Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 3, August, Pp. 429-440

Indonesian pre-trained language models with OCEAN-aware query expansion ranking for personality prediction from social media text

Gede Aditra Pradnyana*[1], I Gede Mahendra Darmawiguna[1]
Department of Informatics, Universitas Pendidikan Ganesha, Indonesia[1]

Article Info
Keywords:
OCEAN Personality Prediction, Indonesian Social Media, Pre-Trained Language Models, OCEAN-aware QER, Lexical Feature Learning, Class-Discriminative Representation

Article history:
Received: June 01, 2026
Accepted: July 22, 2026
Published: August 01, 2026

Cite:
G. A. Pradnyana and I. G. M. Darmawiguna, "Indonesian Pre-trained Language Models with OCEAN-aware Query Expansion Ranking for Personality Prediction from Social Media Text", *KINETIK*, vol. 11, no. 3, Aug. 2026.
https://doi.org/10.22219/kinetik.v11i3.2890

*Corresponding author.
Gede Aditra Pradnyana
E-mail address:
gede.aditra@undiksha.ac.id

Abstract
Personality prediction from social media text has become a critical topic in computational social science, as online posts frequently reflect individual behavioral and psychological characteristics. However, predicting personality from Indonesian social media material remains challenging due to informal language, slang, abbreviations, noisy expressions, and implicit personality-related cues. This study proposed a new OCEAN personality prediction framework that integrates pre-trained language models with an OCEAN-aware Query Expansion Ranking feature learning mechanism. Unlike conventional approaches that rely mainly on contextual embeddings, the proposed framework introduced trait-specific class-discriminative lexical features to strengthen personality-related representation. The prediction task was structured as five separate binary classification problems, with each personality attribute divided into High and Low groups. The experiments were conducted in two stages. First, numerous pre-trained language models, including IndoBERT, IndoBERTweet, Indonesian RoBERTa, and multilingual BERT, were fine-tuned and compared to identify the most suitable contextual representation model. The best baseline model, multilingual BERT, achieved an average accuracy of 75.48% and an average F1-score of 73.58%. Second, multilingual BERT was integrated with class-discriminative lexical features generated by the proposed OCEAN-aware Query Expansion Ranking mechanism. The trait-specific configuration improved the average accuracy to 78.76% and the average F1-score to 76.48%. These results demonstrated that integrating contextual semantic representations with OCEAN-aware lexical feature learning enhanced personality prediction from Indonesian social media material while also offering a clearer representation of trait-relevant language signals.

1. Introduction

Social media has become one of the most accessible and comfortable platforms for individuals to communicate, express opinions, share experiences, and convey emotional states [1], [2], [3]. The rapid growth of user-generated content on social media platforms has resulted in massive amounts of digital behavioral data in the form of text, images, audio, and videos. These digital traces provide valuable opportunities for computational analysis to uncover various aspects of human behavior, including emotional tendencies, social interactions, psychological conditions, and personality characteristics [1]. Consequently, computational psychology and Natural Language Processing (NLP) researchers are increasingly interested in using social media data to predict personality and profile behavior [4].

Personality prediction from social media text has gained considerable attention in recent years because of its potential applications in personalized recommendation systems, mental health monitoring, adaptive learning environments, human-computer interaction, and digital behavioral analytics [5], [6]. In psychology, numerous personality frameworks have been established to characterize individual behavioral traits, such as the Myers-Briggs Type Indicator (MBTI), Dominance Influence Steadiness Conscientiousness (DISC), and the Big Five Personality model [7], [8]. Among these frameworks, the Big Five Personality model, also known as the Five Factor Model or OCEAN model, is one of the most widely adopted approaches in computational personality analysis because of its robustness, psychological validity, and interpretability across different cultural contexts [6], [8], [9].

The Big Five paradigm classifiespersonality into five key traits, namely Openness (OPN), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU) [9]. Openness reflects intellectual curiosity, creativity, imagination, and openness toward new experiences. Conscientiousness is associated with discipline, responsibility, organization, and goal-oriented behavior. Extraversion describes sociability, assertiveness, enthusiasm, and the tendency to seek social interaction. Agreeableness reflects empathy, cooperativeness, trust, and prosocial behavior toward others. Meanwhile, Neuroticism is characterized by emotional instability, anxiety, mood fluctuations,

and vulnerability to psychological stress. These personality factors are frequently implicitly mirrored in language usage patterns, communication styles, emotional expressions, and behavioral inclinations displayed in social media posts.

With the continuous growth of social media platforms such as Twitter/X, textual content generated by users provides valuable behavioral signals that can be utilized to infer personality characteristics automatically [2,3]. The linguistic patterns embedded in social media posts often contain subtle psychological cues that reflect users' habitual behaviors, emotional tendencies, interpersonal attitudes, and cognitive preferences. Therefore, social media text has become an important resource for developing automatic personality prediction systems using machine learning and deep learning approaches.

Recent advances in NLP have significantly improved personality prediction performance through the adoption of pretrained language models (PLMs). Transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT) [10] and its variants [11], [12] are capable of learning contextual semantic representations from large-scale corpora and have demonstrated strong performance across various text classification tasks. In the context of social media analysis, PLMs can capture semantic dependencies, contextual meanings, and latent linguistic patterns more effectively than conventional machine learning approaches [3], [8], [13]. However, their representations may still obscure fine-grained psycholinguistic and trait-specific lexical cues when these cues are sparse, implicit, or unevenly distributed across user posts [14]. Therefore, identifying the most suitable Indonesian pretrained language model becomes an important preliminary step for personality prediction tasks.

Despite their strong contextual representation capability, pretrained language models do not explicitly emphasize class-discriminative lexical information associated with specific personality dimensions. Personality-related linguistic cues are often implicit, sparse, subjective, and embedded within noisy textual patterns, particularly in Indonesian social media environments that contain slang expressions, abbreviations, code-mixing, emojis, and inconsistent writing structures [15], [16]. Consequently, psychologically informative lexical signals may become diluted within contextual embeddings generated by PLMs.

