Instagram influencer classification using fine-tuned BERT model

Ni Putu Sutramiani, Ni Made Dita Dwikasari, I Nyoman Prayana Trisna, I Wayan Agus Surya Darma
Department of Information Technology, Faculty of Engineering, Udayana University, Bali, Indonesia

Article Info

Article history:

Received May 17, 2025
Revised Jan 7, 2026
Accepted Jan 22, 2026

Keywords:

BERT
Fine-tuning
Instagram
Pre-trained model
Social media

ABSTRACT

Influencer marketing has emerged as a powerful strategy in today's digital world, where social media stars can influence how people think about products. However, the rapid growth of influencers and social media users presents novel challenges for brands in identifying suitable influencers for their marketing goals. Traditional approaches that rely on popularity and follower count are no longer the primary metrics for determining an influencer's ability to affect consumer behavior. To address this gap, this study proposed an influencer classification to enhance audience targeting and marketing effectiveness. By utilizing deep learning, specifically fine-tuned bidirectional encoder representations from transformers (BERT), influencer classification was carried out for Instagram users in Indonesia based on their post captions. The multilingual BERT model is optimized through hyperparameter tuning, including learning rate, batch size, and stop word removal variation. With an outstanding 80% accuracy, the model performs best in situations where stop words are not removed. This study on influencer classification using a fine-tuned BERT model has demonstrated the effectiveness of BERT in enhancing influencer selection. It contributes to the digital marketing domain by showcasing the potential of deep learning for social media analysis and content classification, paving the way for future data-driven marketing strategies.

Corresponding Author:

I Nyoman Prayana Trisna
Department of Information Technology, Faculty of Engineering, Udayana University
St. Kampus Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
Email: prayana.trisna@unud.ac.id

1. INTRODUCTION

Technological advancements and shifts in consumer behavior have led to significant transformations in modern marketing strategies. One notable innovation is the emergence of influencer marketing, which has completely changed the aspect of digital marketing by combining the power of word-of-mouth (WoM) marketing and the expansive reach of social media. At the present time, influencer marketing is the top choice of many brands in engaging their target audiences. Influencers are opinion leaders who have the ability and willingness to sway consumer decision-making across social media platforms like YouTube, TikTok, and Instagram [1]. Instagram is the most popular platform for influencers among these platforms [2], providing an excellent opportunity for brands to increase awareness and visibility of their products and services. As a result, influencer marketing is currently one of the best marketing strategies, considering that modern consumers trust recommendations from people they follow and admire on social media more than conventional advertising [3].

The rise of influencers and social media users creates a multifaceted situation for brands. While it offers exciting opportunities, this has left most brands facing new challenges in choosing the right influencers who align with their marketing objectives. Most traditional influencer selection methods are usually based on the number of followers or popularity [4]. However, given today's audiences and consumer behavior diversity, it turns out that relying solely on the number of followers does not accurately reflect an influencer's real potential towards the target market [5]. Moreover, the conventional methods for influencer classification often rely on some simple metrics or manual analysis. All of these take time and may not be accurate. In light of these circumstances, brands and marketers are demanding more targeted and sophisticated approaches to identify those influencers who really driving consumers.

Most influencers tend to have a specific niche or market segment in which they are interested or experts, such as fashion, beauty, gaming, travel, and other niches that represent the social media content they share. These niches give brands the opportunity to reach their target audience through these relevant influencers [6]. Identifying influencer categories will help both brands and marketers in targeting appropriate audiences since the relevance between influencers and the advertised products also affects the effectiveness of marketing campaigns [7]. The right influencer can boost the effectiveness of a marketing campaign, while the wrong one could lead to minimal impact. It follows that identifying influencer categories is essential for marketers to ensure that they are working with the right influencers.

Previous studies have explored the opportunities for enhancing influencer marketing by employing machine learning and deep learning techniques [8]–[11]. A more common approach among the different strategies proposed in the literature is text classification through social media, where text information within users' posts can contribute as a valuable source. The content of influencer posts, including the captions they write, can be better analyzed to provide marketers insights into the interests of the influencer's audience. Consequently, findings relevant influencers will positively impact brand awareness and purchase intentions, as the content shared by influencers is likely to draw attention and effectively influence audiences due to its significant informative value [12].

