Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/1 (2026), 1-9. DOI: http://dx.doi.org/10.21609/jiki.v19i1.1426

Automated Assignment of Community Reports With Early Fusion Multimodal Transformer

Ikhlasul Akmal Hanif[*], Eduardus Tjitrajardja[†], Rahmat Bryan Naufal[‡], Laksmita Rahadianti

Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Email: [*]ikhlasul.akmal@ui.ac.id, [†]eduardus.tjitrahardja@ui.ac.id, [‡]rahmat.bryan@ui.ac.id

Abstract

In the current digital era, city governments require effective and responsive platforms to handle public reports and feedback. One such example is Cepat Respon Masyarakat (CRM) in Jakarta, Indonesia, which allows residents to report various issues to the city government, such as infrastructure damage, traffic accidents, and environmental problems. However, after a report is created, it must be assigned to the appropriate agency. Currently, this assignment process is a challenge, taking an average of nearly two hours. To improve the efficiency and responsiveness of handling public reports through the CRM platform, this research proposes an innovative, multimodal solution for classifying public report data, using both text and images to automatically assign community reports. The proposed method was trained and evaluated using a dataset built from real CRM data. Experiments showed that the multimodal model, using a fusion of the DINOv2 transformer and Multilingual E5 with the Early Fusion method, achieved 80.73% accuracy, an increase from the 68.9% achieved by BERT and ResNet. The results of this research are expected to expedite the issue reporting process and enhance the effectiveness of public services, ultimately contributing to the prosperity of all Indonesian citizens in this era of technological advancement.

Keywords: Public Service, Multimodal Classification, Transformer, Early Fusion

1. Introduction

1.1. Background

In recent years, governments have been continuously striving to be more responsive to public complaints and to improve public services. The public plays a crucial role in identifying issues within a city without requiring the government to conduct direct monitoring, which can be time-consuming and costly. By effectively responding to public complaints, the government not only addresses criticism but also views it as an opportunity for significant improvements [1–4]. In Jakarta, the capital of Indonesia, there is already a system that allows the public to report urban issues they encounter, known as Cepat Respon Masyarakat (CRM)[1], which can be translated as Public Quick Response. Through this system, the public can report issues they encounter by submitting descriptions and supporting images, which are then responded to and followed up by the relevant agencies. It has been recorded that there is an average of four hundred daily users, and more than one million cases have been addressed [5], demonstrating the vital role of this system.

However, based on data analysis conducted by the research team in 3.1, it takes an average of almost two hours to respond to a report on CRM. This delay is due to the need for local authorities to manually determine the appropriate agency to handle each report based on the regulations outlined in the Secretary of the Regional Secretariat's Decree No. 99 of 2022. Response times can be even longer if there are no active local officers, such as at night or on holidays. This poses a challenge, especially for handling reports that require immediate responses, such as cases involving people with mental disorders, or *orang dengan gangguan jiwa* (ODGJ), illegal parking, and other urgent issues.

To address this issue, this study proposes an

[1]https://crm.jakarta.go.id/

automated system to assign reports to the appropriate agencies instantly and accurately in Jakarta's CRM system. By leveraging state-of-the-art transformer models such as DINOv2 (self-**DI**stillation with **NO** labels) for images [6] and Multilingual E5 (**E**mb**E**ddings from bidir**E**ctional **E**ncoder r**E**presentations) for text [7], the system aims to generate meaningful and contextually relevant representations of the text and images in each report. With the integration of this automated system, we believe response times can be reduced, providing better public services to the community and supporting the overall development of the city.

1.2. Research Focus

This study focuses on integrating text and image data within Jakarta's CRM system for accurate and efficient automated report classification. Our primary contributions are twofold: (1) the creation of a novel multimodal dataset of real-world CRM reports from Jakarta, and (2) a comprehensive experimental evaluation exploring various state-of-the-art multimodal models and fusion strategies to determine optimal text and image representations and assess the relative importance of each modality. This evaluation demonstrates the effectiveness of combining text and image information for improved classification accuracy. This approach has the potential for direct implementation on CRM platforms, enhancing public service delivery.

