a designated test dataset. The results indicate that the classifier successfully categorised approximately 68% of the complaints into their correct respective categories. While this accuracy figure may seem fairly robust at first glance, this overall accuracy rate should be interpreted cautiously due to the imbalanced nature of the dataset, which is characterised by a substantial disparity in the number of instances across different categories.

To gain a comprehensive understanding of model performance, it's crucial to examine metrics like precision, recall, and F1-score across different categories. These metrics shed light on the model's strengths and weaknesses, particularly in handling imbalanced datasets. Looking at the details of the value of the evaluation parameters, we can divide the categories by performance into four distinct groups: categories where the model performs well with balanced precision and recall, categories with high precision but low recall, categories with high recall but low precision, and categories where both precision and recall are low.

Categories	precision	recall	f1-Score	Support
Agriculture and Livestock	1	0.1	0.17	94
Economics and Finance	0.55	0.62	0.58	1070
Education and Culture	0.71	0.83	0.77	1594
Employment	0.64	0.86	0.73	1705
Energy and Natural Resources	0.81	0.28	0.42	404
Environment and Forestry	0.86	0.53	0.66	622
Gender Equality and Social Inclusion	0	0	0	39
Health	0.77	0.44	0.56	893
Information and Communication Technology	0.81	0.74	0.77	1308
Politics and Law	0.56	0.28	0.38	562
Population Administration	0.84	0.78	0.81	839
Public Order and Community Protection	0.62	0.65	0.63	1131
Public Works and Spatial Planning	0.59	0.91	0.72	2046
Religion	1	0.21	0.35	229
Social and Welfare	0.64	0.8	0.71	1229
SP4N-LAPOR! Application Problems	0.5	0.04	0.07	526
Transportation	0	0	0	117
Village Development, Rural Areas and Transmigration	0.81	0.06	0.11	214
Accuracy			0.67	14622
macro avg	0.65	0.45	0.47	14622
weighted avg	0.68	0.67	0.64	14622

The classification process exhibited good performance in categorising complaints into the domains of "population and administration", "education and culture", and "information and technology." Among these, the "Population and Administration" category stands out with the highest F1-score of all categories. Of the 785 instances assigned to this category, only 128, or roughly 16%, were misclassified. Additionally, the recall of 0.78 indicates that the model correctly identified 78% of all actual complaints in this category. This is complemented by a high precision of 0.84, reflecting that the model effectively distinguishes relevant instances within this category. The high F1-score suggests that the model's performance is well-balanced, both in terms of precision and recall, making it highly reliable for this category.

Similarly, the category of "information and technology" demonstrated a balance between precision and recall, culminating in an f1-score of 0.77. In terms of precision, the model incorrectly classified only 238 complaints within this category out of a total of 1,207 predictions, indicating a relatively low rate of false positives. The recall metric further underscores the model's

effectiveness in this category, with 969 out of 1,308 relevant complaints correctly identified.

The Naive Bayes classifier also showcased strong performance in the "education and culture" category but with a distinct pattern compared to the preceding two categories Here, the model achieved a recall rate that exceeded its precision. Specifically, only 17% of complaints within this category were misclassified, with most errors occurring in the "Employment" category. This higher recall indicates that the model is effective at capturing the majority of relevant complaints in this category, even if it occasionally misclassifies them into other categories.

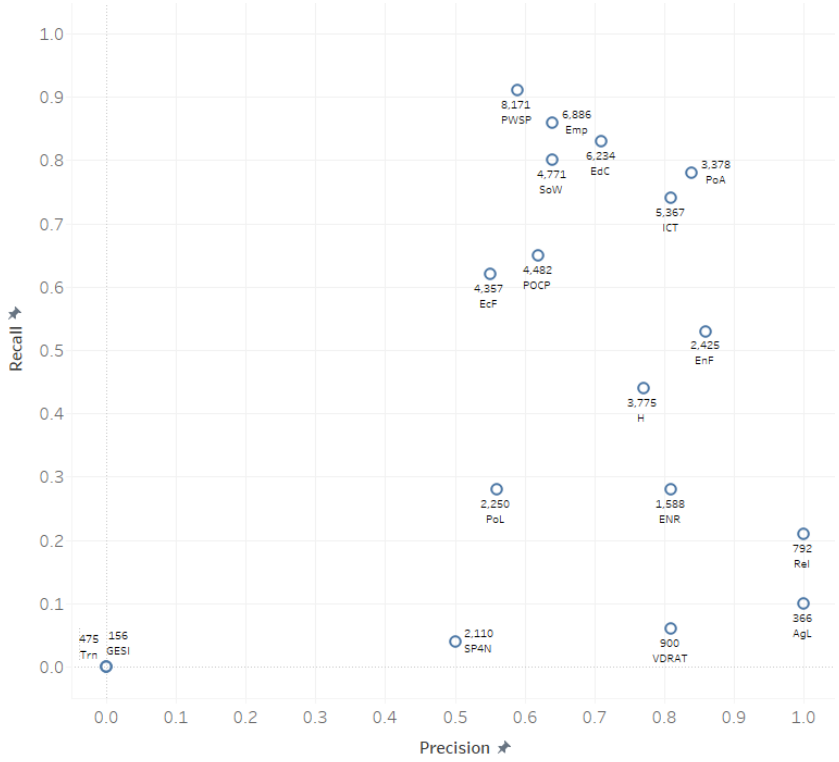

Figure 1: Naïve Bayes Precision, recall, and f1-score

The second group comprises categories where the model exhibits high precision but low recall, with "agriculture and livestock" and "religion" serving as illustrative examples. This indicates that the model excels at correctly identifying instances belonging to these categories, minimising false positives. However, the low recall indicates that the model struggles to capture a substantial portion of the actual cases within these categories, leading to a higher number of false negatives. Essentially, the model adopts a cautious approach to these classes, making predictions only when highly confident,

which results in fewer errors and overlooks many relevant instances. Several factors may contribute to this behaviour, including dataset imbalance, where the model might be biased towards more prevalent categories, consequently exercising greater caution when predicting less frequent ones.