Text classification is the process of categorizing text into one or more predefined classes according to their respective subjects. Text classification through social media can be realized with different modalities, one of which is natural language processing (NLP). In the last couple of years, the most recent one in the line of NLP is using bidirectional encoder representations from transformers (BERT), a model proposed by Google in 2018 [13]. With the ability to understand contextual relationships between words [14], BERT represents a transformer-based architecture that has achieved state-of-the-art results in a variety of NLP tasks, including text classification. BERT's effectiveness in text classification has been well-established across various domains [15]–[17]. According to Merchan *et al.* [18], BERT also outperformed other traditional machine learning methods in text classification, including logistic regression, support vector classifier (SVC), multinomial NB, H2OAutoML, and a few other models.

In regard to the social media influencer classification, studies by Kim *et al.* [11] have proven BERT's excellent performance. The study employed a multimodal approach that combined text and image data using pre-trained BERT and Inception-v3 models, with a post-attention layer to calculate the score of each post in describing the category of the influencer. This approach has a very good performance in influencer profiling. Another study contrasts word embeddings, GloVe and FastText, with fine-tuned BERT, highlighting the superiority of BERT in classifying Instagram user interests based on hashtags and captions, achieving accuracies of 96% and 91%, respectively [19].

Building upon the promising results in [11], [19], this study proposes a fine-tuned BERT model for classifying Instagram influencers in Indonesia based on their post captions. We propose to use multilingual BERT, trained under various hyperparameter tuning scenarios, to find the best performance models. The classification process involves processing the caption text of the influencer's post using the BERT model, which then categorizes influencers into relevant groups based on the result. The proposed method aims to simplify the identification and influencer search that are aligned with the brand marketing objectives. This research is expected to serve as a reference for implementing marketing strategies using deep learning technology. From an academic perspective, the BERT model is expected to classify influencers optimally and can be developed for other digital marketing strategies.

2. METHOD

The study is divided into five stages. Data acquisition, data labeling, pre-processing, fine-tuning the BERT model, and influencer classification. The entire workflow of this study is shown in Figure 1.

2.1. Data acquisition

The dataset we used in this study was obtained through scraping on several Instagram accounts of Indonesian influencers. First of all, we gathered a list of influencer usernames, which was sourced from searches on various platforms, including Instagram and Google. These usernames served as parameters for collecting influencer post data. Further, the Instaloader library was used to scrape Instagram posts to collect data on influencer posts. The datasets acquired comprised 7,500 posts for model building and 19,842 posts for classification tasks, collected from the latest 50 posts of 150 and 400 influencers, respectively. The data obtained included essential text information, such as username, post date, post URL, and caption.