1.3. Research Limitations

The approach presented in this study has several limitations that should be noted:

1) The classification target labels are limited to 8 out of 42 institutions/agencies, based on the highest number of reports, which together account for more than 90% of the total reports.
2) The dataset is only taken from Jakarta's CRM, where each report includes a photo and a description with a minimum of 50 characters and a maximum of 2000 characters.

2. Literature Study

2.1. Text Representation

Encoders using Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) in natural language processing have several limitations.

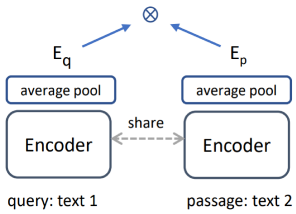


Figure 1. The E5 architecture, where E_p (Embedding passage) and E_q (Embedding query) are compared for their similarity for contrastive learning [16].

RNNs are prone to long-term training issues due to the difficulty in maintaining contextual information in long text sequences. On the other hand, although LSTM addresses this issue by allowing both long-term and short-term memory, the higher computational complexity often poses challenges on a large scale, sometimes resulting in performance that is inferior to simpler approaches like Continuous Bag of Words (CBoW) [8, 9].

Meanwhile, transformer-based models such as Bidirectional Encoder Representations (BERT) [10] and Generative Pre-training Transformer (GPT) [11] have revolutionized natural language processing with the use of self-attention, enabling them to capture more complex relationships between words in text on a large scale in parallel. Additionally, some large-scale encoder-decoder models like mT5 (multilingual Text-To-Text Transfer Transformer) trained on the Cendol dataset have been proven to outperform other large language models of similar size in trials for Indo NLU (Natural Language Understanding) and NLG (Natural Language Generation) [12–15].

E5 (**E**mb**E**ddings from bidir**E**ctional **E**ncoder r**E**presentations) is a state-of-the-art text embedding model that achieves high performance in text representation tasks. This model is trained with weakly-supervised contrastive pre-training on large datasets of text pairs, as shown in Figure 2.1, producing strong single vector representations for information retrieval, clustering, and classification [16]. One of the leading variants of E5 is `multilingual-e5-large-instruct`, which has proven to be highly effective in benchmarks such as the Massive Text Embedding Benchmark (MTEB), optimally supporting multilingual text processing [7, 17].

2.2. Image Representation

Residual Network (ResNet) and Vision Transformer (ViT) are two main approaches to image representation within the fields of machine learning and computer vision. ResNet uses a Convolutional

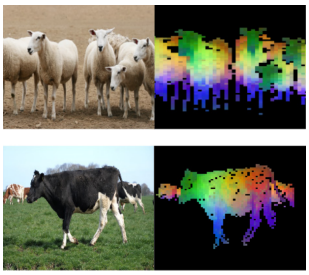


Figure 2. The PCA results of 3 components of the DINOv2 model [6].

Neural Network (CNN) architecture to extract hierarchical features from images with residual connections that facilitate deep network training [18]. In contrast, ViT applies a transformer approach by analyzing image patches as tokens, similar to text processing in transformers [19]. Both models, trained on the ImageNet dataset [20], produce rich image representations useful for multimodal tasks.

DINOv2 is a model that uses self-supervised learning to train ViT by breaking down images into small parts, as illustrated in Figure 2.2, where each component has a specific color, for example, the head components are all colored green. This model allows ViT to understand each part of the image individually and predict the relationships between those parts. The self-supervised learning approach in DINOv2 enables efficient training using unlabeled data, enhancing the generalization of the resulting image representations [6]. With superior performance on ImageNet benchmarks, both in self-supervised and fine-tuning settings, DINOv2 demonstrates its strength in generating broadly applicable image representations.