To address this limitation, this study proposes an OCEAN-aware Query Expansion Ranking (O-QER) framework for learning class-discriminative lexical features in personality prediction from Indonesian social media text. The framework is adapted from the Query Expansion Ranking method introduced in [17], [18], which was originally developed for sentiment classification by ranking lexical features according to their discriminative distributions across classes. Previous studies have shown that QER can improve text classification by prioritizing informative features and reducing the influence of irrelevant terms. QER has also been applied to MBTI personality prediction through the integration of discriminative lexical features, GloVe embeddings, and a BiLSTM classifier, achieving an accuracy of 85.96% [19].

In the present study, QER is reformulated as O-QER and adapted to five independent binary classification tasks corresponding to Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. For each trait, lexical features are scored and ranked according to their discriminative relevance to the high and low personality classes. The proposed framework further incorporates token-level and post-level ranking. Token-level O-QER evaluates the discriminative importance of individual lexical items, whereas post-level O-QER aggregates these signals to represent the overall personality-related information within a social media post. The O-QER is not intended as a spelling-correction or slang-normalization method. Instead, it complements contextual embeddings by emphasizing trait-relevant lexical cues that may be obscured by slang, abbreviations, code-mixing, nonstandard spelling, emojis, and inconsistent writing patterns.

The proposed model is expected to provide several advantages. First, the integration of contextual embeddings and lexical discriminative features may improve the model's ability to capture subtle personality-related language patterns. Second, the token-level and post-level QER mechanisms enable more granular modeling of psychologically relevant lexical signals within noisy Indonesian social media environments. Third, the proposed hybrid representation strategy may enhance classification robustness by complementing transformer semantic representations with interpretable lexical weighting features.

Accordingly, the main contributions of this study can be summarized as follows: it presents a comparative evaluation of several Indonesian pretrained language models for OCEAN personality prediction from social media text to identify the most effective contextual representation model for capturing personality-related linguistic patterns in Indonesian social media environments; proposes a novel OCEAN-aware Query Expansion Ranking (O-QER) feature learning framework to enhance class-discriminative lexical representation by incorporating both token-level and post-level discriminative scoring mechanisms, integrating this framework with the best-performing pretrained language model to enrich contextual semantic representations with psychologically informative lexical cues embedded in noisy Indonesian social media text; and further investigates the impact of different O-QER feature-selection percentages across each OCEAN personality dimension to analyze trait-specific lexical discriminative characteristics and their influence on personality prediction performance.

2. Research Method

This study employed an experimental computational design to develop and evaluate an OCEAN personality prediction model using Indonesian social media text. As illustrated in Figure 1, the proposed methodological framework comprised two main experimental stages. In the first stage, several Indonesian pretrained language models were systematically compared to identify the most effective contextual representation model for personality prediction. In the second stage, the best-performing pretrained language model was integrated with the proposed OCEAN-aware Query Expansion Ranking (O-QER) feature-learning mechanism to strengthen class-discriminative lexical representation.

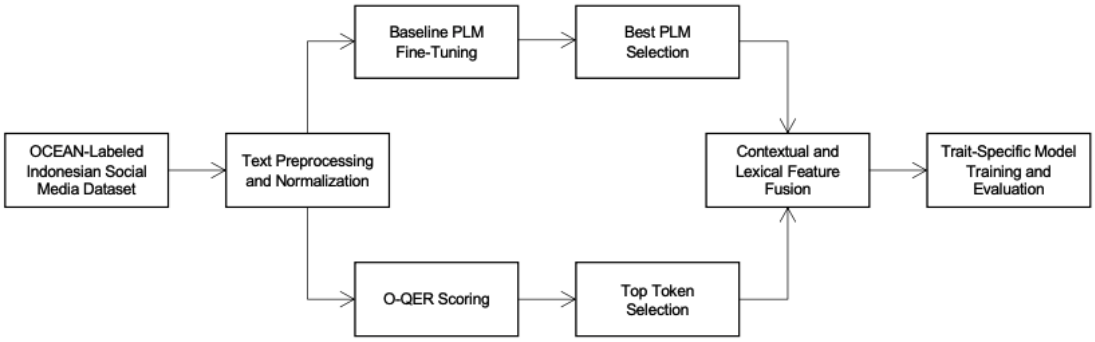

Figure 1. Overall Methodological Workflow for the Proposed OCEAN Personality Prediction

2.1 Dataset

The dataset used in this study was obtained from previous studies by [20], [21] and has also been used in several related studies [3], [8]. The dataset was collected from Twitter, currently known as X, and consists of 508 users, with 100 text posts or tweets in Indonesian language collected from each user. Each user in the dataset was labeled according to the OCEAN personality model, which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. For each personality trait, the class label was categorized as either "High" or "Low," and the distribution of these labels is presented in Table 1.

Table 1. Distribution of High and Low Class Labels Across OCEAN Personality Traits

OCEAN Trait	High (1)	Low (0)
Openness (OPN)	272	236
Conscientiousness (CON)	131	377
Extraversion (EXT)	363	145
Agreeableness (AGR)	278	230
Neuroticism (NEU)	221	287

In the dataset, the labels "High" and "Low" indicate the relative tendency of a user toward each OCEAN personality trait. A "High" label means that the user shows a stronger tendency toward a particular trait, whereas a "Low" label means that the user shows a weaker tendency toward that trait. For computational modeling, each categorical class label was converted into a numerical format, where "High" was encoded as 1 and "Low" was encoded as 0. In this study, the personality prediction task was formulated as five independent binary classification tasks, with one classification task corresponding to each OCEAN trait.