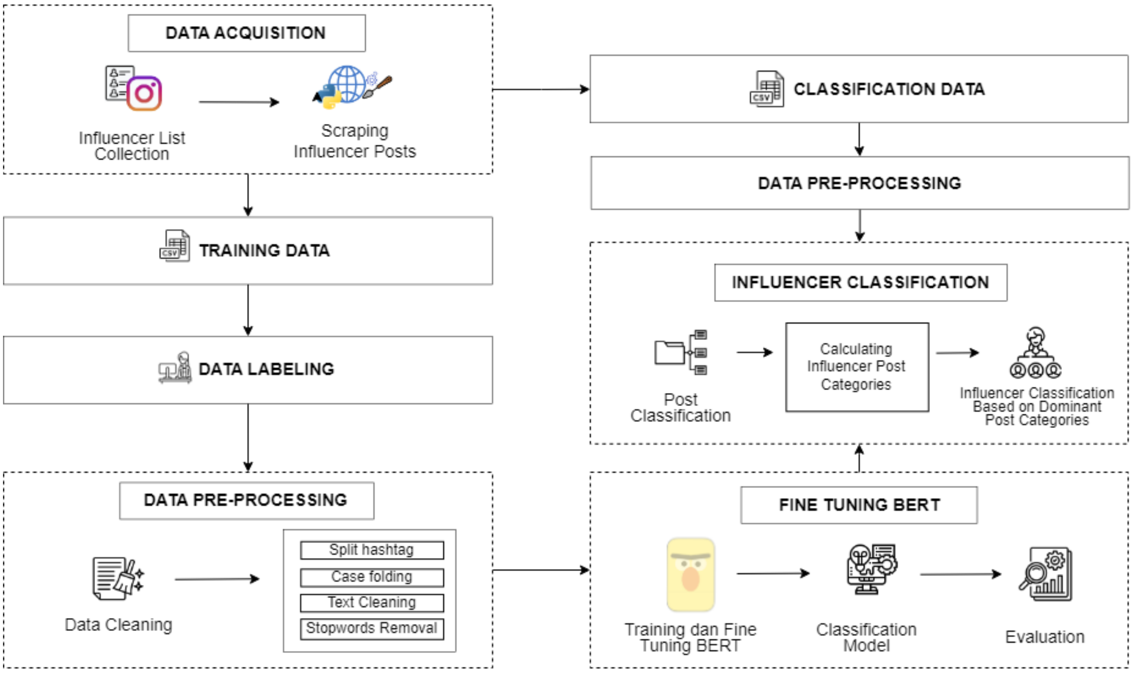

Figure 1. Research overview

2.2. Data labeling

The classification task in this study is a form of supervised learning, where the model is trained to predict the correct label based on input features. This process utilizes a dataset composed of feature-label pairs [20]. Labeling is a critical phase in this study, as it provides the necessary labels for training the model and performing classification. We manually labeled 7,500 posts from 300 influencers, by considering the context of each post caption. The labels were determined based on the top five categories of the most popular types of content produced by Indonesian influencers according to AnyMind: fashion and beauty, entertainment, food, travel, and games and gadgets [21], with one other category for posts that did not fit into these five categories.

2.3. Pre-processing

Scraping results typically yield raw data that requires further processing for efficient analysis. Therefore, the datasets were pre-processed, which included deleting posts without captions, splitting hashtags, case folding, cleaning, and stop words removal. Following the removal of empty entries, English hashtags containing more than one word are segmented using word segmentation from the Ekphrasis library [22] to enhance semantic accuracy. Lowercasing is then done to convert all uppercase letters to lowercase, and cleaning is done to remove URLs, numbers, special characters, or other irrelevant elements. The last step is stop word removal. In this study, we performed two pre-processing scenarios: with and without stop word removal. The removal of stop words depends on language detection to specify an appropriate stop word dictionary, either in Indonesian or English, considering the dataset’s linguistic diversity. Table 1 illustrates examples of the pre-processing data carried out in this study.

2.4. Fine tuning

BERT has become the state-of-the-art for variety tasks such as question answering, language understanding, classification, and other NLP tasks, without substantial changes in the model architecture. BERT is essentially a stack of transformer encoder layers [23] with multiple self-attention “heads”. These heads work together to generate deep bidirectional representations of unlabeled text by observing the context of each layer on both sides (left and right) [13].

This paper uses the multilingual BERT model, specifically the bert-base-multilingual-cased from Hugging face transformers [13]. Multilingual BERT was chosen since it has been pre-trained on 104 languages using the Wikipedia datasets, which include Indonesian and English. It has portrayed excellent capability on cross-language generalization, including code-switching [24]. As a result, this model can be applied to datasets with a wide range of language variations, as is often found in Instagram captions where more than one language is used. The model was then fine-tuned with a new dataset to understand better the context of words appearing in Instagram posts, which enabled it to identify and group influencer posts into their appropriate categories.

Fine-tuning of the BERT model initiates by splitting the labeled datasets into three subsets: training set, validation set, and test set. This shall be done in an 80:10:10 ratio, that is, 80% for training data, 10% for the validation data, and the remaining 10% for test data. The study also experimented with hyperparameter adjustments, including learning rate and batch size, and different stop word removal strategies applied to the data to find the best models. According to Table 2, hyperparameter tuning was tested using these adjustments in 6 scenarios, including variations in batch sizes: 16 and 32, and learning rate: 2e-5, 3e-5, and 5e-5 according to the original BERT paper [13]. Furthermore, an additional experiment is conducted by utilizing stop word removal, resulting in a total of 12 experiments from 6 scenarios. This aims to find the most optimal hyperparameter combination while noticing the impact of using stop word removal on performance for the BERT model. Finally, a linear classification layer is added on top of the BERT output to produce the number of outputs according to the labels used.

The training sessions are defined to run for 25 epochs per scenario, with early stopping used to prevent overfitting. Then, these models will be evaluated using several metrics such as accuracy, precision, recall, and F1-score to determine the best model. The final performance evaluation is conducted on the test data, and the model with the best performance will then be applied to the unlabeled data.