DINOv2 is a model that uses self-supervised learning to train ViT by breaking down images into small parts, as shown in Figure 2.2, where each component is assigned a specific color, for example, all head components are colored green. This model allows ViT to understand each part of the image individually and predict the relationships between these parts. The self-supervised learning approach in DINOv2 enables efficient training using unlabeled data, enhancing the generalization of the resulting image representations [6]. With superior performance in ImageNet benchmarks, both in self-supervised and fine-tuning settings, DINOv2 demonstrates its strength in generating widely applicable image representations

2.3. Combining Text and Images with Early Fusion

Combining text and images within a single model aims to leverage information from both modalities to enhance performance in tasks that require understanding multimodal contexts. One of the common approaches used for this purpose is Early Fusion, where text and image features are merged at an initial stage before being fed into the predictive model. By integrating the features early in the process, this method allows the model to learn from the combined data in a more cohesive manner, capturing the intricate relationships between text and image content. This approach has proven to be effective in several studies, demonstrating improvements in tasks like image captioning, sentiment analysis, and multimedia content retrieval, where understanding the synergy between textual and visual information is crucial [21–24].

3. Methodology

3.1. Dataset

The reports on Jakarta's CRM website can be accessed publicly. We performed web-scraping to build a dataset of 53,885 public reports submitted throughout 2023. Each data point collected contains information such as the report ID, supporting photos, report description, and the time taken by the local authorities to forward the report to the relevant agency. Each data point also includes the information about the agency that followed up on the report. An example of a data point collected from the website can be seen in Table 1.

The dataset was then processed and cleaned to prepare for model training with a classification target for the eight agencies that received the most reports. Although many agencies are involved in the public report handling in Jakarta, we select only eight agencies whose total number of reports cover more than 90% of the total. The distribution of classification targets for these eight agencies can be seen in Table 2.

Instead of splitting the dataset into training, validation, and testing sets, we perform **stratified-5-fold cross-validation** to ensure a more robust evaluation of the model's performance across different subsets of the data. This approach helps mitigate overfitting and provides a better estimate of the model's generalization ability.

Table 1. Example of one data point.

Image	Report (Indonesian)	Report Translation (English)	Label (Indonesian)	Label Translation (English)
	Air tergenang tidak lancar karena ada bangunan rumah di Kalibanglio yg sangat rendah yg menghalangi aliran air. Mohon ditinjau dahulu.	Water stagnation is not flowing properly due to a very low house building in Kalibanglio blocking the water flow. Please inspect it first.	DINAS SUMBER DAYA AIR	DEPARTMENT OF WATER RESOURCES

Table 2. Target departments for classification.

Department (Indonesian)	Department Translation (English)	Count
Dinas Bina Marga	Department of Public Works	10,526
Satuan Polisi Pamong Praja	Department of Civil Service Police	10,424
Dinas Perhubungan	Department of Transportation	9,351
Kelurahan	Sub-District Office	8,914
Dinas Pertamanan dan Hutan	Department of Parks and Forestry	5,405
Dinas Sumber Daya Air	Department of Water Resources	1,872
Dinas Cipta Karya, Tata Ruang, dan Pertanahan	Department of Public Works and Spatial Planning	1,519
Badan Pembinaan Badan Usaha Milik Daerah	Department of Regional-Owned Enterprises Development	1,403
Instansi lain	Other Departments	4,471

3.2. Metrics

We selected weighted F_1 as the primary evaluation metric for our machine learning models classifying community reports. This choice is particularly relevant given the potential for class imbalance in our dataset. weighted F_1 provides a balanced evaluation of the model's performance by considering both precision and recall for each class and then calculating a weighted average, where the weight is the number of true instances for each class. This approach ensures that the performance on larger classes has a greater impact on the overall score. The calculation involves first computing the F_1 score for each class i:

$$F_1^{(i)} = \frac{2 \cdot \text{Precision}^{(i)} \cdot \text{Recall}^{(i)}}{\text{Precision}^{(i)} + \text{Recall}^{(i)}} \tag{1}$$

Then, the Weighted F_1 score is calculated as the weighted average of these individual F_1 scores:

$$F_1^{\text{weighted}} = \sum_{i=1}^{C} \frac{N_i}{N} \cdot F_1^{(i)} \tag{2}$$

where C is the number of classes, N_i is the number of true instances for class i, and N is the total number of instances.

weighted F_1 offers a more comprehensive evaluation than accuracy alone, as it accounts for both false positives and false negatives across all classes, particularly in cases of class imbalance. This metric is especially useful when the dataset exhibits unequal class distributions, as it emphasizes the model's ability to perform well across all classes, giving more importance to larger classes. Using weighted F_1 ensures a more accurate assessment of the model's overall classification performance in the presence of class imbalance.

