Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-023-01512-3

ORIGINAL RESEARCH

Explainable cross-lingual depression identification based on multi-head attention networks in Thai context

Vajratiya Vajrobol[1] · Nitisha Aggarwal[1] · Unmesh Shukla[1] · Geetika Jain Saxena[2] · Sanjeev Singh[1] · Amit Pundir[1,2]

Received: 3 June 2023 / Accepted: 5 September 2023

Abstract Depression is a significant global mental health challenge, and its early detection is crucial for effective treatment. Social media platforms are intricately linked with users' emotions, and thus, on many levels, reflect the users' personal lives through written content. Researchers have access to English data regarding depression detection on social media. However, identifying Depression in low-resource languages can be challenging due to the limited availability of annotated data and language models, especially in the Thai language. This study introduces an approach to tackle the scarcity of resources in low-resource languages. It proposes knowledge transfer from English to Thai as a viable strategy. This approach takes a specific focus on the domain of depression detection, a growing global concern. Additionally, the study delves into the topic analysis of depression within the Thai context. Furthermore, an attempt has been made to determine the most effective architectures used for cross-lingual Thai-English applications for depression detection. Results show that RoBERTa achieved the highest accuracy with 77.97%, recall 77.81%, precision 77.97%, and F1-score 77.86%. Explainable NLP also indicated that RoBERTa has the highest prediction probabilities to capture the context in Depression and non-depression classes.

Keywords Natural language processing · Machine translation · Depression · Transformers · Cross-lingual · Explainable NLP

1 Introduction

Low-resource languages, also known as minority or under-resourced languages, have limited digital resources, such as text corpora, dictionaries, and language models, available for machine learning (ML) and natural language processing (NLP). Cross-lingual learning is leveraging knowledge learned from one language to improve performance in another [1]. Developing effective NLP techniques for low-resource languages can significantly advance cross-lingual learning. For example, machine translation systems trained in high-resource languages like English can be used to improve translations for low-resource languages. Similarly, cross-lingual transfer learning techniques can transfer knowledge from high-resource to low-resource languages, allowing for more effective training of NLP models with a wider range of languages and improving cross-lingual communication and understanding. In addition, researchers can overcome the data deficiency issues in the target languages by transferring knowledge [2]. In this research, translation is conducted from Thai to English to overcome limited language support issues in topic modeling and to investigate the cross-lingual model further to capture the context in Thai when transferred to English. The main benefit of translating data from a low-resource language to English is increased training data that results in improved model performance of language models. By translating data into English, language models can be trained on a larger and more diverse corpus of text, improving accuracy and understanding and generating text in both languages. This can help improve

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41870-023-01512-3.

✉ Amit Pundir
amitpundir@mac.du.ac.in

[1] Institute of Informatics and Communication, University of Delhi, Delhi, India

[2] Maharaja Agrasen College, University of Delhi, Delhi, India

Published online: 07 October 2023

language-based technology and applications for speakers of low-resource languages.

Depression, according to the American Psychiatric Association (APA) [3], is defined as a "negative affective state, ranging from unhappiness and discontent to an extreme feeling of sadness, pessimism, and despondency, that interferes with daily life. Various physical, cognitive, and social changes also tend to co-occur, including altered eating or sleeping habits, lack of energy or motivation, difficulty concentrating or making decisions, and withdrawal from social activities. In Thailand, an estimated 1.5 million Thais struggle with depression. According to the World Health Organization (WHO) [4], depression is more prevalent in females than males, at 2.9% and 1.7%, respectively. Recently, social sensors have been developed to detect major depressive disorder via social networking sites (SNSs), providing a platform to share information efficiently. Social media platforms provide people with an easy way to communicate their thoughts and opinions, allowing academics to look into a variety of psychological issues and human behavior [5].

There are some challenges to accurately identifying mental illnesses on social media networks. Numerous researchers have attempted to find crucial markers in the literature using various NLP techniques. One needs to learn enough about the specific field of research to extract the essential elements and create an appropriate predictive model. Even though these traits were retrieved, this does not guarantee that they are the main causes of the enhanced accuracy [5]. In 2014, studies on "Neural Machine Translation by Jointly Learning to Align and Translate" were introduced and attention was drawn by Dzmitry Bahdanau et al. It is an obvious progression from their earlier work on the encoder-decoder concept [6]. This investigation on transformers led to the establishment of the well-known research Attention Is All You Need by Vaswani et al. [7]. It introduced the concept of word parallel processing, revolutionizing deep learning (DL) [8]. Attention-based models have several benefits over conventional DL models that make them outperform others in certain tasks [9, 10]. Self-attention enables the model to focus on various input components to produce context-aware representations. As a result, the model may better represent the semantic connections among the various input components, which can be helpful for tasks like NLP. Our primary objective in this study was to detect depression using the most effective deep neural architecture from two of the most popular DL approaches in the field of NLP: convolutional neural networks (CNN) and long short-term memory (LSTM) and other attention mechanisms such as Bidirectional Encoder Representations from Transformers (BERT), distilled version of BERT (DistilBERT), Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), and Robustly Optimized BERT-Pretraining Approach (RoBERTa). The study also aims to experiment with cross-lingual settings to investigate depression in low-resource languages (Thai) by inferring with high-language mode. Our approach and key contributions can be summarized as follows:

- The depression dataset was built by translating Thai texts into English utilizing Google Translate API. The dataset will be available on request.
- The studies used BERT pre-trained in a high-resource language (English) to infer outcomes from texts in low-resource languages (Thai).
- Topic-level attention mechanism: a topic modeling framework based on sentence transformers was applied to optimize identifying the possible cause of depression based on Thai contexts.
- Comparative evaluation: The performance of several DL architectures commonly used in NLP was investigated, mainly to detect depression in cross-lingual settings (from Thai to English).
- Explainable NLP for depression dataset using LIME (Local Interpretable Model-agnostic Explanations)

The novelty of this study lies in the innovative application of Explainable NLP, specifically employing LIME prediction using the RoBERTa model. This approach addresses a critical need in the field of mental health detection by making complex black-box models more interpretable. In the context of sensitive domains like mental health, where trust and acceptance are paramount, this research takes a pioneering step towards enhancing transparency and understanding in AI predictions. By shedding light on the underlying reasons for model predictions, the bridge the gap between advanced AI techniques and the human need for comprehensible and trustworthy insights, makes this study a valuable contribution with significant implications for both NLP and mental health applications. The rest of the paper is organized as follows: Sect. 2 describes a few previous works in this domain. Section 3 provides the dataset description, topic modeling techniques based on BERT, and several BERT-based methods. Next, in Sect. 4, we describe the experimental results and analysis. Finally, we conclude our work and define the scope of future work in Sect. 5.

2 Literature review

According to Bel et al. [11], cross-lingual categorization is used when labeled training documents are available only in one language to classify documents written in another. Several studies have focused on cross-lingual text classification; for example, Lee et al. [12] proposed a model that translates a post written in the target language (i.e., Korean) into English and Chinese and then uses the separate suicidal-oriented

word embeddings developed for English and Chinese, respectively. By applying an ensemble approach for different languages, the model achieves an accuracy of over 87%. Another major work by Conneau et al. [13] introduced a benchmark called the Cross-lingual Natural Language Inference Corpus (NLI), or by extending these NLI (XNLI) corpora to 15 languages. XNLI consists of 7500 human-annotated development and test examples in NLI three-way classification format in English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu, making a total of 112,500 annotated pairs. This corpus is designed to evaluate cross-lingual sentence understanding [13] as mentioned in Table 1.