Table 1. Example of pre-processing data

Language	Caption	Pre-processing phase	Clean caption
English	What happened in Nepal? Full episode available in YouTube. Link in bio 👆 like, share and comment please 🙏 #nepal #jungle #hikingadventures #travelstories #indonesianbackpacker #backpacker #indonesia	Split hashtag	What happened in Nepal? full episode available in YouTube. Link in bio 👆 like, share and comment please 🙏 nepal jungle hiking adventures travel stories indonesian backpacker backpacker indonesia
		Case folding	what happened in nepal? full episode available in youtube. link in bio 👆 like, share and comment please 🙏 nepal jungle hiking adventures travel stories indonesian backpacker backpacker indonesia
		Text cleaning	what happened in nepal full episode available in youtube link in bio like share and comment please nepal jungle hiking adventures travel stories indonesian backpacker backpacker indonesia
		Stop word removal	happened nepal full episode available youtube link bio like share comment please nepal jungle hiking adventures travel stories indonesian backpacker backpacker indonesia
Indonesian	Available now: 20.08.23 ✅ *Koleksi jaket* #KembaliMerdeka *sudah rilis secara* official *di* website cardinal.co.id 🌐 *Cuma ada 50pcs yang bisa kamu dapatkan di* link *ini:* https://shorturl.at/dxBDS Get yours! ID 🔥	Split hashtag	Available now: 20.08.23 ✅ *Koleksi jaket kembali merdeka sudah rilis secara* official *di* website cardinal.co.id 🌐 *Cuma ada 50pcs yang bisa kamu dapatkan di* link *ini*: https://shorturl.at/dxBDS Get yours! ID 🔥
		Case folding	available now: 20.08.23 ✅ *koleksi jaket kembali merdeka sudah rilis secara* official *di* website cardinal.co.id 🌐 *cuma ada 50pcs yang bisa kamu dapatkan di* link *ini*: https://shorturl.at/dxbds get yours! ID 🔥
		Text cleaning	available now *koleksi jaket kembali merdeka sudah rilis secara* official *di* website cardinal co id *cuma ada pcs yang bisa kamu dapatkan di* link *ini* get yours
		Stop word removal	available now *koleksi jaket merdeka rilis* official website cardinal co id *cuma pcs kamu dapatkan* link get yours

Table 2. BERT hyperparameter tuning scenario

Scenario	Batch size	Learning rate
1	16	2e-5
2	16	3e-5
3	16	5e-5
4	32	2e-5
5	32	3e-5
6	32	5e-5

2.5. Influencer classification

To classify influencers based on their content, a fine-tuned multilingual BERT model was first used to categorize unlabeled posts from influencers. Then, a simple algorithm is used to assign a dominant category for each influencer. This algorithm calculates the frequency of the pre-defined categories based on the BERT model's post classifications, and the category with the highest frequency will be considered as the influencer's primary niche. It is important to note that the other category was excluded as it lacked a specific focus and limited insights into influencer and audience preferences. The focus remains on five well-defined categories: fashion and beauty, entertainment, food, travel, and games and gadgets.

3. RESULTS AND DISCUSSION

This section will explain the results of model selection, model evaluation, and influencer classification that have been carried out in this study. Model selection is the first step in determining the best model on training and validation data, while model evaluation examines the performance of the best previously selected model on test data. Finally, the best models are applied to unlabeled influencer data. The discussions of each are presented as follows.

3.1. Model selection

Several multilingual BERT training scenarios were run with different parameter combinations to find the best model with regard to an influencer post classification task. This has been performed for 12 experiments depending on the parameters described in Table 2. Table 3 shows the experimental results of BERT performance on validation data for each fine-tuning scenario. According to the findings, model scenario 5 with batch size of 32 and a learning rate of 3e-5 performs the best on data without stop word removal. This model reached the most outstanding performance compared to other scenarios, with an accuracy of 80% and a loss of 0.029.

Table 3. Comparison of BERT performance on validation data

Scenario	Batch size	Learning rate	Stop words removal		Without stop words removal	
			Loss	Accuracy (%)	Loss	Accuracy (%)
1	16	2e-5	0.065	77.02	0.060	76.61
2	16	3e-5	0.066	77.02	0.062	77.84
3	16	5e-5	0.068	75.10	0.063	76.74
4	32	2e-5	0.034	75.38	0.031	76.88
5	32	3e-5	0.035	76.06	**0.029**	**80.03**
6	32	5e-5	0.031	74.97	0.029	79.07

Furthermore, there was a general trend that models trained on a batch size of 32 had significantly different loss values compared to those trained with a batch size of 16. Another thing to be observed is there were significant differences in performance between models trained on data with and without stop word removal. The findings show that models trained on data without stop word removal consistently outperformed models trained on data with stop word removal. Usually, in classification tasks, stop words will be removed in order to allow focus on those words that are more important to the text, to improve text analysis with computational efficiency [25]. However, in this case, it was found that removing stop words slightly dropped the performance of the BERT model compared to training from completely original text without stop word removal. The reason may be related to the mechanism of the BERT, which requires the context of the whole sentence [14] in order to understand the proper sense of it. It means that removing stop words reduces the amount of context and, correspondingly, the model's accuracy becomes less in text classification.

