

Original article

Supervised learning algorithms for subject- and text-level classification of depression-related context among Thai-language Facebook users

Keito R. Yoneyama^{a,*} , Peerapon Vateekul^{b,c} , Solaphat Hemrungronj^{c,d} 

^aMedical Sciences Program, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

^bDepartment of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

^cCenter of Excellence in Cognitive Fitness and Biopsychiatry Technology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

^dDepartment of Psychiatry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

Abstract

Background: Depression is an increasing concern worldwide that requires urgent attention, particularly in Thailand. Without timely intervention, it can lead to serious consequences, such as health complications, reduced social participation, and even suicide. Consequently, research has increasingly focused on developing algorithms to identify and classify social media data into depression severity.

Objectives: This study aimed to examine label mismatches between self-assessment forms and social media behaviors and to evaluate the feasibility of using experts' opinions to classify Thai Facebook text into depression categories.

Methods: The methodology involved data collection through surveys and Facebook scraping, followed by data preprocessing, feature engineering, and model training and validation. The models employed included the support vector machine, logistic regression, and decision tree, applied at the subject- and text-level. Labels were derived from the Thai Depression Inventory (TDI) and experts' evaluations.

Results: Only 30 out of 60 participants actively used Facebook during the study period, with substantial variation in posting behavior. Subject-level classification using TDI labels yielded better, more balanced performance metrics than classifications based on experts' opinions at the subject- and text-level. Part-of-speech tagging showed slight correlations with depression severity, whereas emotions did not.

Conclusion: Subject-level classification using self-assessment labels performed better than expert-derived labels and text-level classification when compared to a dummy baseline model. This suggests a slight correlation between the self-assessment forms and social media behaviors. In contrast, the poor agreement between labeling approaches suggests that text data alone may be insufficient as a reliable clinical indicator of depression.

Keywords: Depression severity, Facebook text, machine learning, natural language processing, Thai.

Depression is one of the most common and concerning mental health issues worldwide, with the number of affected individuals increasing by 59.3%, or 290 million cases, in 2019. Beyond its widespread prevalence, it also contributes substantially to various areas, including cardiovascular diseases, increased suicide rates, and economic impact.⁽¹⁾ Seeking help from psychiatrists, family

members, or friends can reduce the severity of this condition. Recently, another common approach is the use of social media to express feelings, thoughts, and moods in an open and anonymous environment.⁽²⁾

Over the past two decades, Thailand has undergone rapid demographic and economic changes, which have placed a considerable burden on the healthcare system and affected the mental health of its citizens.⁽³⁾ According to health organizations in Thailand, the number of depression diagnoses has increased from approximately 355,000 in 2020 to 358,000 in 2021. Moreover, suicide rates have increased from 5 to 7 per 100,000 individuals in 2021.⁽⁴⁾ Timely intervention, particularly consultation with a

***Correspondence to:** Keito R. Yoneyama, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
E- mail: 6678019030@student.chula.ac.th

Received: December 3, 2025

Revised: April 10, 2026

Accepted: May 17, 2026

psychiatrist, can help reduce the severity of depression; however, most hospitals face challenges such as insufficient staff, underfunding, and patient overload.⁽⁵⁾ The World Health Organization stated that the average number of mental health professionals worldwide is 10.15 per 100,000 people. As of 2022, Thailand has approximately 2.85 per 100,000, with most concentrated in Bangkok and other major cities.

Unfortunately, more than 70% of individuals do not seek help due to stigma, discrimination, and negative stereotypes created by society, including societal pressure, cultural norms, and anticipated discrimination.^(6,7) Therefore, various questionnaires are becoming more widely available as preliminary screening tools for self-assessment, with the aim of promoting self-awareness and empowering individuals to take proactive steps toward their well-being. Well-known examples include the Beck Depression Inventory (BDI), Hamilton Rating Scale for Depression (HRSD), and Zung Self-Rating Depression Scale (SDS).⁽⁸⁾ However, most of these tools are largely derived from Western concepts of depression, which may limit their reliability and validity when applied to other cultural contexts, particularly in Asia. This is because Western ethnocentric frameworks often fail to capture the culturally specific expressions.⁽⁹⁾ As a result, Lotrakul M, et al. developed the Thai Depression Inventory (TDI), which is a scale specifically designed to measure the severity of depression within the Thai population. The TDI was adapted from the BDI, SDS, and HRSD, and it incorporates culturally specific expressions as well as common manifestations of distress among Thai individuals with depression.⁽¹⁰⁾ Despite the availability of these tools, their reliance on manual administration remains a limitation. This has motivated a growing interest in computational approaches.

One of the most transformative and adaptable technologies of the 21st century is machine learning (ML), along with natural language processing (NLP).⁽¹¹⁾ By integrating these two fields, researchers have developed innovative tools for classification tasks. Such tools are scalable, cost-effective, and proactive, as well as applicable to languages other than English. For example, Sho T, et al. conducted a study to evaluate the effectiveness of using social media activity, combined with a support vector machine (SVM) model, for estimating depression severity.

Their methodology involved participants completing a self-assessment questionnaire to measure the degree of depression, which was then used to label Japanese data collected from Twitter.⁽¹²⁾ Similarly, Boriarn K, et al. compared the performance of convolutional neural networks, long short-term memory networks, and a hybrid of both to determine the most effective combination for Thai text.⁽¹³⁾ Another similar study was conducted by Dhiaa M, et al., which focused on widely used ML algorithms, including random forest, logistic regression (LR), and Naïve Bayes, applied to Arabic Twitter data.⁽¹⁴⁾ The latter two studies collected data and labeled it based on depression-related keywords.

Even though numerous studies have been dedicated to this field, most have primarily focused on evaluating model performance using labels from self-assessment questionnaires or depression-related keywords. While this is a promising approach, it raises an important question about whether identifying the best-performing model or using keyword-based data collection truly reflects the practical effectiveness of ML models in diagnosing depression in real-world scenarios. Therefore, this study aimed to investigate whether ML models are a suitable approach for classifying depression into different categories based solely on Facebook text. It examined label mismatches between self-assessment forms and social media behaviors. In addition, it evaluated whether these data alone could provide sufficient information for this task by analyzing them at the subject- and text-level, using labels derived from self-assessment forms and experts' opinions. This article contributes to the growing body of research on mental health classification by evaluating the feasibility of applying ML to classify depression-related categories based on text, providing insights into the extent to which social media data can serve as a reliable indicator of depression, and offering suggestions for future research in this field. The findings are intended to inform researchers of the potential and limitations of ML for text-based depression screening. Unlike traditional approaches that rely solely on self-assessment tools or experts' opinions, the methodology used in this research integrates both evaluations, with classification conducted at the subject- and text-level. In this research, expert opinion refers to the evaluation conducted by a board-certified psychiatrist experienced in diagnosing mental illnesses. Furthermore, subject-level classification involves concatenating all texts posted by each participant

during the study period, with each data point representing a participant and potentially including multiple texts. In contrast, text-level classification treats each post as a separate data point. In addition, this study recruited participants with a history of seeking help from a psychiatrist due to depression.

Materials and methods

Figure 1 provides an overview of the methodology used in this research. The four main steps, indicated within the dotted black box, are as follows: 1) data collection; 2) preprocessing and feature engineering; 3) data labeling; and 4) model selection, training, and validation. In the "Preprocessing and feature engineering" section, the colored boxes indicate the workflows used for the subject- (pink) and text-level (purple) classification procedures. Similarly, in the "Data labeling" section, the green boxes represent the experts' labeling procedures. These main steps are explained in detail in the following subsections.

Data collection

Before data collection, full board review and ethical approval were obtained from the Institutional Review Board of the Faculty of Medicine, Chulalongkorn University (COA no. 0229/2025). Next, a Google Form was created, consisting of two sections:

1. General questions comprised of five short-answer questions for participants who had never visited a psychiatrist and six for those who had (**Figure 2A**). It collected demographic information as well as the participant's Facebook account name.

2. TDI self-assessment questions comprised of 20 multiple-choice questions (**Figure 2B**).

These forms were primarily distributed via Line groups to friends and family members. In addition, during data scraping, a new Facebook account was created to ensure that only public profiles were being used.

Data preprocessing and feature engineering

The two essential steps were data preprocessing and feature engineering. Data preprocessing involved cleaning the text by removing noise, such as emojis, hashtags, and special characters, and performing spell checks. This was followed by tokenization, which breaks down the text into smaller units called "tokens", then vectorization. The tokens were mapped into vectors in a numerical space, which can be interpreted and processed by machines. Another important step was feature engineering, which focused on extracting meaningful attributes that aid classification and differentiation. Therefore, the selected features included linguistic structure and emotional expression, as these are indicators that are commonly evaluated by psychiatrists.^(15, 16)

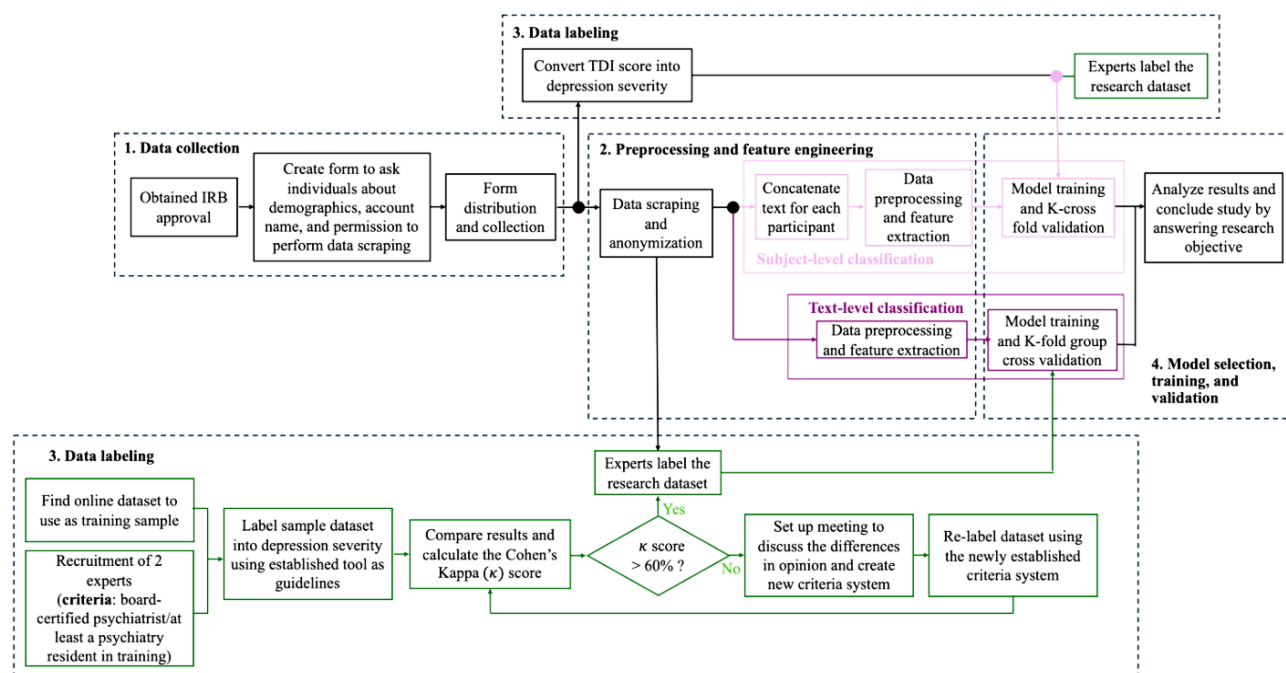


Figure 1. Methodology overview.

ส่วนที่ 1: ข้อมูลเกี่ยวกับอาสาสมัคร/ผู้เข้าร่วมการวิจัยและบัญชีเฟซบุ๊ก 1. เพศ: 2. อายุ: 3. การศึกษา: 4. คุณเคยไปพบจิตแพทย์ที่โรงพยาบาลจุฬาลงกรณ์มหาวิทยาลัยหรือไม่: ☐ เคย (ถ้าผู้เข้าร่วมตอบว่า "เคย" ค่าถามต่อไปจะเป็นคำถาม 4.1 แล้วค่อยคำถาม 5) ☐ ไม่เคย (ถ้าผู้เข้าร่วมตอบว่า "ไม่เคย" ค่าถามต่อไปจะเป็นคำถาม 5) 4.1 ถ้าเคย มาพบในเรื่องอะไร: 5. ชื่อบัญชีเฟซบุ๊ก: หากไม่พบข้อมูลของอาสาสมัคร/ผู้เข้าร่วมการวิจัย กรุณาแนบรูปถ่ายหน้าบัตรประชาชนของผู้เข้าร่วมการวิจัย	ส่วนที่ 2: แบบสอบถามเพื่อประเมินความรุนแรงของภาวะซึมเศร้า คำถามที่จะถามต่อไปนี้ เป็นคำถามที่เจาะลึกถึงข้อมูลส่วนบุคคลที่มีความละเอียดอ่อน ซึ่ง อาสาสมัคร/ผู้เข้าร่วมงานวิจัย จะเลือกตอบหรือไม่ตอบคำถามใด กรุณาเลือกข้อที่ตรงกับท่าน หรือใกล้เคียงมากที่สุด ในช่วง 1 สัปดาห์ที่ผ่านมา: 1. ☐ ท้อแท้ใจมาก หมดหวังในอนาคต ☐ ท้อแท้ใจ มองอนาคตในแง่ร้าย ☐ ท้อแท้ใจบ้างบางครั้ง ☐ ไม่หมดหวัง 2. ☐ คิดวางแผนฆ่าตัวตาย ☐ คิดอยากตาย ☐ ชีวิตนี้ไม่น่าอยู่ คิดถึงเรื่องความตายบ่อยๆ แต่ไม่ถึงกับอยากตาย ☐ ไม่ได้คิดถึงเรื่องการตาย 3. ☐ รู้สึกมากกว่าตัวเองเป็นคนไม่ดี หรือไม่มีค่าเลย ☐ รู้สึกผิด คิดแต่เรื่องความผิดของตัวเองในอดีต ☐ รู้สึกตัวเองเป็นคนไม่ดีอยู่บ่อยๆ ☐ ไม่รู้สึกผิด หรือคิดว่าตัวเองไม่ดี
A	B

Figure 2. Google Form used for data collection. (A) section 1 (general questions) and (B) section 2 (TDI).⁽¹⁰⁾ TDI, Thai Depression Inventory.