In cross-linguistic settings, there are some studies in the Thai context. Angskun et al. [14] investigate the detection of depression in Twitter by translating the extracted terms from Thai to English using the Google Translation API using ML techniques, including support vector machines (SVM), decision trees, naive Bayes, random forests, and DL techniques. The experimental results show that the Random Forest technique outperformed other techniques for detecting depression. In addition, there is a study on the possibility of detecting depression in the Thai community through Facebook social media and developing a detection algorithm as a new psychiatric instrument. This study examined whether individuals were depressed by first evaluating their Facebook posts by translating them into English and then applying ML algorithms such as SVM, random forests, and DL. The depressed class achieved the highest accuracy of 85% with DL [15]. On several occasions when data was unavailable in Thai, researchers resorted to machine translation for assistance. Mookdarsanit and Mookdarsanit [16] translated the corpus from English to Thai to train the fake news detection models and Transfer learning models consisting of BERT, ULMFiT, and GPT have been applied and the results indicated that the ULMFit performs best. Apart from the Thai language, several languages have adopted cross-lingual techniques due to the lack of language resources. Wan [17] leveraged an English corpus as training data for Chinese sentiment classification. Lee et al. [12] proposed that suicidal-oriented word embeddings created for English and Chinese, respectively, can be used to translate a post written in the target language (i.e., Korean) into English and Chinese. The proposed model also uses the suicide dictionaries created for other languages (i.e., English and Chinese) in word embedding. The model achieves an accuracy of over 87% by using an ensemble technique for various languages [12]. Some studies have looked into the Thai dataset without using translator. Inrak and Sinthupinto [18] have proposed models to classify 6 emotions in Thai texts, such as anger, disgust, fear, happiness, sadness, and surprise, based on latent semantic analysis of nouns and verbs in the sentences. The results showed that Naïve Bayes performs best among the models investigated [18]. Furthermore, Chirawichitchai [19] suggested emotion classification in Thai texts using terms weighing and the SVM algorithm with the Information gain feature selection yielded the best performance with an accuracy of 77.86%. In a recent study in 2021, Hämäläinen et al. collected post data from two leading Thai blog platforms ("Storylog'') and personal blogs posted via two blog platforms ("bloggang.com" and "blogspot.com."). After labeling the posts as depression or non-depression and applying four models to detect depression in Thai, such as LSTM, LSTM with word2vec embeddings, multilingual BERT, and Thai BERT, they discovered that ThaiBERT achieved the highest overall accuracy of 77.53% [20].

In recent years, there has been a notable upsurge in the study of emotion detection, coinciding with a heightened awareness of mental health issues and their profound impact on individuals' lives across various cultures and societies. Researchers have proposed A novel approach combining static facial images and speech modulation, achieving an outstanding accuracy of 94.26%. This surpasses the individual recognition rates of 89% for voice and 91.49% for facial expressions. Notably, the framework also incorporates auto-suggestions to support individuals experiencing depression, thus promoting mental well-being [21]. Similarly, another study introduces a robust emotion recognition system that utilizes Spectral, Prosodic, and Discrete Wavelet Transform features. Through hyper-parameter optimization, a support vector machine classifier is modified, resulting in a language-independent model with 90.02% accuracy on the EmoDB dataset and 71.66% on the SAVEE dataset across seven distinct emotion types. The increasing recognition of the significance of mental health has sparked a surge in research aimed at understanding and addressing a number of challenges, including conditions like depression and suicide [22]. Pandey et al. have developed an AI web-based chatbot named "Ted," leveraging natural language processing and deep learning techniques. This chatbot exhibits an impressive success rate of 98.13% in accurately responding to mental health-related queries, providing an accessible and stigma-free avenue of support for individuals who might be hesitant to engage with conventional mental healthcare providers. As the paramount importance of mental health gains further acknowledgment, a dedicated effort has been made to delve into its multifaceted dimensions, encompassing issues such as suicide [23]. Kancharapu and Ayyagari have utilized Twitter data and various Word Embedding Techniques including Word2Vec, Glove, and FastText. These techniques were employed to train Artificial Neural Network (ANN) Models, with Valence Aware Dictionary and Sentiment Reasoner (VADER) applied to detect signs of suicide inclinations in tweets. Through a comparison of different word embedding techniques, this research provides valuable

Table 1 Summary of some prominent research on multilingual dataset

Author	Dataset	Methodology	Achievements	Detection
[12] 2020	Social media post on suicide	Implemented suicidal-oriented word embeddings in Korean, English, and Chinese and developed suicide detection models	The best performance was achieved by an ensemble cross-lingual model with 87.5% accuracy	Suicide or non-suicide label
[13] 2018	Cross-lingual Natural Language Inference corpus,	Used machine translation systems and parallel data for aligned multilingual bag-of-words and LSTM encoders	The best performance was achieved on BiLSTM-max by directly translating the test data	Entailment, Neutral, and Contradition
[14] 2022	Patient Health Questionnaire-9 survey	This study investigated five distinct machine learning techniques: Support Vector Machine, Decision Tree, Naïve Bayes, Random Forest, and Deep Learning	The experimental findings indicate that the Random Forest technique outperformed the other methods in accurately detecting depression	Depression levels such as moderately depressed, slightly depressed, and, no sign of being depressed
[15] 2014	Depression posts from Facebook	Developed depression detection tool based on 3 algorithms such as SVM, Random Forest, and Deep learning	The best model is deep learning with an F1-score of 88.9%	Depressed or non-depressed
[16] 2021	Global open COVID-19 datasets	Designed a model for thai fake news detection such as BERT, ULMFiT, and GPT	The best accuracy was achieved 72.93% with ULMFiT	Fake news or non- fake news
[17] 2009	Chinese sentiment dataset	Used machine translation for developing Chinese sentiment dataset and applied co-training from both languages for SVM models	A co-training algorithm was used to get a high-performance accuracy of 81.3%	Positive and negative sentiments
[18] 2010	Thai text with emotion labels	Naïve Bayes, SVM, and Decision Tree have been applied	Naïve Bayes performed the best with 90% accuracy	Emotion: anger, disgust, fear, happiness, sadness, and surprised labels
[19] 2014	Thai webboard posts with emotion labels	Term weighting technique with Support Vector Machine approach	Support Vector Machine algorithm, demonstrated the highest accuracy of 77.86%	Emotion: anger, disgust, fear, happiness, sadness, and surprised labels
[20] 2021	Depression dataset from Thai blog posts	Developed a model for the detection of depression using LSTM, ThaiBERT, and Multilingual BERT	ThaiBERT achieved the best performance with 77.53% accuracy	Depression or Non-depression
[21] 2022	Facial recognition and speech dataset	Proposed the fusion of deep classifier	89% and 91.49% respectively for voice signal and facial expression with proposed integrated framework	emotion recognition
[22] 2022	EMODB and SAVEE dataset	Developed speech emotion system based on Language-independent hyperparameter approach	EMODB 90.22% and SAVEE achieves 71.66% performance	Emotions: Neutral, Happy, Sad, Angry, Disgust, Surprise, and Fear
[23] 2022	Mental health queries and answers	Developed AI web-based chatbot called "Ted" to assist people with mental health-related queries using ANN	Achieved 98.13% accuracy	Answer of queries

insights into predicting and addressing feelings of suicide, particularly during the pandemic [24]. It's worth noting that, to the best of current knowledge, no study within the Thai context has explored topic-level attention mechanisms. This research aims to bridge this gap by conducting topic modeling to explore the theme of depression in the Thai context. Moreover, the research seeks to identify the most suitable model based on attention mechanisms, thereby facilitating a deeper understanding of depression in cross-lingual settings.