Conversely, the third group includes categories demonstrating a contrasting trend of high recall but low precision. For instance, categories with relatively high numbers of complaints, such as "Public Works and Spatial Planning", fall into this category. While the model has a decent f-1 score and effectively identifies instances belonging to these categories, it also incorrectly assigns many instances from other categories to them. This suggests that the model is sensitive to the presence of issues related to them but lacks specificity in distinguishing these issues from other related concerns. The high recall indicates that the model is good at capturing relevant instances, but the low precision reveals that it struggles to differentiate these instances accurately, leading to a higher rate of false positives.

Finally, there are categories where the model exhibits poor performance, characterised by both low precision and recall. These categories present the most significant challenges for the model and indicate areas where further refinement is needed. This situation happens in several categories, but the worst can be found in the categories of "Transportation" and "Gender Equality and Social Inclusion", where the f-1 scores are both 0. The test dataset contains 117 complaints from the "Transportation" category; 61 of them are categorised as "Public Works and Spatial Planning", and 29 complaints are predicted in "Public Order and Community Protection". Out of 39 complaints from the "Gender Equality and Social Inclusion" category, 12 are also predicted to be in "Public Order and Community Protection", and 10 are predicted to be in "Employment".

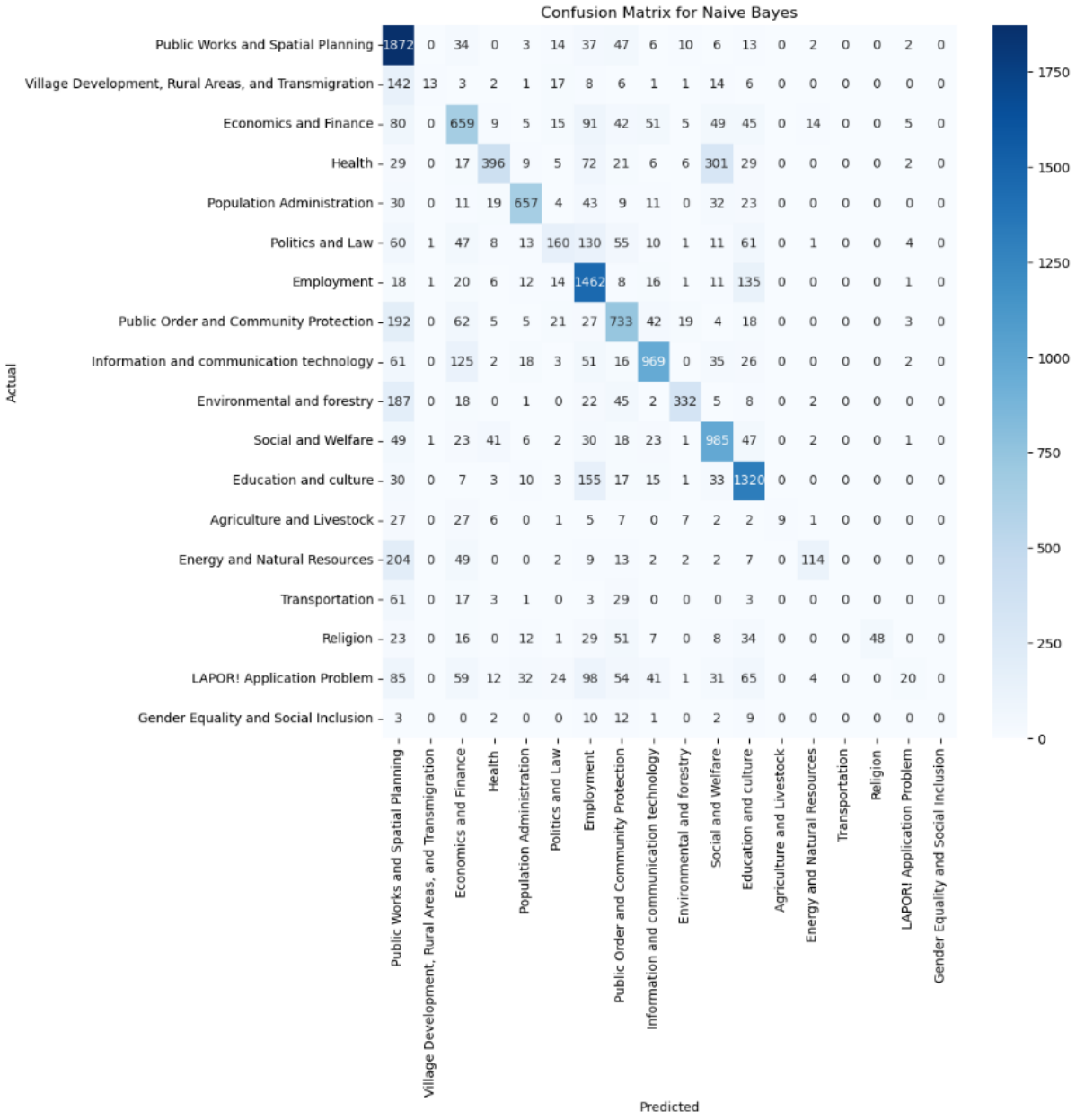

Figure 2: Naïve Bayes Confusion Matrix

4.2. Support Vector Machine

The Support Vector Machine (SVM) model underwent a training process that began with selecting the most suitable kernel, which is foundational to the model's performance. The choice of the kernel in an SVM is analogous to choosing the proper lens through which the data is viewed and classified. In this case, four kernel types were evaluated: linear, polynomial, radial basis function (RBF), and sigmoid. Each kernel type was tested using 5-fold cross-validation to ensure the model generalises well to unseen data and avoids overfitting. This thorough comparison revealed that the radial basis function

(RBF) kernel was the most effective for this particular dataset. The RBF kernel, known for its ability to handle non-linear relationships by mapping the input features into a higher-dimensional space, achieved an overall accuracy of 72.62%. This accuracy was slightly superior to that produced by the linear kernel, which achieved a 72.28% accuracy. Additionally, the polynomial kernel has the lowest accuracy, far behind other kernels.

