

Original Article

Determination of Users' Sentiments through Posts on Social Media

**Giang Ma¹, Bao Chau¹, Anh Le¹, Quang Cao¹, Nhan Doan¹,
Le Nong¹, Thanh Nguyen², Hai Tran^{1*}**

¹Faculty of Information Technology, Ho Chi Minh University of Education (HCMUE), HCM, Vietnam.

²Faculty of Applied Sciences, Ho Chi Minh University of Technology and Education (HCMUTE), HCM, Vietnam.

*haits@hcmue.edu.vn

Received: 08 October 2024; Revised: 07 November 2024; Accepted: 01 December 2024; Published: 24 December 2024;

Abstract - In the booming technology industry nowadays, social media platforms are widely popular and used by almost everyone. With an enormous number of users, these platforms fully deliver a variety of posts on distinctive topics. As a result, many enterprises have decided to examine the feedback and analyze the sentiments of their customers through posts, videos, or casual conversations about their products or some service to capture the market's actual needs. This article will show the research on a model created by the collaboration of Deep Learning (DL) and Recurrent Neural Network (RNN) algorithms, and Vietnamese texts and images are the datasets of this research. The performance was evaluated by Precision, Recall, Accuracy, and F1-Score, with the accuracy score for the text model at 94.45% and the image model at 90.87%.

Keywords - Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Deep Learning, Bidirectional Long-Short Term Memory (Bi-LSTM).

1. Introduction

User demand is inevitable in most business strategies because it is one of the leading factors for brand development [1, 2]. Based on that, they can identify target groups and optimize user needs, efficiently handling complaints. Therefore, many businesses have chosen emotional analysis models, also known as user status analysis, to personalize their brands, from which these models can be promoted. Gain advantages in exploiting psychology and, at the same time, detect and overcome limitations in the operations of businesses.

Artificial intelligence has applied machine learning and deep learning algorithms to analyze human emotions with scores of huge achievements [3]. In the study of Priya and Udayan, the model was trained using the convolutional neural network algorithm with three datasets: International Affective Picture System (IAPS), Artistic Photos, and Emotion-Image dataset, respectively, creating an analysis and evaluation model [4]. Human emotions are expressed through images with the accuracy of all three datasets at approximately 98%.

In the study by Yang, the authors researched text and image models trained using the TumEmo, MVSA-Single, and MVSA-Multiple datasets to create a model based on a Multi-View Attentional Network (MVAN), yielding an accuracy of 72.98%, 72.36%, and 66.46% [5]. Another model from Nguyen et al. analyzing student feedback is trained on the UIT-VSFC dataset [6]. The study using Maximum Entropy and Naïve Bayes classification models obtained accuracy results of 87.9% and 86.1% in tasks related to emotions and topics, respectively. From there, tools



to determine emotional states also contribute to developing the quality of education at universities, and they are widely applied in many areas of life. The combined text and image model by Li et al. was trained using logistic regression [7]. By training on various text-image pair datasets (Flickr, Flickr-ML, Twitter, Visual Sentiment Ontology), they achieved accuracy metrics exceeding 84%. The problem is collecting user opinions and how practical the evaluation must be in the business market. The team's research will provide a model to evaluate and classify user emotional states through social media posts using a convolutional neural network and recurrent neural network algorithms. With the input being a Vietnamese post, the proposed model will visually analyze and assess the user's status through two status labels: positive or not positive. From there, with the predicted results given by the model, businesses can rely on it to analyze and develop correct strategic plans.

This article is divided into four sections. Section 1 introduces the topic, while Section 2 presents the research subjects and methodology, including the research subjects, proposed model, research data, research methods, and deep learning model analysis. Section 3 covers the results and discussion, including experimental results and their discussion. Finally, Section 4 provides the conclusion for this research.

2. Research Subjects and Research Methods

2.1. Research Subjects

The research subjects are Vietnamese posts on social networks, including text and images. The posts can be on any topic that users share their thoughts on. Data is collected from diverse sources, ensuring diversity and widely reflecting the opinions and sentiments of the user community in Vietnam and foreign communities on social networks.

2.2. Proposed Model for Analyzing User Sentiment through Social Media Posts

In Figure 1, the method for analyzing user states from posts has inputs consisting of two types of data: text strings and image paths. However, if only one of the two data types is provided as input, the model can still process it quickly. The next step involves pre-processing the data to ensure a suitable structure for the model, converting text into feature form, resizing images, and providing results with two types of labels: positive and not positive.

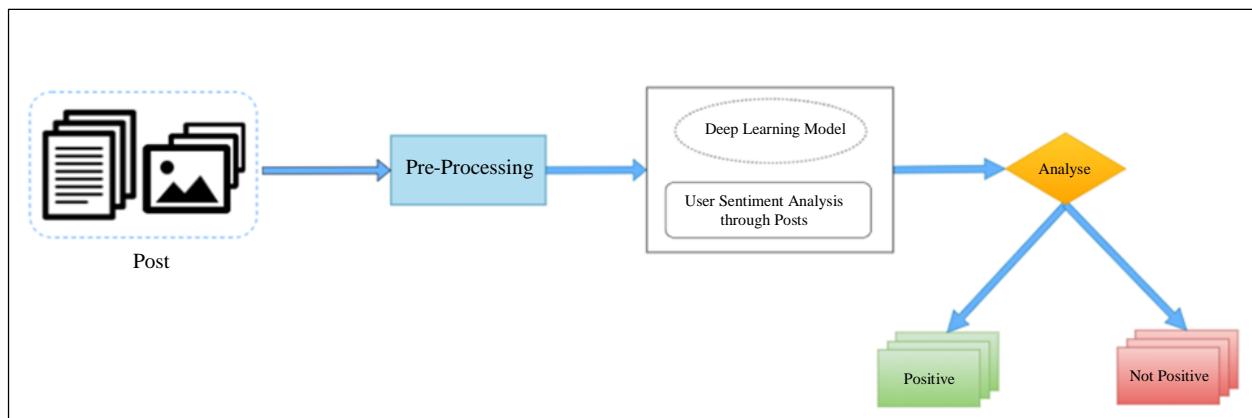


Fig. 1 User sentiment analysis method through social media posts

Figure 2 describes the working mechanism of the post-based status classification model by combining two deep learning models that analyze text and image data, divided into two main stages:

- Stage 1: The model operates based on the mechanism of two text and image labelling models trained from two found datasets. Particularly with text models, because classification is based on three labels, so neutral labels are further processed into positive or not positive labels. After training, the structure is saved for use in stage 2.

- Stage 2: The learned structure of the model is downloaded to process the input posts, then extract the characteristics of the data type and finally classify the post results as positive or not positive.