3 Dataset

The dataset was obtained from existing studies that have collected data from Thai blog platforms ("Storylog") as well as personal blogs posted on two blog platforms ("bloggang.com" and "blogspot.com"). The keywords "depressed," "depression," "depressive disorder," "useless," "fail," "death," "overdose," "suicide," "cut," and "self-harm" appear in each post in the depression category [25]. In terms of non-depression posts have also been crawled from posts that do not have those words of depression sign from the Story log platform. There are 12,837 depressed posts and 12,240 non-depressed posts in the training dataset. Regarding the testing dataset, there are 2567 and 2448 depressed and non-depressed labels, respectively, as illustrated in Table 2 [20]. The average length of sentences is 17 words in a sentence. The dataset can be downloaded from Zenodo [25]. After retrieving data, the dataset was translated from Thai to English using the Google Translation API, as illustrated in Fig. 1 In this experiment, the dataset is divided into a training set of 75%, a validation set of 10%, and a testing set of 15%. The distribution of the dataset can be seen in Table 3.

4 Methods

The pre-processed training data is then forwarded to topic modeling and Hierarchical Clustering. Regarding the detection depression process, the pre-processed training dataset has been trained with several models, such as CNN, LSTM, BERT, DistilBERT, ELECTRA, and RoBERTa as mentioned in Algorithm 1. Consequently, the models have been evaluated using evaluation metrics and generating the results. To enhance comprehension, a concise overview of each model is presented below.

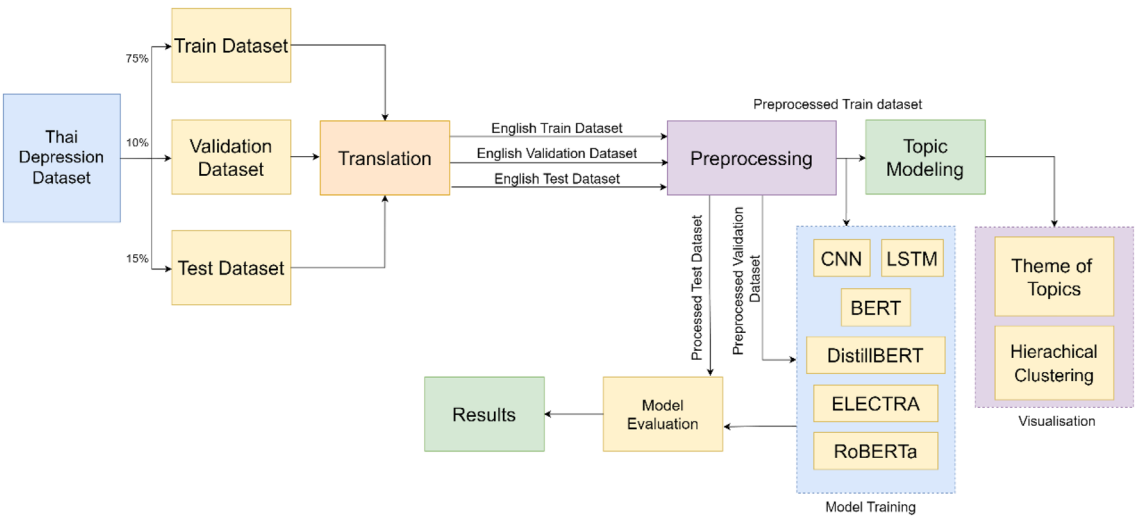


Fig. 1 Cross-lingual depression detection framework

Algorithm 1

Preprocessing:
Input: Thai depression text: thai_text
Output: processed text: processed_eng_text
01: translate thai_text into eng_txt using google API
02: remove punctuations from eng_text
*03: define an array of punctuation punt_text=[!()-[]{};:'"\,<>./?@#$%^&*_~◦ 'space']*
04: ***if*** *eng_text contains punt_text*
05: processed_eng_text = eng_txt.subtract(punt_text)
06: ***end if***
07: remove hypertext links from eng_text
08: define an array of punctuation html_text=[<.?>]*
09: ***if*** *processed_eng_text contains html_text*
10: processed_eng_text = processed_eng_text .subtract(html_text)
11: ***end if***
12: remove URL from processed_eng_text (if any)
13: remove emoji from processed_eng_text (if any)
14: remove abbreviations from processed_eng_text (if any)
15: ***return*** *processed_eng_text*
DoParallel:
Classification and Explainability:
Input: processed_eng_text
Output: best-performing classification model for depression identification on test dataset
17: training deep_learning models (CNN, LST, BERT, DistillBERT, ELECTRA, RoBERTa) on processed_eng_text
18: model evaluation on performance metrics (accuracy, precision, recall, F1 score)
19: for each model:
20: prediction_probability = LIME(processed_eng_text)
*21:****return*** *best_performancing model*
DoParallel:
topic_modeling:
Input*:* processed_eng_text
Output*:* theme analysis of text
22: linkage of a topic-based approach = hierarchical clustering approach(processed_eng_text)
23: cosine distance matrix between topic embeddings = distance(linkage of a topic-based approach)
24: most_frequent_themes = frequency(topics)
25: ***return*** *most_frequent_themes*

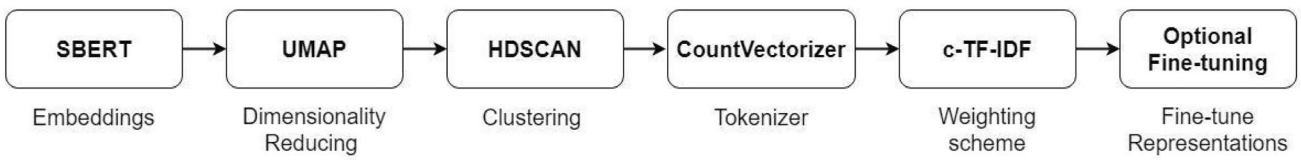


Fig. 2 The process of BERTopic

Table 2 The number of training, validation, and testing sets

Class	Train	Valid	Test
Depressed	12,837	1712	2567
Non-depressed	12,240	1632	2448

Table 3 The example of a dataset after translation

Sentences	Categories
record the devil (mental illness 19) deep regret	Depression
I'll tell her that I've finished reading her favorite book. Do you have any recommendations for me?	Non-Depression

4.1 Topic-level based on attentive mechanism

We investigated the theme analysis of depression posts using the combination of BERT Topic and Sentence Transformers. BERTopic is a topic modeling method that uses the BERT language model to create excellent topic representations for text data. BERTopic employs a clustering strategy to group related documents together into coherent topics, in contrast to classic topic modeling techniques like Latent Dirichlet Allocation (LDA), which uses statistical methods to infer topic distributions [26]. Here, a dataset is recursively divided into smaller and smaller clusters using the hierarchical clustering approach, creating a hierarchy of clusters. Until all data points are contained within a single cluster, the algorithm merges the two closest clusters at each stage. A dendrogram, which displays the hierarchy of clusters and the separation between them, can be used to visualize this method. In hierarchical clustering, the linkage calculates the separation between groups. The distance between clusters is determined using one of several distinct linkage approaches. For a default setting, a ward linkage function performs the hierarchical clustering based on the cosine distance matrix between topic embeddings.

Each word's weight in a particular dataset is shown visually as a bar chart of word scores. Term frequency-inverse document frequency (TF-IDF) or the word frequency in the dataset can be used to calculate the scores. Users may immediately determine which words in the dataset are the most crucial by looking at the bar chart, which displays the top words and their accompanying scores. To help users read and comprehend the resultant topics, this technique can be used in concert with topic modeling techniques like BERTopic to show the top words related to each topic [27].