Kernel	Accuracy
Linear	72.28%
Polynomial	62.33%
RBF	72.62%
Sigmoid	71.89%

After the kernel selection, the SVM model was trained on the entire dataset. The core principle of SVM is to identify the optimal hyperplane that distinctly separates data points into their respective classes while maximising the margin between them. When the SVM model's performance was compared to that of the Naive Bayes model, it was clear that the SVM approach outperformed Naive Bayes across all key evaluation metrics. Notably, the SVM model accurately classified 5% more complaints into their correct categories, underscoring its superior data handling capability. Furthermore, the SVM model achieved a weighted average f1-score of 0.71, demonstrating its robust performance. While the SVM model and Naïve Bayes showed comparable precision and recall on a weighted average basis, a more significant disparity was observed in the micro average. This suggests that several classes with a low number of instances tend to have higher precision compared to their recall.

However, the primary contributors to this imbalance differ between the two models. In the case of Naive Bayes, "Agriculture and Livestock" and "Religion" had the widest gaps. In contrast, for SVM, the categories "Gender Equality and Social Inclusion" and "Transportation" showed the widest gaps; it is important to notice that these categories have zero f1-scores in the Naive Bayes model. Meanwhile, "Religion" and "Agriculture and Livestock"

significantly reduced the gap between precision and recall in the SVM model, even though the gap in the "Religion" class is still relatively high, at 0.46. Interestingly, there were no instances where recall was significantly higher than precision in the SVM model, which is found in the previous model. The largest gap was observed in the "Public Works and Spatial Planning" category, where the recall was 0.88, 0.13 points higher than precision. This suggests that the model is adapting at identifying true positives in this task.

Categories	precision	recall	f1-Score	Support
Agriculture and Livestock	0.75	0.56	0.64	94
Economics and Finance	0.58	0.62	0.60	1070
Education and Culture	0.74	0.84	0.79	1594
Employment	0.75	0.83	0.79	1705
Energy and Natural Resources	0.74	0.75	0.74	404
Environment and Forestry	0.81	0.75	0.78	622
Gender Equality and Social Inclusion	1	0.05	0.10	39
Health	0.71	0.64	0.67	893
Information and Communication Technology	0.77	0.82	0.79	1308
Politics and Law	0.57	0.38	0.46	562
Population Administration	0.8	0.85	0.82	839
Public Order and Community Protection	0.66	0.67	0.67	1131
Public Works and Spatial Planning	0.75	0.88	0.81	2046
Religion	0.88	0.42	0.57	229
Social and Welfare	0.74	0.74	0.74	1229
SP4N-LAPOR! Application Problems	0.42	0.11	0.18	526
Transportation	0.88	0.3	0.45	117
Village Development, Rural Areas and Transmigration	0.5	0.37	0.43	214
Accuracy			0.72	14622
macro avg	0.72	0.59	0.61	14622
weighted avg	0.71	0.72	0.71	14622

Examining the F1-scores across various categories reveals the SVM model's good performance, particularly in certain areas where it consistently achieved F1-scores exceeding 70%. These categories include "Population Administration", "Public Works and Spatial Planning", "Information and Communication Technology", "Employment", "Education and Culture", "Environment and Forestry", "Energy and Natural Resources", and "Social and Welfare." The "Population Administration" category emerged as the best, where the SVM model excelled with an F1-score of 0.82, the highest across all categories. This score was achieved by accurately predicting 80% of the

instances in this category while also capturing 85% of all actual instances. This high level of performance underscores the model's ability to distinguish and correctly classify complaints within this domain, reflecting a well-balanced precision and recall. The score indicates that the SVM model effectively recognised patterns specific to this category, contributing to its high prediction accuracy.

Following closely was the 'Public Works and Spatial Planning' category, where the SVM model achieved an F1-score of 0.81. This score was primarily due to the model's good recall rate of 88%, successfully identifying 1,809 out of 2,046 instances in this category. This high recall rate, the best among all categories, underscores the model's sensitivity to detecting complaints related to public works and spatial planning issues. However, the F1-score slightly decreased due to the lower precision, with only 75% of all predictions classified under this category being correct. This decrease in precision suggests that while the model was highly effective at identifying relevant instances, it misclassified a significant number of complaints into this category.