Data labeling

Data labeling was derived from two sources: self-assessment and experts' opinion. Once the participants completed the TDI form, their total scores were calculated, and a depression severity level was assigned according to the following ranges: (1) no depression: 0-20; (2) mild depression: 21-25; (3) major depression: 26-40; and (4) severe depression: 41-60.

For the expert's opinion, two board-certified psychiatrists were recruited to assist in labeling the text and subject. Before labeling the research dataset, they were asked to independently label a sample dataset consisting of 30 social media posts, as shown in **Figure 3.**⁽¹⁷⁾

Text	Depression classification
"และฉันก็ไม่มีทางที่จะเข้าใจความรู้สึกของเธอได้เลย แม้แต่ชนิดเดียว ทุกคำพูดของเธอที่พร่ำพรูออกมา และความรู้สึกของฉันที่ติดอยู่ในใจ"	
"ไม่ต้องเข้าใจทุกอย่างที่เรา รู้สึกก็ได้ แต่ที่รู้ว่ามีเพื่อนที่ไม่ไปไหน"	
"เสียงของลูกสาววัยสิบหกปีของเธอ ดังขึ้นและนั่นเป็นคำพูดสุดท้ายที่ชายหนุ่มได้ยินเป็นคำสุดท้ายที่เขาได้ยินเมื่ออยู่บนโบกี้ลิฟต์นี้ ขยายหนุ่มประคองสติที่เหลือน้อย น้อยเต็มทีเอาไว้ก่อนจะบินไทย ประดิษฐ์กระบอกเก่า ๆ ที่เขาเคยซื้อไว้สมัยเป็นเรียนชั้นมา ฟังอย่างเลื่อนลอย ในหัวของเขา คิดเพียงว่าคงจะง่ายกว่าถ้าเมื่อกระสุนสักนัดแล่นผ่านเขาไป จากนั้นเขาก็อาจจะไม่ต้องคิดอะไรอีก"	
"เหมือนเขาเปิดโลกให้เธอเข้ามาซะแล้วสิ หรืออาจจะ เป็นทั้งจักรวาลของเขาเลยก็ได้! ในขณะที่เขาเล็ก ๆ พยายามเดินตามอย่างขั่นเข้ง มันคงจะดีกว่าเล็ก ๆ นั้นก้าวทันเขา เขาจะได้ไม่ต้องคอยหยุดเดินเพื่อรอเธอ"	
"ความสัมพันธ์ระหว่างประเทศ / เศรษฐกิจ เช่น ทำให้การที่จีนสั่งห้ามหมู่เกาะอันเหมียวอย่างอลังการ แต่ UN กลับไม่สามารถทำอะไรด้วย"	
"ผลที่ได้คือ ไม่มีใครเชื่อ โดนคำกลัมน่าว่าเราใส่ร้ายคนอื่น เป็นบาป และคำเราว่าเป็นโรคประสาท ไม่มีคำปลอมโยน มีแต่หน้าตาเราที่ไหลออกมาไม่รู้จบ ตาขวามือ บัดทั้งตัว บัดทั้งใจเลย ไม่กล้าบอกเรื่องที่เคยหยุดลมหายใจตัวเองเลย"	
"และยากแค่ไหนกว่าจะถึงบ้างล่ะ" หัวผมไร้ซึ่งข้อมูล มีแต่ความห้าวและวิธีกลิ้งตกเขา ยังใจให้จับกิ่งไม้ทัน" นี่ยังไม่ถึง 1 / 4 ของทั้งหมด"	
"แล้วทางยากประมาณนี้ มีกี่ครั้ง"	
"ผมนอนหลับตาฟังเสียงสายฝนที่ตกกระทบลงมาอยู่ราวครึ่งชั่วโมง แต่ไม่มีท่าทีว่าจะเข้าสู่ห้วงนิรหาในคืนนี้" โอเค นอนไม่หลับ ก็ไม่ต้องนอน"	
"ผมตื่นร่างกายตัวเองให้ลุกขึ้นจากเตียงนอนก่อนก้าวขาตรงไปยังประตูระเบียง แสงไฟสลัวจากภายนอกห้องนอนกับสายฝนที่ปรากฏบนผ้านสายตาของผม ในตอนนี้อยู่คนเดียวคนเดียวอยู่คนเดียว" อาแปะกับดิฉันครับ ที่ฤดูร้อนมีฝนตก" " และสำหรับคนชอบฤดูฝนเป็นทุนเดิมอย่างผม" " จะคิดซะว่าเป็นกำไร" ผมเอื้อมมือเปิดประตูระเบียงพร้อมกับพาวางกายที่ถูกจัดแจงด้วยเสื้อโอเวอร์ไซส์สีขาวกับกางเกงนอนสีขาว สิวเทาออกไปยืนดื่มด่ำบรรยากาศที่คุ้นชินในความชื้นซอบจ่อมองหยดฝนที่หลังวันลงมานั้นที่ไม่มีท่าทีจะหยุดลงนั้นทำให้ผมรู้สึกพอใจอยู่ไม่น้อย"	
"อยากให้ช่วงเวลาที่ดีแบบนี้เกิดขึ้นกับผมอีก เพื่อให้ผมได้มีแรงผลักดันในการใช้ชีวิตเพื่อเป็นแรงบันดาลใจในยามที่ท้อและเหนื่อยล้าและเพื่อเป็นกำลังใจในการก้าวเดินไปข้างหน้าเมื่อความคิดเหล่านี้จบลงเหลือไว้เพียงรอยยิ้มบนใบหน้าของผม ให้เธอได้ยิ้มตาม"	
"ผมเรียบเรียงอะไรไม่ได้เลยได้แต่บอกหมอว่า" ก็น่าจะเครียดแล้วรู้สึกว้าวเปอร์เกินไปจนเหนื่อยแต่ถ้า 1 / 4 นั้นน้อยไปจนไม่หายเหนื่อย" ผมถามหมอว่าหมอจะเปลี่ยนยาไหม"	
"สายลมรักได้พัดมาที่เธออีกครั้ง เจสันก็รู้สึกดีที่เขาได้รับการปล่อยใจจากแม่หมดทั้งสองด้าน เป็นที่พึงใจให้กันและกัน แต่แม่ก็รับรู้ได้ว่าเจสันนั้นมิใช่สิ่งง่ายอยู่ด้วยความรักที่มีให้อัคริน เจสันเองก็ถามความปรารถนาว่าตอนนี้เจสันต้องการสิ่งใด เจสันอยากพบครอบครัวอยากกลับบ้านเมื่อได้ยินเช่นนั้น แม่จึงให้ของวิเศษหาหาหาหนึ่งที่สามารถพาเจสันไปหาคนที่รักได้ เจสันสัญญาว่าเมื่อเขาไปหาเขาจะกลับมาหาเธอ เจสันใช้ของวิเศษนั้นไปหาครอบครัวเจสันเหมือนได้พักผ่อนเขาได้แล้วว่าชีวิตนี้เขาอยู่เพื่อครอบครัวเขาจึงจะกลับไปทักอัครินเพื่อฝึกต่อให้จบแต่แล้วเรื่องไม่คาดฝันก็เกิดขึ้น"	
"ฉันจะตาเป็นประกายเสมอฉันกลับบ้านครั้งล่าสุด"	
"สิ่งนี้ทำให้ฉันได้คิดอะไรหลาย ๆ อย่างได้ หรือจะเพื่อนอีกคนที่หันไปเอาดีทางด้านธุรกิจส่วนตัว"	
"ส่วนเรานั้นโรคทางกายนั้นไม่มีชีวิตแต่ไม่ยากมีชีวิตอยู่"	
"บอกหน้าหมาหมอดูให้คนอื่นที่มิเข้าใจตราหน้าว่ายุ่งยากก็เพราะว่าอยากรู้เรื่องความรัก เรื่องเนื้อคู่ที่รอมานานนักหนา ราวกับว่ากำลังเดินทางมาจากโลกอื่น เป็นอีกเรื่องในชีวิตนอกจากเรื่องเสียใจแล้วที่รู้เลยว่าไม่มีดวงจริงๆ ยังพอหมอดูพูดอีกเสียงก็ยังมั่นใจ" มีเงินก่อนแล้ว"	
"พอดได้อ่านเจอเพจของหมอเจ ในเฟซที่เขียนถึงอาการของ anxiety disorder ที่: https://www.facebook.com/D2JED/ ในเฟซที่เขียนถึงอาการของ anxiety disorder ที่มีทั้งหมด 13 ข้อ"	

Figure 3. Example of the sample dataset provided separately to experts for the inter-rater reliability test⁽¹⁷⁾, with the "Depression classification" column left blank for the experts to complete.

Table 1. Summary of the models, hyperparameters, and rationale used in this research.

Model	Hyperparameters	Rationale
SVM	max_iter = 1000, loss = 'hinge', C = 0.001,	Suitability in handling high-sparse data commonly seen after vectorization and used in text classification tasks due to its strong generalization performance.
LR	penalty = 'l1', max_iter = 1000, loss = 'log_loss', alpha = 2.64e-05, max_depth = 8,	Computational efficiency and ability to process large-scale sparse text representations.
DT	min_samples_leaf = 1, min_samples_split = 4, criterion = 'entropy'	Used as benchmark model to compare classification performance against linear-based approaches. It can also capture non-linear relationships between linguistic features and depression levels.

Their results were then used to compute the weighted Cohen's Kappa (κ) score. If the κ value exceeded 0.60, the experts would proceed to label the actual research dataset. If the κ value was below 0.60, they would undergo additional training and repeat the reliability assessment until the threshold was achieved.⁽¹⁸⁾ During the labeling of the research dataset, if the two experts disagreed on a label, the more severe label was assigned to prioritize early detection. For subject-level labeling, if more than one text instance was labeled with a higher severity, the final assigned subject-level label was assigned based on the highest severity classification.

Model selection, training, and validation

Based on previous studies that also used social media text, the best-performing algorithms were SVM, LR, and decision tree (DT).^(12, 14, 17) The SVM model exhibited strong generalization performance. The variant used was the LinearSVC classifier, which was selected for its suitability in handling high-dimensional, sparse data commonly produced after vectorization. The LR model, specifically the Stochastic Gradient Descent (SGD) classifier, was used as it has high computational efficiency and capable of processing large-scale data. This optimizer updates the model's weight incrementally using small random batches of training data rather than the entire dataset, thereby allowing efficient handling of high-dimensional and sparse inputs. Lastly, the DT model, DecisionTreeClassifier was used as a benchmark because it can capture nonlinear relationships. As for the hyperparameters, "RandomizedSearchCV" from "scikit-learn sklearn.model_selection" was used to identify the optimal combinations for maximizing model accuracy. A summary of the models, hyperparameters, and rationale is presented in **Table 1**.

The data were then split into training and testing sets using 5-fold cross-validation for the subject-level and 5-fold group cross-validation for the text-level classification to maximize the use of the small dataset. Participant Id.'s were not used as model inputs but rather as reference parameters. This ensured that data points belonging to the same participant were kept together, thereby preventing data leakage and overly optimistic performance estimates when the data are not independent and identically distributed. For model evaluation, "classification_report" from the sklearn.metrics library was used to automatically calculate the precision, recall, F1-score, support, and accuracy for the model.

Results

Overview of the collected data

A total of 60 forms were collected, comprising 42 female and 18 male participants. Among them, 45 individuals had never visited a psychiatrist, while the remaining 15 had. The primary reasons for these visits, in addition to depression, included anxiety, insomnia, and stress. The participants' ages ranged from 18 to 88 years, with a mean age of 46.0 ± 17.0 years. Most of the participants were either currently enrolled in or had completed a bachelor's degree (35), followed by a master's degree (17), a PhD (4), a vocational degree (2), a postgraduate degree (1), or a nursing degree (1). Participants with "mild depression" contributed 6 months of data collected between January 1 and June 30, 2025. In contrast, those with "severe depression" contributed 9 months of data collected between September 1, 2024, and June 30, 2025.^(15, 16)

From the collected data, some participants either did not post during the study period (6–9 months), did not have an active Facebook account, or had their accounts set to private. These individuals were therefore excluded from the study, as data scraping

could not be conducted. The removal of these participants introduced selection bias toward individuals with higher levels of social media activity. In addition, it limited the generalizability and applicability of the ML models, as the analysis was restricted to active Facebook participants only. Among these, the majority of users (16) suffer from major depression, followed by mild (11), severe (2), and no depression (1), based on their TDI responses.

Data preprocessing

The primary platform and tools used for model development were Jupyter Notebook, with Python's built-in "re" library and PyThaiNLP.⁽¹⁹⁾ The "re" library was used to remove noise such as hashtags, URLs, special characters, and emojis, while the latter was used for spell checking, tokenization, and vectorization.