Focusing on the process of BERTopic, as illustrated in Fig. 2, five steps make up the BERTopic technique—first sentence BERT, which produces high-quality sentence embeddings. Vector representations of sentences called sentence embeddings capture the semantic content of the sentences. Sentence BERT allows BERTopic to produce superior sentence embeddings that can be used to group together comparable sentences [28]. In the second step, Sentence BERT reduces the dimensionality of the sentence embeddings using the Uniform Manifold Approximation and Projection (UMAP) technique. The semantic meaning of the sentence embeddings can still be maintained when the dimensionality of the embeddings is reduced through UMAP [29]. As a result, grouping together comparable sentences is made simpler. In the third stage, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is introduced, which aims to group related words together using the clustering algorithm [30]. is based on the reduced-dimensional embeddings produced by UMAP. Sentence clusters thick and isolated from other clusters can be found using HDBSCAN. Count Vectorizer is frequently used as a pre-processing stage in the fourth phase. Count Vectorizer turns a group of text documents into a matrix of token counts. Each column in the collection of documents represents a distinct word, and each row in the corpus represents a document. The values in the matrix correlate to the word count for each respective document. In the fifth stage, the c-TF-IDF formula is improvised on the standard TF-IDF formula used in topic modeling. It stands for class-based Term Frequency-Inverse Document Frequency and is employed to determine the relative relevance of phrases within a subject. The c-TF-IDF formula, as illustrated in Eq. (1)

$$W_{x,c} = ||tf_{x,c}||xlog\left(1 + \frac{A}{F}\right) \tag{1}$$

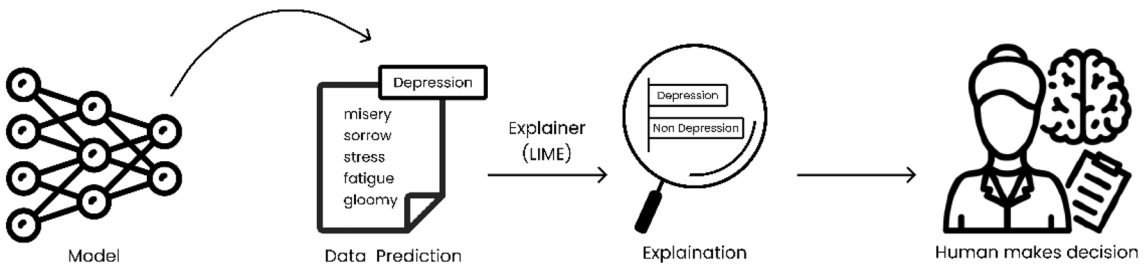


Fig. 3 The process of LIME

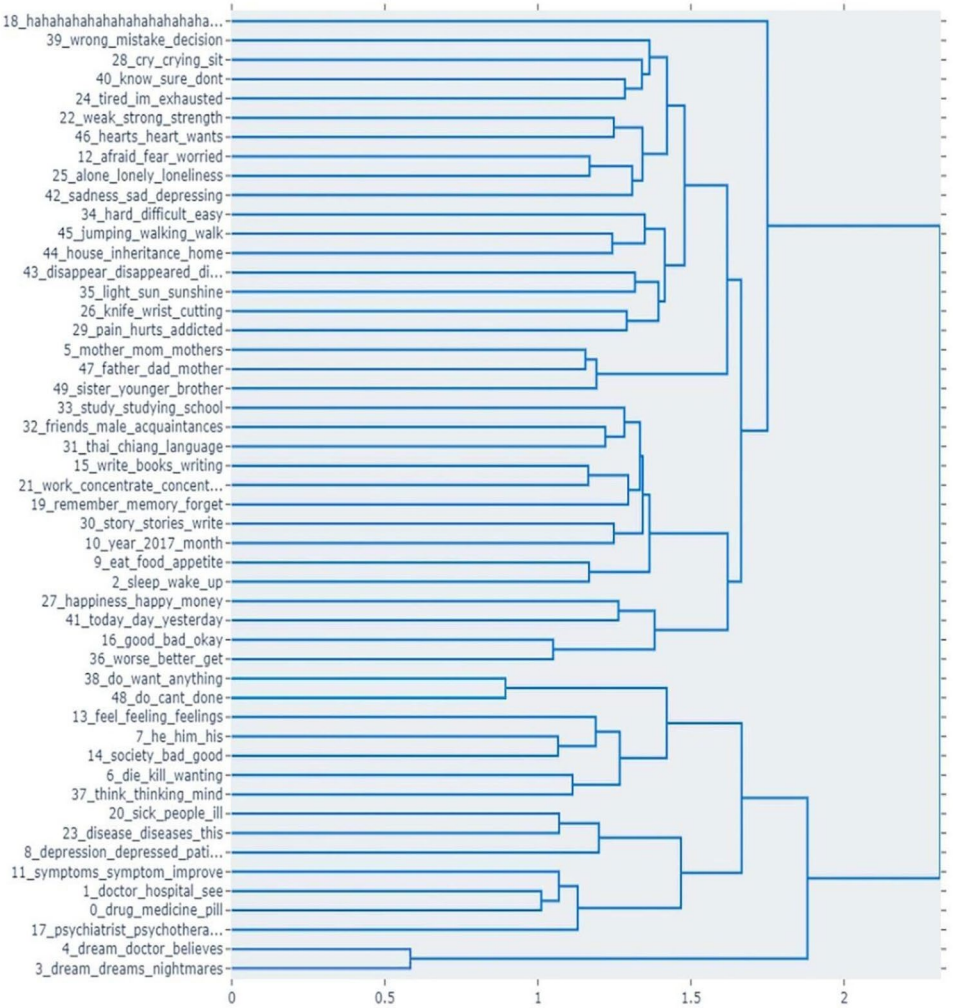


Fig. 4 The hierarchical clustering of each topic in the depression category

where f_x represents the frequency of word x across all classes, $tf_{x,c}$. is the frequency of word x in class c, and A implies an average number of words per class. c-TF-IDF is a TF-IDF formula used for multiple classes that joins all documents from each class. As a result, each class is transformed into a single document. Each class c's frequency of each word x is retrieved. This constitutes the term frequency. After that, the term frequency is multiplied by IDF, calculated by taking the logarithm of 1 and multiplying it by the average number of words in class A divided by the frequency of word x across all classes. In BERTopic, the most crucial words in each sentence cluster produced by HDBSCAN are determined via c-TF-IDF. It makes it easier to pinpoint each cluster's important subjects. The final phase in the BERTopic process is fine-tuning the topic's representation using a ML model. To do this, an ML model must be trained on the collections of sentences produced by HDBSCAN and then used to foretell the themes of fresh sentences. The topic modeling method at BERTopic can be made more accurate by fine-tuning the topic representation.