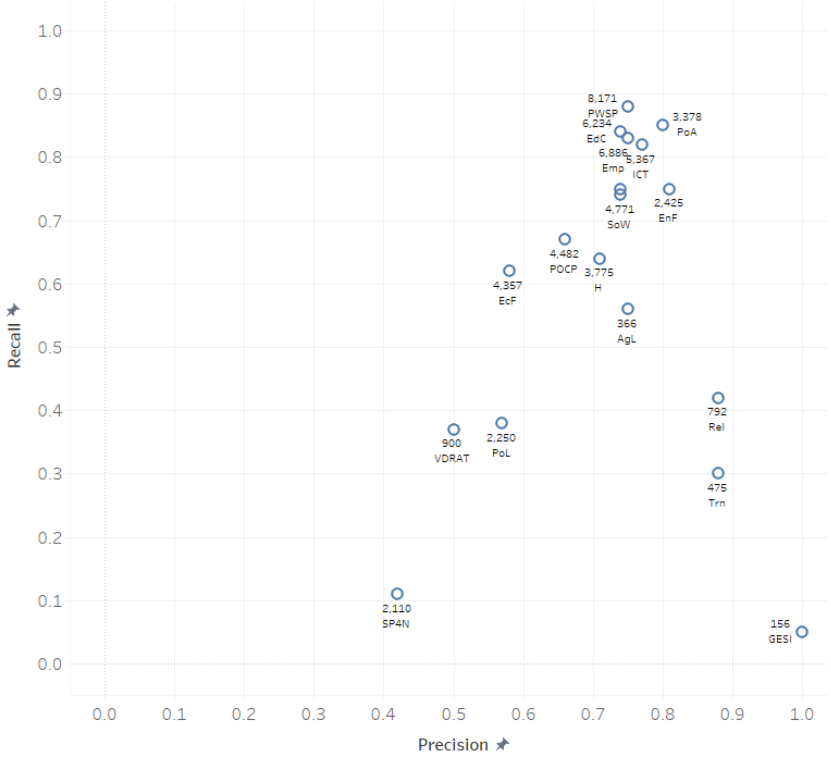

Figure 3: SVM Precision, recall, and f1-score

However, not all categories benefited equally from the SVM model's strengths. The model performed poorly in the "SP4N-LAPOR! Application

Problems" and "Gender Equality and Social Inclusion" categories, where the F1 scores were significantly lower. The "Gender Equality and Social Inclusion" categories likely present challenges due to a lack of sufficient training data, leading to underfitting. Only two instances were predicted in the category, and even though both were correct, they only covered 0.05% of the total actual instances. Meanwhile, it seems the features used to define the "SP4N-LAPOR! Application Problems" category were not as distinctive as those in other categories, making them challenging for the model to classify accurately.

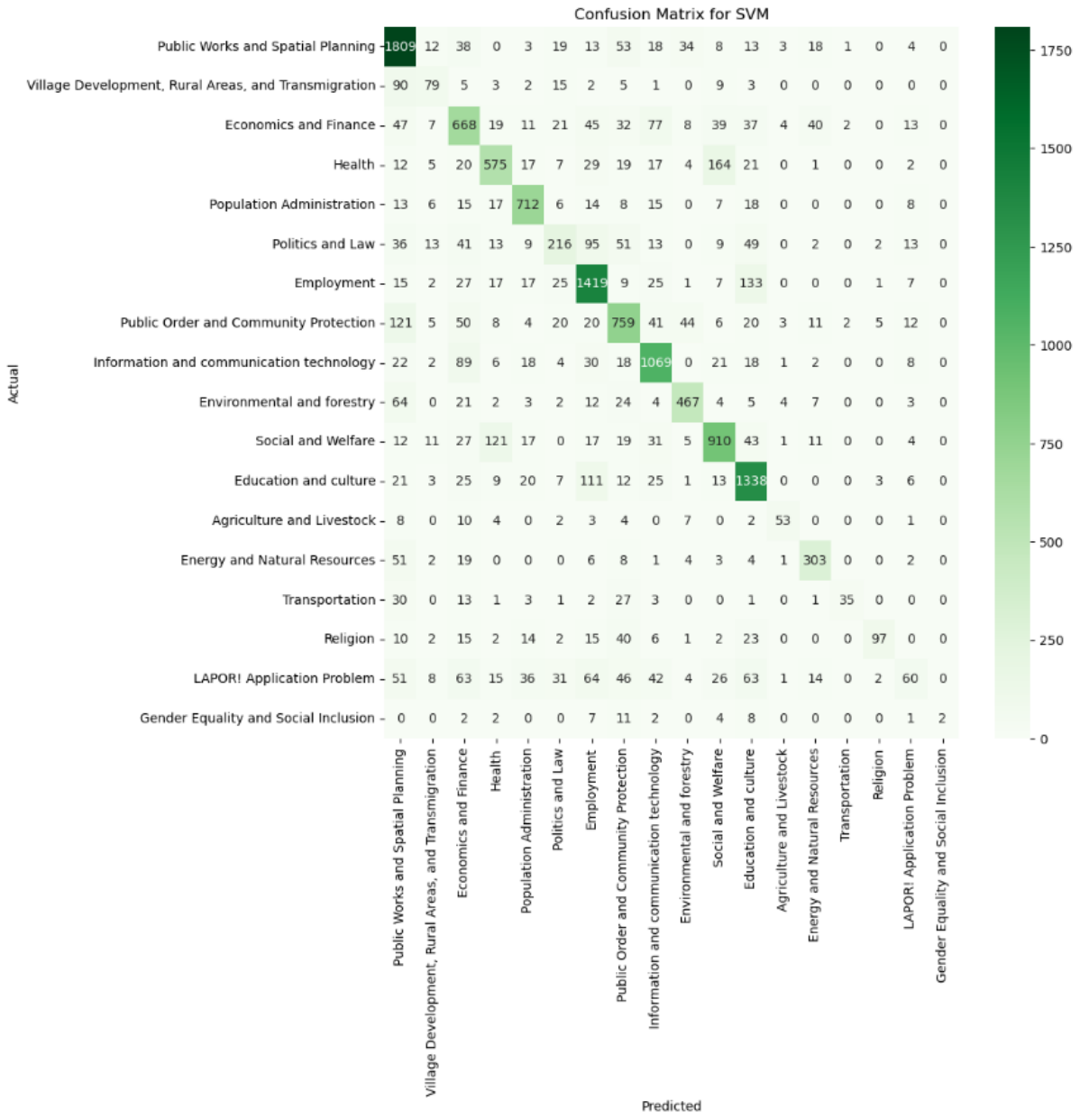

Figure 4: SVM confusion matrix

4.3. Multilayer Perceptron

The Multi-Layer Perceptron (MLP) classifier builds the model to fit the training data through an iterative process involving forward pass, loss computation, backpropagation, and optimisation (Aggarwal, 2015). The MLP model in this research was initiated with two hidden layers, which contain 512 and 256 neurons. During the forward pass, the input features undergo a series of transformations as they traverse through the hidden layers. These transformations are governed by weights, biases, and ReLU (Rectified Linear Unit) activation functions, which introduce non-linearity into the model. The

output layer produces predictions, which are compared to actual labels using a cross-entropy loss function.