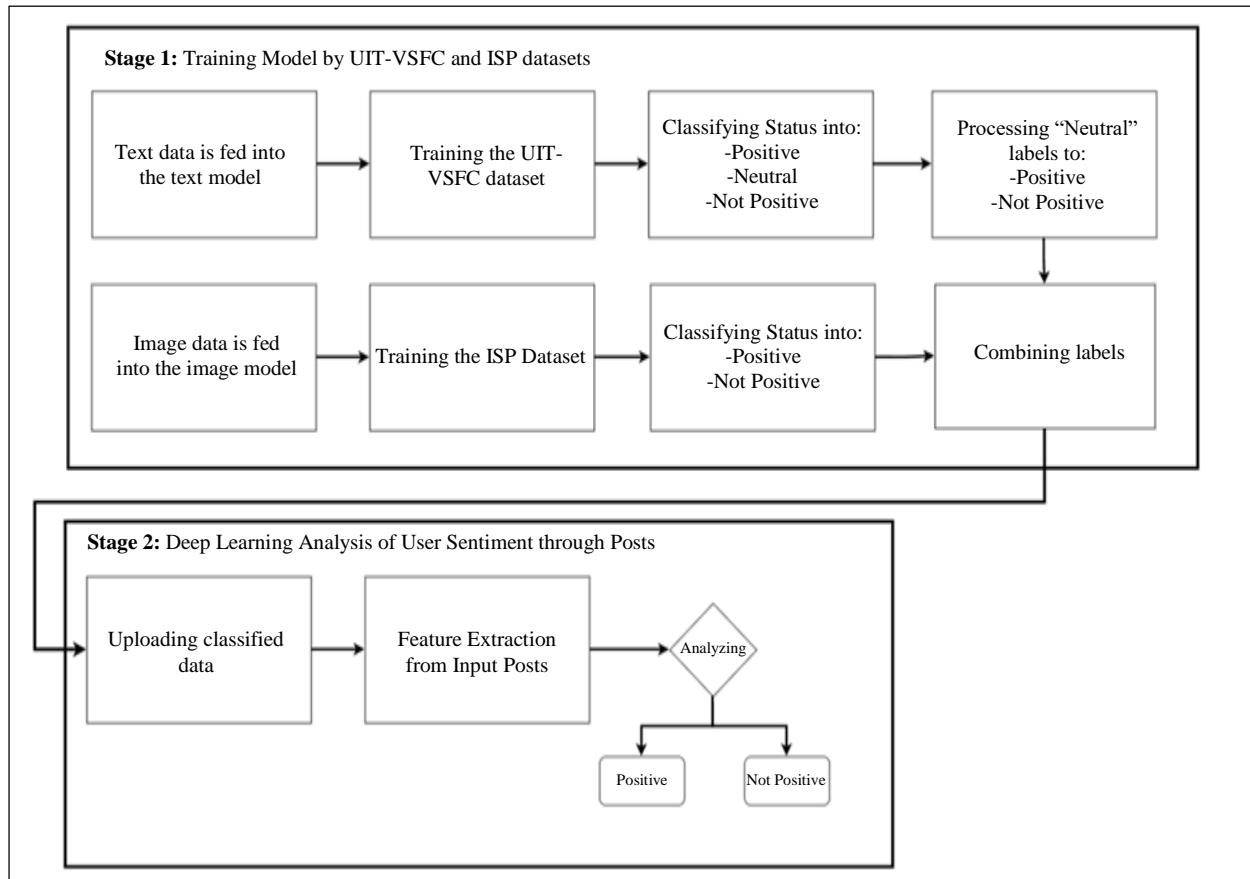


Fig. 2 Structure of deep learning model

2.3. Data and Research Methods

2.3.1. Research Data

The Image Sentiment Polarity image dataset is a public dataset by CrowdFlower on the Data.world website that we use to research image processing and classification models [8]. The dataset consists of an extensive collection of image paths classified and labelled as positive and not positive, totalling 2716 images, 1394 positive images, and 1322 not positive images.

The Vietnamese Students' Feedback Corpus (UIT-VSFC) text dataset contains student feedback on the teaching quality of UIT teachers [6]. The dataset is used through the Hugging Face site [9] with 11,426 training data, 1,583 evaluation data, and 3,166 data used to evaluate the text processing and classification model. Both the above datasets will be processed and analyzed through the proposed model, from which the researched model will be able to identify and categorize posts as positive or not positive.

2.3.2. Research Methods

This research employs quantitative methods to precisely measure the influence of specific factors or events on social network users, as reflected in their posts. The model, once trained, analyses these posts to provide a comprehensive understanding of the user's emotional state. Based on the model's implemented indicators, this is done by assigning two distinct labels, positive and not positive. This method is crucial in gaining a nuanced understanding of social network users' emotional responses.

Method of Selecting Research Datasets

Research model results are based on different available datasets and text and image data sources. The dataset contains student responses. Vietnamese Students' Feedback Corpus (UIT-VSFC) is used for the text dataset. [6] With 11,426 training data, 1,583 evaluation data, and 3,166 testing data, the dataset is suitable for training a model on text-based state classification.

The Image Sentiment Polarity dataset by CrowdFlower, [8] a dataset with many pre-labelled image URLs suitable for image and sentiment analysis, was also used in the model training process. After filtering out inappropriate paths from the positive and not positive sets, another small dataset is created from the original dataset consisting of 2716 images with 1394 positive and 1322 not positive images. From the data collection and evaluation perspective, the dataset must ensure stability and reliability during training to draw conclusions and meet the problem requirements.

Method of Analyzing Data

The team will construct a user-state analysis model once suitable data is obtained. This deep learning model takes in the input post, which is a string of text and the address path of the image. The data then undergoes a pre-processing stage for normalization. Subsequently, it is analyzed and evaluated using two separate models for images and text. These models are learned from convolutional neural networks and Bi-LSTM hybrid convolutional neural network algorithms. The outputs of each model are then combined to produce a comprehensive state prediction of the user.

Evaluation Method

Several formulas are used in the experimental model to measure accuracy and related parameters. The calculation formulas are described as follows research from Dam, Ngo and Han:

Confusion Matrix with TP as true positives, FP as false positives, TN as true negatives and FN as false negatives. [10],

$$\text{Confusion Matrix} = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \quad (1)$$

Accuracy (abbreviated Ac) is calculated based on the formula:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision (abbreviated as Pr) is the outcome of data prediction during training. When equal, it signifies a favourable prediction outcome.:

$$Pr = \frac{TP}{TP+FP} \quad (3)$$

Detection rate or Recall (abbreviated RC) if the value is close to 1, the results obtained are positive:

$$RC = \frac{TP}{TP+FN} \quad (4)$$

Finally, the F1-score (abbreviated F1S) is the average of the Detection rate, and Precision is used to evaluate the identification rate of imbalanced datasets:

$$F1S = \frac{2*Pr*DR}{Pr+DR} \quad (5)$$

2.4. Proposed Model of Deep Learning

2.4.1. Image-Based User State Analysis Model

The two most popular deep learning neural network models are CNN and LSTM. CNN can be used for visualization in deep learning [11]. Our team employs a Convolutional Neural Network (CNN) model for image data to classify sentiment through images into positive or not positive labels. This model is the most suitable for image data processing as it extracts features from each convolutional layer, allowing it to make predictions based on these extracted features from the image dataset. Specifically, the input image data undergoes processing and analysis according to the following scheme:

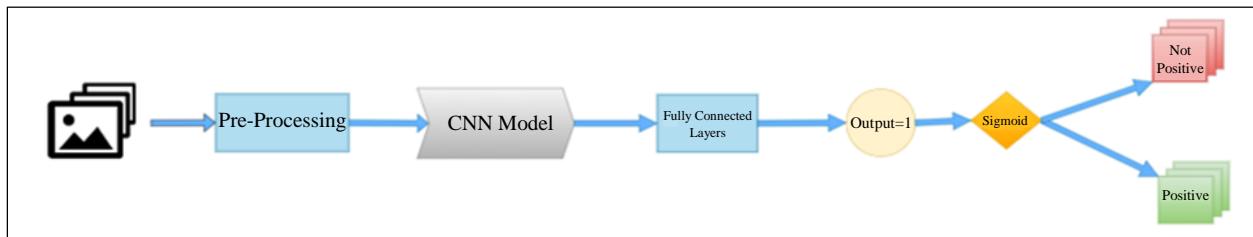


Fig. 3 Steps to determine user status through images

The pre-processing phase makes the back-end training more convenient and more accessible. Images are adjusted to the default size of 256x256 pixels. At the same time, drag the colour values in the image from 0 - 255 to 0 - 1 because, during the learning process, the presence of tremendous values in the image data may impede computational operations and subsequently impact the efficiency of the model, so replacing with values from 0 to 1 by dividing by 255 to optimize and speed up the processing speed of the image model [12].

$$f = \text{lambda } x, y: x = \frac{x}{255}, y \quad (6)$$

With f , a function object that receives and stores the result of an expression, lambda is a method that takes an argument as an expression.

x is the range of colour belonging to the image (0 – 255).

y is the status label.

The convolutional neural network model extracts image features through multiple Convolution2D convolutional layers. The scope is then reduced via Max Pooling layers to mitigate computational complexity while retaining the distinctive image features [13].

Specifically, the Sequential Layers, Conv2D, MaxPooling2D, and Batch Normalization modules are utilized to construct the structure of the convolutional neural network, comprising successive Convolutional Layers and Max Pooling Layers to extract the necessary feature data for classification. Through each combined convolutional layer, a set of new features is extracted, enabling the model to produce the final extracted data for the classification process [14].

Next, the output data from the convolutional neural network model is flattened using Flatten, followed by Fully Connected layers responsible for classifying the status with a single output label, either positive or not positive. In addition to the available Dense layers in this module, Batch Normalization is also employed to normalize the feature data (output of each layer after passing through activations) to a zero-mean state with a standard deviation of 1, helping to minimize the risk of overfitting (a phenomenon where the model performs very well on the training dataset but poorly on the validation and test datasets [13, 15].

Moreover, the Adam algorithm, one of the optimization algorithms (optimizers), is also used in optimization problems. [16] Fundamentally, the optimization algorithm serves as the foundation for building a neural network model to “learn” the features of the input data. This process enables the identification of a suitable pair of weights and biases to optimize the model [17].

Early Stopping will halt the training process when signs of uncontrolled learning occur. This optimization method saves time and effort if the model shows inferior performance [18].

The input parameters of the image used for the proposed convolutional neural network model are 256x256 pixels in size. The output is a dense layer with 1 type of label and a sigmoid activation function (probability less than 0.5 is not positive, otherwise positive). The total trained parameters are 213,441. The number of epochs is 40, divided into three batches of 15-10-15 epochs, and the average time per epoch is 250s. This model is trained and evaluated with batch_size = 32.

The number of output labels of the model depends on the number of key label types marked in the dataset used. In the Image Sentiment Polarity dataset, [8] there are two main types of labels: positive and not positive, suitable for the research problem.

2.4.2. Text-Based User State Analysis Model

For text data, the team employs a deep learning model that combines Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (Bi-LSTM) to address the classification problem based on Vietnamese text, with three output labels: positive, neutral, and not positive. The Bi-LSTM model, a variant of the recurrent neural network algorithm, demonstrates excellent performance, hence the team’s decision to integrate it with the CNN model to produce the best prediction results. The analysis and training process is illustrated in the following diagram:

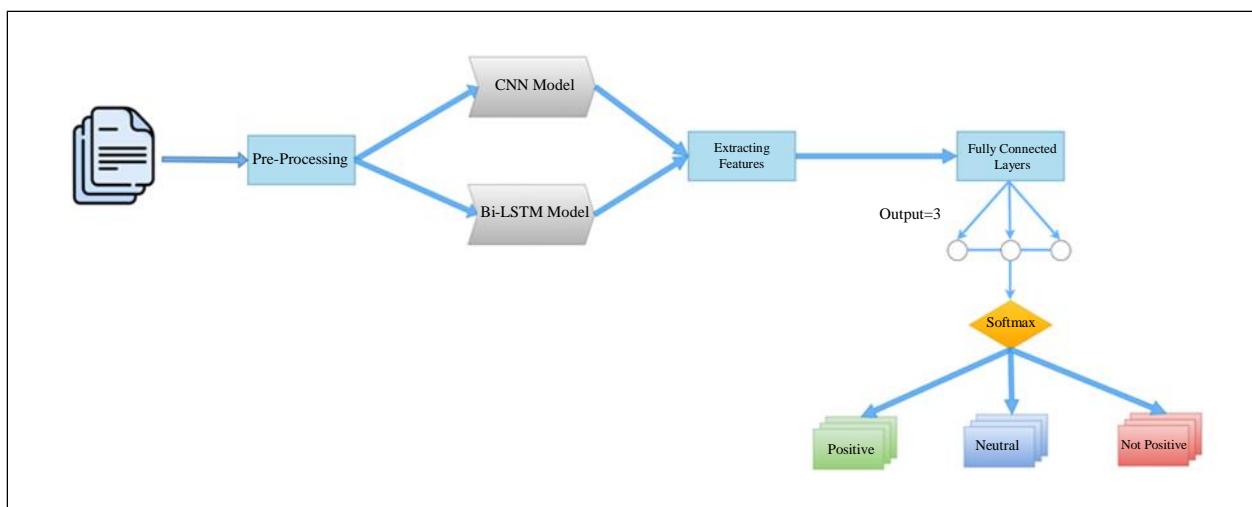


Fig. 4 Steps to determine user status through images

Pre-processing is a step to standardize data in the model [19]. After removing special characters, the text is processed into unaccented and accented lowercase text. Creating an unsigned version of the input text helps the model learn more forms than usual. At the same time, when creating a dictionary through training data, the number of words in the dictionary will also be more significant, leading to a broader source of extracted data. Utilizing a convolutional neural network, the input text is meticulously traversed by the window (kernel) from left to right, with each slide representing a phrase. This precision allows the model to thoroughly analyze each cluster and

extract the features of the text after synthesis [20]. At the same time, Bi-LSTM - an extended version of LSTM, operates based on two separate LSTM networks. If LSTM analyses and learns each word in one direction, Bi-LSTM learns both directions from left to right and vice versa. [21] Both LSTM networks return a probability vector as output, and the final output is a combination of both these probability vectors based on the following formula:

$$p_x = p_t^f + p_t^b \quad (7)$$

Where p_t is the final probability vector of the network.

p_t^f is the probability vector from the feedforward LSTM network.

p_t^b is the probability vector from the inverse LSTM network [22].

Combining two convolutional neural network models and Bi-LSTM, we get a model that inherits the ability to extract features through each phrase in a sentence of a one-dimensional convolutional neural network and can also extract features from each phrase in a sentence. Bi-LSTM, whose ability to analyze and learn each word in a sentence in 2 dimensions, achieves the best results more effectively than LSTM [23].