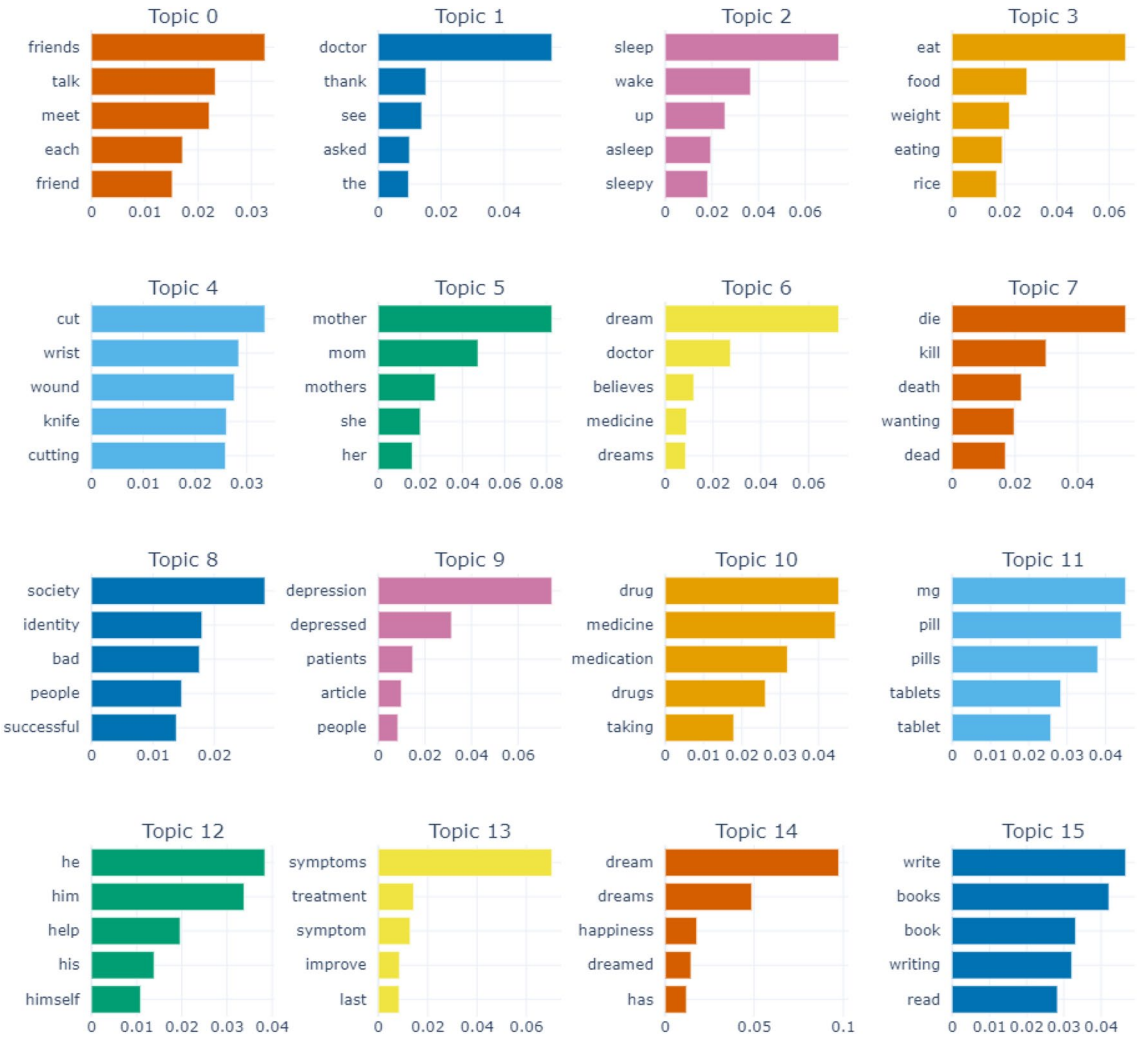


Fig. 5 Bar charts show word scores for each topic

Table 4 The results of each algorithm

Algorithms	Accuracy	Precision	Recall	F1-score
CNN	0.7211	0.7196	0.7149	0.7169
LSTM	0.7450	0.7439	0.7440	0.7438
BERT	0.7767	0.7771	0.7747	0.7753
DistilBERT	0.7695	0.7693	0.7682	0.7685
ELECTRA	0.7725	0.7720	0.7717	0.7718
RoBERTa	0.7797	0.7797	0.7781	0.7786

4.2 Classification model

In the experiment, the deep learning algorithms used, namely CNN, and LSTM, for classification purposes in addition to the multi-head attention approach, for instance, BERT, DistilBERT, ELECTRA, and RoBERTa. The details of each algorithm are illustrated in the supplementary material.

4.3 Evaluation metrics

The models were evaluated for binary classification using the performance metrics of accuracy, precision, recall, and F1 score. A situation in which the model accurately identified the depression class is considered a true positive. An outcome when the model accurately predicts the non-depression class is known as a true negative. A false positive is a result when the model forecasts the depression class inaccurately. A false negative is a result where the model forecasts the non-depressive class inaccurately.

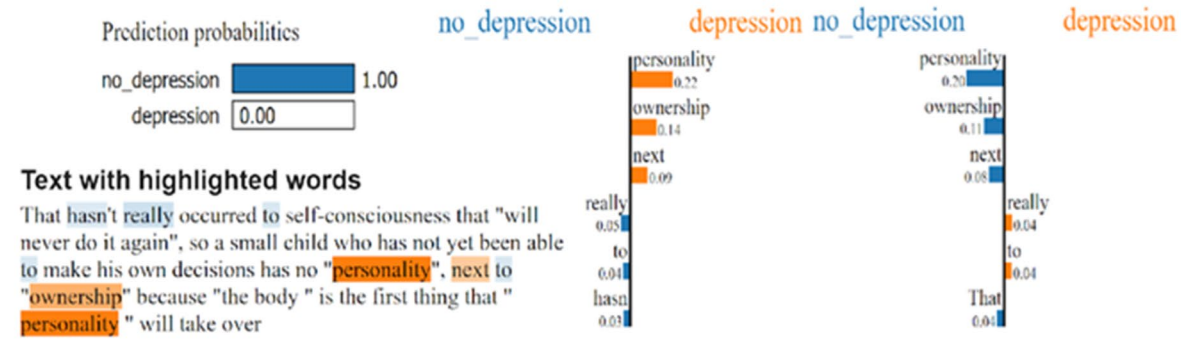
4.4 Explainable NLP

The following section explainable NLP and LIME explanations will be introduced. Explainable NLP is employed to make NLP models more transparent and understandable. It focuses on creating ML algorithms that illuminate the factors and patterns affecting NLP models' predictions and explain how they make decisions. LIME (Local Interpretable Model-Agnostic Explanations) is a well-known method in the explainable NLP framework that can explain predictions of any ML model, including those used in NLP. It functions by building a less complex, more understandable model that closely resembles the behaviors of the original model. Using this basic model, explanations for specific predictions are generated, highlighting the key characteristics and trends in the input and creating regional justifications for specific projections data as follows in Fig. 3. To achieve this, the input data around the point of interest are perturbed and the output is then examined for changes. To capture the critical features and patterns in the data, the perturbations are made using a sampling technique, such as random or perturbation-based sampling. Following the perturbations, LIME utilizes the altered data to fit a simpler model to the data and then uses this model to explain the initial forecast. The LIME explanations are displayed as feature weights, emphasizing the most important elements in the supplied data.

A key benefit of LIME is that it promotes NLP model transparency and trust and can assist in finding biases and flaws in NLP models. In addition, it can identify patterns or relationships that might lead the model to provide inaccurate predictions by highlighting the most important aspects of the input data. As a result, business choices, marketing plans, and other applications where it's critical to comprehend text data can all benefit from LIME insights [31].