Once the loss is computed, backpropagation comes into play, which is a critical phase in training neural networks. In this stage, the gradients of the loss with respect to each weight are calculated. These gradients indicate the direction in which each weight should be adjusted to minimise the loss. The Adam optimizer, a widely-used optimisation algorithm, leverages these gradients to update the model's weights iteratively. This process is performed batch by batch across multiple epochs until the model converges, meaning that the loss shows minimal improvement or stabilises at a specified threshold. This signifies that the model has learned to generalize well on the training data and is ready to make predictions on unseen data. Additionally, to ensure that the model does not overfit, 20% of the training dataset is randomly selected and used for validation during each epoch. This validation helps in monitoring the model's performance on a separate set of data, which is crucial for generalisation.

The final model was achieved after 21 repetitions. The MLP model obtained an overall accuracy of 69% on the test dataset, with a weighted average F1-score of 0.68. While surpassing Naive Bayes in overall performance, the MLP model fell short of SVM in several vital aspects. Similar to SVM, the model demonstrated proficiency in classifying categories such as "Population Administration", "Public Works and Spatial Planning", "Information and Communication Technology", "Employment", "Education and Culture", "Environment and Forestry", "Energy and Natural Resources", and "Social Welfare." However, across these categories, SVM consistently outperformed MLP in terms of evaluation parameters (precision, recall, and f1-score). The only exception was observed in the "Information and Communication Technology" category, where the MLP model achieved slightly higher precision compared to SVM. However, SVM maintained a superior F1-score, indicating a better balance between precision and recall.

The MLP model achieved the highest F-1 score in the "Public Works and Spatial Planning" category. It accurately identified 1,686 complaints in this

category, which represents 82% of the actual complaints. The precision of predictions in this category was 75%, with the highest number of incorrect predictions coming from complaints that were actually in the "Public Order and Community Protection" category (110 complaints). This performance can be attributed to the fact that the "Public Works and Spatial Planning" category had the highest number of instances in the training dataset, with 8,171 data points, allowing the SVM model to learn its characteristics effectively.

Categories	precision	recall	f1-Score	Support
Agriculture and Livestock	0.64	0.63	0.63	94
Economics and Finance	0.61	0.53	0.57	1070
Education and Culture	0.74	0.79	0.76	1594
Employment	0.73	0.8	0.77	1705
Energy and Natural Resources	0.74	0.68	0.71	404
Environment and Forestry	0.77	0.71	0.74	622
Gender Equality and Social Inclusion	0.39	0.18	0.25	39
Health	0.62	0.66	0.64	893
Information and Communication Technology	0.78	0.77	0.77	1308
Politics and Law	0.46	0.42	0.44	562
Population Administration	0.77	0.8	0.78	839
Public Order and Community Protection	0.61	0.62	0.61	1131
Public Works and Spatial Planning	0.75	0.82	0.79	2046
Religion	0.81	0.52	0.64	229
Social and Welfare	0.71	0.69	0.70	1229
SP4N-LAPOR! Application Problems	0.24	0.2	0.22	526
Transportation	0.51	0.35	0.41	117
Village Development, Rural Areas and Transmigration	0.43	0.44	0.44	214
Accuracy			0.69	14622
macro avg	0.63	0.59	0.6	14622
weighted avg	0.68	0.69	0.68	14622

A noteworthy observation from the MLP model's performance is the general balance it maintained between precision and recall across most categories. The most significant imbalance was detected in the "Religion" category, which has the highest precision with 0.81, but recall lagged behind by 2.9 points, resulting in an F1-score of 0.64. Despite these strengths, the MLP model faced significant challenges in certain categories, notably in "Gender Equality and Social Inclusion" and "SP4N-LAPOR! Application Problems." In the "Gender Equality and Social Inclusion" category, the model struggled considerably, correctly predicting only 39% of the instances. The model's recall

in this category was even more concerning, covering just 18% of the total complaints, suggesting a substantial number of relevant instances were missed. Many complaints that actually belonged to this category were misclassified, frequently falling into the "Public Order and Community Protection" category. Nevertheless, despite these difficulties, the F1-score indicated that the MLP model still performed slightly better in this challenging category than other traditional models, reflecting some marginal improvement.