3.2. Model evaluation

Evaluation on the best model was conducted to assess the model's performance in classifying text into the appropriate categories. The evaluation process was performed on 731 test data, using confusion matrix metrics such as accuracy, precision, recall, and F1-score. The results of this evaluation are shown in Table 4.

Table 4. Model performance evaluation

Labels	Accuracy	Precision	Recall	F1-score
Fashion and beauty	80%	0.82	0.85	0.84
Entertainment		0.63	0.45	0.53
Food		0.83	0.93	0.88
Travel		0.78	0.88	0.82
Games and gadget		0.92	0.81	0.86
Others		0.77	0.72	0.74

The evaluation results showed that the model maintained good performance with an accuracy of 80% on the test data, suggesting that the model is quite effective at classifying the majority of text captions into the appropriate categories. Furthermore, based on the precision, recall, and F1-score metrics used to evaluate the model's performance for each category, variations in the model performance were found in each category. Based on these findings, we conducted a qualitative analysis of the model's performance for each category. Figure 2 shows the confusion matrix, that compares the model's predicted values with the actual test data values.

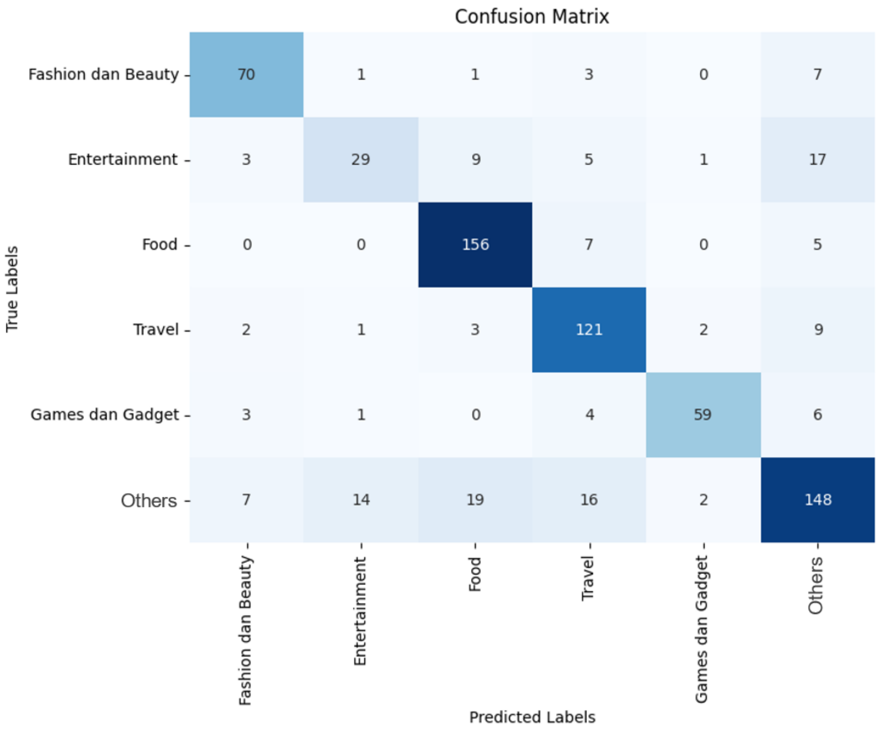

Figure 2. Confusion matrix of model classification results

When examining each category, "food" and "games and gadget" performed the best, with the majority of accurate predictions, resulting in high accuracy, precision, recall, and F1-scores. This indicates that the model is very effective at recognizing and classifying posts related to food and gadgets. However, some categories showed lower performance and inaccurate labels, affecting the overall model accuracy. Examples of incorrect model predictions on test data can be seen in Table 5.

The confusion matrix in Figure 2 highlights significant challenges in the classification process. It shows that the "entertainment" category experiences a high rate of false negatives and significant confusion with the "others" category, indicating that the model struggled to accurately identify and distinguish within these types of content. Entertainment posts frequently lack explicit captions and instead rely on visual content with captions that use common words without clear patterns, as shown in several examples in Table 5. Moreover, because entertainment is subjective and depends on individual perception, it may be difficult to fully describe this content using just words. Conversely, the "others" category had issues with both high false negatives and false positives, showing that the model also struggled to classify content in this category. This category covers a wide range of topics, making it more challenging for the model to classify them correctly. The overlap between these two categories is also due to their use of common words, which further complicates accurate classification. These challenges contribute to the model's overall accuracy rate of around 80%, reflecting the inherent challenges in classifying subjective and diverse content with textual cues alone.