3.3. Experimental Design

The experimental flow is depicted in Figure 3.3. The process began with data collection, acquisition, and cleaning, as detailed in Section 3.1. Subsequently, as explained in Sections 2.2 and 2.3, text and image data were transformed into embedding vectors.

The pre-trained weights for the text and image models, detailed in Tables 3 and 4, respectively, are all open-source and accessible via Hugging Face.

The embedding process is defined as follows: Given a text input x_{text} and an image input x_{image}, separate encoders generate embedding vectors e_{text} and e_{image}, respectively. These encoders are pre-trained models with frozen parameters during training to preserve their learned feature representations,

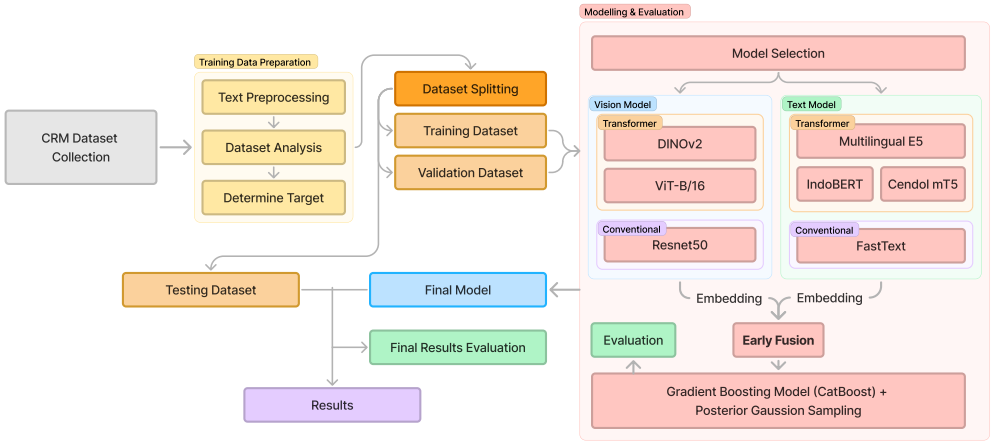


Figure 3. Experimental design.

Table 3. Model details for images.

Model Name	Model Checkpoint
ResNet50	`microsoft/resnet-50`
ViT-B/16	`google/vit-base-patch16-224-in21k`
DINOv2	`facebook/dinov2-large`

Table 4. Model details for text.

Model Name	Model Checkpoint
FastText	`facebook/fasttext-id-vectors`
IndoBERT	`cahya/bert-base-indonesian-522M`
Cendol MT5	`indonlp/cendol-mt5-large-inst`
Multilingual E5	`intfloat/multilingual-e5-large-instruct`

ensuring that the embeddings remain fixed while the classifier is trained.

$$e_{\text{text}} = f_{\text{text}}(x_{\text{text}}) \tag{3}$$

$$e_{\text{image}} = f_{\text{image}}(x_{\text{image}}) \tag{4}$$

The embeddings from both modalities are then combined using early fusion. The combined embedding e_{combined} is obtained by concatenating the text and image embeddings:

$$e_{\text{combined}} = \text{concat}(e_{\text{text}}, e_{\text{image}}) \tag{5}$$

This combined embedding is then input to the classifier. In this experiment, we used CatBoost with default parameters for classification. To address potential class imbalance, we employed weighted class settings. The weight w_l for each class l is calculated as:

$$w_l = \frac{N}{C \cdot N_l} \tag{6}$$

where N is the total number of instances, C is the number of unique classes in the `dinas_final` column, and N_l is the number of instances associated with class l. As shown in Equation (6), the class weight depends on the total number of instances, the number of classes, and the number of instances in each class.

Finally, the trained model's performance was evaluated using accuracy and other relevant metrics on the test dataset to determine the best-performing model.

4. Experiment Results and Analysis

4.1. Multimodal Model Comparison

Initially, we built a baseline multimodal model by combining more conventional models. Subsequently, to improve performance, further experiments were conducted using transformer-based models. Table 5 presents the weighted F_1 scores achieved by training with the Early Fusion method using the CatBoost model with various combinations of image and text models.