During training based on text data, the problem also applies several algorithms and relevant libraries such as Embedding (for representing input text), Dense (fully connected neural network layer with ReLu activation function), Dropout (to prevent overfitting), Bidirectional LSTM (to extract and learn features in both directions), Bidirectional GRU (used to optimize learning time due to fewer parameters than Bi-LSTM), Input, GlobalMaxPooling1D (simplifies computation), Layer Normalization (normalizes feature layers and limits overfitting), Conv1D (extracts text features in one direction) to facilitate the construction of the convolutional neural network model combined with Bi-LSTM.

Bidirectional GRU helps with timing, as GRU uses fewer training parameters and less memory than LSTM. LSTM has an advantage in accuracy and set size data. LSTM is also used in modelling to solve large sequences and requires higher accuracy [24].

In addition, the ViTokenizer and ViUtils modules were used in the pre-processing stage to process Vietnamese vocabulary [25]. The maximum text input parameters are 512 words, and the dense_4 output includes three labels: positive, neutral, and not positive. The total trained parameters are 249,585, the number of epochs is 15, and the time Average per epoch is 350s. This model is trained with batch_size = 128.

The number of output labels of the model depends on the number of main labels marked in the dataset used. The Vietnamese Students' Feedback Corpus has three main labels: positive, neutral, and not positive. [6] Attempting to change the "neutral" status label before use into one of the remaining two labels might adversely affect the results of the text model. Therefore, the model predicts based on three status labels and then adjusts based on the remaining two status indices (the SoftMax activation function helps ensure that the sum of the outputs equals 1) to provide appropriate results. Consequently, the output of the text model returns positive or not positive, including the image model, laying the groundwork for combining text and image models.

2.4.3. The Model Analysis User Status through Posts (Text, Images) on Social Networks

With the text and image state analysis model completed, the proposed model is built based on a combination of state assessment of both types of input data, thereby providing the ability to analyze the state of text and images. Ability to figure out a user's status through social media posts. To classify post status, the model needs to complete two critical processes: pre-processing and combining analysis results of text and images to produce the final status result.

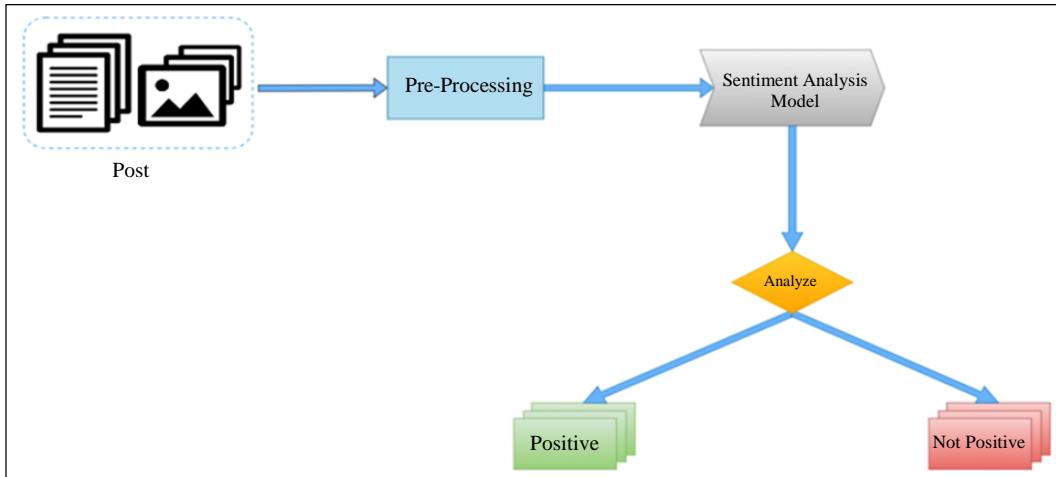


Fig. 5 Steps to analyze user status

In this study, the `text_prediction` function is a key component. It encompasses text pre-processing (`preprocessing_raw_input`) and a pivotal function called `inference_model`. The latter is responsible for predicting the final sentiment label of the text, along with its corresponding index (ranging from 0 to 1). Additionally, the `inference_model` function adjusts the three labels of the text model (positive, neutral, and not positive) to just two labels (positive and not positive) based on the index of each label. This adjustment is made according to the following formula:

$$Negative_{point} = \frac{Neutral_{point}}{2} + Negative_{point} \quad (8)$$

$$Positive_{point} = \frac{Neutral_{point}}{2} + Positive_{point} \quad (9)$$

With $Negative_{point}$ is the non-positive index of the text.

$Neutral_{point}$ is the neutral index of the text.

$Positive_{point}$ is the positive index of the text.

After obtaining the values of the two new labels, they are compared to decide the label with the highest value. This is also the result of the user sentiment analysis through text. Next, `image_prediction` is built, including pre-processing functions and analyzing user status through images. In the function, the image is processed to match the model's input and then fed into the image model to produce the results. The returned result is a ratio ranging from 0 to 1. If greater than or equal to 0.5, the image is positive; otherwise, it is not positive.

After obtaining the processed user state analysis function through images and text, proceed to construct a prediction function for user state through posts, with parameters including the input text (`text_input`), the dictionary built from the previous text training data (`tokenizer_data`), the text processing model file (`text_model`), the image path (`image_input`), and finally, the image processing model file (`image_model`). The function considers the type of data contained in the input, from which it provides appropriate analysis. Specifically, if there is only text or images, the function will call the function to analyze the image or text corresponding to the input separately.

When both data types are present, the model analyses both and compares their suitable labels. If both types yield the same final state result, that result will be the final. Conversely, every kind of data produces a different outcome. In that case, the model will classify by using the difference between the two detailed indices of each data, comparing it with a coefficient of 0.1 (the coefficient chosen by the consensus of the entire team through multiple

experiments with different coefficients, suitable for making label classification decisions) to determine the final state. Specifically, the processing steps of the model are as follows:

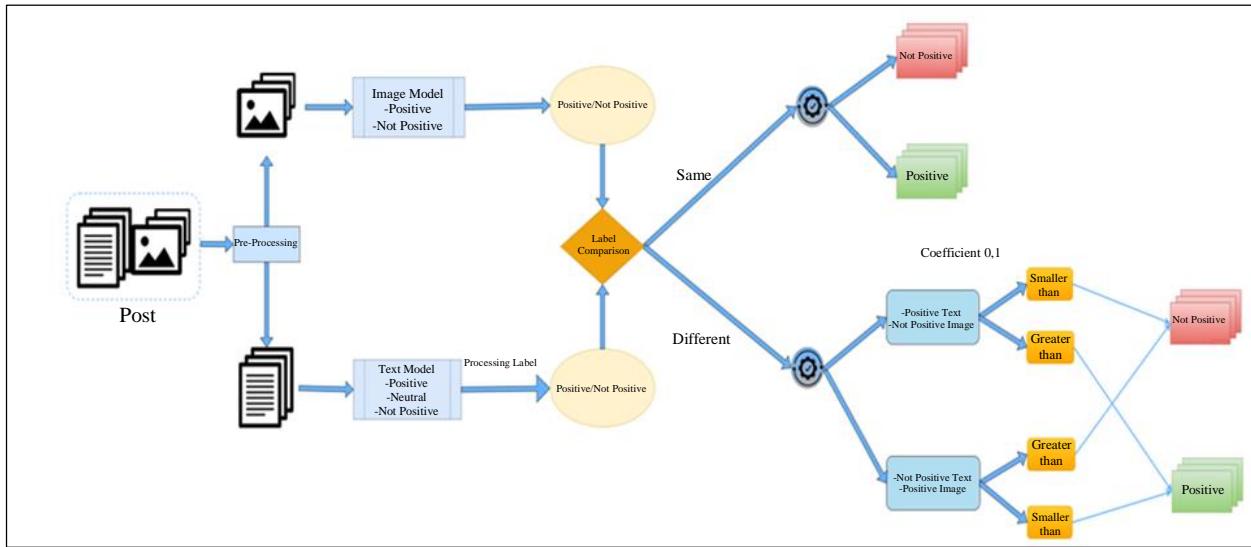


Fig. 6 Diagram of the classification

3. Results and Discussion

3.1. Experimental Results

3.1.1. Result with Visual Data

After training on the training dataset, experiment on the test dataset and the results: Precision is 86.2%, Recall is 97.6%, F1-score is 91.57%, and accuracy is 90.87%. In the test dataset, a confusion matrix was applied to visually represent the model's accuracy. Out of 252 image data points, including 128 non-positive and 124 positive images, the model correctly identified 89% (114 images) in the non-positive group and 97% (120 images) in the positive group.

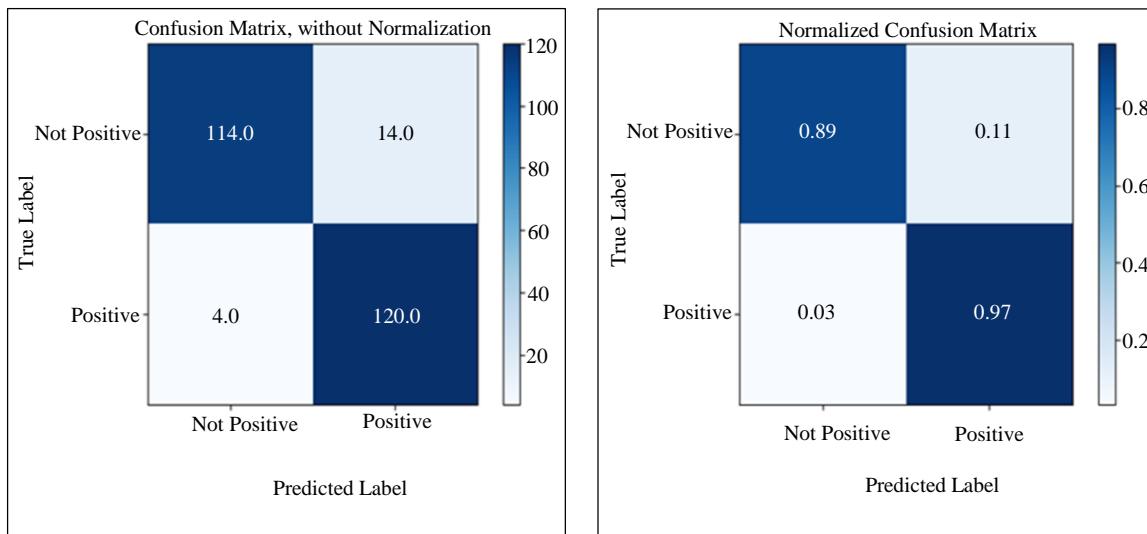


Fig. 7 Confusion matrix for the image test dataset

Several images were tested during the training process. For example, specific results for positive and non-positive outcomes are presented in Figures 8 and 9.