Once the data had been scraped from Facebook and stored in an Excel sheet (**Figure 4A**), the first step was to remove noise using the "re" library by deleting a set of predefined characters and emojis, which resulted in the output shown in **Figure 4B**. The cleaned text then underwent spell check and tokenization (**Figure 4C**) before being vectorized. Tokenization was performed using the "pythainlp.tokenize.word_tokenize" function with the default "NewMM" engine. This engine is a dictionary-based Thai word segmentation algorithm that combines maximal-matching approaches with Thai character clusters. It works by inserting spaces between the longest possible matching words found in the Thai dictionary. Spell checking also relied heavily on the "Thai_words" function, which identifies the closest correct spelling for a given word by using a frozenset of approximately 62,000 common Thai words. Finally, vectorization was performed using the "word_vector" module with the default "thai2fit_wv" model. This model is based on the Word2Vec approach, a shallow neural network that learns word associations from bodies of large text. It contains more than 51,000 words, each represented by a 300-dimensional vector. The tokenized text was input into the model and matched against the "thai2fit_wv" vocabulary. If the token was present, its corresponding vector was stored. The vectors for all tokens within a sentence were then averaged to produce a single fixed-size vector representing the entire text.⁽¹⁹⁾ This averaging process enabled texts of varying lengths to be represented as fixed-size semantic vectors, allowing consistent input into the model while preserving the

contextual information. **Figure 4D** presents examples of 10 random words from the dataset and their corresponding vectors.

Labels derived from the TDI score and experts' opinions for the usable data

Based on the completed TDI scores, 16 individuals were classified as having "major depression", 11 with "mild depression", 2 with "severe depression", and 1 with "no depression". A total of 1,496 text posts were collected. However, due to the extremely limited number of samples in the "no depression" and "severe depression" categories, separate analysis was not statistically feasible. In addition, the corresponding TDI scores were borderline between these two classes. Therefore, these categories were merged into the "mild depression" and "major depression" classifications to minimize the effects of underrepresented classes and improve the stability of the classification analysis.

The inter-rater reliability test was calculated using the "irr" library in RStudio and the "kappa2" function, which computes the weighted κ score. This accounts for the degree of disagreement by assigning different weights to varying levels of discrepancy. The initial score was relatively low at 14.0%, prompting a discussion session in which both experts explained their labeling processes and collaboratively developed a new set of labeling criteria. Once the new criteria were established, the experts re-labeled the sample dataset, resulting in a substantially improved score of 76.0%. Subsequently, they were instructed to label the research dataset, which had already been preprocessed and anonymized. Each expert reviewed all text data for every participant and assigned text- and subject-level labels, with the latter based on all participants' posts.

Table 2 summarizes the number of participants and text entities in each depression severity category, based on the TDI score and experts' opinions, which were used as input for the ML model. In total, the experts labeled 1,382 rows of text, rather than the originally scraped 1,496, because some rows became empty after data preprocessing and were therefore removed. In addition, there was no TDI score for individual text entries because the TDI assesses a participant's overall emotional state. In contrast, text data are largely instantaneous and reflect how a participant feels at a specific moment. As a result, text data is more sensitive to short-term emotional fluctuations.

A	0 เมื่อความฝันของเธอ คือความฝัน และ หวังเดียวของ... 1 ✨งานมหกรรมหนังสือระดับชาติรอบนี้...\\nแอดฟอยยากขว... 2 📖❤️ รวมเกมบริหารกิจการ สนุก เล่นเพลิน ให้ฟลราย... 3 https://youtu.be/MsfqBPbe4H8?si=DLk6gfeCD8GAGjVx 4 ใครไลค์ขอให้มีแฟนแต่ถ้าแชร์ต่อจะมีพุ้นนี้เลย 5 ใครไลค์ขอให้มีแฟนแต่ถ้าแชร์ต่อจะมีพุ้นนี้เลย 6 มาแล้วววว ✨ \nญธงาน COSCI Alumni Job Fair 2025... 7 🚀 ศูนย์จำหน่ายรถไฟฟ้า ✨สาขา ราชพฤกษ์📍KT-EV05📍 ... 8 ขอแสดงความยินดีกับBG ที่ได้เกียรติยืนมอันดับ ... 9 ไลฟ์คะ ที่นี้คะ \nบ่ายสามครึ่งวันอาทิตย์นี้ ...
B	0 เมื่อความฝันของเธอ คือความฝัน และ หวังเดียวของ... 1 งานมหกรรมหนังสือระดับชาติรอบนี้...\\nแอดฟอยยากขว... 2 รวมเกมบริหารกิจการ สนุก เล่นเพลิน ให้ฟลราย... 3 https://youtu.be/MsfqBPbe4H8?si=DLk6gfeCD8GAGjVx 4 ใครไลค์ขอให้มีแฟนแต่ถ้าแชร์ต่อจะมีพุ้นนี้เลย 5 ใครไลค์ขอให้มีแฟนแต่ถ้าแชร์ต่อจะมีพุ้นนี้เลย 6 มาแล้วววว \nญธงาน COSCI Alumni Job Fair 2025 ... 7 ศูนย์จำหน่ายรถไฟฟ้าสาขา ราชพฤกษ์KTEV05 เปิดแล... 8 ขอแสดงความยินดีกับBG ที่ได้เกียรติยืนมอันดับ ... 9 ไลฟ์คะ ที่นี้คะ \nบ่ายสามครึ่งวันอาทิตย์นี้ ...
C	0 [เมื่อ, ความฝัน, ของ, เธอ, คือ, ความฝัน, และ, ... 1 [งาน, มหกรรม, หนังสือ, ระดับชาติ, รอบ, นี้, แอ... 2 [รวม, เกม, บริหาร, กิจการ, สนุก, เล่น, เพลิน, ... 3 [nan] 4 [ใคร, ไลค์, ขอ, มี, แฟน, แต่, ถ้า, แชร์, ต่อ, ... 5 [ใคร, ไลค์, ขอ, มี, แฟน, แต่, ถ้า, แชร์, ต่อ, ... 6 [มา, แล้ว, ญธ, งาน, โดย, มา, พบ, กับ, เรา, ที่... 7 [ศูนย์, จำหน่าย, รถไฟฟ้า, สาขา, ราชพฤกษ์, เปิด... 8 [ขอแสดงความยินดี, กับ, ที่, ได้, เกียรติยืนม, ... 9 [ไลฟ์, คะ, ที่นี้, คะ, บ่าย, สาม, ครั้ง, วัน...
D	Word: ข้าเงา -> : [0.004759 -0.018662 -0.088027 0.008484 -0.024828 -0.027423 -0.016357] Word: ข้าเง -> : [-0.342918 0.280995 -0.208891 -0.163247 0.201959 -0.451034 -0.347697] Word: คุณครู -> : [0.052834 0.047166 0.247386 -0.087843 -0.030104 0.15188 -0.020633] Word: ประสิทธิ์ -> : [-0.010478 0.07826 -0.089248 -0.236044 -0.104712 -0.022505 -0.093837] Word: เหนื่อยอ่อน -> : [0.000047 -0.016536 -0.082356 0.01152 -0.025415 -0.011029 -0.015677] Word: นิ่ง -> : [-0.174175 -0.092187 0.066448 0.096926 -0.324222 -0.098734 0.030752] Word: บอล -> : [0.335247 -0.354551 0.379746 0.157264 0.380625 0.046399 -0.164222] Word: ผู้ปกครอง -> : [0.070706 0.095563 -0.169842 0.101871 0.21644 -0.039019 -0.167663] Word: บรรจุภัณฑ์ -> : [-0.016636 -0.001467 -0.077151 0.006055 -0.018888 -0.031713 -0.026633] Word: หอระพีง -> : [-0.01152 -0.025724 -0.091054 0.0064 -0.039774 -0.049601 -0.022726]

Figure 4. (A) Original Facebook text, (B) results after noise removable, (C) results after spell check and tokenization, and (D) examples of 10 random vectorized words.

Table 2. The number of participants and text entities for each depression severity level based on TDI score and experts' opinion.

Depression severity	TDI score		Experts' opinion	
	No. participants*	No. participants	No. text entities	No. text entities
No relevance	-	10		1,206
No depression	-	17		126
Mild depression	12	3		50
Major depression	18	-		-

TDI, Thai Depression Inventory.; No. participants*, Number of participants; No. text entities, Number of text entities; No, No depression; Mild, Mild depression; Major; Major depression.

*The number of participants in the TDI column excludes the “no depression” and “severe depression” groups because they were merged.

Table 3. Comparison between depression severity labels derived from TDI score and experts' opinion for subject-level classification.

		TDI score			
		No	Mild	Major	Total
Experts' opinion	No relevance	0	6	4	10
	No	0	6	11	17
	Mild	0	0	3	3
	Major	0	0	0	0
	Total	0	12	18	

TDI, Thai Depression Inventory; No, no depression; Mild, mild depression; Major, major depression; Total; total number of labels for each category.

Table 3 presents a confusion matrix comparing depression-related labels across the TDI scores and expert evaluations at the subject-level. The TDI categorized depression severity into four levels: no, mild, major, and severe depression, whereas the experts classified participants into three categories: no relevance, no depression, and mild depression. The term "no relevance" indicates that the text was ambiguous and did not provide sufficient evidence to determine the individual's mental state.

As shown in **Table 3**, there was limited agreement between the two labeling methods, with no participants receiving identical labels from both sources. The agreement rate was 0.0%, and the calculated κ was extremely low at 11.0%, indicating weak consistency between the TDI-based and experts-derived labels. This discrepancy may be attributed to a time-window mismatch between the two labeling approaches. The TDI questionnaire captures depression-related context over two weeks, whereas expert labels were derived from Facebook data spanning several months. Nevertheless, three participants were labeled as having "major depression" according to their TDI scores, but experts assigned them as having "mild depression". This may suggest that certain depression-related indicators could still be observed from text, even without prior background knowledge or detailed clinical history. In addition, self-presentation bias on Facebook may contribute to this gap, as users can curate their idealized online personas. This results in a divergence between an individual's cyber-self and their real-life mental state.

Feature extraction

The selected features included part-of-speech (POS) tagging and emotion, both of which have been shown in numerous studies to be effective in supporting the classification process.⁽¹⁹⁾ POS tagging covers a broad range of grammatical categories; however, the

features chosen for this study were the number of nouns and verbs used, the number of word tokens, and the average token length. The word count and average token length were computed using standard Python functions, where the word count was obtained from the length of the token list, and the average token length was calculated by dividing the sum of token lengths by the total number of tokens. Noun and verb counts were derived using the "pos_tag" module by counting words tagged as nouns ("N") or verbs ("V") according to the Perceptron model that was trained on the ORCHID dataset.⁽²⁵⁾

Emotion detection was performed using the Sentiment library, which is a pretrained model within the PyThaiNLP corpus. This model was developed by native Thai speakers, combined with deep learning approaches (e.g., DeepCut and neural networks) and dictionary-based feature models designed and trained for this task. The engine used was a lexicon-based approach. Each token in a sentence was assigned a positive (pos) or negative (neg) score based on the lexicon, and these scores were summed to determine the overall text sentiment.

Figure 5A shows that the word count for the 30 participants varied substantially, with an average value of $2,683 \pm 5,160$. Some participants posted very little (e.g., Participant Id. 6 posted only five words), whereas others generated considerably more text (e.g., Participant Id. 11 produced 17,215 words). The average noun count was $1,100 \pm 2,270$, whereas the average verb count was $686 \pm 1,236$. Regarding the number of nouns and verbs, nouns appeared more frequently, which aligns with their common usage in Thai text. For emotions (**Figure 5B**), approximately 90% were labeled as positive, with only Participant Id.'s 6, 12, and 26 receiving negative labels. These participants were also among those who posted fewer than 10 words.

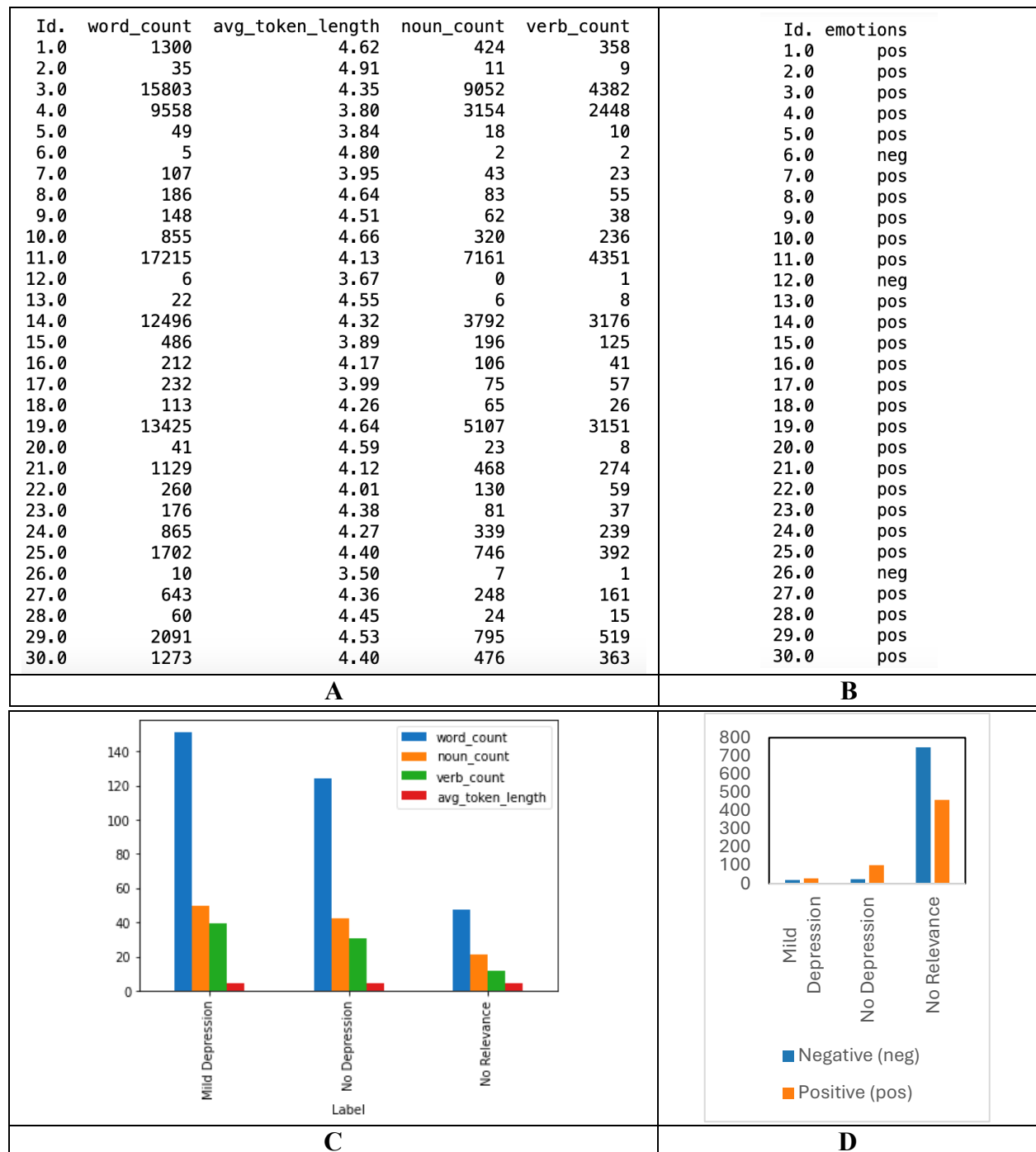


Figure 5. (A) POS tagging and (B) assigned emotions for subject-level classification. Bar charts showing (C) POS tagging and (D) assigned emotions for text-level classification, grouped based on experts' opinion. POS, part-of-speech.