Table 5 LIME explanation with CNN model

Non-depression class



Depression class

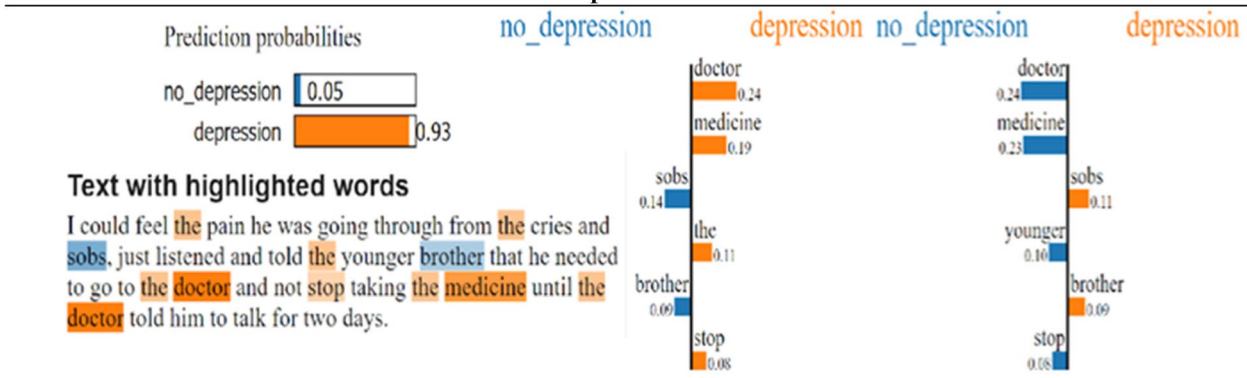
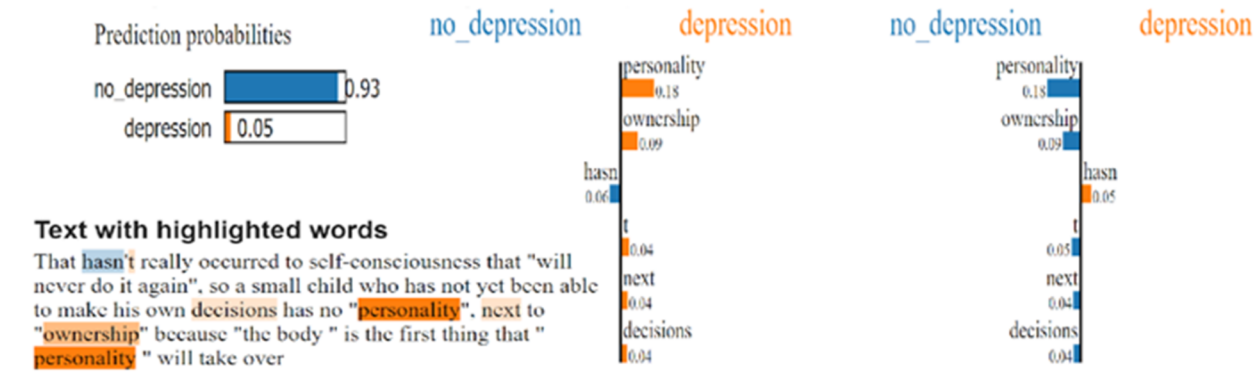
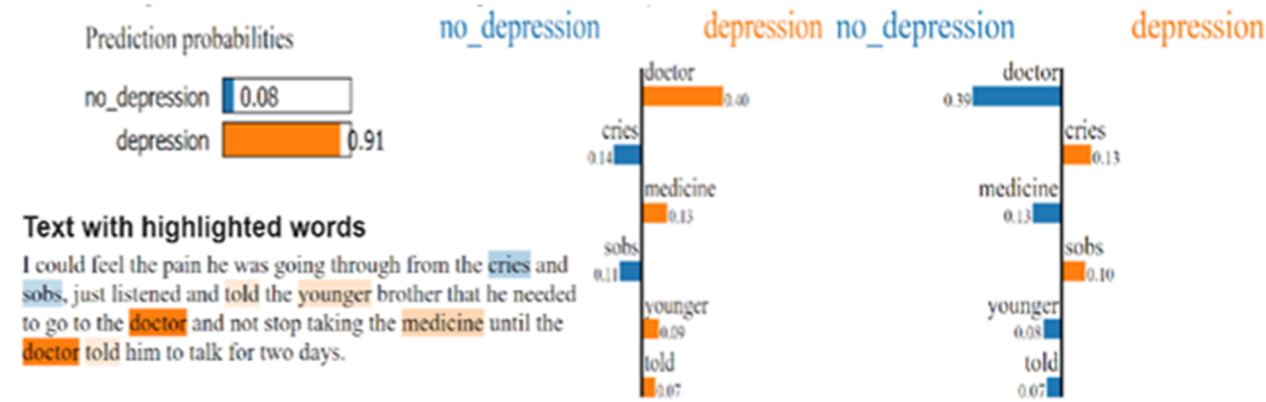


Table 6 LIME explanation with LSTM model

Non-depression class



Depression class



5 Results and discussion

In this section, the results from topic modeling, text classification models, and explainable NLP will be illustrated as follows.

5.1 Topic modeling results

We used BERTopic and sentence transformers to generate topic modeling, as shown in Fig. 4. The hierarchical clustering represents the linkage of a topic-based approach, with 50 topics presented. The clustering result is presented on the cosine distance matrix between topic embeddings using a tree-structured graph. By examining the dendrogram's first level (level 0), we can observe that topics have been grouped together. For example, Topic 6 (die, kill, wanting) and 37 (think, thinking, mind), Topic 4 (afraid, fear, worry) and 3 (alone, lonely, loneliness), and Topic 12 (dream, doctor, believes) and 25 (dream, dreams, nightmare) are grouped based on their similarity. Furthermore, several topics highlight the symptoms of depression, such as feeling tired and exhausted, crying, fearing and worrying, being alone and lonely, feeling sad and depressed, and feeling like disappearing. There are also exhibits of self-harm, such as using a knife and cutting. Some analyses have shown the treatment methods for depression, such as seeing a psychiatrist and receiving psychotherapy.

The bar chart, Fig. 5 illustrates the word score in each topic. Visualizing the selected terms for 16 topics by creating bar charts out of the c-TF-IDF scores. Each bar chart contains the probability that the word belongs to the topic. Topic 0 depicts a communication channel with friends.

Table 7 LIME explanation with BERT model

Non-depression class

Depression class

Topic 1 depicts a doctor's visit; Topic 2 depicts a sleep-related theme; and Topic 3 is related to food and diet. Topic 4 depicts self-harm action. The mother–child relationship is depicted in Topic 5. Topic 6 talks about dreams and medicine. Topic 7 was identified as involving death. Topic 8 is related to society, and Topic 9 is about depression. Topic 10 involves medication. Topic 11 is about tablets. Topic 12 represents using the pronoun "he" for males. Topic 13 is related to treatments and symptoms; Topic 14 talks about dreams; and Topic 15 is related to reading and writing a book.