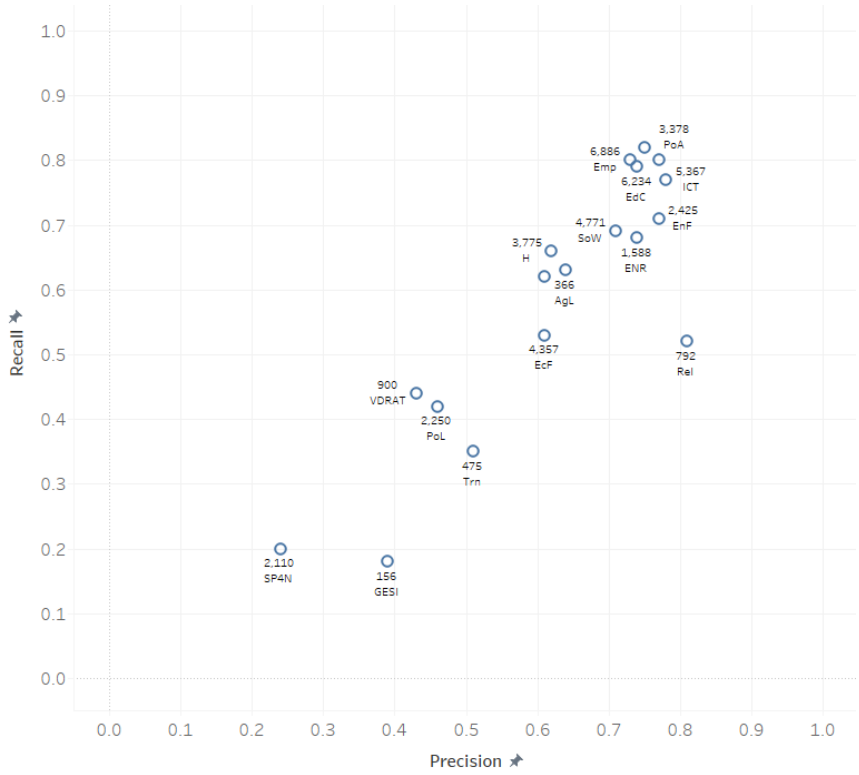

Figure 5: MLP precision, recall, and f1-score

A problematic situation was also exhibited in the "SP4N-LAPOR! Application Problems" category, where the MLP model's performance was particularly poor. Out of 526 instances in this category, the model correctly identified only 106, yielding a mere 20% recall. The remaining instances were misclassified into a wide range of other categories (62 to "Employment", 53 to "Public Works and Spatial Planning", 48 to "Education and Culture", etc.), underscoring the lack of distinctive features that could help the model accurately identify complaints belonging to this category.

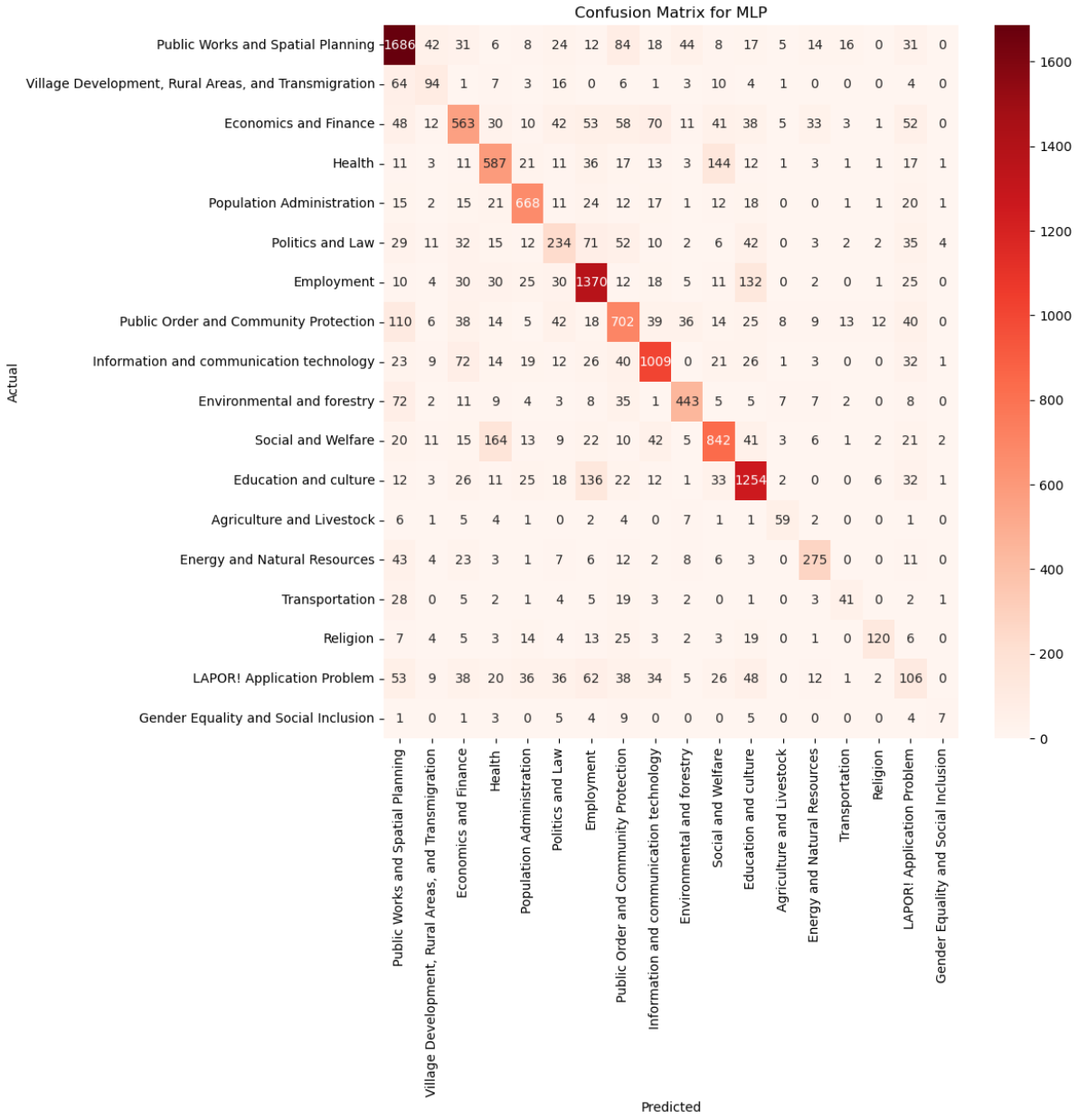

Figure 6: MLP confusion matrix

4.4. Bidirectional Encoder Representations from Transformers

The research employed a fine-tuning approach to adapt a pre-trained BERT language model, specifically IndoBERT, to the text classification task. This process involved iterative model updates through multiple training epochs. During each epoch, the model's parameters were adjusted based on the computed loss gradients to optimise performance. The majority of BERT's layers were frozen to preserve the pre-trained model's linguistic knowledge while tailoring it to the specific classification task, allowing for the training of only the classification layer and the newly added final layers (Devlin et al., 2018).