Figure 5C illustrates that participants labeled with "mild depression" tend to write more and therefore use more nouns and verbs than the other two categories. For emotions (**Figure 5D**), the "mild depression" group was associated with positive emotions in contrast to the "no relevance" group, which was mostly associated with negative emotions. This pattern was consistent across all participants when grouped by experts' labels.

A Kruskal–Wallis H test was conducted on the word count, average token length, normalized noun

and verb counts, and emotion to examine whether these features differed across labels at the subject-level. Due to the correlations among the extracted features and the multiple statistical comparisons, the *P*-values were adjusted using the Benjamin-Hochberg procedure to control the false discovery rate. This analysis was not conducted at the text-level, as it could introduce nonindependence, as some participants may post identical text multiple times. **Figure 6A** shows that only the word count was significant. In contrast, when the expert labels were grouped at the subject-

A	Kruskal-Wallis H-test results for extracted features at Subject-level grouped by TDI labels: word_count H-statistic = 6.6685 Original p-value = 0.009813 BH corrected p-value = 0.049067 → Statistically significant after BH correction avg_token_length H-statistic = 1.8938 Original p-value = 0.168778 BH corrected p-value = 0.210973 → Not significant after BH correction noun_count (rate per 100 words) H-statistic = 3.6306 Original p-value = 0.056725 BH corrected p-value = 0.094541 → Not significant after BH correction verb_count (rate per 100 words) H-statistic = 4.6613 Original p-value = 0.030850 BH corrected p-value = 0.077125 → Not significant after BH correction emotions H-statistic = 0.9547 Original p-value = 0.328518 BH corrected p-value = 0.328518 → Not significant after BH correction
B	Kruskal-Wallis H-test results for extracted features at Subject-level grouped by expert labels: word_count H-statistic = 0.8458 Original p-value = 0.655134 BH corrected p-value = 0.818917 → Not significant after BH correction avg_token_length H-statistic = 3.0019 Original p-value = 0.222913 BH corrected p-value = 0.818917 → Not significant after BH correction noun_count_rate H-statistic = 1.8244 Original p-value = 0.401648 BH corrected p-value = 0.818917 → Not significant after BH correction verb_count_rate H-statistic = 0.2349 Original p-value = 0.889168 BH corrected p-value = 0.889168 → Not significant after BH correction emotions H-statistic = 1.1717 Original p-value = 0.556628 BH corrected p-value = 0.818917 → Not significant after BH correction

Figure 6. Kruskal-Wallis H-test, along with the BH correction for extracted features at the subject-level grouped based on (A) TDI labels and (B) expert labels. BH, Benjamin-Hochberg; TDI, Thai Depression Inventory.

level (**Figure 6B**), no significant differences were observed in the word count. In addition, average token length, noun count rate, verb count rate, and emotion scores did not differ significantly across groups for either labeling approach. These findings suggest that text length may be associated with depressive symptoms, whereas other linguistic and emotional features demonstrated limited discriminatory capability at the subject-level.

Input for the subject-level model

The input data for the subject-level classification included the participant Id.'s, labels from TDI (**Figure 7A**) and expert opinions (**Figure 7B**), vectorized and concatenated Facebook text, and the extracted features.

Subject-level classification with TDI labels

For the supervised learning approach, three models were employed using input from the TDI and expert labels. From the box plot, the SVM model achieved moderate training accuracy across all five iterations, with values ranging from 50.0% to 79.0% and a median

of 75.0%, where most iterations achieved an accuracy above 70.0% (**Figure 8A**). The overall accuracy was 50.0%, indicating limited discriminative capability. In addition, the macro-average metrics were similar, with a precision and recall of 53.0% and an F1-score of 50.0% (**Figure 8B**). The confusion matrix further demonstrated this, with 15 misclassifications, particularly where the 11 individuals with "major depression" were labeled as "mild depression" (**Figure 8C**).

The box plot for the LR model also demonstrated limited classification capability, with training accuracy ranging from 54.0% to 67.0% and a median of 63.0%. However, the results were more variable and spread out than those of the SVM model (**Figure 9A**). The overall accuracy was slightly higher at 60.0%, although the macro-average metrics were lower. Precision was the lowest at 30.0%, followed by an F1-score of 38.0% and a recall of 50.0% (**Figure 9B**). The confusion matrix further revealed that the model predominantly predicted the "major depression" class, resulting in 12 misclassifications and a complete failure to identify any participants with "mild depression" (**Figure 9C**).

A	Id.	Self_label	vector	emotions	word_count	avg_token_length	noun_count	verb_count
	0	1 Major Depression	[-4.98842746e-02 -9.89691243e-02	pos	1300	4.62	424	358
	1	2 Mild Depression	[5.61361872e-02 -5.51094785e-02	pos	35	4.91	11	9
	2	3 Major Depression	[-2.15627477e-02 -5.90430833e-02	pos	15803	4.35	9052	4382
	3	4 Major Depression	[-6.91094920e-02 -1.02478467e-01	pos	9558	3.80	3154	2448
	4	5 Mild Depression	[-3.66947502e-02 -1.22042269e-01	pos	49	3.84	18	10
	5	6 Major Depression	[-0.25844422 -0.10744081 0.0263i	neg	5	4.80	2	2
	6	7 Major Depression	[-4.79369909e-02 -8.91898572e-02	pos	107	3.95	43	23
	7	8 Mild Depression	[-4.56390642e-02 -5.47117889e-02	pos	186	4.64	83	55
	8	9 Mild Depression	[-0.04737718 -0.07560657 0.1039i	pos	148	4.51	62	38
	9	10 Major Depression	[-3.90405171e-02 -6.25812784e-02	pos	855	4.66	320	236

B	Id.	experts	vector	emotions	word_count	avg_token_length	noun_count	verb_count
	0	1.0 No Depression	[-4.98842746e-02 -9.89691243e-02	pos	1300	4.62	424	358
	1	2.0 No Depression	[5.61361872e-02 -5.51094785e-02	pos	35	4.91	11	9
	2	3.0 No Depression	[-2.15627477e-02 -5.90430833e-02	pos	15803	4.35	9052	4382
	3	4.0 Mild Depression	[-6.91094920e-02 -1.02478467e-01	pos	9558	3.80	3154	2448
	4	5.0 No Relevance	[-3.66947502e-02 -1.22042269e-01	pos	49	3.84	18	10
	5	6.0 No Relevance	[-0.25844422 -0.10744081 0.0263i	neg	5	4.80	2	2
	6	7.0 No Relevance	[-4.79369909e-02 -8.91898572e-02	pos	107	3.95	43	23
	7	8.0 No Relevance	[-4.56390642e-02 -5.47117889e-02	pos	186	4.64	83	55
	8	9.0 No Relevance	[-0.04737718 -0.07560657 0.1039i	pos	148	4.51	62	38
	9	10.0 No Relevance	[-3.90405171e-02 -6.25812784e-02	pos	855	4.66	320	236

Figure 7. Input data including participant's Id, labels from the (A) TDI form and (B) experts' opinion, vectorized concatenated Facebook text, assigned emotions, and POS tagging for subject-level classification. POS, part-of-speech; TDI, Thai Depression Inventory.

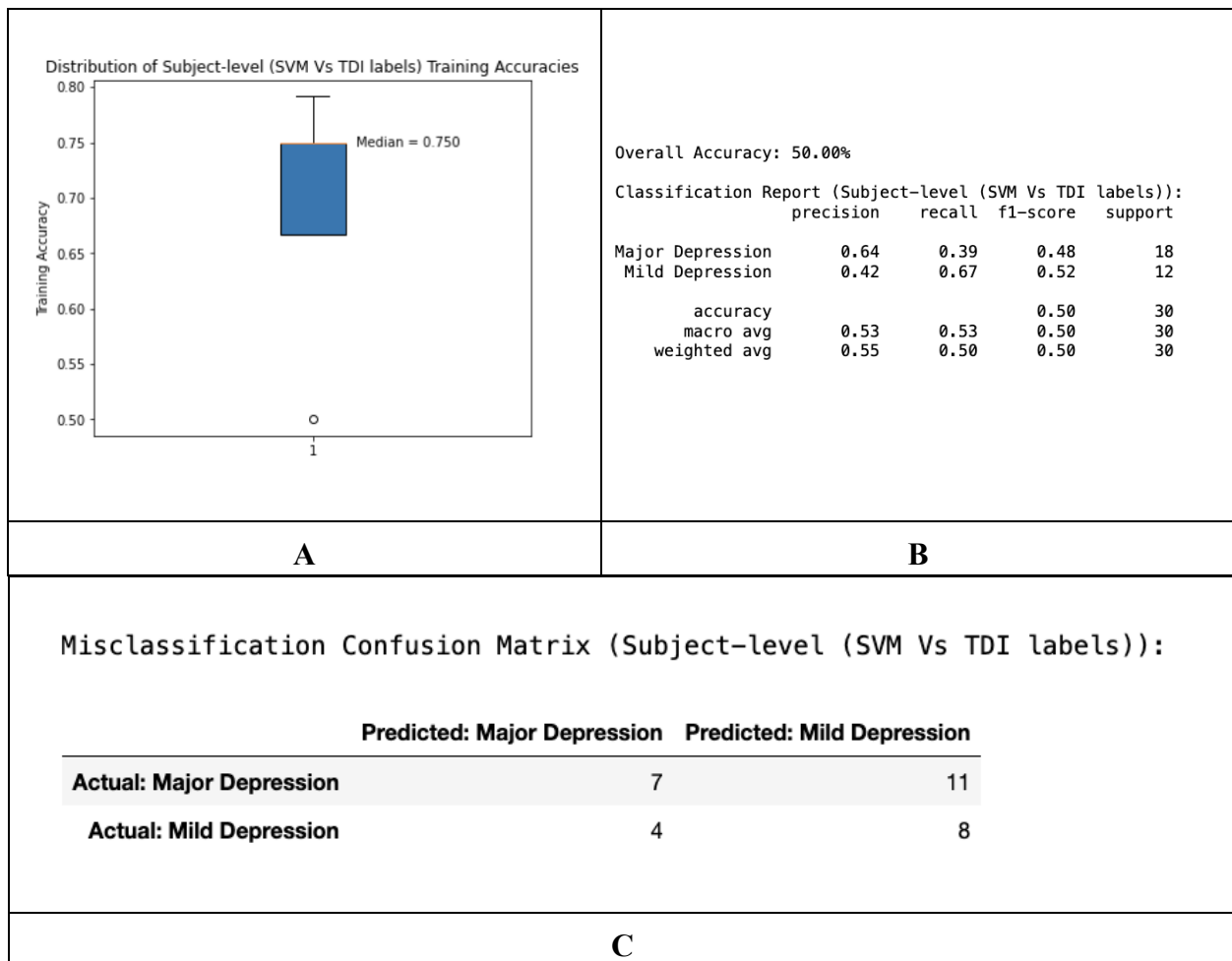


Figure 8. Box plot of SVM's (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against TDI labels at the subject-level. SVM, support vector machine; TDI, Thai Depression Inventory.