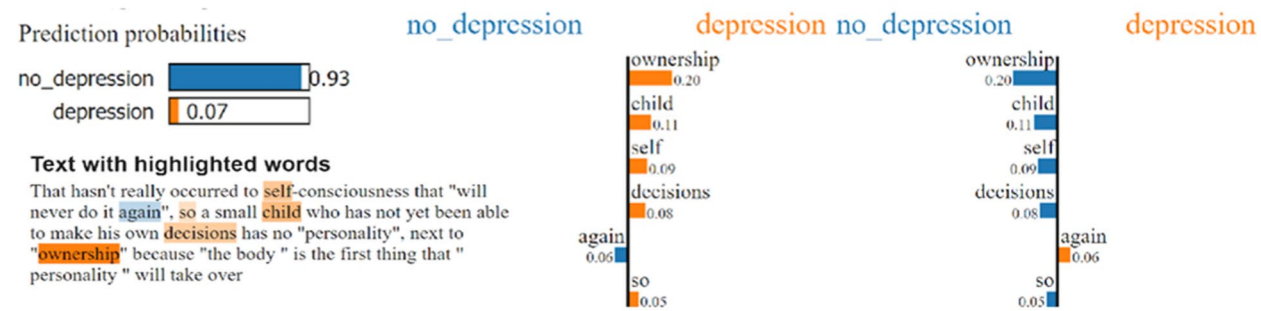
5.2 Text classification results

Focusing on the detection of depression tasks. Several algorithms based on deep learning have been applied, such as CNN and LSTM. Another group of algorithms from the attention neural network model includes BERT, DistilBERT, ELECTRA, and RoBERTa. Focusing on Table 4, results showed that LSTM outperformed CNN with 2.39% accuracy. Due to their superior ability to handle input sequences of varying lengths, LSTMs perform better in this case than CNNs. The reason might be that the length of the input text in tasks involving text classification can differ significantly from one case to the next. CNNs require padding or truncating the input sequences to a fixed length, whereas LSTMs can manage variable-length input sequences by dynamically adjusting their memory cell size. Contrarily, attention models perform better than traditional deep learning models such as LSTM and CNN, as attention-based approaches give the model the ability to selectively pay attention to portions of the input sequence relevant to the current job. As a result, attention models can better extract the most important information from the input sequence and utilize it to generate more accurate predictions. Furthermore, among attention network models, we discovered that RoBERTa performs the best in accuracy, precision, recall, and the F1-score with 77.97%, 77.97%, 77.81%, and 77.86%, respectively.

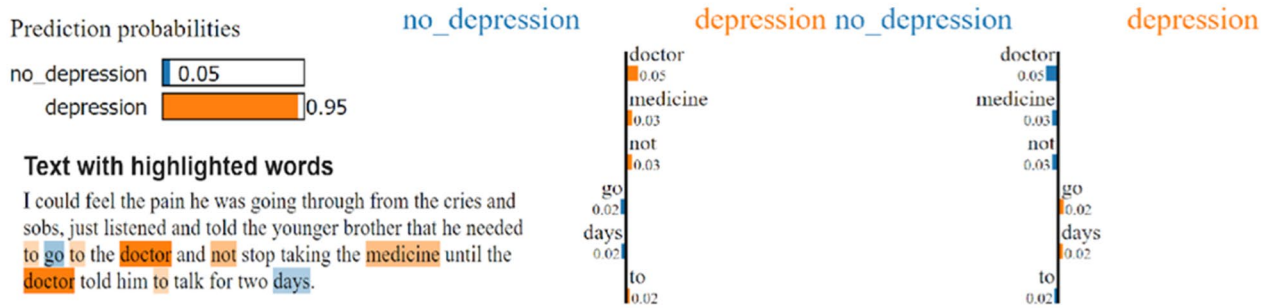
In addition, the comparison of the previous study on the same dataset was conducted in the Thai language.

Table 8 LIME explanation with DistilBERT model

Non-depression class



Depression class



Hämäläinen et al. [20] achieved the highest accuracy with 77.53% using Thai BERT while in this study achieved 77.97% accuracy. The results are increasing a bit, but this study aims to propose low-resource language transfer from English to Thai as an alternative approach for dealing with low-resource language datasets is a valuable contribution. By demonstrating that transferring knowledge from English, a high-resource language, to Thai, a low-resource language, can yield significant results comparable to using only Thai language data, this study could be used to address the scarcity of labeled data in Thai and the potential benefits of leveraging the abundance of English resources to improve performance.

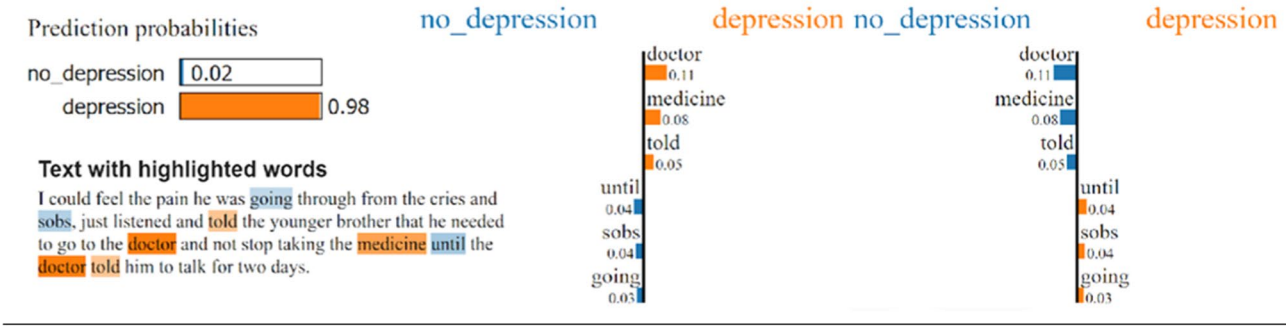
5.3 Explainable NLP results

When using LIME with a classification model, it can tell how likely each class will be classified. This can be useful in understanding why a particular class was chosen over others and identifying potential areas for improvement in the model. Applying all algorithms with LIME, using 1 sample from each depression class and another sample in the non-depression class randomly, as shown in Tables 5, 6, 7, 8, 9 and 10. Eventually, CNNs have been shown to be particularly effective at capturing complex patterns and relationships within the no-depression class with probabilities of 1. Regarding the overall highest probabilities for depression and non-depression classes, it was found that RoBERTa has the highest probability when compared to other algorithms. Regarding ELECTRA, it has been shown that the prediction probabilities in the no-depression class are only 85%, which is among the lowest probabilities among other methods. ELECTRA predicts probabilities of depression class with 98%, which is almost the highest probability. However, this is only one example of text in the dataset. The objective of explainable NLP is to illustrate how the model works inside a black box, not to conduct a comparison of prediction probabilities.

Table 9 LIME explanation with ELECTRA model

Non-depression class

Depression class



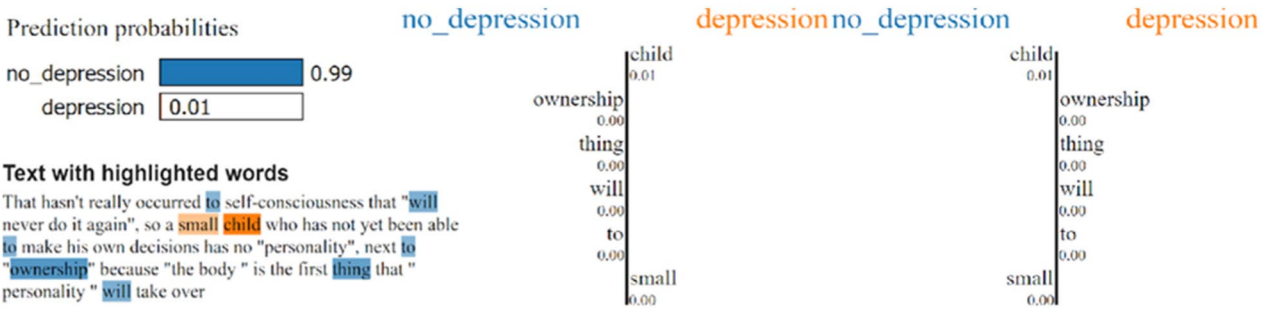
6 Conclusion

We have presented a comparative study of various attention mechanism approaches to optimize depression detection in a cross-lingual Thai context. This study employs high-language models to detect depression in low-resource languages (Thai-English). Topic-level identification based on attention mechanisms by combining BERTopic and Sentence Transformers has been performed. The results show that several markers of depression, such as self-harm, relationships, sleep-related issues, diet, and psychotherapy, were expressed in the posts. A comparative evaluation of some of the widely used deep learning and attention network models for depression detection from blog posts at the user level was conducted. The results showed that RoBERTa achieved the highest accuracy, recall, precision, and F1-score with 77.97%, 77.81%, 77.97%, and 77.86%, respectively. RoBERTa, pre-trained in a high-resource language (English), has the highest capability to detect depression outcomes in low-resource languages (Thai) compared to other algorithms. Thus transferring knowledge from English to Thai can be seen as a solution for low-resource languages. It shows that using English resources can produce comparable results to using Thai data alone, benefiting from the abundance of English resources and addressing the scarcity of labeled Thai data. Overall, it offers a valuable contribution to overcoming data limitations in low-resource language settings.