To fine-tune the model, the AdamW optimiser was employed with a learning rate of 5e-5 and batch size of 8, and the input text length was capped at a maximum of 300 words per instance. The training dataset was passed through the model three times, allowing the model to learn and refine its predictions gradually. This iterative training resulted in the model achieving a classification accuracy of 74% on the test dataset. Additionally, the model exhibited a weighted average F1-score of 0.73, precision of 0.73, and recall of 0.74, underscoring its effectiveness level in the classification task.

However, it is noteworthy that the model's performance was not uniformly distributed across all categories. The macro-average of the F1-score, which provides an average measure across all categories, was lower than the weighted average. This discrepancy highlights the model's tendency to perform better in categories with a higher number of instances. Specifically, the F1-score macro average for the top five categories, in terms of the number of instances, was 0.80, while the bottom five categories had a significantly lower F1-score of 0.54.

Q2

Categories	precision	recall	f1-Score	Support
Agriculture and Livestock	0.62	0.7	0.66	94
Economics and Finance	0.64	0.62	0.63	1070
Education and Culture	0.77	0.85	0.81	1594
Employment	0.79	0.82	0.81	1705
Energy and Natural Resources	0.75	0.83	0.79	404
Environment and Forestry	0.75	0.79	0.77	622
Gender Equality and Social Inclusion	0.33	0.28	0.31	39
Health	0.71	0.71	0.71	893
Information and Communication Technology	0.79	0.8	0.80	1308
Politics and Law	0.51	0.46	0.48	562
Population Administration	0.83	0.86	0.84	839
Public Order and Community Protection	0.68	0.66	0.67	1131
Public Works and Spatial Planning	0.8	0.86	0.83	2046
Religion	0.72	0.63	0.67	229
Social and Welfare	0.74	0.74	0.74	1229
SP4N-LAPOR! Application Problems	0.36	0.13	0.19	526
Transportation	0.59	0.54	0.57	117
Village Development, Rural Areas and Transmigration	0.55	0.48	0.51	214
Accuracy			0.74	14622
macro avg	0.66	0.65	0.65	14622
weighted avg	0.73	0.74	0.73	14622

In terms of specific category performance, the model excelled in nine categories, where the F1-score exceeded the 0.70 threshold. These categories included "Population Administration", "Public Works and Spatial Planning", "Education and Culture", "Employment", "Information and Communication Technology", "Energy and Natural Resources", "Environment and Forestry", "Social and Welfare", and "Health." The highest performance was observed in the "Population Administration" category, where the model accurately predicted 83% of instances assigned to this category and captured 87% of the actual cases, resulting in the highest F1-score among all categories.

The model's performance was generally balanced between precision and recall across most categories with average gap was only 0.055. However, a notable exception was the "SP4N-LAPOR! Application Problems" category, where a significant disparity between precision and recall was observed, with a difference of 0.23. This category also had the lowest F1-score among all classes. The model correctly predicted only 70 out of 526 complaints in this category, with the remaining predictions being incorrectly classified into several other categories, such as "Economics and Finance" (54 complaints), "Education and Culture" (65 complaints), "Employment" (56 complaints), "Public Order and Community Protection" (46 complaints), and "Public Works and Spatial Planning" (50 complaints).

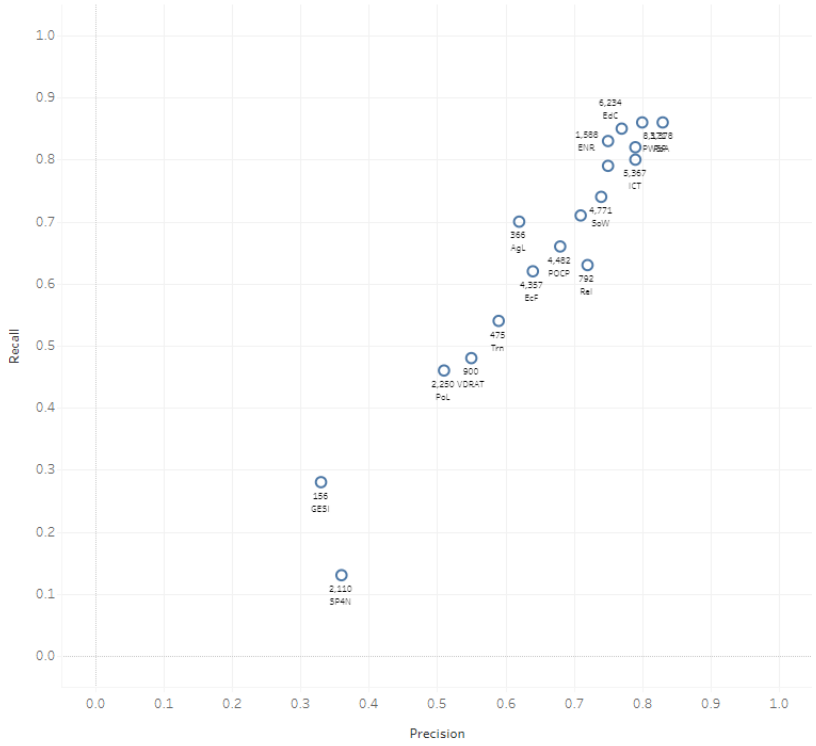

Figure 7: BERT precision, recall, and f1-score

In addition to the "SP4N-LAPOR! Application Problems" category, two other categories—"Politics and Law" and "Gender Equality and Social Inclusion"—also exhibited F1-scores lower than 0.5. The "Politics and Law" category had an F1-score of 0.48, with the model correctly predicting 256 instances out of 526 in this category, which accounted for only 46% of the total actual complaints. This resulted in a precision of 0.51 and a recall of 0.46. The "Gender Equality and Social Inclusion" category had the lowest performance, with a precision of 0.33 and a recall of 0.28. Despite these lower scores, it is important to note that the BERT model's performance in these categories was still the best among all the models that were tested.