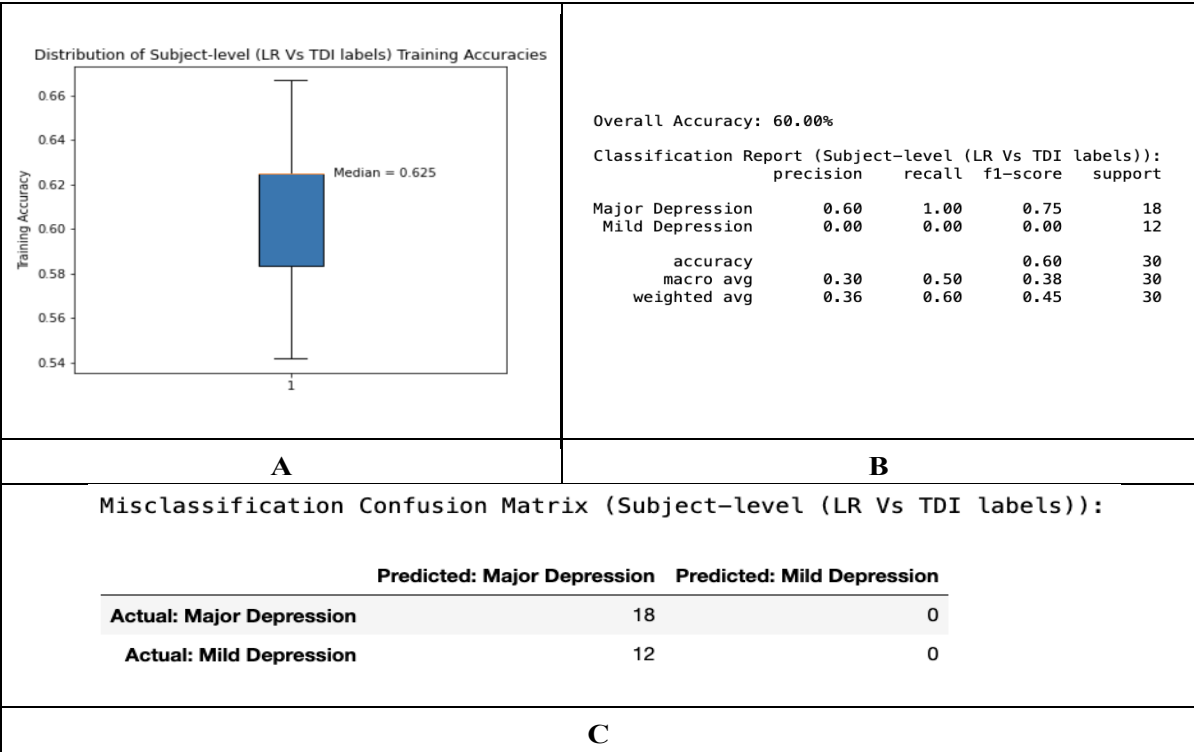


Figure 9. Box plot of LR’s (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against TDI labels at the subject-level.
LR, logistic regression; TDI, Thailand Depression Inventory.

Finally, for the subject-level classification using TDI labels, the DT box plot demonstrated the best overall classification performance, with training and median accuracy of 100.0% across all five iterations (Figure 10A). The overall accuracy was also the highest among the three models at 70.0%. The macro-average metrics were all greater than 60.0%, with

the highest being precision at 69.0%, followed by recall and F1-score at 67.0% (Figure 10B). The confusion matrix showed nine misclassifications, with the majority (six) occurring when "mild depression" cases were incorrectly labeled as "major depression" (Figure 10C).

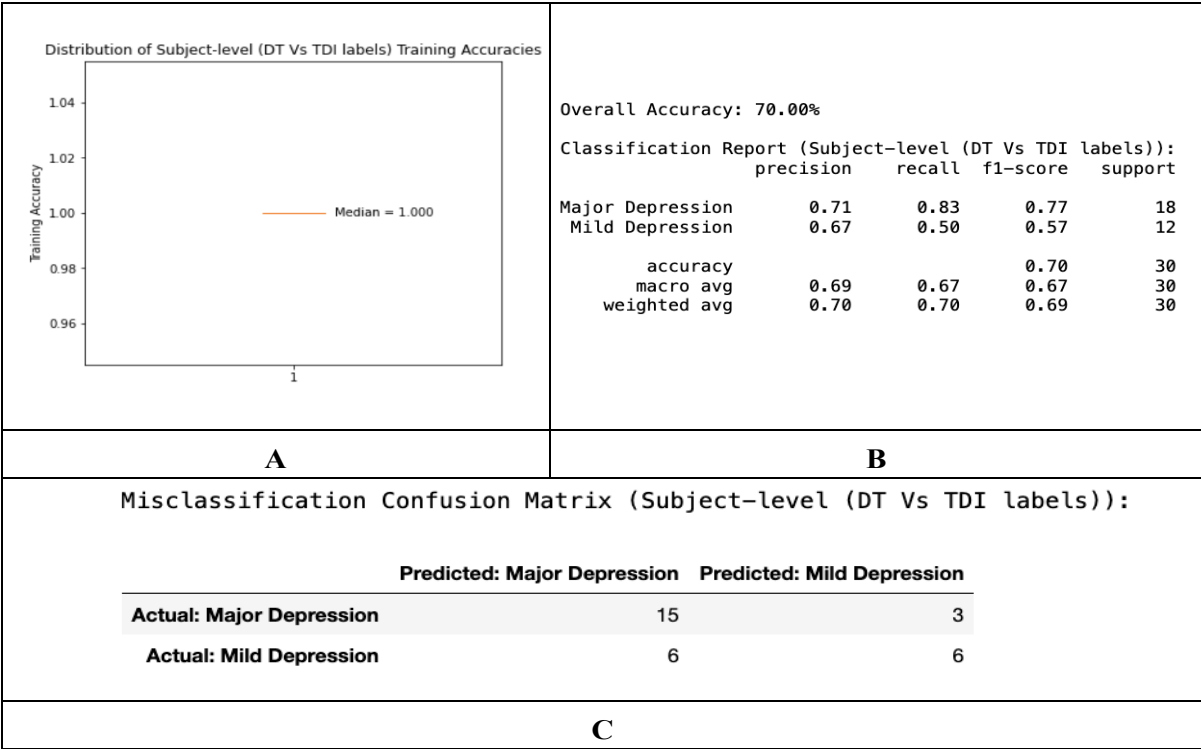


Figure 10. Box plot of DT’s (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against TDI labels at the subject-level.
DT, decision tree; TDI, Thai Depression Inventory.

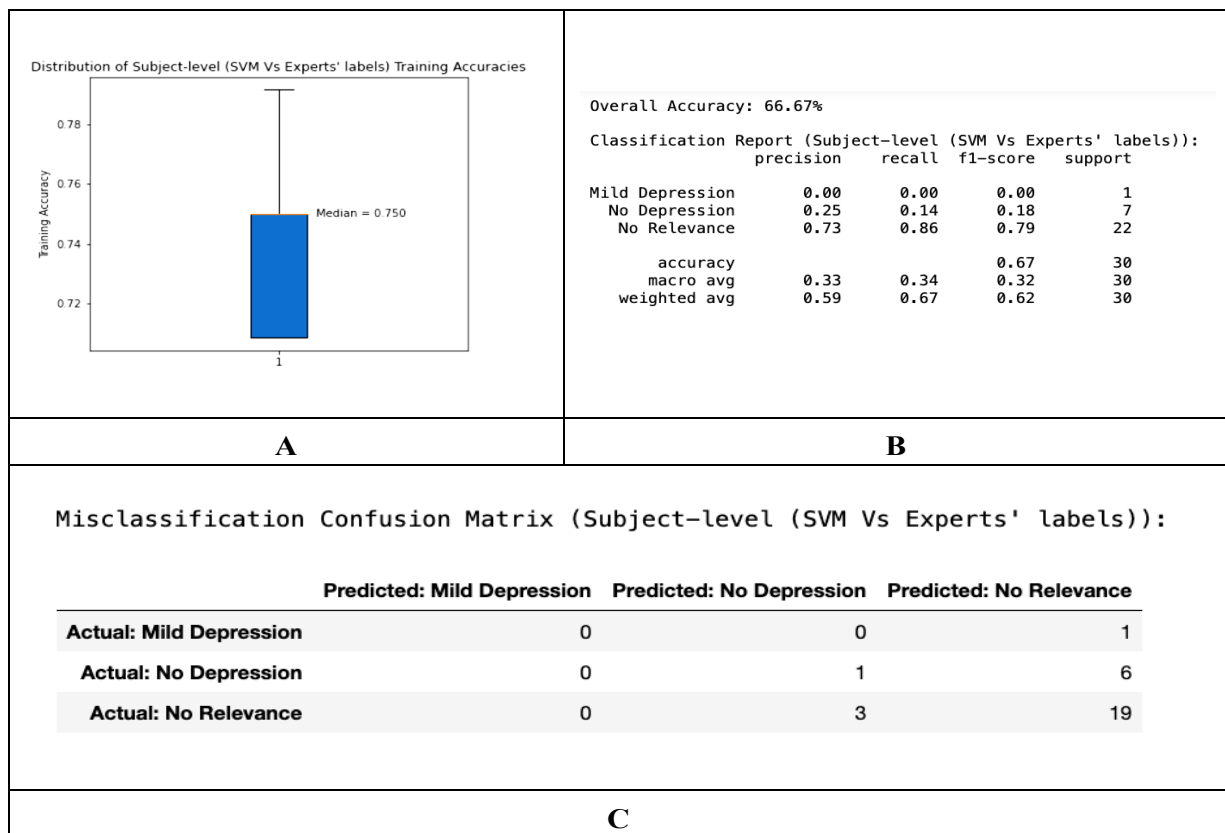


Figure 11. Box plot of SVM's (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against expert-derived labels at the subject-level.

SVM, support vector machine.

Subject-level classification with expert-derived labels

When the SVM model was trained using experts' opinions as labels, the training accuracy ranged from 71.0% to 79.0%, with a median of 75.0%, and was mostly concentrated near the lower boundary, as illustrated in the box plot (Figure 11A). The overall accuracy was 67.0%; however, the macro-average metrics were considerably lower. Recall was the highest at 34.0%, followed by precision at 33.0% and an F1-score of 32.0% (Figure 11B). The confusion matrix revealed that the model produced only 10 misclassifications. However, the model failed to predict any "mild depression" cases and predominantly classified samples into the dominant "no relevance" category (Figure 11C).

The LR model achieved a training accuracy range similar to that of the SVM model, ranging from 71.0% to 80.0%. However, the median accuracy was approximately 71.0%, placing it closer to the lower end of the distribution (Figure 12A). The overall accuracy was slightly higher at 73.0%, although the macro-average metrics were lower. Precision was the lowest at 24.0%, followed by F1-score at 28.0% and recall at 33.0% (Figure 12B). Based on the confusion matrix, the model produced only eight

misclassifications but demonstrated severe majority-class prediction bias. In other words, the model only predicted the "no relevance" class (Figure 12C).

Finally, the DT model was also used for the subject-level classification based on experts' labels. The box plot showed findings similar to those obtained with TDI labels, with training accuracy ranging from 96.0% to 100.0% and a median of 100.0% (Figure 13A). However, the overall accuracy was the lowest among the models using experts' labels, at 60.0%. The lowest macro-average metric was precision at 29.0%, followed by F1-score at 30.0%, and recall at 31.0% (Figure 13B). Furthermore, it had the highest number of misclassifications of 12, according to the confusion matrix. The model predominantly predicted either "no depression" or "no relevance," with the latter being more frequent (Figure 13C).

Input for the text-level model

The input data for the text-level classification included the participant Id.'s, labels, Facebook text, and extracted features, as illustrated in Figure 14. The "Label" column was derived only from the expert evaluations. Furthermore, the input data substantially increased from 30 rows (corresponding to 30 participants) to 1,382 rows.

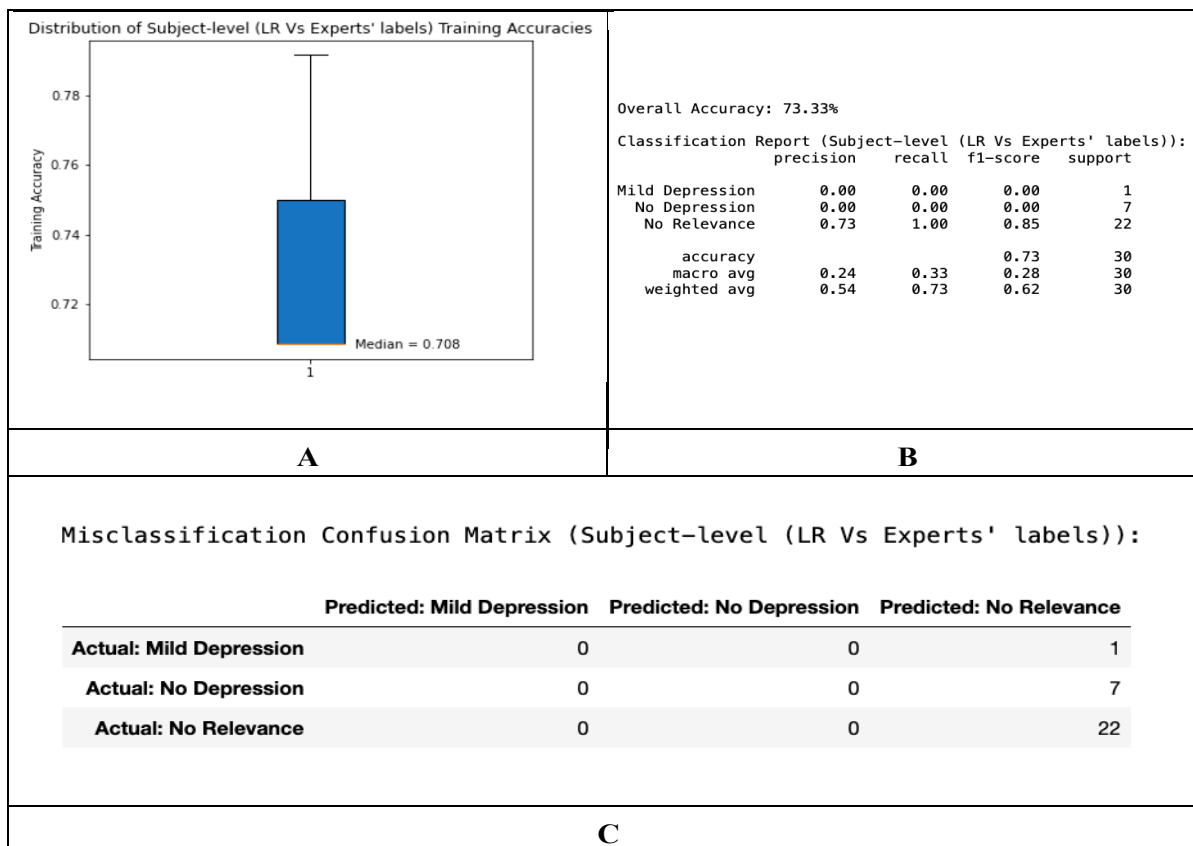


Figure 12. Box plot of LR's (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against expert-derived labels at the subject-level. LR, logistic regression.