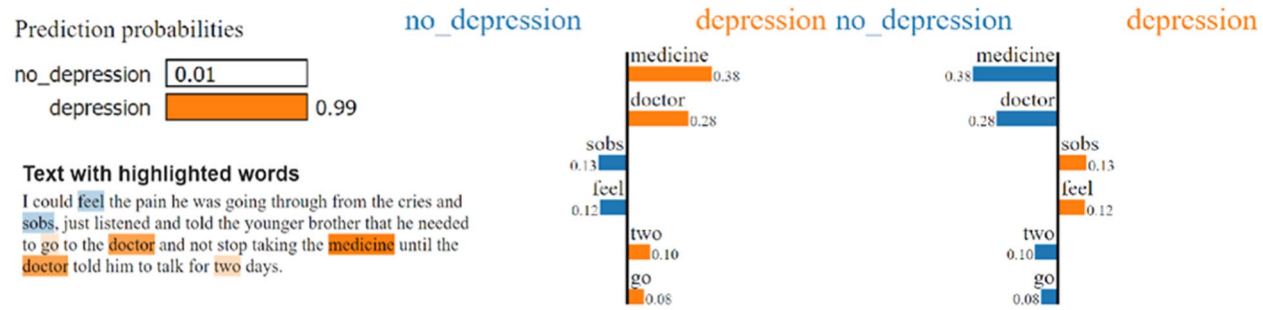
According to Explainable NLP, LIME prediction using RoBERTa had the highest probabilities (99%) in both the depressed and non-depressive classes. However, one limitation of this work is the limited dataset, which may affect the model's accuracy and generalizability. We will evaluate hybrid attention neural network models for future work, focusing on other mental disorders such as post-traumatic stress disorder (PTSD) or suicidal ideation. This research could benefit psychologists in assessing depression based on linguistic patterns and emotional cues in text data with low-resource languages. Utilizing a cross-lingual approach from high-resource languages can overcome limitations in data and resources for low-level languages, enabling more effective NLP applications.

Table 10 LIME explanation with RoBERTa model

Non-depression class



Depression class



Acknowledgements The authors are thankful to Project SAMARTH, an initiative of the Ministry of Education (MoE), Government of India, at the University of Delhi South Campus (UDSC), for their support. I would also like to thank Harshita Sharma and Rinshal Kumar for designing all the figures.

Funding Authors have not received any funding for conducting this research.

Data availability Dataset used in this research is publicly available at https://zenodo.org/record/4734552.

Declarations

Conflict of interest Authors declare no conflict of interest.

References

1. Pikuliak M, Šimko M, Bieliková M (2021) Cross-lingual learning for text processing: a survey. Expert Syst Appl 165:113765. https://doi.org/10.1016/j.eswa.2020.113765
2. Huang K-H, Ahmad W, Peng N, Chang K-W (2021) Improving zero-shot cross-lingual transfer learning via robust training. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1684–1697
3. American Psychological Association APA Dictionary of Psychology. https://dictionary.apa.org/depression
4. World Health Organization Creating awareness on prevention and control of depression. https://www.who.int/thailand/activities/creating-awareness-on-prevention-and-control-of-depression
5. Husseini Orabi A, Buddhitha P, Husseini Orabi M, Inkpen D (2018) Deep learning for depression detection of Twitter users. In: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 88–97
6. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
7. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing System, pp 6000–6010
8. Singh C (2021) Attention mechanism in deep learning, explained. In: KDnuggets
9. Chinnalagu A, Durairaj AK (2022) Comparative analysis of BERT-base transformers and deep learning sentiment prediction models. In: 2022 11th international conference on system modeling & advancement in research trends (SMART). IEEE, pp 874–879

10. Eke CI, Norman AA, Shuib L (2021) Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model. IEEE Access 9:48501–48518. https://doi.org/10.1109/ACCESS.2021.3068323
11. Bel N, Koster CH, Villegas M (2003) Cross-lingual text categorization. In: Research and advanced technology for digital libraries: 7th european conference, ECDL 2003. Lecture notes in computer science, vol 2769. Springer, Berlin, Heidelberg, pp 126–139. https://doi.org/10.1007/978-3-540-45175-4_13
12. Lee D, Park S, Kang J, Choi D, Han J (2020) Cross-lingual suicidal-oriented word embedding toward suicide prevention. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 2208–2217
13. Conneau A, Rinott R, Lample G, Williams A, Bowman S, Schwenk H, Stoyanov V (2018) XNLI: evaluating cross-lingual sentence representations. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 2475–2485
14. Angskun J, Tipprasert S, Angskun T (2022) Big data analytics on social networks for real-time depression detection. J Big Data 9:69. https://doi.org/10.1186/s40537-022-00622-2
15. Katchapakirin K, Wongpatikaseree K, Yomaboot P, Kaewpitakkun Y (2018) Facebook social media for depression detection in the Thai community. In: 2018 15th international joint conference on computer science and software engineering (JCSSE). IEEE, pp 1–6
16. Mookdarsanit P, Mookdarsanit L (2021) The COVID-19 fake news detection in Thai social texts. Bull Electr Eng Inform 10:988–998. https://doi.org/10.11591/eei.v10i2.2745
17. Wan X (2009) Co-training for cross-lingual sentiment classification. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, vol 1. Association for Computational Linguistics, USA, pp 235–243
18. Inrak P, Sinthupinyo S (2010) Applying latent semantic analysis to classify emotions in Thai text. In: 2010 2nd international conference on computer engineering and technology. IEEE, pp V6-450–V6-454
19. Chirawichitchai N (2014) Emotion classification of Thai text based using term weighting and machine learning techniques. In: 2014 11th international joint conference on computer science and software engineering (JCSSE). IEEE, pp 91–96
20. Hämäläinen M, Patpong P, Alnajjar K, Partanen N, Rueter J (2021) Detecting depression in Thai blog posts: a dataset and a baseline. In: Proceedings of the seventh workshop on noisy user-generated text (W-NUT 2021). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 20–25
21. Jayanthi K, Mohan S (2022) An integrated framework for emotion recognition using speech and static images with deep classifier fusion approach. Int J Inf Technol 14:3401–3411
22. Thakur A, Dhull SK (2022) Language-independent hyperparameter optimization based speech emotion recognition system. Int J Inf Technol 14:3691–3699
23. Pandey S, Sharma S, Wazir S (2022) Mental healthcare chatbot based on natural language processing and deep learning approaches: ted the therapist. Int J Inf Technol 14:3757–3766
24. Kancharapu R, Ayyagari SNA (2023) A comparative study on word embedding techniques for suicide prediction on COVID-19 tweets using deep learning models. Int J Inf Technol 15:3293–3306
25. Hämäläinen M (2021) Thai depression detection dataset and baseline models. In: Zenodo. https://doi.org/10.5281/zenodo.4734552
26. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
27. Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794
28. Reimers N, Gurevych I, (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084
29. McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426
30. McInnes L, Healy J, Astels S (2017) hdbscan: hierarchical density based clustering. J Open Source Softw 2:205
31. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144. https://doi.org/10.1145/2939672.2939778