2.2 Text Preprocessing and Normalization

The dataset consists of Indonesian Twitter/X user-generated text in which multiple posts from each user are concatenated into a single user-level document. Therefore, preprocessing was designed to preserve the user-level structure of the data. Two text representations were generated: one for the pretrained language model branch and one for the OCEAN-aware Query Expansion Ranking (O-QER) branch, as illustrated in Figure 1. The 100 tweets associated with each user were concatenated into a single user-level document, resulting in 508 classification instances. The preprocessing procedures transformed the textual content but did not remove any tweets or users. Consequently, all 50,800 tweets and 508 user-level documents were retained after preprocessing.

For the transformer-based branch, the original tweet separator was replaced with the tokenizer-specific separator "[SEP]" token to preserve tweet-level boundaries within each user document. Light normalization was applied to preserve contextual, stylistic, and affective cues. HTML entities, Unicode artifacts, excessive whitespace, URLs, retweet markers, and dataset-specific markers such as "[UNAME]", "[HASHTAG]", and "[pic]" were normalized into stable textual

tokens. Emoji were converted into Indonesian semantic tokens, such as “emoji_senyum”, “emoji_tertawa”, and “emoji_sedih”, to preserve their affective meaning.

For the O-QER branch, stronger lexical normalization was applied to reduce token sparsity and support stable class-discriminative scoring. Common Indonesian abbreviations and informal forms were normalized using a slang dictionary, for example, “utk” into “untuk”, “krn” into “karena”, and “bgt” into “banget”. However, stylistically meaningful pronouns and discourse particles were preserved when they were considered potentially relevant to personality expression.

Sentence boundary markers, including periods, exclamation marks, and question marks, were retained in the QER preprocessing stage to support sentence-level QER scoring. These punctuation marks were removed only during the QER calculation stage after sentence segmentation had been completed. Thus, the preprocessing strategy balanced linguistic normalization with the preservation of natural Indonesian social media expression.

2.3 Baseline PLM Fine-Tuning

At this stage, several pretrained language models (PLMs) trained on or supporting Indonesian were fine-tuned for the OCEAN personality prediction task. The pretrained models evaluated in this study included IndoBERT, IndoBERTweet, Indonesian RoBERTa, and multilingual BERT (mBERT). IndoBERT [12], IndoBERTweet [11], Indonesian RoBERTa [22], and mBERT[23] are transformer-based language models derived from the Bidirectional Encoder Representations from Transformers (BERT) architecture [10]. The pretrained language models evaluated in this study were selected based on their demonstrated effectiveness in Indonesian and multilingual text-classification tasks. Previous studies have reported favorable performance for these models in sentiment analysis, text classification, and social media language processing [1], [8], [24], [25]. Their inclusion therefore provides an empirically grounded comparison of monolingual, domain-specific, and multilingual contextual representations for OCEAN personality prediction.

IndoBERT is a monolingual Indonesian BERT model pretrained on large-scale Indonesian corpora. In this study, two IndoBERT variants were used, namely IndoBERT-base and IndoBERT-large. IndoBERT-base consists of 12 transformer encoder layers, a hidden dimension of 768, 12 attention heads, and approximately 124.5 million parameters. In contrast, IndoBERT-large consists of 24 transformer encoder layers, a hidden dimension of 1024, 16 attention heads, and approximately 335.2 million parameters. These two variants were included to examine the effect of model capacity on personality prediction performance from Indonesian social media text.

Unlike the general-domain IndoBERT models, IndoBERTweet [11] is a domain-specific pretrained language model developed for Indonesian Twitter/X data. It was introduced by Koto, Lau, and Baldwin as the first large-scale pretrained language model specifically designed for Indonesian Twitter. IndoBERTweet was developed by extending monolingual IndoBERT through the incorporation of a domain-specific vocabulary that better captures the linguistic characteristics of Twitter/X, including informal expressions, abbreviations, hashtags, mentions, and other social media-specific lexical patterns. Architecturally, IndoBERTweet follows the BERT-base configuration, consisting of 12 hidden layers, a hidden dimension of 768, and 12 attention heads. Indonesian RoBERTa is an Indonesian pretrained language model based on Robustly Optimized BERT Pretraining Approach (RoBERTa). Unlike BERT, RoBERTa [22] removes the Next Sentence Prediction (NSP) objective and focuses on optimizing the Masked Language Modeling (MLM) objective with improved pretraining strategies.

On the other hand, multilingual BERT (mBERT) is a multilingual BERT model developed by Google. It was pretrained using Wikipedia corpora from 104 languages, including Indonesian, with the Masked Language Modeling (MLM) objective [26]. In this study, mBERT was included as a multilingual baseline to examine whether a general multilingual representation model can perform competitively compared with monolingual Indonesian PLMs. This comparison is important because multilingual models are often more general in linguistic coverage, whereas monolingual and domain-specific models may provide more precise representations for Indonesian social media text.

After each pretrained, model was successfully loaded, the fine-tuning process was conducted by exploring several hyperparameter combinations, including the number of epochs, batch size, and learning rate. The range of hyperparameter values used in this study was determined based on previous studies [8], [10] and adjusted to the computational requirements of transformer-based language model fine-tuning. Specifically, the evaluated hyperparameter values were as follows: learning rate (LR) of 5e-5, 3e-5, and 2e-5; number of epochs (EP) of 2, 3, and 4; and batch size (BZ) of 16 and 32.

Overall, the fine-tuning of these pretrained language models was performed to identify the most effective contextual representation backbone for OCEAN personality prediction from Indonesian social media text. The comparison among monolingual, domain-specific, and multilingual models provides empirical evidence for selecting the most suitable pretrained language model before integrating it with the proposed OCEAN-aware Query Expansion Ranking (O-QER) feature learning mechanism.