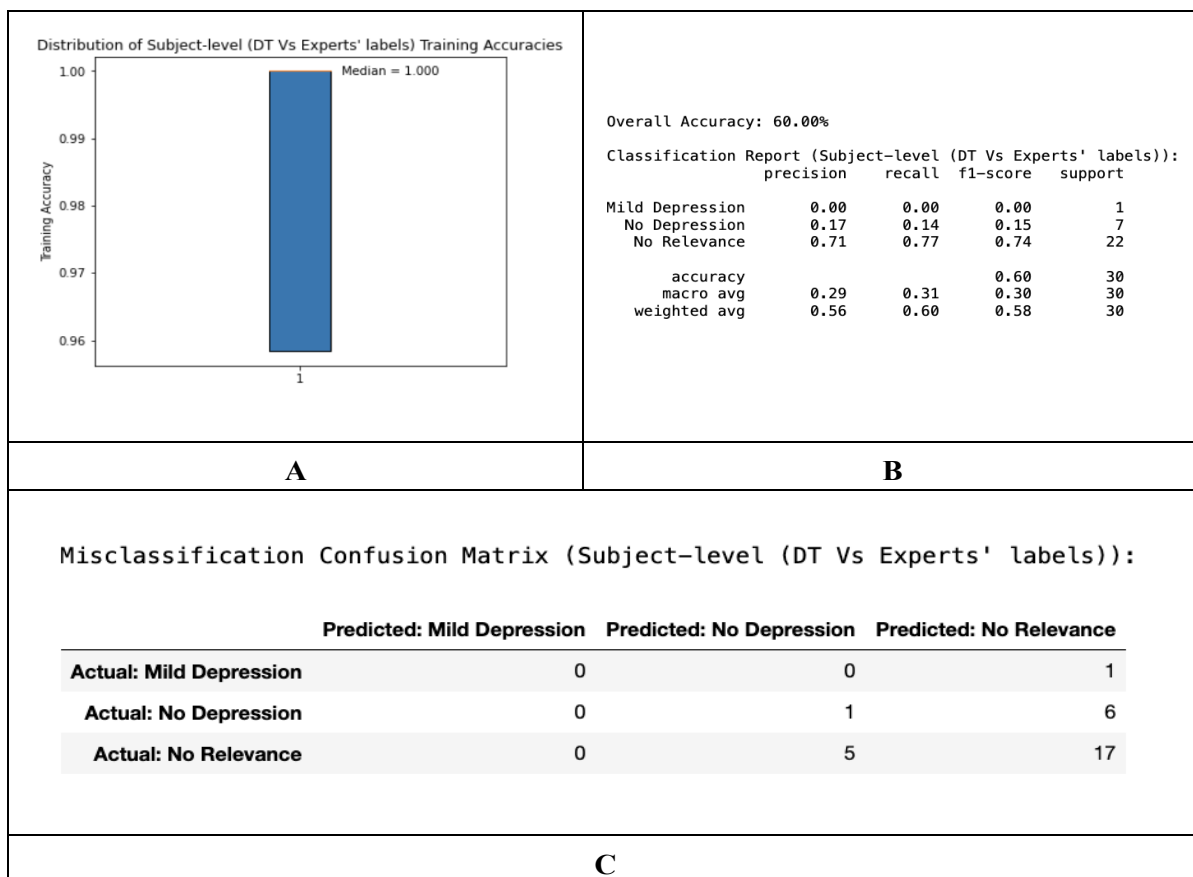


Figure 13. Box plot of DT's (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against expert-derived labels at the subject-level. DT, decision tree.

Id.	Label	vector	emotions	word_count	avg_token_length	noun_count	verb_count
0	1	No Depression [5.77670e-02 -5.12330e-02 1.64826e-01 -1.21894e-01 -1.57074e-01\...	neg	1	3.00	1	0
1	1	No Depression [-2.13970e-02 -1.40640e-02 -6.60840e-02 -5.94000e-04 -2.31120e-02\...	neg	1	6.00	1	0
2	1	No Depression [5.77670e-02 -5.12330e-02 1.64826e-01 -1.21894e-01 -1.57074e-01\...	neg	1	3.00	1	0
3	1	Mild Depression [-3.43288928e-02 -7.63790235e-02 1.21513501e-01 7.81601444e-02\...	pos	204	4.09	54	56
4	1	No Depression [-6.52510598e-02 -1.06766336e-01 1.18217535e-01 3.83929126e-02\...	pos	705	4.91	210	196
5	1	No Depression [-0.03459614 -0.14197513 0.10434332 0.05954731 -0.08966398 -0.00\...	pos	148	4.20	44	39
6	1	No Depression [-0.023738 0.037523 0.064625 -0.004121 -0.015204 0.005439 -0.03\...	neg	1	5.00	1	0
7	1	No Depression [-3.50265019e-02 -8.38399976e-02 -2.94350013e-02 1.09700501e-01\...	neg	2	6.00	2	0
8	1	No Relevance [-4.41870019e-02 -1.24310739e-01 1.40782207e-01 3.14802490e-02\...	neg	28	4.39	9	8
9	1	No Relevance [-0.07256795 -0.11095355 0.10533956 0.00565076 0.12225766 -0.05\...	pos	20	4.05	11	4
10	1	No Relevance [-1.65705726e-01 -8.69260952e-02 1.40863359e-01 -1.11674644e-01\...	neg	11	3.64	8	1
11	1	No Relevance [0.00734889 0.04296667 0.03869 0.065876 -0.03878144 -0.04\...	neg	9	5.89	5	4
12	1	No Relevance [-5.63538261e-02 -6.80816695e-02 1.18302211e-01 1.12692183e-01\...	pos	19	4.11	5	6
13	1	No Relevance [0.00833739 -0.09856502 0.03759618 -0.039257 -0.11861628 -0.01\...	pos	29	5.28	16	7
14	1	No Relevance [-9.77960005e-02 -8.15910101e-02 4.56133336e-02 -2.37783328e-01\...	pos	4	5.75	3	0
15	1	No Depression [-1.29915494e-02 -5.09500764e-02 8.52388814e-02 6.23606332e-02\...	pos	116	4.48	55	32
16	2	No Depression [5.61361872e-02 -5.51094785e-02 9.12459269e-02 1.82389673e-02\...	pos	35	4.91	11	9
17	3	No Relevance [-1.02249399e-01 6.16134033e-02 8.90111029e-02 3.43342535e-02\...	neg	20	4.00	13	5
18	3	No Relevance [-2.67554708e-02 -1.23873822e-01 7.40136802e-02 7.36576319e-02\...	neg	36	4.53	13	10
19	3	No Relevance [2.45002825e-02 -4.24242998e-03 -1.04995705e-01 -7.18522146e-02\...	pos	14	3.79	10	2

Figure 14. Input data including participant's Id, labels from experts' opinion, vectorized Facebook text, assigned emotions, and POS tagging for text-level classification.

POS, part-of-speech.

Text-level classification with expert-derived labels

For the supervised learning approach at the text-level, the same three models were employed, using labels derived solely from the expert evaluations. Moreover, the overall test accuracy exceeded 70.0% due to the substantial increase in the amount of available data.

The SVM box plot showed that the training accuracy, as well as the median, clustered around 88.0% with

marginal fluctuations (**Figure 15A**). The overall accuracy was only slightly lower at 87.0%, however, the macro-average metrics were substantially lower, with a precision of 29.0%, followed by an F1-score of 31.0% and recall of 33.0% (**Figure 15B**). According to the confusion matrix, 176 of 1,382 samples were misclassified. However, the model only predicted the "no relevance" class (**Figure 15C**).

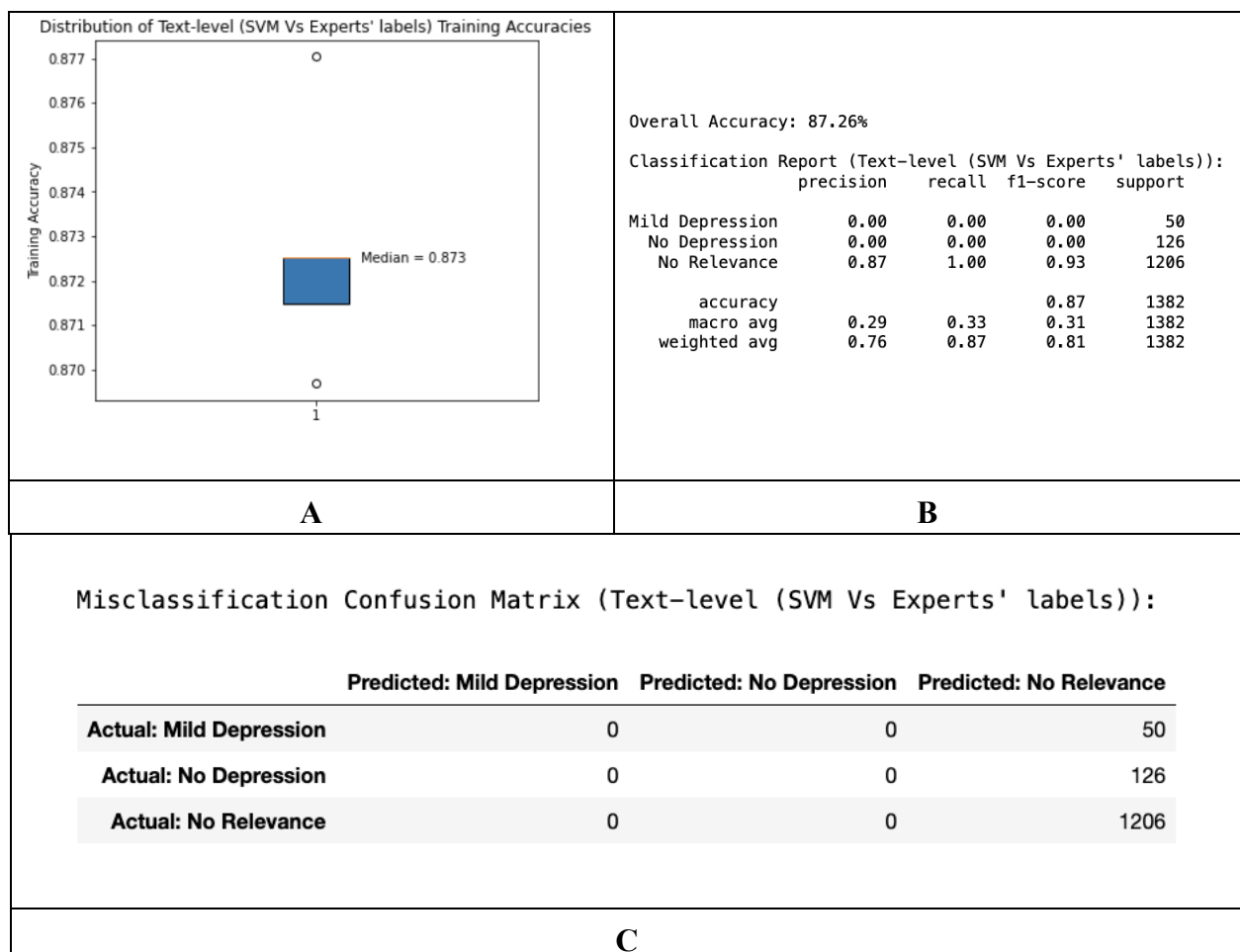


Figure 15. Box plot of SVM's (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against expert-derived labels at the text-level.

SVM, support vector machine.