The experimental results demonstrate that combinations of conventional models, such as the ResNet50 + FastText multimodal model, achieved an weighted F_1 score of only 0.6903. In contrast, using combinations of transformer models yielded significantly higher weighted F_1 scores. As shown in Table 5, the DINOv2 + Multilingual E5 multimodal model achieved the highest weighted F_1 score of

Table 5. Model weighted F_1 Scores for image + text.

Model	**Weighted** F_1
ResNet 50 + IndoBERT	0.6892
ResNet 50 + FastText	0.6903
ResNet 50 + Cendol mT5	0.6972
ResNet 50 + Multilingual E5	0.7941
ViT-B/16 + IndoBERT	0.6911
ViT-B/16 + FastText	0.6974
ViT-B/16 + Cendol mT5	0.7049
ViT-B/16 + Multilingual E5	0.7965
DINOv2 + IndoBERT	0.7349
DINOv2 + FastText	0.7381
DINOv2 + Cendol mT5	0.7423
DINOv2 + Multilingual E5	**0.8073**

0.8073. This combination consistently outperformed the other combinations, indicating the superiority of transformer models, particularly DINOv2 and Multilingual E5, in capturing relevant and robust features and representations from image and text data, as explained in Sections 2.1 and 2.2.

DINOv2's exceptional performance can be attributed to its strong generalization capabilities, which stem from its semi-supervised training approach. Furthermore, Multilingual E5 is trained with objectives that enhance its performance in robust text understanding and classification tasks, including scenarios resembling zero-shot classification. This allows the model to function as an effective classifier without requiring extensive task-specific fine-tuning, further contributing to the superior performance of the multimodal model.

Conversely, models such as ResNet and BERT, while powerful in their respective domains, may lack sufficient generalization as standalone encoders. These models typically require fine-tuning to achieve competitive performance, as they are not inherently optimized for general-purpose feature extraction. This limitation likely contributes to their comparatively lower weighted F_1 scores when used as components in multimodal frameworks.

4.2. Ablation Test

Table 6 presents the results of the ablation test, evaluating the performance of different models using text-only, image-only, and combined modalities. Among the single-modality models, Multilingual E5 achieves the highest weighted F_1 score (0.7893), while DINOv2 achieves 0.6270. This difference highlights the greater relevance of text features for this classification task, as text often contains explicit information directly correlated with the target labels.

Table 6. Ablation Test: Single Modality (Text or Image) and Multimodal (Image + Text) Models. **Modality Key**: **T** refers to text embeddings, **I** refers to image embeddings, and **T+I** combines both image and text embeddings.

Model	**Modality**	**Weighted** F_1
DINOv2	I	0.6270
ResNet-50	I	0.5185
ViT-base	I	0.5125
Cendol-mT5	T	0.6613
FastText	T	0.6472
IndoBERT	T	0.6410
Multilingual E5	T	0.7893
DINOv2 + Multilingual E5	**T+I**	**0.8073**

Image-based models face additional challenges due to their reliance on contextual interpretation.

4.3. Error Analysis

The combination of DINOv2 and Multilingual E5 achieves an weighted F_1 score of 0.8073, demonstrating the benefit of integrating image and text modalities. While the improvement over the text-only Multilingual E5 model is relatively small, the multimodal approach effectively addresses cases where text information is ambiguous or insufficient. For instance, as shown in Table 7, the text "Tolong dibenerin atau diikat bahaya kalo kena orang" (Please fix or tie it; it's dangerous if it hits someone) does not specify the issue, causing the text-only model to misclassify it as "DINAS SUMBER DAYA AIR" (Water Resources Agency) instead of "DINAS BINA MARGA" (Public Works Agency). The inclusion of image features in the multimodal approach resolves this misclassification by providing crucial context.

The performance gain from multimodal fusion varies depending on the strength of the underlying text model. Stronger text models, such as Multilingual E5, see smaller gains from incorporating images compared to weaker text models such as IndoBERT. For example, IndoBERT's weighted F_1 score of 0.6410 improves to 0.7349 when combined with DINOv2, a larger relative improvement than the increase from 0.7893 to 0.8073 observed with Multilingual E5.