2.4 The Proposed O-QER Feature Learning

Query Expansion Ranking (QER) was originally introduced by Parlar et al. [17], [18] as a feature selection method for sentiment analysis. The method was adapted from query expansion in information retrieval, where candidate terms are ranked based on how useful they are for improving retrieval performance [17]. In QER, a feature is considered important when its probability of occurrence differs clearly between positive and negative documents. In other words, features that appear more strongly in one class than in the other are treated as more discriminative. A later study by [19] showed that QER can also be effectively used for feature selection in MBTI personality prediction. This suggests that QER is not only useful for sentiment analysis, but also for personality-related text classification, where discriminative lexical patterns are important for distinguishing psychological characteristics.

In this study, QER is modified into an OCEAN-aware Query Expansion Ranking (O-QER) mechanism for personality prediction from Indonesian social media text. Unlike the original QER, which was designed for binary sentiment classification, the proposed O-QER is adapted to five independent binary classification tasks corresponding to the OCEAN personality dimensions. Each personality trait is represented using two classes, namely High(1) and Low(0). For each OCEAN dimension $o \in \{OPN, CON, EXT, AGR, NEU\}$ each token t is assigned a class-discriminative score by comparing its occurrence in the target class *c* and the opposite class ¬*c*. The proposed token-level O-QER score is defined in Equation 1.

$$OQER(t, o, c) = \log\left(\frac{TF(t, o, c) + \alpha}{TF(t, o, \neg c) + \alpha}\right) \times IDF(t) \quad (1)$$

where $TF(t, o, c)$ is the frequency of token *t* in class *c* for OCEAN dimension *o*, $TF(t, o, \neg c)$ is the frequency of the same token in the opposite class, and *α* is a smoothing factor used to avoid division by zero. The inverse document frequency term is computed as Equation 2.

$$IDF(t) = \log\left(\frac{N + 1}{DF(t) + 1}\right) + 1 \quad (2)$$

where N is the total number of training documents and $DF(t)$ is the number of documents containing token t. The IDF component is incorporated to reduce the influence of overly common tokens and to increase the contribution of more informative lexical units.

2.4.1 Token-Level O-QER Selection

At the token level, O-QER is used to identify discriminative lexical units for each OCEAN dimension. After computing the O-QER score for each token, tokens are ranked according to their absolute discriminative strength by using Equation 3:

$$Strength(t, o) = \max(|OQER(t, o, 0)|, |OQER(t, o, 1)|) \quad (3)$$

The absolute score is used because both positive and negative values may carry discriminative information. A strongly positive score indicates that a token is more associated with the target class, while a strongly negative score indicates that the token is more associated with the opposite class. Therefore, both directions are informative for distinguishing High and Low personality traits. A rank-based percentage selection strategy is then applied. Given a selection percentage r, the number of retained tokens for an OCEAN dimension is calculated using Equation 4.

$$K_o = \lceil r \times |V_o| \quad (4)$$

where $|V_o|$ is the size of the O-QER-ranked vocabulary for trait *o*, and $r \in \{0.50, 0.70, 0.90, 1.00\}$ represents the proportion of selected tokens. For example, r = 0.90 means that the top 90% of tokens are retained after sorting by O-QER rank. This rank-based percentage strategy is used instead of percentile-threshold selection to ensure that each experimental setting retains a controlled proportion of lexical units.

2.4.2 Post-Level O-QER Representation

After token-level selection, the selected O-QER tokens are used to construct a post-level representation. For a social media post d_i let T_i denote the set of tokens appearing in the post and S_o denote the selected O-QER vocabulary for OCEAN dimension *o*. The selected tokens in post d_i are defined as Equation 5.

$$T_i^o = \{t \in T_i \mid t \in S_o\} \quad (5)$$

The post-level O-QER representation is then obtained by aggregating the O-QER scores of all selected tokens in T_i^o. In this study, the top 50%, 70%, 90%, and 100% settings refer to vocabulary-level token selection within each post. Therefore, once the selected vocabulary S_o has been determined, all matching tokens from S_o that appear in a post are used to compute the post-level O-QER features. This design enables the model to capture the cumulative discriminative lexical evidence contained in a post. Rather than relying only on the strongest individual token, the post-level representation summarizes how much personality-relevant lexical information is present in the text. This is important for personality prediction because personality cues are often distributed across multiple lexical units, especially in informal social media language. The post-level O-QER features are computed using several statistical summaries as shown in Table 2.

Table 2. The Post-level O-QER Features

Post-Level Feature	Formula
Mean	$MeanOQER(d_i, o) = \frac{1}{\lvert T_i^o \rvert} \sum_{t \in T_i^o} OQER(t, o, c)$
Mean Absolute	$MeanAbsOQER(d_i, o) = \frac{1}{\lvert T_i^o \rvert} \sum_{t \in T_i^o} \lvert OQER(t, o, c) \rvert$
Sum Absolute	$SumAbsOQER(d_i, o) = \sum_{t \in T_i^o} \lvert OQER(t, o, c) \rvert$
Maximum Absolute	$MaxAbsOQER(d_i, o) = \max_{t \in T_i^o} \lvert OQER(t, o, c) \rvert$
Selected Token Count	$SelectedCount(d_i, o) = \lvert T_i^o \rvert$
Selected Token Ration	$SelectedRatio(d_i, o) = \frac{\lvert T_i^o \rvert}{\lvert T_i \rvert}$

Overall, the post-level O-QER representation transforms the selected token-level discriminative scores into compact numerical features. These features summarize the intensity, direction, and coverage of OCEAN-relevant lexical evidence in each post. The resulting post-level O-QER vector is then standardized and concatenated with the IndoBERTweet contextual embedding to form the final hybrid representation for personality prediction.

2.4.3 Integration with Pre-trained Language Model

After the token-level and post-level O-QER feature learning processes, the resulting post-level O-QER representation is integrated with the best-performing pretrained language model identified in the previous experimental stage. The model that achieved the strongest overall performance was then selected as the contextual representation backbone for the proposed O-QER integration.