Table 5. Example of incorrect model prediction

Caption	True labels	Predicted labels
Hidup lagi capek-capeknya malah kena prank *gembel*!! *Untung gue tetep* enjoy *soalnya ada* @pucukharumid 🤪 *#SegerinDiManaAja #KesegaranUntukSemua #DitemeninTehPucukHarum*	Entertainment	Food
Jangan ditonton sampe akhir aja 😭	Entertainment	Others
vlog pt2! ❤️	Entertainment	Others
Lagi ngobrol apa hayo sama pak bos Surabaya.. @royadidharma	Others	Entertainment
Begini nih lika liku untuk pengiriman jualannya @gratu.id, *untung ada solusinya*! #GoSendCar #BestSellerGoSend	Others	Food
Kalo aja jaman itu udah ada Grabjian, *mungkin gak akan ada istilah* "*cinta ku berat di ongkos*". *Jalan-jalan kemanapun naik* GrabBike, *diskon* 50% *dengan kode promo* GRABJIAN. @grabid	Others	Travel
Spontaneous night stroll in SG ✨	Travel	Others

3.3. Influencer classification results

Using our fine-tuned BERT model, we classified 400 Indonesian influencers with a total of 19,842 unlabeled posts to identify their categories. This classification process resulted in category predictions for each post and a final category for each influencer. The results demonstrate that our model performed well in classifying influencer posts into relevant categories. The post classification results are shown in Figure 3.

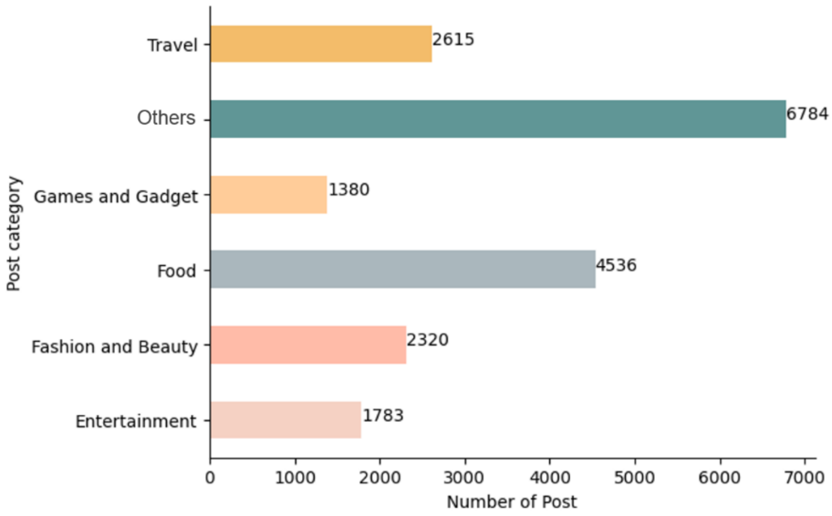

Figure 3. Influencer post classification results

Based on the post classification results in Figure 3, the majority of posts fall into the "others" category, followed by "food", "travel", "fashion and beauty", "entertainment", and "games and gadget" categories. The dominance of the "others" category is due to the variation in influencer content, as most influencers have posts outside their main niche, such as about their personal or more general topics. From these post classification results, we can then determine the category for each influencer. The distribution of 400 Indonesian influencers across each category can be seen in Figure 4.