These findings suggest that while text-based models are effective at capturing explicit information, the multimodal approach enhances performance by providing complementary information, particularly in cases of ambiguity. This reinforces the value of leveraging both modalities for robust classification.

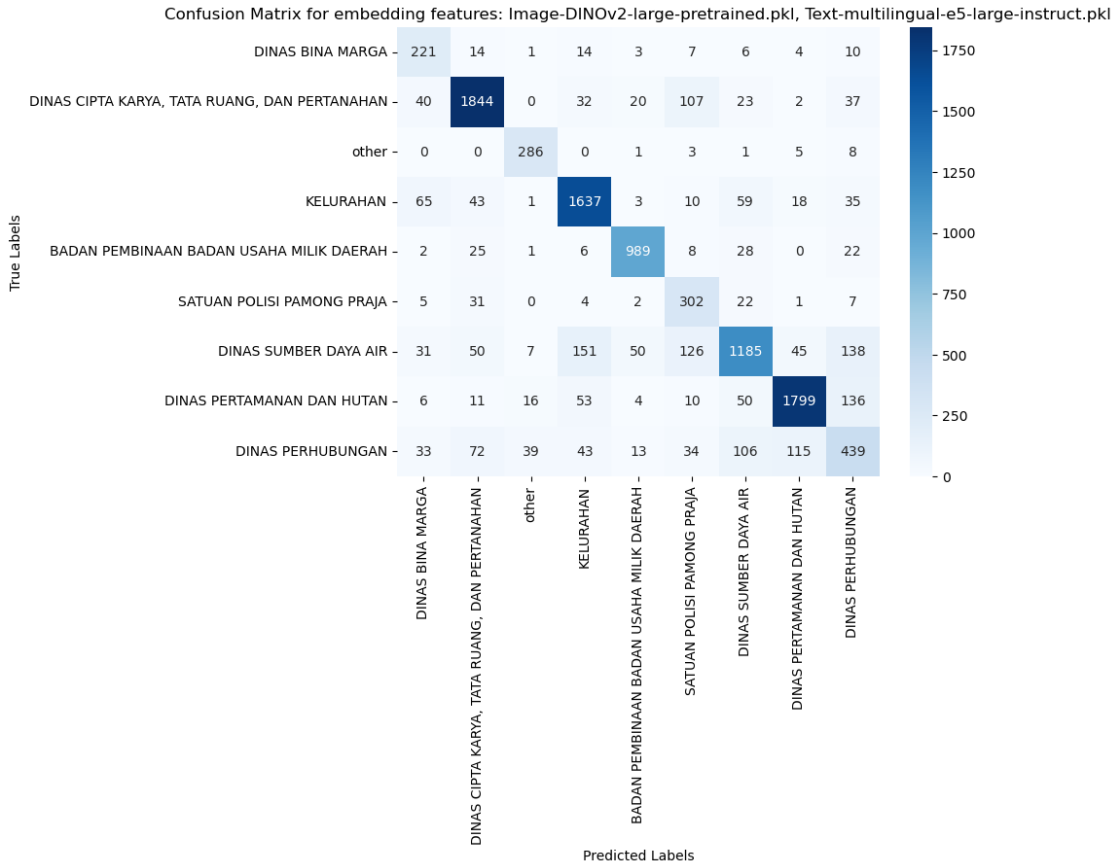


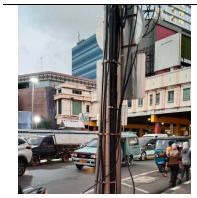
Figure 4. Confusion matrix of DINOv2 + Multilingual E5.

Figure 4.3 shows that the model encountered the most difficulty with the ”DINAS PERHUBUNGAN” (Department of Transportation) class. For this class, the model achieved a precision of 0.5158, meaning that 51.58% of instances predicted as ”DINAS PERHUBUNGAN” were correctly classified. The recall for this class was 0.5277, indicating that 52.77% of actual ”DINAS PERHUBUNGAN” instances were correctly identified.

We now analyze classification errors that occurred when using only text features and demonstrate how these errors were corrected by incorporating image features.