The proposed O-QER mechanism is not used to remove or filter posts before transformer encoding. Instead, it functions as a complementary lexical feature learning module. The selected PLM encodes the full preprocessed social media post to capture contextual semantic information, while O-QER provides OCEAN-aware class-discriminative lexical evidence derived from the training corpus. The final fused representation is obtained by concatenating the contextual embedding generated by the best-performing PLM with the standardized post-level O-QER feature vector in Equation 6.

$$H_i = \left[PLM_{\text{best}}(d_i); OQER_{post}(d_i) \right] \tag{6}$$

where PLM_{best} denotes the contextual embedding produced by the selected best-performing pretrained language model for post d_i, $OQER_{post}(d_i)$ denotes the standardized post-level O-QER feature vector, and H_i represents the final hybrid representation used for OCEAN personality prediction.

This integration strategy allows the model to use two complementary types of information. The PLM provides contextual semantic representations that help capture the meaning of informal Indonesian social media text. Meanwhile, O-QER adds explicit class-discriminative lexical cues that are specific to each OCEAN personality dimension. As a result, the proposed hybrid representation retains the contextual strength of the selected PLM while enriching it with personality-aware lexical evidence from O-QER.

2.5 Trait-Specific Model Training and Evaluation

The experimental comparison was carried out in two main phases. In the first phase, several pretrained language models were fine-tuned and compared as baseline models to identify the best-performing PLM. In the second phase,

the selected PLM was integrated with O-QER features using different top-token selection percentages. The integrated model was evaluated under five O-QER feature-selection settings, namely 50%, 60%, 70%, 80%, and 90%. The resulting fused representation was then used as input for the final classification model.

Training and evaluation were conducted separately for each OCEAN personality dimension. This trait-specific strategy was used because each personality trait may be associated with different linguistic cues and discriminative lexical patterns. Model performance was assessed using accuracy, precision, recall, and F1-score. These metrics were computed for each OCEAN trait to evaluate trait-level performance. In addition, macro-averaged performance across the five traits was calculated to provide an overall evaluation of the model. Macro averaging was applied because it assigns equal weight to each personality dimension, regardless of differences in class distribution or sample size.

3. Results and Discussion

This section presents the experimental results and discussion of the proposed framework for OCEAN personality prediction from Indonesian social media text. The evaluation was conducted in two main stages. First, several Indonesian and multilingual pre-trained language models were fine-tuned and compared to identify the most effective contextual representation model. Second, the best-performing model was integrated with the proposed OCEAN-aware Query Expansion Ranking feature learning mechanism to evaluate the contribution of class-discriminative lexical features at different QER feature-selection percentages.

3.1 Performance Comparison of Pre-Trained Language Models

This section presents the baseline performance comparison of several pre-trained language models for OCEAN personality prediction from Indonesian social media text. The purpose of this experiment was to identify the most suitable contextual representation model before integrating it with the proposed OCEAN-aware QER feature learning mechanism. Four models were evaluated, namely IndoBERT, IndoBERTweet, IndoRoBERTa, and mBERT, using accuracy and F1-score across the five OCEAN traits.

Table 3. Baseline Performance Comparison of Pre-Trained Language Models for OCEAN Personality Prediction

Pre-trained Language Model	Mean Accuracy	Mean F1-score
IndoBERT [12]	0.7304	0.7086
IndoBERTweet [11]	0.7444	0.7152
IndoRoBERTa [22]	0.7136	0.6922
mBERT [23]	0.7548	0.7358

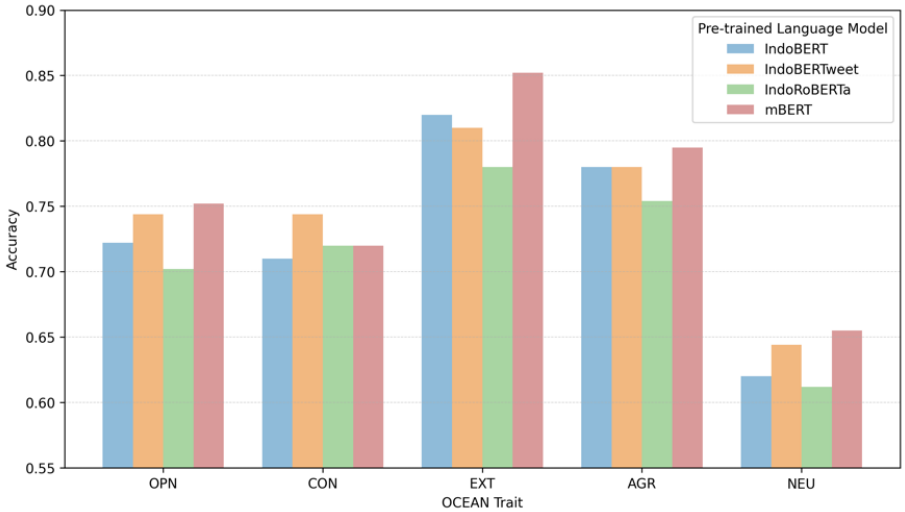

Figure 2. Accuracy Comparison of Pre-trained Language Models Across OCEAN Traits

Table 3 compares the baseline performance of four pre-trained language models for OCEAN personality prediction from Indonesian social media text. Overall, mBERT achieved the best results, with a mean accuracy of 0.7548 and a mean F1-score of 0.7358. As shown in Figure 2, mBERT [23] performed best on most traits, including Openness, Extraversion, Agreeableness, and Neuroticism. In contrast, IndoBERTweet achieved the highest performance for Conscientiousness, suggesting that a domain-specific model trained on Indonesian Twitter data can still be useful for

certain personality dimensions. The variation in performance across traits also indicates that personality prediction is not a uniform classification problem. Each OCEAN trait may be reflected through different linguistic and psychological cues, so a model that performs well on one trait may not achieve the same level of performance on another. This finding suggests that each personality dimension has trait-specific lexical and contextual characteristics in Indonesian social media text. Based on its strongest overall performance across the five traits, mBERT was selected as the contextual representation backbone for integration with the proposed OCEAN-aware Query Expansion Ranking feature learning mechanism.