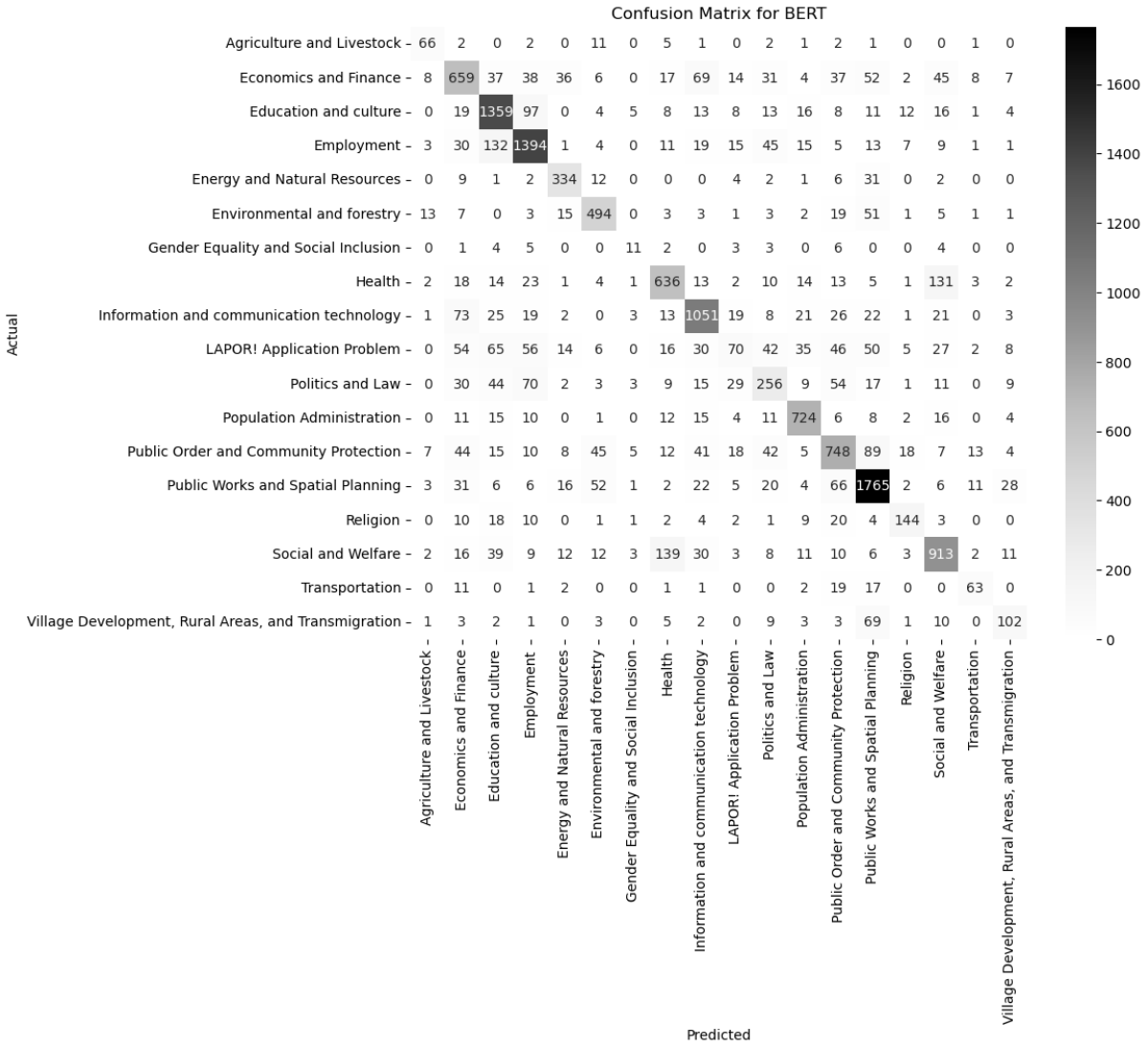

Figure 8: BERT confusion matrix

The confusion matrix above provides a detailed breakdown of the BERT model's performance across all categories, illustrating how well it could differentiate between categories and where it struggled. The matrix reveals that while the model was highly effective in certain categories, there were notable misclassifications in others and mapping where the correct and wrong predictions lie. This analysis highlights the areas where the model could be further refined to improve its overall performance, particularly in categories with lower F1-scores.

5. Discussion:

The results highlight a clear advantage of deep learning models like BERT for complex, real-world NLP tasks in Bahasa Indonesia. While traditional machine learning models are simpler and faster to train, they lack the contextual understanding needed for nuanced text classification.

BERT's ability to capture semantic relationships through attention mechanisms enables it to differentiate even subtle category differences—critical in a dataset like SP4N-LAPOR! where complaints can involve overlapping issues (e.g., public works vs. environmental concerns).

However, computational costs are higher for BERT, requiring greater infrastructure for deployment.

This study also revealed the importance of proper preprocessing, balanced sampling, and tailored evaluation metrics for multi-class text classification projects involving citizen feedback.

6. Conclusion:

Automating the categorization of citizen complaints is both feasible and highly beneficial. It can significantly reduce the burden on administrative staff, minimize human errors, and accelerate the response cycle in public service delivery.

BERT demonstrates superior performance and robustness, suggesting that e-government platforms like SP4N-LAPOR! could greatly benefit from integrating deep learning models. Future work may explore hybrid approaches, ensemble models, or even institution assignment automation to further streamline the complaint resolution process.