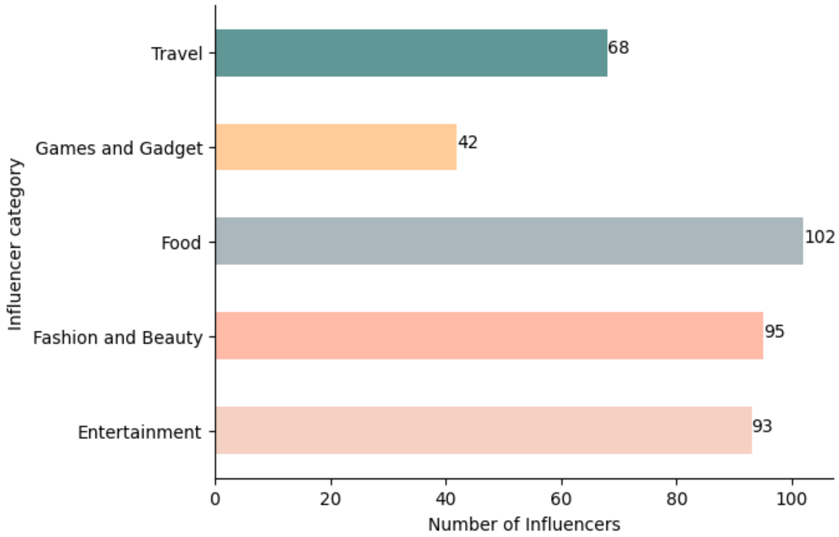

Figure 4. Influencer classification results

4. CONCLUSION

This study proposed an influencer classification method using BERT. The model development process, which included hyperparameter tuning and stop word removal experiment, achieved significant accuracy in categorizing posts into six relevant categories: fashion and beauty, entertainment, food, travel, games and gadget, and others. Our best-performing model, with an 80% accuracy in post classification, was trained with a batch size of 32, learning rate of 3e-5, and data without stop words removed. This study also demonstrates that using BERT for text classification effectively categorizes 400 Indonesian influencers based on their post captions. This aligns with our initial goal of categorizing diverse influencer content, offering a sophisticated and targeted approach to influencer identification. Looking forward, the findings of this research present significant opportunities for further development and application. However, because captions only provide a limited amount of information, they may not be sufficient. Future studies should consider incorporating additional data sources, such as images, video content, or engagement metrics, to create a more comprehensive influencer classification system.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ni Putu Sutramiani	✓	✓	✓		✓	✓	✓		✓	✓		✓	✓	
Ni Made Dita Dwikasari	✓	✓	✓			✓		✓	✓		✓		✓	
I Nyoman Prayana Trisna	✓	✓		✓	✓		✓	✓		✓	✓	✓	✓	
I Wayan Agus Surya Darma	✓			✓					✓	✓				

C	:	**C**onceptualization	I	:	**I**nvestigation	Vi	:	**Vi**sualization
M	:	**M**ethodology	R	:	**R**esources	Su	:	**Su**pervision
So	:	**So**ftware	D	:	**D**ata Curation	P	:	**P**roject administration
Va	:	**Va**lidation	O	:	Writing - **O**riginal Draft	Fu	:	**Fu**nding acquisition
Fo	:	**Fo**rmal analysis	E	:	Writing - Review & **E**diting			

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [INPT], upon reasonable request.

REFERENCES

[1] H.-C. Lin, P. F. Bruning, and H. Swarna, "Using online opinion leaders to promote the hedonic and utilitarian value of products and services," *Business Horizons*, vol. 61, no. 3, pp. 431–442, 2018, doi: 10.1016/j.bushor.2018.01.010.

[2] HypeAuditor, "State of influencer marketing 2024 Indonesia," *HypeAuditor*. Accessed: Feb. 11, 2024. [Online]. Available: https://hypeauditor.com/resources/whitepapers/state-of-influence-marketing-indonesia-2024/

[3] R. Edelman, *In brands we trust? 2019 edelman trust barometer special report*. Chicago, United States: Edelman, 2019.

[4] S. Wies, A. Bleier, and A. Edeling, "Finding goldilocks influencers: how follower count drives social media engagement," *Journal of Marketing*, vol. 87, no. 3, pp. 383–405, 2022, doi: 10.1177/00222429221125131.

[5] M. D. Veirman, V. Cauberghe, and L. Hudders, "Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude," *International Journal of Advertising*, vol. 36, no. 5, pp. 798–828, 2017, doi: 10.1080/02650487.2017.1348035.

[6] C. Guan and E. Y. Li, "A note on influencer marketing in social media," *International Journal of Internet Marketing and Advertising*, vol. 15, no. 2, pp. 123–130, 2021.

[7] L. Janssen, A. P. Schouten, and E. A. J. Croes, "Influencer advertising on Instagram: product-influencer fit and number of followers affect advertising outcomes and influencer evaluations via credibility and identification," *International Journal of Advertising*, vol. 41, no. 1, pp. 101–127, 2021, doi: 10.1080/02650487.2021.1994205.

[8] B. E. Elbaghazaoui, M. Amnai, and Y. Fakhri, "Data profiling and machine learning to identify influencers from social media platforms," *Journal of ICT Standardization*, vol. 10, no. 2, pp. 201–218, May 2022, doi: 10.13052/jicts2245-800X.1026.