The finding that mBERT [23] achieved strong performance is consistent with the study by [3], which reported similar effectiveness of mBERT for Indonesian text classification. However, its performance should not be attributed solely to the size of its multilingual pretraining corpus. The dataset used in this study contains informal Indonesian, code-mixed expressions, English loanwords, abbreviations, nonstandard spellings, emojis, and other social media-specific patterns. Because mBERT was pretrained across multiple languages, its shared multilingual vocabulary and contextual representations may provide broader coverage of heterogeneous lexical patterns than models trained predominantly on monolingual or more formal Indonesian text.

Another possible explanation concerns the contextual variability of personality-related expressions. The same lexical item may convey different psychological meanings depending on its surrounding words, while similar personality tendencies may be expressed through Indonesian, English, or mixed-language forms. mBERT's bidirectional contextual representation may therefore be advantageous for capturing these context-dependent and cross-lingual patterns. Nevertheless, this interpretation should be regarded as a plausible explanation rather than a confirmed causal mechanism because vocabulary coverage, token fragmentation, and code-mixing performance were not directly measured in this study.

Despite its superior baseline performance, mBERT achieved a mean F1-score of 0.7358, indicating that contextual representations alone did not fully capture all personality-discriminative information. Personality-related lexical cues may be sparse, implicit, or diluted within high-dimensional contextual embeddings. This remaining limitation provides the rationale for integrating mBERT with O-QER, which explicitly identifies and weights lexical features according to their discriminative relevance to the high and low classes of each OCEAN trait. Based on its strongest aggregate performance, mBERT was selected as the contextual representation backbone for the subsequent experiments.

3.2 Effect of OCEAN-Aware QER Feature Learning

Based on the results of the previous stage, mBERT was selected as the baseline PLM because it achieved the strongest overall performance across the five OCEAN traits. In the next experiment, mBERT was used as the contextual representation backbone and integrated with the proposed OCEAN-aware QER (O-QER) feature learning mechanism. To evaluate the effect of different lexical feature-selection levels, six O-QER feature-selection percentages were examined: 50%, 60%, 70%, 80%, 90%, and 100%.

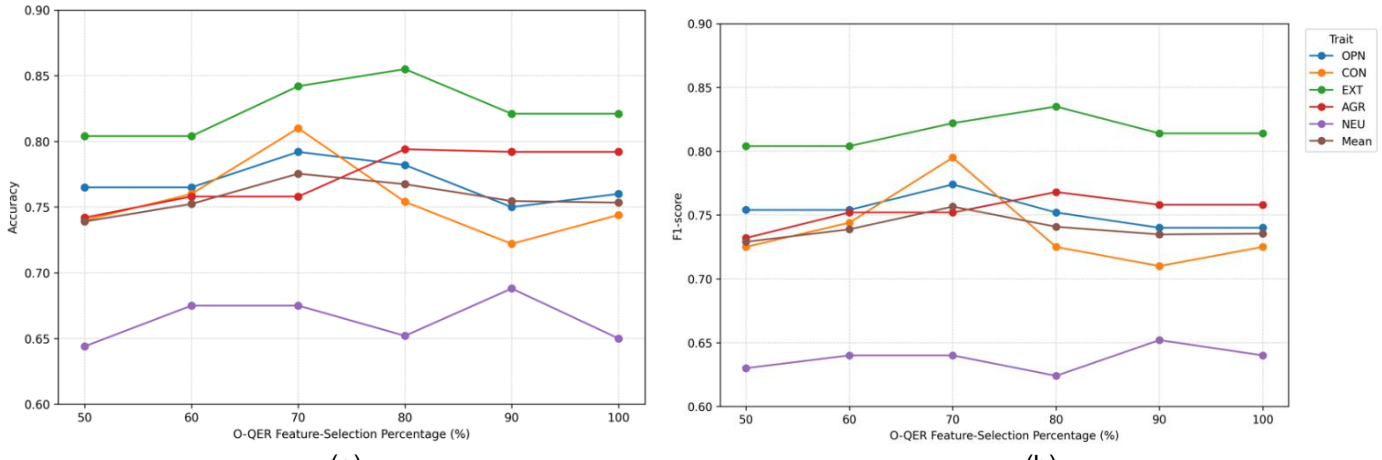

Figure 3. Performance Trends of the mBERT + O-QER Model Under Different Feature-Selection Percentages. (a) Accuracy Trends Across the Five OCEAN Traits. (b) F1-score Trends Across the Five OCEAN Traits

Figure 3 shows the performance trends of the proposed mBERT + O-QER model across different O-QER feature-selection percentages. In general, adding O-QER features improved the baseline mBERT performance in several settings. The best uniform setting was obtained with 70% O-QER, which achieved a mean accuracy of 0.7754 and a

mean F1-score of 0.7566. These results are higher than the baseline mBERT scores of 0.7548 for mean accuracy and 0.7358 for mean F1-score. This improvement suggests that O-QER adds useful class-discriminative lexical information that complements the contextual representations produced by mBERT.

At the trait level, Figure 3 shows that the best O-QER percentage was different for each OCEAN dimension. Openness and Conscientiousness achieved their highest performance at 70% O-QER, with accuracy and F1-score values of 0.792 and 0.774 for Openness, and 0.810 and 0.795 for Conscientiousness. Extraversion and Agreeableness performed best at 80% O-QER, reaching 0.855 accuracy and 0.835 F1-score for Extraversion, and 0.794 accuracy and 0.768 F1-score for Agreeableness. Neuroticism obtained its best result at 90% O-QER, with an accuracy of 0.688 and an F1-score of 0.652. These findings show that each personality trait requires a different amount of discriminative lexical information.