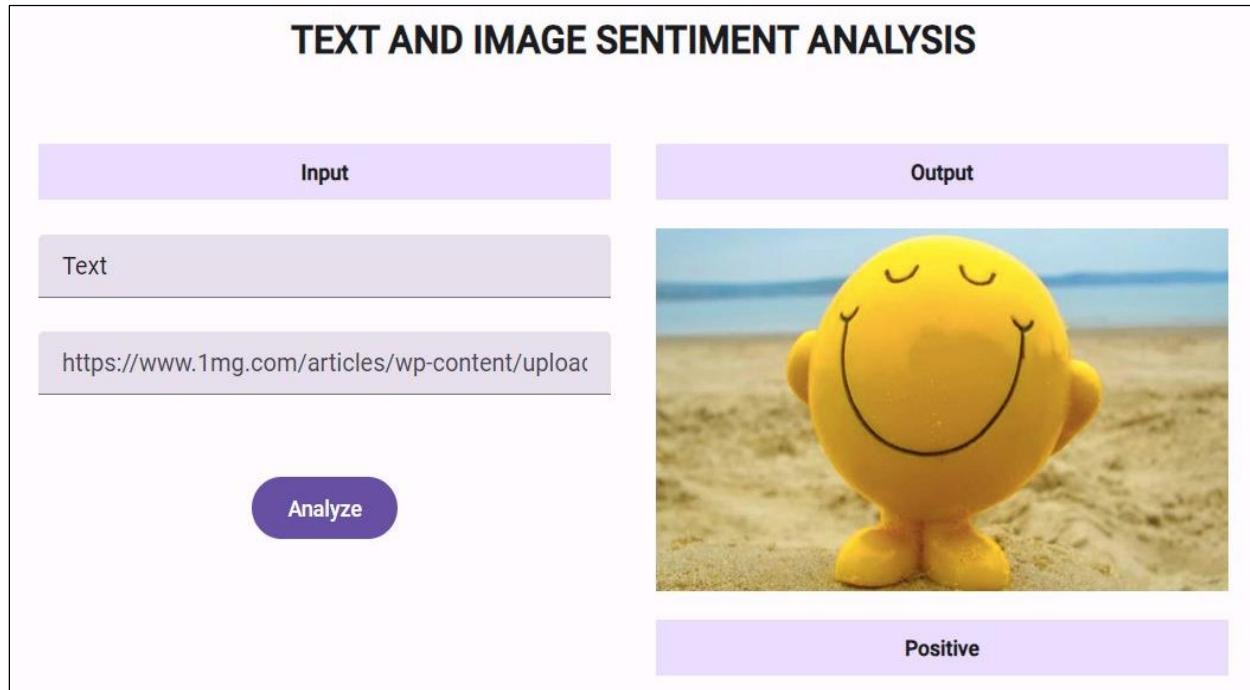


Fig. 8 Example of a positive image

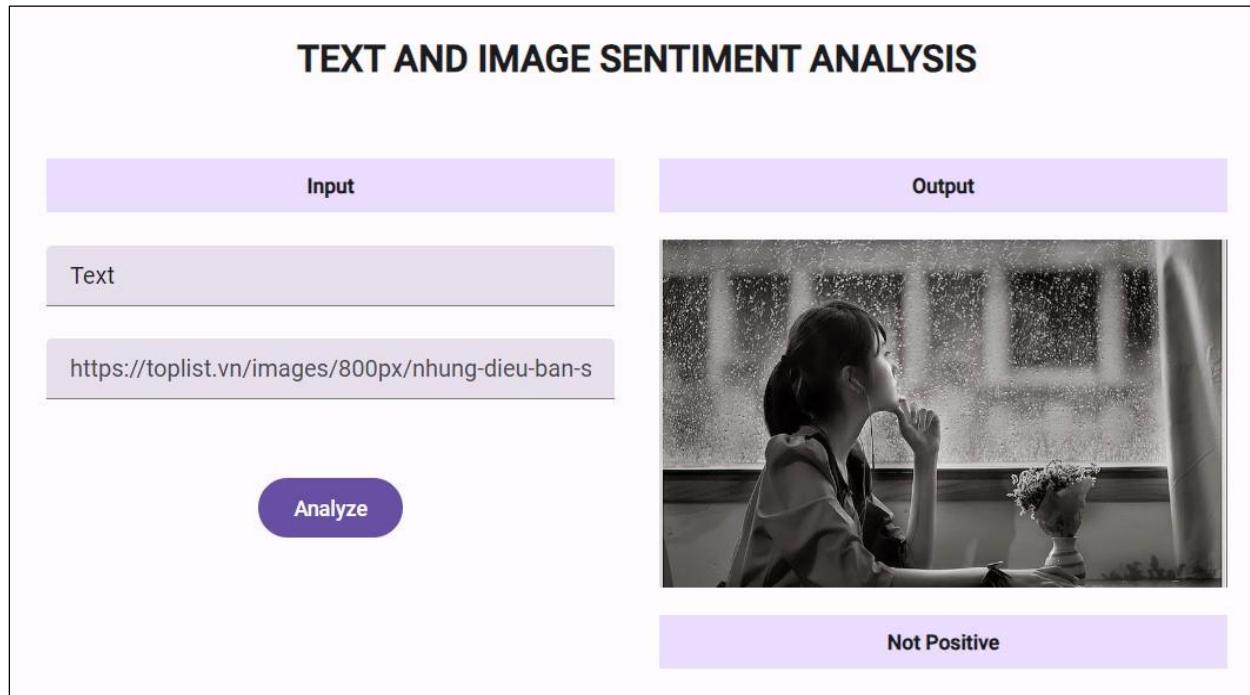


Fig. 9 Example of not positive image

3.1.2. Result with Textual Data

After training on the training dataset, the test dataset was experimented on, and the results were as follows: Precision is 92.1%, Recall is 91.1%, F1-score is 91.6%, and accuracy is 94.4%. Figures 10 and 11 present some of the post-texts tested for the model, which were positive and not positive, respectively.

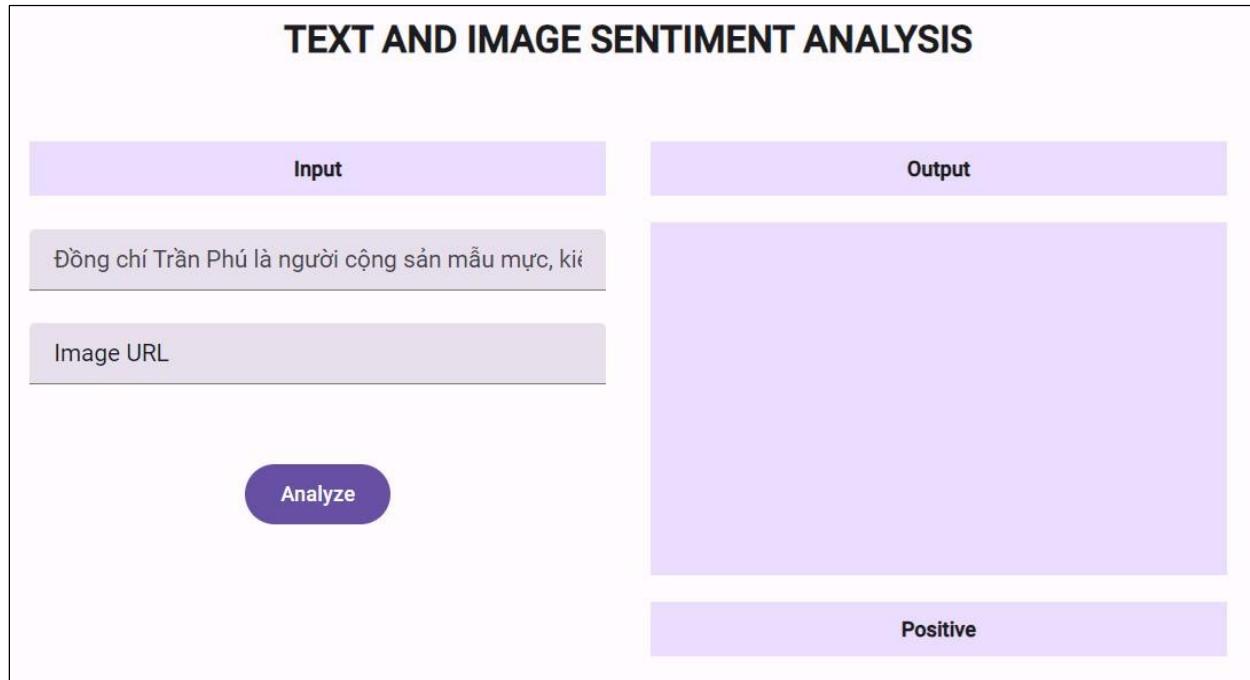


Fig. 10 Example of positive text

(Đồng chí Trần Phú là người cộng sản mẫu mực, kiêu cường, bất khuất, đã hiến dâng trọn đời mình cho sự nghiệp cách mạng của đảng và của nhân dân, luôn nêu cao phẩm chất đạo đức cách mạng trong sáng, thủy chung, trung thành vô hạn với sự nghiệp cách mạng của đảng, của giai cấp và của dân tộc, lạc quan tin tưởng vào tương lai và sự tất thắng của cách mạng.)