[9] S. S. Rani and S. S. Baby, "Real-time influencer detection in Twitter using a hybrid approach," *Procedia Computer Science*, vol. 215, pp. 461–470, 2022, doi: 10.1016/j.procs.2022.12.048.

[10] B. Bashari and E. F.-Ersi, "Influential post identification on Instagram through caption and hashtag analysis," *Measurement and Control*, vol. 53, no. 3–4, pp. 409–415, 2020, doi: 10.1177/0020294019877489.

[11] S. Kim, J.-Y. Jiang, M. Nakada, J. Han, and W. Wang, "Multimodal post attentive profiling for influencer marketing," in *Proceedings of The Web Conference 2020*, 2020, pp. 2878–2884, doi: 10.1145/3366423.3380052.

[12] C. Lou and S. Yuan, "Influencer marketing: how message value and credibility affect consumer trust of branded content on social media," *Journal of Interactive Advertising*, vol. 19, no. 1, pp. 58–73, Feb. 2019, doi: 10.1080/15252019.2018.1533501.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186.

[14] R. A. Rajagede, "Improving automatic essay scoring for Indonesian language using simpler model and richer feature," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 6, no. 1, pp. 11–18, 2021, doi: 10.22219/kinetik.v6i1.1196.

[15] B. Juarto and Yulianto, "Indonesian news classification using IndoBert," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2, pp. 454–460, 2023.

[16] H. S. Wicaksana, R. Kusumaningrum, and R. Gernowo, "Determining community happiness index with transformers and attention-based deep learning," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 1753–1761, 2024, doi: 10.11591/ijai.v13.i2.pp1753-1761.

[17] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641–5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[18] E. C. G. Merchan, R. G. Brizuela, and S. G.-Carvajal, "Comparing BERT against traditional machine learning models in text classification," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 352–356, 2023, doi: 10.47852/bonviewjcce3202838.

[19] S. Hamdi, A. Hamdi, and S. B. Yahia, "BERT and word embedding for interest mining of Instagram users," in *Advances in Computational Collective Intelligence*, 2022, pp. 123–136, doi: 10.1007/978-3-031-16210-7_10.

[20] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[21] AnyMind, "Influencer marketing in Indonesia market overview 2021," *AnyMind*. Accessed: Sep. 26, 2024. [Online]. Available: https://anymindgroup.com/

[22] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 task 4: deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 747–754, doi: 10.18653/v1/s17-2126.

[23] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017*, 2025, pp. 1–11, doi: 10.65215/2q58a426.

[24] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4996–5001, doi: 10.18653/v1/p19-1493.

[25] D. J. Ladani and N. P. Desai, "Stopword identification and removal techniques on TC and IR applications: a survey," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 466–472, doi: 10.1109/icaccs48705.2020.9074166.

BIOGRAPHIES OF AUTHORS

Ni Putu Sutramiani received a bachelor's degree in Computer Systems from STIKOM Bali in 2011, a master's degree in Information Systems and Computer Management from Udayana University in 2015, and a doctoral degree in Computer Science at the Department of Informatics, Institut Teknologi Sepuluh Nopember in 2022. Her research interests include computer vision, image processing, and artificial intelligence. She can be contacted at email: sutramiani@unud.ac.id.

Ni Made Dita Dwikasari holds a bachelor's degree in Information Technology from the Faculty of Engineering, Udayana University, Indonesia, earned in 2024. Her research interests include the digital economy, data science, and natural language processing, with a focus on using data-driven solutions to enhance economic processes and drive innovation. She can be contacted at email: dita.dwikasari08@student.unud.ac.id.

I Nyoman Prayana Trisna received the bachelor's and master's degree in Computer Science and Electronics from the Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2017 and 2020 respectively. He is currently an assistant professor in the Information Technology Study Program, Faculty of Engineering, Udayana University, Bali, Indonesia. His current research interests include machine learning, evolutionary computation, natural language modelling, and text mining. He can be contacted at email: prayana.trisna@unud.ac.id.

I Wayan Agus Surya Darma holds a Doctor of Computer Science degree from Institut Teknologi Sepuluh Nopember, Indonesia, with the Dissertation “Graph feature fusion for the recognition of highly varying and complex Balinese carving motifs”. He is currently an assistant professor of Computer Science at Department of Information Technology, Udayana University, Bali, Indonesia. He has published more than 10 scopus indexed journal articles and conference papers. His research interests are in artificial intelligence, computer vision, and computational intelligence. He can be contacted at email: agussurya@unud.ac.id.