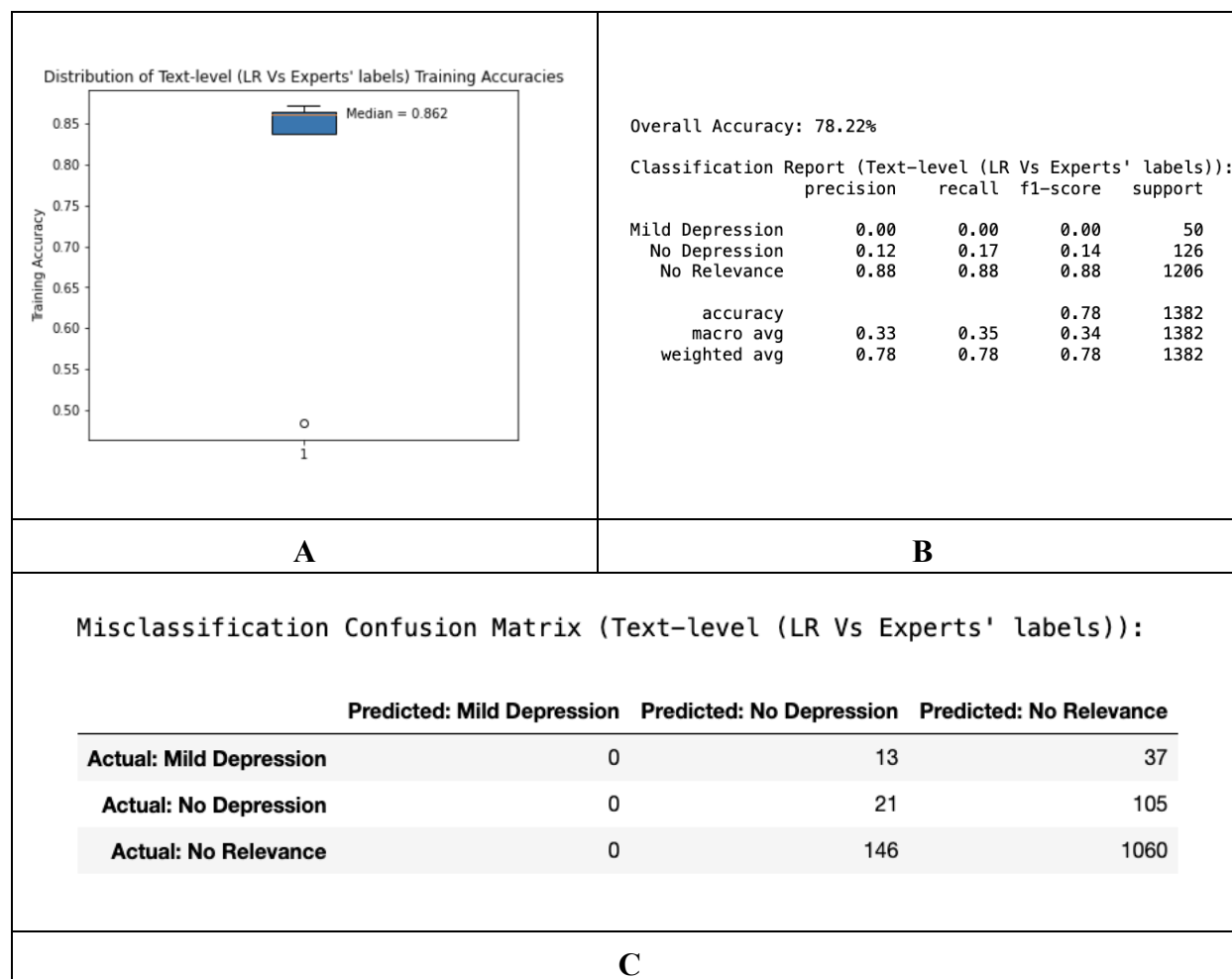


Figure 16. Box plot of LR's (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against expert-derived labels at the text-level. LR, logistic regression.

The training accuracy for the LR model ranged from 48.0% to 87.0%, with most values clustering at the higher end and a median of 86.0% (**Figure 16A**). The overall accuracy was lower than that of the SVM model at 78.0%, but with slightly higher macro-average values. Recall was the highest at 35.0%, followed by an F1-score of 34.0%, and a precision of 33.0% (**Figure 16B**). From the confusion matrix, the model produced 301 misclassifications, primarily by labeling "no depression" when the correct label was "no relevance" (**Figure 16C**).

Lastly, the DT model exhibited training accuracy ranging from 96.0% to 99.0%, with a median of 99.0% (**Figure 17A**). A slight drop in overall accuracy to 79.0% was observed. This model achieved the highest macro-average performance among the three models. Recall was 42.0%, followed by an F1-score of 41.0%, and precision of 40.0% (**Figure 17B**). A total of 291 misclassifications were observed in the confusion matrix, with most misclassifications occurring when "no relevance" was labeled as "mild depression".

However, this model predicted all three labels (**Figure 17C**).

Some examples of subject-level misclassifications when the model predictions were compared with the actual labels derived from TDI scores (**Figure 18A**) and experts' opinions (**Figure 18B**). Neither of the original texts contained linguistic cues indicating mental health issues. Therefore, from a linguistic perspective, these texts alone were insufficient to indicate depression. Similar observations can be made in **Figure 18C**, which presents examples of text-level misclassifications when the model predictions are compared with the expert-assigned labels. Moreover, the distinction between the "no depression" and "no relevance" classes was inherently vague, which may have contributed to these misclassifications.

Due to the severe class imbalance and small sample size, a "most_frequent" DummyClassifier was implemented as a baseline comparator. This baseline was used to determine whether the ML models performed beyond the majority-class prediction.

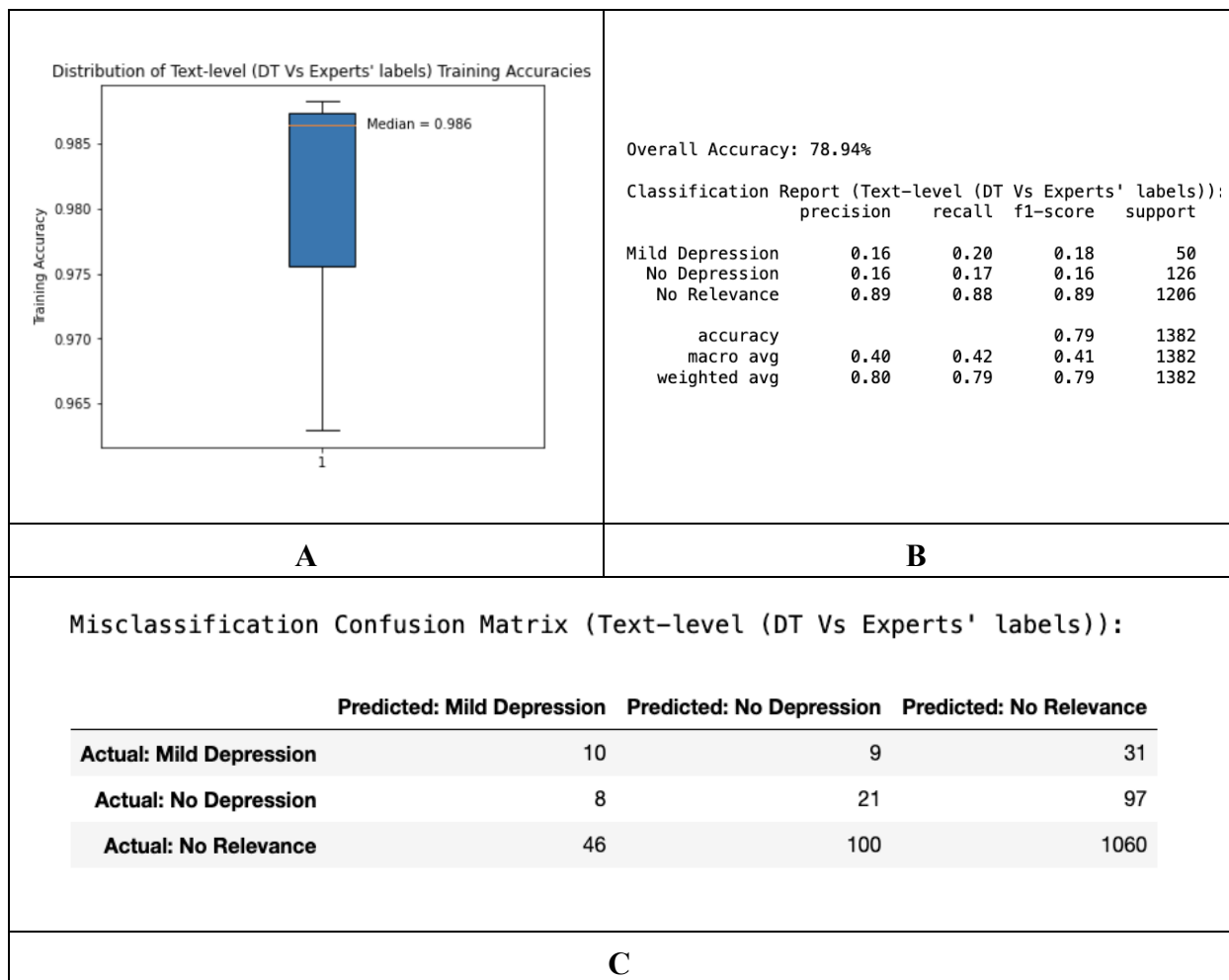


Figure 17. Box plot of DT's (A) training accuracy, (B) overall testing accuracy and classification report, and (C) confusion matrix against expert-derived labels at the text-level. DT, decision tree.

Overall, the TDI labels outperformed the expert opinions for subject-level classification, with the DT model achieving the best performance. In fact, it performed better than the dummy baseline model across all metrics, which suggests that it learned meaningful patterns beyond those of the majority-class prediction. In contrast, when the experts' opinion labels were used, the best-performing model (LR) produced results identical to those of the dummy baseline model. This indicates that the model may have primarily predicted the majority class rather than learning meaningful patterns.

At the text-level, the dummy baseline achieved the highest accuracy, reflecting the dataset's severe class imbalance. Although all the ML models demonstrated slightly higher macro-precision, recall, and F1-score than the dummy baseline, the results suggest limited improvement beyond the majority-class prediction. The summary of these results is presented in **Table 4**.

Discussion

To evaluate the suitability of ML models for the preliminary classification of Facebook text into depression-related categories, this study adopted workflows consistent with previous research, using self-assessment tools and expert labeling at the subject^(12, 17) and text-levels.^(13-15, 20) Following a workflow comparable to that of other studies enabled assessment of whether previous findings, primarily identifying the most accurate models or optimal ML combinations, truly advance ML as a screening tool. Furthermore, unlike traditional approaches that rely solely on public data collection through questionnaires or keyword searches, this study leveraged data from individuals who sought psychiatric care, thereby providing a more clinically grounded dataset.⁽²¹⁾

A	TEXT: เนินหมาทอบ ละคนจะไปเหลืออะไร ขอบคุนตากล้องค่าบ อยากไปเที่ยวครั้งหน้าแล้ว อุทยานแห่งชาติเขาหลวง... กำลังไปหานั้น 😊 ACTUAL LABEL: Mild Depression PREDICTED LABEL: Major Depression TEXT: POV: เมื่อคุณเป็นตัวละครในเกม ACTUAL LABEL: Major Depression PREDICTED LABEL: Mild Depression
B	นยาทงน้นนอกจากความช้นแล้วก็อยู่ทงหะของชีวิตด้วยละ ของใหม่ที่อยู่ฟอรัมคะ ไม่น่าจะม่อะไรเพิ่มอีกไหมบางอย่างต้องลุ้นว่าจะทันงานหรือเราจะจัด ตารางมาทำได้ไหมแปะลิงก์ฟอรัมในเน้นห้นะคะ ขอบคุนอีกคร้งนะคะ อยากรู้สติกเกอร์แสดมปีนกว่าดที่ขอบมาแปะลงสมุดทำเป็นสมุดสะสมแสดมปี ช่วงนี้จัดการอะ ไรช้ห้น้อยเพราะตาแห้งแล้วมันเจ็บตาพยายามพิกตากับหอยค่น้ำตาเทียมเยะอยู่ตื้นแต่ยังไม่หายอะ วันสุดท้ายส้าหรับการพริ้อเคอร์นะคะ น้องลิ เสนคุนใบเคย พบกันวัน เสาร์นะคะ บางอันไม่ใช่คอมมิแต่ก็เอาจมรวมด้วยคะ กำลังนึ่งแพคของเตรียมงานพ้งนี้คะของใหม่ของเก๋าเพิ่มแพคเกจใหม่มีรายการแบบฉบับของเก๋าใส่ แพคเกจใหม่แต่ราคาเดิมแบบแยกชิ้นกับของก็ยังมีขายนะคะตัวสินค้าจะมีทั้งแบบสีดิสคิมและเหลืองคะก้นนะคะพอดีเราปรับสีลองมาหลายร้านแต่คุณภาพจะเป็นตามเ ดิมนะคะสติกเกอร์สีน้องลิเวอร์ช้นแม่คสาวกับแมวตัวทำกับสติกเกอร์สีน้องลิอีกเหมือนเดิมแต่มีลิตละองควาด้วยอันนี้ทำกับพบกันได้ที่วัน เสาร์วันเดียวค่า ACTUAL LABEL: No Relevance PREDICTED LABEL: Mild Depression TEXT: เมื่อคุณเป็นตัวละครในเกม ACTUAL LABEL: No Relevance PREDICTED LABEL: No Depression TEXT: ตอนเด็กเจ้สอนมาบอกว่าถ้าใครตาเราให้ตบกลับทันทีวันเด็กแห่งชาติวันเด็ก วันเด็ก กราบลาหลวงตากกราบขอบพระคุณที่ดูแลและซัพพอร์ตหนูมาตลอดตั้งแต่แรกเกิด จนโต ลายมือหลวงตา เก่งกันสุดสุดเลยยอม ขอบพระคุณมากนะคะ เดิมช้าคะ อนุโมทนาบุญด้วยนะคะ สวัสดิวันศุกรคะ วิสักรานต์ เกลียดแค่ไหนก็ไม่บอกรอก จะให้เห็นความสุขเมื่อมีคนด้น พรอมฉลองโทรศัพท์เครื่องใหม่คิดถึงนะเอียวรี ยอมอดทนทำสิ่งที่เราไม่ชอบเพื่อเก็บเงินวันข้างหน้าเราจะได้ทำในสิ่งที่เราชอบได้ ACTUAL LABEL: No Depression PREDICTED LABEL: No Relevance
C	TEXT: คากาหัวใจหัวเวสสุวรรณสะพุสะผู้ใดสวดบริกรรมภาวนาบูชาคาถาหัวใจหัวเวสสุวรรณนี้แล้วคลาดปลอดภัยและยังอุทมนไปด้วยโภคทรัพย์หัวใจเศษหัวเวสสุวรรณ หัวเวสสุวรรณป้า ACTUAL LABEL: Mild Depression PREDICTED LABEL: No Relevance TEXT: หนาวบิณีไปไหนไม่ได้เปิดแอร์องศาแต่งหน้านอนระก้นจบ ACTUAL LABEL: No Relevance PREDICTED LABEL: Mild Depression TEXT: เหล่าอาจารย์มาคุยกันพูดคุยเรื่องเสริมจุมทุกแบบ ACTUAL LABEL: No Depression PREDICTED LABEL: No Relevance

Figure 18. Examples of subject-level misclassification of the original text when compared with (A) TDI scores, (B) experts' opinion, and (C) examples of text-level misclassification when compared with expert's opinion. TDI, Thailand Depression Inventory.