The different trends across O-QER percentages indicate that OCEAN personality prediction is trait-dependent. Openness and Conscientiousness seem to benefit from a moderate number of selected lexical cues, while Extraversion and Agreeableness require broader lexical coverage to capture more diverse social and interpersonal expressions. Neuroticism shows a slightly different pattern, as its best performance was obtained when more O-QER features were included. This may be because expressions related to emotional instability are more varied in Indonesian social media text. However, using 100% O-QER features did not produce the best overall performance, suggesting that adding all lexical features may introduce redundant or noisy information.

Table 4. Best Trait-specific Performance of the Proposed mBERT + O-QER Model Across OCEAN Personality Dimensions

OCEAN Trait	Optimal O-QER (%)	Accuracy	F1-score
Openness	70	0.792	0.774
Conscientiousness	70	0.810	0.795
Extraversion	80	0.855	0.835
Agreeableness	80	0.794	0.768
Neuroticism	90	0.688	0.652
Average Performance	**—**	**0.788**	**0.765**

Table 4 further summarizes the best O-QER percentage for each OCEAN trait. Using the best trait-specific configuration, the proposed model achieved an average accuracy of 0.7878 and an average F1-score of 0.7648. Compared with the baseline mBERT model, this represents an improvement of 0.0330 in average accuracy and 0.0290 in average F1-score. This result shows that selecting the most suitable O-QER percentage for each trait is more effective than applying one fixed percentage across all traits. Overall, these findings support the role of O-QER as a selective lexical feature learning mechanism. By emphasizing trait-relevant and class-discriminative cues, O-QER strengthens the contextual representation learned by mBERT. The results also suggest that OCEAN traits are expressed through different linguistic patterns in Indonesian social media text, making a trait-specific O-QER configuration a more suitable strategy for personality prediction.

The proposed framework was compared with the closely related study by Pradnyana et al. [8], which applied fine-tuned IndoBERT models to Big Five personality prediction using a comparable Indonesian social media dataset. IndoBERT-base achieved an accuracy of 0.72 and an F1-score of 0.71, whereas IndoBERT-large achieved 0.78 and 0.74, respectively. In comparison, the proposed mBERT–O-QER model achieved a mean accuracy of 0.7754 and a mean F1-score of 0.7566 using a uniform 70% feature-selection setting. Although its accuracy was slightly lower than that of IndoBERT-large, its F1-score was higher. When the optimal O-QER percentage was selected separately for each trait, the proposed model achieved an average accuracy of 0.7878 and an average F1-score of 0.7648, exceeding the IndoBERT-large results by 0.0078 and 0.0248, respectively.

These results suggest that the improvement was not solely due to the use of a different pretrained language model, but also to the integration of contextual representations with trait-specific class-discriminative lexical features. Unlike the previous approach, which relied exclusively on IndoBERT fine-tuning, the proposed framework incorporates token-level and post-level O-QER information to emphasize lexical cues associated with the High and Low classes of each OCEAN dimension. Nevertheless, the numerical comparison should be interpreted cautiously because differences in preprocessing, model configuration, data partitioning, and reporting precision may affect direct comparability. The main contribution of this study therefore lies in demonstrating that trait-specific lexical ranking can complement transformer-based representations for OCEAN personality prediction from Indonesian social media text.

4. Conclusion

This study proposed an OCEAN personality prediction framework for Indonesian social media text by integrating pretrained language models with an OCEAN-aware Query Expansion Ranking (O-QER) feature learning mechanism. The research was carried out in two main stages. The first stage evaluated several pretrained language models to identify the most suitable contextual representation backbone for the task. The second stage integrated the best-performing model with O-QER features to strengthen the model's ability to capture class-discriminative lexical cues. The results show that pretrained language models provide a strong foundation for understanding informal Indonesian social media text. However, relying only on contextual representations may not fully capture lexical patterns that are closely related to personality traits. The integration of O-QER helps address this limitation by identifying tokens that are more strongly associated with the High and Low classes of each OCEAN dimension. As a result, the proposed framework combines the semantic strength of pretrained language models with explicit personality-aware lexical evidence.

An important finding of this study is that the most effective proportion of selected O-QER features differs across personality traits. This indicates that each OCEAN dimension has its own lexical characteristics and does not necessarily benefit from the same feature-selection setting. Some traits may be better represented by a smaller set of highly discriminative tokens, while others require a wider range of lexical information. These findings suggest that trait-specific feature selection is important when modeling personality from noisy and informal social media data. Overall, the proposed O-QER mechanism offers a promising way to enhance OCEAN personality prediction from Indonesian social media text. By combining contextual embeddings and class-discriminative lexical features, the framework provides a more comprehensive representation than using pretrained language models alone. The findings also highlight the value of explicit lexical feature learning in personality prediction, especially in low-resource and informal language settings.

Future research can extend this work in several directions. First, the proposed framework should be evaluated on larger and more diverse Indonesian social media datasets to examine its generalizability across different platforms, topics, and user populations. Second, the O-QER mechanism can be further developed by incorporating semantic similarity, contextual token weighting, or attention-based lexical scoring, so that it can capture not only discriminative token occurrence but also contextual meaning. Third, more advanced fusion strategies, such as gated fusion or attention-based fusion, can be explored to balance contextual and lexical representations more adaptively. Fourth, future studies may compare the proposed framework with recent large language models to assess its competitiveness in low-resource and informal language contexts. Finally, further interpretability analysis can be conducted to examine how specific O-QER features contribute to the prediction of each OCEAN personality trait.