In the example shown in Table 7, the text-only model classified the report as ”DINAS SUMBER DAYA AIR” (Department of Water Resources), while the multimodal model correctly classified it as ”DINAS BINA MARGA” (Department of Public Works). This demonstrates that text information alone can be insufficient. In this specific case, the text is ambiguous, requiring additional information from the image for accurate classification.

Table 7. Example of an error when using only text.

Image	Report	Ground Truth
	Tolong dibenerin atau diikat bahaya kalo kena orang. (Please fix or tie it, it’s dangerous if it hits someone.)	DINAS BINA MARGA (Public Works Agency)

The integration of image representations with DINOv2 is crucial due to DINOv2’s ability to effectively represent images, as illustrated by the PCA visualization of image vector representations in Figure 5 (related to the image in Table 7). As shown in Figure 5, DINOv2 highlights the poles in the image, thereby assisting the classifier in making more accurate predictions.

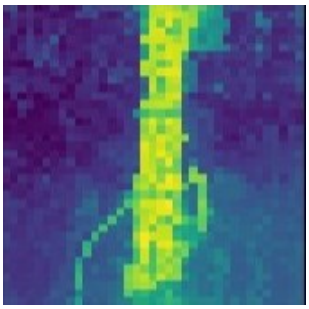


Figure 5. PCA visualization with 3 components from the DINOv2 model applied to the image in Table 7.

5. Conclusion

This study demonstrates that using Multilingual E5 for text representation and DINOv2 for image representation achieved the highest accuracy of 80.73%, outperforming the combination of BERT and ResNet, which reached only 68.9%. This 11.83% performance improvement highlights the effectiveness of using DINOv2 for image representation and Multilingual E5 for text representation. The multimodal approach, by combining these advanced models, delivers more accurate predictions, demonstrating clear advantages over traditional models like BERT and ResNet. This superior performance, evaluated through cross-validation, suggests that the proposed method is more suitable for tasks such as automatic assignment systems on CRM platforms, where accuracy and efficiency are critical.

References

[1] M. Hume, K. Hobart, L. Briz, S. Amara, S. D. Cleary, and P. J. Candilis, "Ethics oversight in psychiatry: Data from a model of organizational monitoring," *Psychiatric Clinics of North America*, vol. 44, no. 4, pp. 563–570, 2021, ethics in Psychiatry.

[2] A. Singh, S. Saha, M. Hasanuzzaman, and A. Jangra, "Identifying complaints based on semi-supervised mincuts," *Expert Systems with Applications*, vol. 186, p. 115668, 2021.

[3] I. Siret and W. Sabadie, "Public complaining: A blessing in disguise? educational calling as a benevolent process that gives consumers voice on brands' social media," *Journal of Business Research*, vol. 150, pp. 476–490, 2022.

[4] A. Filip, "Complaint management: A customer satisfaction learning process," *Procedia - Social and Behavioral Sciences*, vol. 93, pp. 271–275, 2013, 3rd World Conference on Learning, Teaching and Educational Leadership.

[5] [Online]. Available: https://crm.jakarta.go.id/data-harian

[6] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[7] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," *arXiv preprint arXiv:2402.05672*, 2024.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://aclanthology.org/Q17-1010/

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] S. Cahyawijaya, H. Lovenia, F. Koto, R. Putri, W. Cenggoro, J. Lee, S. Akbar, E. Dave, N. Nuurshadieq, M. Mahendra, R. Putri, B. Wilie, G. Winata, A. Aji, A. Purwarianti, and P. Fung, "Cendol: Open instruction-tuned generative large language models for Indonesian languages," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 14 899–14 914. [Online]. Available: https://aclanthology.org/2024.acl-long.796/

[13] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and

A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, K.-F. Wong, K. Knight, and H. Wu, Eds. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. [Online]. Available: https://aclanthology.org/2020.aacl-main.85

[14] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8875–8898. [Online]. Available: https://aclanthology.org/2021.emnlp-main.699/

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[16] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.03533*, 2022.

[17] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. [Online]. Available: https://aclanthology.org/2023.eacl-main.148

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[22] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, 2020, pp. 1–6.

[23] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *2005 IEEE international conference on systems, man and cybernetics*, vol. 4. IEEE, 2005, pp. 3437–3443.

[24] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.