Table 4. Summary of subject- and text-level model performance based on labels from TDI scores and experts' opinion.

Labels	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Subject-level classification					
TDI	SVM	50	53	53	50
	LR	60	30	50	38
	DT	70	69	67	67
	Dummy baseline	60	30	50	38
Experts' opinion	SVM	67	33	34	32
	LR	73	24	33	28
	DT	60	29	31	30
	Dummy baseline	73	24	33	28
Text-level classification					
Experts' opinion	SVM	87	29	33	31
	LR	78	33	35	34
	DT	78	40	42	41
	Dummy baseline	87	29	33	31

In terms of POS tagging and emotions, only word count was significantly correlated with TDI labels at the subject-level, suggesting that text length may be associated with self-perceived depressive symptoms, as depressed individuals may post more.⁽²²⁾ However, significant differences were not observed when labels were derived from an expert's opinion, which suggests that expert-based assessment may not correspond as strongly with text data alone due to its ambiguity and need for more contextual information.^(23,24) In addition, when comparing labels derived from self-assessment forms and experts' opinions, there was limited agreement. One possible reason is that individuals tend to under- or over-rate their symptoms.⁽²⁵⁾

Subject-level classification performed better than text-level classification when compared against the dummy baseline model. More specifically, the DT model at the subject-level, using TDI-derived labels, outperformed expert-opinion-based classification. In addition, a similar study conducted by Siranuch H, et al. concluded that per-person experimental data were superior to text data in terms of precision and recall.⁽²⁶⁾ Although the text-level classification performed poorly against the baseline models, when comparing the three ML models, SVM performed the best, followed by LR and DT, which is consistent with current research.^(13, 27)

Several limitations should be acknowledged for this research. First, there were biases and time-window mismatches between the two-week TDI assessment and the 6-9 months of Facebook data collected. As depressive symptoms may fluctuate over

time, linguistic patterns extracted from older posts may not accurately reflect the participants' psychological states at the time of assessment. This inconsistency may have weakened the associations between linguistic features and depression labels, contributing to the limited significance observed in several features and the moderate ML performance. In addition, self-selection and negativity bias may have affected the dataset, as participants willing to share Facebook data may present themselves more positively online. In contrast, private self-report forms may capture more distressing emotions. Second, the Kruskal-Wallis H test was not performed at the text-level due to violations of the assumption of independent and identically distributed observations, arising from repeated posts by the same participants. This reduced the amount of data available for statistical testing and may have limited the detection of finer linguistic patterns. Furthermore, the substantial improvement in the inter-rater reliability may have been influenced by re-labeling the same dataset using revised criteria. Repeated exposure to the same data could increase the raters' familiarity and introduce recall bias, thereby potentially inflating the agreement score. Lastly, despite Facebook's widespread use in Thailand as a popular platform,^(28, 29) only half of the participants were active users, and much of the text was ambiguous, requiring expert labeling as "no relevance". The small dataset and class imbalance significantly affected the ML models, as evidenced by relatively high training and overall accuracy alongside the low macro-average performance metrics.

Supervised models offer several advantages for mental health classification. They can enable early intervention by identifying individuals who may be at risk of developing depression, and they allow for the efficient analysis of large and complex datasets in a shorter amount of time. Some improvements include using more recent NLP algorithms, such as those developed for social media purposes, as the meaning and language patterns differ substantially from those of formal written language. Using large language models, such as OpenAI's GPT or Google's Gemini family, could enhance the model's performance. In addition, incorporating other forms of media, such as images, vocal acoustics, or video data, could provide additional contextual and historical information. This may improve model performance and enhance the accuracy and reliability of depression detection. Furthermore, partnering with hospitals or clinics to recruit more participants would increase the sample size while providing more reliable "true" cases. Incorporating data from private accounts may also help capture more negative or emotionally expressive posts. Lastly, an important factor to consider is understanding participants' Facebook behavior, whether they use the platform for personal or promotional activities. This distinction is crucial, as promotional content may lead to more posts unrelated to their mental health status.

The proposed system should be interpreted as a supportive screening framework rather than as a standalone depression classification tool. Model predictions based solely on text may help identify individuals who exhibit distinct language patterns associated with depressive symptoms; however, these outputs should not be considered definitive clinical diagnoses. In practical applications, individuals identified by the model as potentially at risk of depression could be referred for further psychological assessment. Therefore, this system may be beneficial for preliminary large-scale monitoring, early risk identification, or decision support in settings where direct clinical assessment is limited.

Conclusion

This study evaluated three supervised ML models (SVM, LR, and DT) for classifying Facebook text into depression-related categories at the subject- and text-levels. The objective was to examine mismatches between self-assessment forms and social media behaviors and to determine whether experts can differentiate Thai Facebook text into depression-related categories.

The findings suggested that the subject-level classification using TDI labels outperformed the expert-derived labels and text-level classification when evaluated against the dummy baseline model. This may indicate a slight correlation between the self-assessment forms and social media behaviors. In contrast, text data alone appeared insufficient for expert-based depression classification. The poor agreement between the two labeling approaches further indicates that text alone may not provide a sufficiently reliable clinical signal. In other words, when only text data are used, labels derived from self-assessment forms appear to be more applicable. In contrast, if expert-derived labels are to be used, additional contextual information would be required to maximize the ML model's potential.

In the future, collecting larger datasets for each participant and incorporating multiple data modalities may improve the model's performance and generalizability. In addition, testing larger language models such as OpenAI GPT and WangchanBERTa may provide a deeper contextual understanding of Thai-language texts. Finally, a proactive feature could be implemented to provide warnings and recommendations for individuals identified as potentially experiencing depressive symptoms. With further development, this approach may serve as a preliminary screening tool to help psychiatrists gain additional insight into patients' mental health and extend outreach to individuals with limited access to healthcare services.

Author contributions

KRY designed the study, collected data, analyzed and interpreted the results, and drafted the manuscript. PV and SH supervised and reviewed the first draft. All authors approved the final version submitted for publication and take responsibility for statements made in the published article.

Acknowledgments

The authors also express appreciation to Dr. Teeravut Wiwattarangkul and Dr. Sorawit Wainipitapong in the Department of Psychiatry at Chulalongkorn University for their assistance with data labeling and for their valuable feedback. Lastly, sincere thank are extended to all the participants who completed the TDI form.

This research received full IRB approval (COA no. 0229/2025) from the Faculty of Medicine and was funded by the Ratchadapiseksompotch Research Fund, Chulalongkorn University.

Conflict of interest statement

All authors have completed and submitted the International Committee of Medical Journal Editors Uniform Disclosure Form for Potential Conflicts of Interest. All authors declare that they have no conflicts of interest.

Data sharing statement

Aggregated data and statistics are included; de-identified text samples are available upon reasonable request under IRB-compliant data use conditions. Further details are available for noncommercial purposes from the corresponding author upon reasonable request.

ORCID

Keito R. Yoneyama  <https://orcid.org/0009-0007-2397-9118>

Peerapon Vateekul  <https://orcid.org/0000-0001-9718-3592>

Solaphat Hemrungronj  <https://orcid.org/0000-0003-0196-6790>

References

1. Wu Y, Fan L, Xia F, Zhou Y, Wang H, Feng L, et al. Global, regional, and national time trends in incidence for depressive disorders, from 1990 to 2019: an age-period-cohort analysis for the GBD 2019. *Ann Gen Psychia* 2024;23:28.
2. Tuarob S, Noraset T, Tawichsri T. Using large-scale social media data for population-level mental health monitoring and public sentiment assessment: a case study of Thailand. Bangkok: Puey Ungphakorn Institute for Economic Research; 2021.
3. Siriwanarangsana P, Liknapichitkul D, Khandelwal SK. Thailand mental health country profile. *Int Rev Psychiatry* 2004;16:150–8.
4. N, Klaywong K, Piwsaart R, Inrueng S. A study of stress and depression in health science students at a university in Bangkok, Thailand. *J Health Sci Med Res* 2025;43:1–9.
5. Udomratn P. Mental health and psychiatry in Thailand. *Int Psychiatry* 2007;4:11–4.
6. Doll CM, Michel C, Rosen M, Osman N, Schimmelmann BG, Schultze-Lutter F. Predictors of help-seeking behaviour in people with mental health problems: a 3-year prospective community study. *BMC Psychiatry* 2021;21:432.
7. Henderson C, Evans-Lacko S, Thornicroft G. Mental illness stigma, help seeking, and public health programs. *Am J Public Health* 2013;103:777–80.
8. Santor DA, Gregus M, Welch A. Eight decades of measurement in depression. *Measurement* 2006;4: 135–55.
9. Kalibatseva Z, Leong FTL. Depression among Asian Americans: review and recommendations. *Depress Res Treat* 2011;2011:1–9.
10. Lotrakul M, Sukanich P. Development of the Thai Depression Inventory. *J Med Assoc Thai* 2000;82: 1200–7.
11. Brandao PR. The impact of artificial intelligence on modern society. *AI* 2025;6:190.
12. Tsugawa S, Kikuchi Y, Kishino F, Kosuke N. Recognizing depression from Twitter activity. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015)*; 2015 Apr 18–23; Seoul, Republic of Korea: Association for Computing Machinery; 2015. p. 3187–96.
13. Kumnunt B, Sornil O. Detection of depression in Thai social media messages using deep learning. *Proceedings of the 1st International Conference on Deep Learning Theory and Applications (DeLTA 2020)*; 2020 Jul 8–10; Online. Setúbal (PT): SciTePress; 2020.
14. Musleh D, Alkhales TA, Almakki RA, Elnajim SE. Twitter Arabic sentiment analysis to detect depression using machine learning. *Comput Mater Continua* 2021;71:3463–77.
15. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preotiuc-Pietro D, Asch DA, Schwartz HA. Facebook language predicts depression in medical records. *Proc Natl Acad Sci U S A* 2018;115:11203–8.
16. National Collaborating Centre for Mental Health (UK). Appendix 12, The classification of depression and depression rating scales/questionnaires. In: National Collaborating Centre for Mental Health (UK). *Depression in adults with a chronic physical health problem*. Leicester: British Psychological Society; 2010.
17. Hämäläinen M, Patpong P, Alnajjar K, Partanen N, Rueter J. Detecting depression in Thai blog posts: a dataset and a baseline. In: Xu W, Ritter A, Baldwin T, Rahimi A, editors. *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*; 2021 Nov 11; Online. Association for Computational Linguistics; 2021.
18. McHugh LM. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22:276–82.
19. Phatthiyaphaibun W, Chaovavanich K, Polpanumas C, Suriyawongkul A, Lowphansirikul L, Chormai P, et al. PyThaiNLP: Thai natural language processing in Python. In: Tan L, Milajevs D, Chauhan G, Gwinnup J, Rippeth E, editors. *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*; 2023 Dec 6; Singapore: Association for Computational Linguistics; 2023. p. 25–36.

20. Xiao M, Liu B, Li H, Huang J, Xie Q, Zong X, et al. MoodAngels: a retrieval-augmented multi-agent framework for psychiatry diagnosis. In: Belgrave D, Zhang C, Lin H, Pascanu R, Koniusz P, Ghassemi M, Chen N, editors. *Advances in Neural Information Processing Systems 38 NeurIPS 2025*; 2025 Dec 2–7; San Diego, CA: Neural Information Processing Systems Foundation; 2025.
21. Hendra H, Deris MM, Windiarti IS. Mental health classification using machine learning with pca and logistics regression approaches for decision making. *Eng Proc* 2025;84:47.
22. Chasombat P, Karuchit W. Digitally depressed: the effects of digital media usage in regard to depression and approaches for alleviation for individuals and the society. *Commun Media Asia Pac* 2020;3:1–15
23. Mahasiriakalayot S, Senivongse T, Taephant N. Predicting signs of depression from twitter messages. *Proceedings of the 19th International Joint Conference on Computer Science and Software Engineering (JCSSE 2022)*; 2022 Jun 22–25; Bangkok, Silpakorn University; 2022. p. 1–6.
24. Hemtanon S, Kittiphattanabawon N. An automatic screening for major depressive disorder from social media in Thailand. *Proceedings National & International Conference* 2019;10:103–13.
25. Wongpakaran N, Wongpakaran T, Wannarit K, Saisavoey N, Pinyopornpanish M, Lueboonthavatchai P, et al. Level of agreement between self-rated and clinician-rated instruments when measuring major depressive disorder in the Thai elderly: a 1-year assessment as part of the THAISAD study. *Clin Interv Aging* 2014;9:377–82.
26. Hemtanon S, Aekwarangkoon S, Kittiphattanabawon N. Proactive depression detection from Facebook text and behavior data. *Int J Electr Comput Eng* 2022; 12:5027–35.
27. Vajrobol V, Shukla U, Pundir A, Singh S, Saxena GJ. Depression detection in Thai language posts based on attentive network models. *Proceedings of the 19th International Conference on Natural Language Processing (ICON 2022)*; 2022 Dec 15–18; IIIT Delhi, India: WNLPe-Health; 2022
28. Electronic Transactions Development Agency (Public Organization), Ministry of Information and Communication Technology. Thailand internet user profile 2015. Bangkok: Electronic Transactions Development Agency (Public Organization), Ministry of Information and Communication Technology; 2016.
29. Khumsri J, Yingyeun R, Manwong M, Hanprathet N, Phanasathit M. Prevalence of Facebook addiction and related factors among Thai high school students. *J Med Assoc Thai* 2015;98:51–60.