References

[1] F. Celli et al., "Twenty Years of Personality Computing: Threats, Challenges and Future Directions," ACM Comput. Surv., vol. 58, no. 11, pp. 1–37, Aug. 2026. https://doi.org/10.48550/arXiv.2503.02082

[2] G. A. Pradnyana, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "An explainable ensemble model for revealing the level of depression in social media by considering personality traits and sentiment polarity pattern," Online Soc. Netw. Media, vol. 46, May 2025. https://doi.org/10.1016/j.osnem.2025.100307

[3] G. Z. Nabiilah and D. Suhartono, "Personality Classification Based on Textual Data using Indonesian Pre-Trained Language Model and Ensemble Majority Voting," Revue d'Intelligence Artificielle, vol. 37, no. 1, pp. 73–81, Feb. 2023. https://doi.org/10.18280/ria.370110

[4] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," Artif. Intell. Rev., vol. 53, no. 4, pp. 2313–2339, Apr. 2020. https://doi.org/10.1007/s10462-019-09770-z

[5] A. Bruno and G. Singh, "Personality Traits Prediction from Text via Machine Learning," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), 2022. https://doi.org/10.1109/aic.2022.99

[6] H. Ahmad, M. Z. Asghar, A. S. Khan, and A. Habib, "A Systematic Literature Review of Personality Trait Classification from Textual Content," Open Computer Science, vol. 10, no. 1, pp. 175–193, Jul. 2020. https://doi.org/10.1515/comp-2020-0188

[7] W. Kang, F. Steffens, S. Pineda, K. Widuch, and A. Malvaso, "Personality traits and dimensions of mental health," Sci. Rep., vol. 13, no. 1, p. 7091, May 2023. https://doi.org/10.1038/s41598-023-33996-1

[8] G. A. Pradnyana, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "Fine-Tuning IndoBERT Model for Big Five Personality Prediction from Indonesian Social Media," in 2023 International Seminar on Intelligent Technology and Its Applications (ISITIA), IEEE, Jul. 2023, pp. 93–98. https://doi.org/10.1109/ISITIA59021.2023.10221074

[9] M. L. Smith, D. Hamplová, J. Kelley, and M. D. R. Evans, "Concise survey measures for the Big Five personality traits," Res. Soc. Stratif. Mobil., vol. 73, Jun. 2021. https://doi.org/10.1016/j.rssm.2021.100595

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018. https://doi.org/10.48550/arXiv.1810.04805

[11] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," Sep. 2021. https://doi.org/10.48550/arXiv.2109.04607

[12] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics, Online, 2020, pp. 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

[13] R. L. Vásquez and J. Ochoa-Luna, "Transformer-based Approaches for Personality Detection using the MBTI Model," in Proceedings - 2021 47th Latin American Computing Conference, CLEI 2021, Institute of Electrical and Electronics Engineers Inc., 2021. https://doi.org/10.1109/CLEI53233.2021.9640012

[14] E. Kerz, Y. Qiao, S. Zanwar, and D. Wiechmann, "Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features," in Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 182–194. https://doi.org/10.18653/v1/2022.wassa-1.17
[15] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer," Soc. Netw. Anal. Min., vol. 15, no. 1, p. 30, Mar. 2025, https://doi.org/10.1007/s13278-025-01444-9
[16] L. Hu, H. He, D. Wang, Z. Zhao, Y. Shao, and L. Nie, "LLM vs Small Model? Large Language Model Based Text Augmentation Enhanced Personality Detection Model," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, pp. 18234–18242, Mar. 2024. https://doi.org/10.1609/aaai.v38i16.29782
[17] T. Parlar and S. A. Ozel, "A new feature selection method for sentiment analysis of Turkish reviews," in Proceedings of the 2016 International Symposium on Inovations in Intelligent Systems and Applications, INISTA 2016, Institute of Electrical and Electronics Engineers Inc., Sep. 2016. https://doi.org/10.1109/INISTA.2016.7571833
[18] T. Parlar, S. A. Özel, and F. Song, "QER: a new feature selection method for sentiment analysis," Human-centric Computing and Information Sciences, vol. 8, no. 1, Dec. 2018. https://doi.org/10.1186/s13673-018-0135-8
[19] G. A. Pradnyana, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "Enhancing MBTI Personality Trait Prediction from Imbalanced Social Media Data Using Hybrid Query Expansion Ranking and Glo Ve- BiLSTM," in 2023 IEEE International Conference on Fuzzy Systems (FUZZ), IEEE, Aug. 2023, pp. 1–6. https://doi.org/10.1109/FUZZ52849.2023.10309718
[20] V. Ong, A. D. S. Rahmanto, W. Williem, N. H. Jeremy, D. Suhartono, and E. W. Andangsari, "Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model," International Journal of Intelligent Engineering and Systems, vol. 14, no. 2, pp. 248–261, 2021, https://doi.org/10.22266/ijies2021.0430.22
[21] H. Lucky, Roslynlia, and D. Suhartono, "Towards Classification of Personality Prediction Model: A Combination of BERT Word Embedding and MLSMOTE," in Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 346–350. https://doi.org/10.1109/ICCSAI53272.2021.9609750
[22] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, https://doi.org/10.48550/arXiv.1907.11692
[23] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 833–844. https://doi.org/10.18653/v1/D19-1077
[24] D. Sebastian, H. D. Purnomo, and I. Sembiring, "BERT for Natural Language Processing in Bahasa Indonesia," in 2022 2nd International Conference on Intelligent Cybernetics Technology and Applications, ICICyTA 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 204–209. https://doi.org/10.1109/ICICyTA57421.2022.10038230
[25] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: A survey," Pattern Recognit. Lett., vol. 167, pp. 204–211, Mar. 2023, https://doi.org/10.1016/j.patrec.2023.02.016
[26] 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, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4996–5001. https://doi.org/10.18653/v1/P19-1493