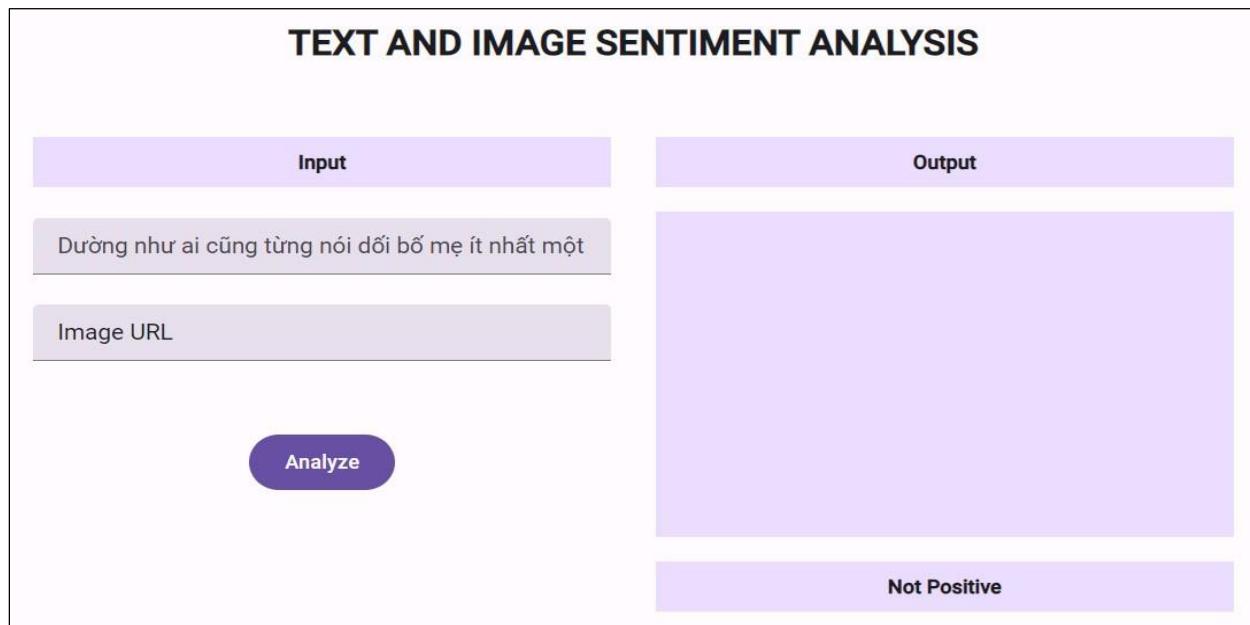


Fig. 11 Example for not positive text

(Dường như ai cũng từng nói dối bố mẹ ít nhất một lần trong đời, dù được dạy là không tốt. Một phần vì chúng ta có tâm lý phản kháng (càng bị cấm, bạn càng muốn làm), một phần bởi đây là cách dễ nhất giúp chúng ta thực hiện một nhu cầu nào đó của bản thân, mà ta gặp khó khăn khi giao tiếp với bố mẹ.)

3.1.3. Results with Combined Image and Text Data

With both text and image inputs, the two models are combined to create a function capable of processing and analyzing both text and images, producing an overall result. Specifically, below are some illustrative examples:

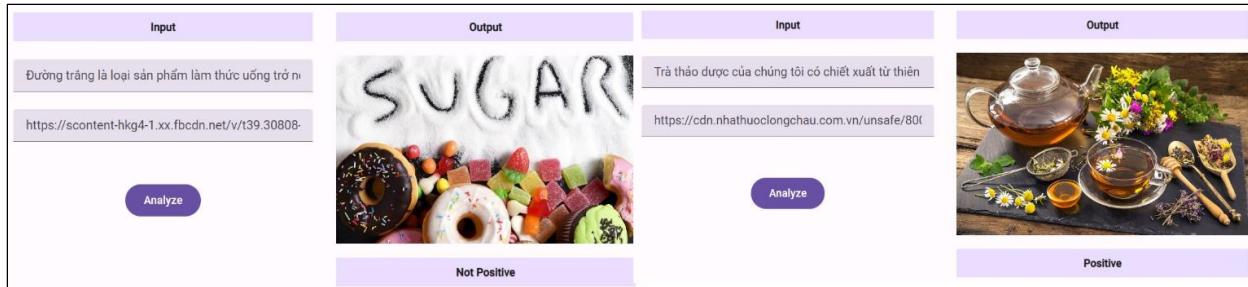


Fig. 12 Predictions when text and image have the same label

(Đường trắng là loại sản phẩm làm thức uống trớn nên ngọt miệng hơn, tuy nhiên nếu sử dụng với liều lượng không đúng sẽ gây tác hại nguy hiểm cho sức khỏe.) (Trà thảo dược của chúng tôi có chiết xuất từ thiên nhiên, an toàn và lành tính cho sức khỏe.)

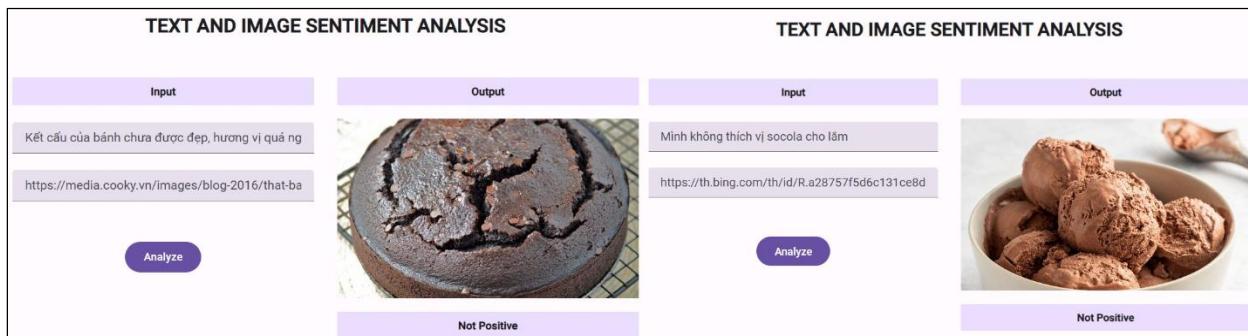


Fig. 13 Predictions when text and image have the different label

(Kết cấu của bánh chưa được đẹp, hương vị quá ngọt dẫn đến nhanh chán.) (Mình không thích vị socola cho lắm.)

3.2. Discuss

In the study, analyzing user sentiment through social media posts met real-time needs, employing deep learning technology and achieving relatively high accuracy results. In the convolutional neural network model applied to image data, not only the quality and compatibility of the model are essential, but also the diversity of images, colours, and shades, along with the meaning of the pictures, plays a significant role in the analysis and classification of sentiment. These aspects also represent challenging elements in our team's research process. Although the results of the image model are not yet satisfactory, this constitutes a step forward and a foundation for developing and improving future models.

Table 1. Model comparison table of research by Priya & Udayan and Li, Zhu, Gao, Cao & Wang

Research	Datasets	Accuracy
Priya & Udayan	International Affective Picture System (IAPS), Emotion-Image, Artistic Photos.	~98%
Li et al.	Flickr	73.2%
Proposed model	Image Sentiment Polarity (ISP)	90.87%

The model by Li et al. used the Flickr text-image dataset [7]. Their study presented performance metrics using the model on separate text, image, and combined text-image datasets. Our team only used the accuracy of the image dataset to compare with our proposed model. Our proposed model can be compared to the mentioned model because the dataset our team used is a smaller subset of the Flickr dataset. Therefore, the comparison results are relevant to each other. Compared with the logistic regression model and the proposed convolutional neural network model, the accuracies were 73.2% and 90.87%, respectively, showing that the proposed model yields more accurate results.

Although our model Priya & Udayan also uses the network neuron accumulation convolution but produces higher accuracy with all three datasets, the model based on the IAPS and Artistic Photos datasets gives better results than the Emotion-Image dataset [4]. The article's authors also stated that their model has better performance when compared to network models of neuron accumulation. Other convolutions are mentioned. As for the proposed model, although the accuracy is lower than that of Priya and Udayan, there is still a negligible difference. Therefore, the proposed model still works effectively and performs similarly to this model. Furthermore, the proposed model will also focus more on research and development in the realm of image data in the future.

Next, the convolutional neural network model + Bi-LSTM analyses the student's feedback text and produces extremely positive indicators:

Table 2. Model comparison table of Nguyen et al.

Research	Datasets	Accuracy
Nguyen et al. - Maximum Entropy	UIT-VSFC	87.9%
Nguyen et al. - Naïve Bayes	UIT-VSFC	86.1%
Proposed model	UIT-VSFC	94.45%

The research team utilized a training dataset as the one used in our proposed model, but they employed entirely different algorithms. Maximum Entropy and Naïve Bayes are both part of the probabilistic classifier family. In contrast, the algorithm of our proposed model involves learning recognition features, followed by using convolution layers and pooling layers to extract feature vectors for each word. Regarding the output results, there is a difference in accuracy: the proposed model achieved 94.45%, while the accuracies for Maximum Entropy and Naïve Bayes were 87.9% and 86.1%, respectively, with Maximum Entropy achieving higher accuracy than Naïve Bayes [6]. In conclusion, the proposed model demonstrates higher accuracy than the two models presented.

By surveying the results of the research work, the assessed accuracy of the mentioned model and the proposed model showed the differences in each sub-problem handle. As for the image through tissue, the proposed model has an elevated level of accuracy (90.87%), higher than Li, and the difference is tiny compared to tissue images Priya & Udayan. Still, in general, when combining both text and images, thanks to the extremely high accuracy of the model proposed text image (94.45%), the error rate will be reduced. The survey shows that the proposed model has a practical and extremely objective learning rate.

4. Conclusion

Through this study, the research team investigated two main areas: firstly, proposing a model for analyzing user sentiment through images using a deep learning convolutional neural network algorithm with the Image Sentiment Polarity dataset, which includes positive and not positive labels; secondly, suggesting a deep learning model combining convolutional neural networks and Bi-LSTM to analyze human emotional states through text using the Vietnamese Students' Feedback Corpus, where data is labelled as positive, not positive, and neutral. The two models were integrated into a comprehensive model for analyzing user states through posts that include

images and text on social media. The evaluation process was based on accuracy, precision, recall, and F1-score criteria, with the image training accuracy result being 90.87% and the text being 94.45%. These results have been applied effectively to determine human emotions accurately.

Regarding images, the model has managed quite well with datasets with clear differentiation in colour tones, using bright and dark colour schemes to categorize positive and not positive sentiments, respectively. However, this also presents a limitation in that it is difficult to discern the true meaning of an image, leading to potential undesired inaccuracies.

The proposed model in the article can address issues with Vietnamese data. However, training on Vietnamese posts presented challenges due to the diversity of grammar and context, requiring critical time to manage spelling obstacles to achieve the highest accuracy. Furthermore, more than the UIT-VSFC dataset is needed to predict a broader range of topics, thus limiting the model's applicability to the themes present within that dataset.

We envision a collaborative effort to develop the classification model further in the future. This includes diversifying the labelling topics of the dataset and expanding the analysis scope by considering and utilizing various topics, text datasets, and images. In addition to analyzing posts on social media platforms, we are also focusing on user interactions with a particular post through reactions or shares to classify and improve the model for better emotional state recognition. These advancements will significantly enhance the accuracy and applicability of sentiment analysis, paving the way for a more nuanced understanding of human emotions.

References

- [1] David J. Teece, "Business Models and Dynamic Capabilities," *Long Range Planning*, vol. 51, no. 1, pp. 40-49, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Jonathan T. Eckhardt, Michael P. Ciuchta, and Mason Carpenter, "Open Innovation, Information, and Entrepreneurship within Platform Ecosystems," *Strategic Entrepreneurship Journal*, vol. 12, no. 3, pp. 369-391, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Oswald Campesato, *Artificial Intelligence, Machine Learning and Deep Learning*, Mercury Learning and Information, 2020. [[Google Scholar](#)] [[Publisher Link](#)]
- [4] D. Tamil Priya, and J. Divya Udayan, "Affective Emotion Classification Using Feature Vector of Image Based on Visual Concepts," *International Journal of Electrical Engineering & Education*, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Xiaocui Yang et al., "Image-Text Multimodal Emotion Classification via Multi-View Attentional Network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014-4026, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Kiet Van Nguyen et al., "UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis," *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, Ho Chi Minh City, Vietnam, pp. 19-24, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Mengyao Li et al., "Joint Sentiment Part Topic Regression Model for Multimodal Analysis," *Information*, vol. 11, no. 10, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] CrowdFlower, data.world, 2016. [Online]. Available: <https://data.world/crowdflower/image-sentiment-polarity>
- [9] HuggingFace, vietnamese_students_feedback. [Online]. Available: https://huggingface.co/datasets/uitnlp/vietnamese_students_feedback
- [10] Dam Minh Linh, Ngo Xuan Thoai, and Han Minh Chau, "Face Mask Detection Using Deep Learning," *Journal of Science*, vol. 20, no. 11, pp. 1931-1942, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Lubab Ahmed Tawfeeq et al., "Predication of Most Significant Features in Medical Image by Utilized CNN and Heatmap," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 12, no. 4, pp. 217-225, 2021. [[Google Scholar](#)] [[Publisher Link](#)]
- [12] A. Lazarev, GitHub, 2016. [Online]. Available: <https://github.com/Arsey/keras-transfer-learning-for-oxford102/issues/1>

- [13] Zewen Li et al., "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999-7019, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Nguyen Thanh Tuan, *Basic Deep Learning Book*, 2020. [[Publisher Link](#)]
- [15] Pham Van Toan, [AI Interview] 12 Super Cool Deep Learning Interview Questions You Can't Miss, VIBLO, 2019. [Online]. Available: <https://viblo.asia/p/ai-interview-12-cau-hoi-phong-van-deep-learning-sieu-hay-khong-the-bo-quan-LzD5djvEZjY>
- [16] TensorFlow, TensorFlow v2.16.1, tf.keras.optimizers.Adam, 2024. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam
- [17] Tran Trung Truc, Optimizer - Deep Understanding of Optimization Algorithms (GD, SGD, Adam,..), VIBLO, 2020. [Online]. Available: <https://viblo.asia/p/optimizer-hieu-sau-ve-cac-thuat-toi-uu-gdsgdadam-Qbq5QQ9E5D8>
- [18] TensorFlow, TensorFlow v2.16.1, tf.keras.callbacks.EarlyStopping, 2024. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping
- [19] Hao Tuan Huynh et al., "Vietnamese Text Classification with TextRank and Jaccard Similarity Coefficient," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 6, pp. 363-369, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Aston Zhang et al., "Dive into Deep Learning," *arxiv Preprint*, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Md. Arif Istiak Sunny, Mirza Mohd Shahriar Maswood, and Abdullah G. Alharbi, "Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model," *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, Giza, Egypt, pp. 87-92, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] GeeksForGeeks, Bidirectional LSTM in NLP, 2023. [Online]. Available: <https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/>
- [23] Shuai Ma et al., "An Improved Bi-LSTM EEG Emotion Recognition Algorithm," *Journal of Network Intelligence*, vol. 7, no. 3, pp. 623-639, 2022. [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Vijaysinh Lendave, Difference between LSTM Vs GRU in Recurrent Neural Network, AI Mysteries, 2021. [Online]. Available: <https://analyticsindiamag.com/ai-mysteries/lstm-vs-gru-in-recurrent-neural-network-a-comparative-study/>
- [25] V.T. Tran, Python Vietnamese Toolkit, Github, 2021. [Online]. Available: <https://github.com/trungtv/pyvi/blob/master/README.rst>