Vietnamese Text Detection, Recognition and Classification in Images

Tuan Le Xuan
People's Security Academy
Hanoi, Vietnam
tuanlx.psa@gmail.com

Hang Pham Thi
People's Security Academy
Hanoi, Vietnam
phamthihang78@gmail.com

Hai Nguyen Do
People's Security Academy
Hanoi, Vietnam
nguyendohai@gmail.com

2022 14th International Conference on Knowledge and Systems Engineering (KSE) | 978-1-6654-5281-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/KSE56063.2022.9953789

***Abstract*—Detecting and recognizing text in images is a task that has received a lot of attention recently due to its high applicability in many fields such as digitization, storage, lookup, authentication,... However, most current research works and products are focusing on detecting and extracting text from images but not paying very much attention to analyzing and exploiting semantics and nuances of those extracted texts. In this study, we propose a three-in-one system to detect, recognize and classify Vietnamese text in images collected from social media to help authorities in monitoring tasks. The system receives as input images containing Vietnamese text, uses the Character-Region Awareness For Text detection (CRAFT) model to perform background processing to produce areas containing text in the image; these text containers will then be rearranged in the same order as in the original image, and the text in the image will also be extracted out according to the text container. Next, we use VietOCR model to convert these text images into text fragments. Finally, these texts will be classified using an ensemble of machine learning models. Preliminary results show that the proposed model has an accuracy of up to 88.0% in detecting and recognizing text and 94% in classifying text nuances on the collected data set.**

***Index Terms*—OCR, Vietnamese Text, CRAFT, VietOCR, Voting Clasifier.**

I. Introduction

Text detection and recognition in images has been a topic of interest in recent years as it has been used to a wide variety of applications such as automatic reading of license plate [1], signboards [2], book covers [3], identity card [4], passport [5], receipts [6], etc. However, most of the works focus on detecting and extracting text from images while not paying very much attention to analyzing and exploiting semantics and nuances of those extracted texts. This aspect plays an important role in the context of ensuring social security and safety, especially when many online social networks are rapidly growing up nowadays where many individuals or organizations can easily disseminate articles, images, and information that might be harmful to society. For example, there are numerous malicious photographs have been shared on popular social media platforms such as Zalo, Facebook, and Tik Tok, among others, carrying messages that falsify ethnicity, human rights, religious legislation or slandering government leaders, etc. Thus, in this study, we developed a system that can detect, recognize and classify Vietnamese text in images in order to assist authorities in their social media monitoring task.

In building such a system, the first and most important thing is to have a suitable dataset. Since the dataset containing images with anti-social (or even reactionary) text, we have to use expertise knowledge from social security protection domain to decide whether the text in one image, and thus that imgage is anti-social or not. We have collected around 2000 photos and 1700 paragraphs of text in this study, so actually we have two datasets: one for text detection and recognition task and one for text classification task. Basically, we employ CRAFT technology [7] to find text in photos via bounding boxes, then use the VietOCR model to convert these bounding boxes to text. Finally, we use a text classification model based on a combination of machine learning methods to determine if the text has anti-social content. The three machine learning methods used are Multinomial Naive Bayes [8], Linear Support Vector Machine [9], and Voting Classifier [10]. Preliminary results are promising in terms of detecting and extracting Vietnamese text from images. The main contributions of this paper are:

1) Build two datasets: one contains 2000 images, the other contains 1700 vietnamese texts which were labeled using expertise knowledge from social security domain.
2) Propose an effective three-in-one model to deal with vietnamese text detection, recognition and classification in images.

The rest of the paper is organized as follows. Section II reviews the related work, Section III presents the proposed model, Section IV discusses the experiments and results and Section V concludes the work and discuss some future works.

II. Related works

Text detection and recognition in images has been received much of interest recently. In [11], S. Akopyan, O.V. Belyaeva, T.P. Plechov, and D.Y. Turdakov developed a text extraction pipeline for extracting text from photos of varying quality collected from social media. Their work is primarily concerned with classifying the incoming photos and then doing preprocessing on each class. Following that, the text is recognized using the OCR engine. This work makes use of a dataset collected from social media. [12] considers photos with a colorful backdrop and describes a preprocessing technique that enhances the performance of the Tesseract Optical Character Recognition (OCR) engine. The text is initially segmented to isolate it from the vibrant backdrop by separating the original image into k images. The picture containing text is then recognized by a classifier. Preprocessing resulted in a 20% boost in performance when compared to Tesseract OCR performance. The authors of [13] suggested a method for text extraction from scanned documents. In their study, Otsu's

technique was employed for segmentation and the Hough transform was used for skew detection. Additionally, the OCR technology was used to recognize characters. They conducted trials and verified the suggested method using a variety of photos from a variety of sources. The average accuracy was determined to be 93%. K. Karthick, K.B. Ravindrakumar, R. Francis, and S.Ilankannan [14] have covered in depth the many phases involved in text detection, highlighting the various strategies employed. Additionally, they have focused handwritten text recognition, which is a challenging subject. According to their research, the best results may be obtained with a shorter calculation time, and it is feasible to partition multilingual characters and improve the character recognition rate. Anupriya Shrivastava, Amudha J., Deepa Gupta, and Kshitij Sharma [15] built a system utilizing Convolutional Neural Networks and Long Short Term Memory in their study. The created approach recognizes text in horizontal, curved, or oriented pictures. The model is composed of four components. The first component does low-level feature extraction. The second component extracts high-level characteristics through a common convolutional method. The third component disregards irrelevant characteristics. The fourth component forecasts the sequences of characters.

In case of vietnamese language, there are also plenty of works. In [3], Nga et al. demonstrated a new approach for extracting Vietnamese text from photos of scanned book covers. The suggested method took a photo of the book covers, filtered the image for better quality, found the text-filled sections, and then extracted the text using an optical character recognizer (OCR). To obtain the final text output, the extracted text was filtered in conjunction with a dictionary. Experiments with the suggested approach utilizing our dataset yielded promising results. Van Hoai, D. P., Duong, H. T., and Hoang, V. T looked towards developing a deep features network-based approach for recognizing Vietnamese identity cards [4]. On the character level and word level, it obtained accuracy of more than 96.7% and 89.7%, respectively, on various main data fields of identity cards. Hung, P. D., and Loan, B. T. created an application that recognizes Vietnamese passports automatically [5]. The method was developed through research into image processing, OCR, and language processing. In the application, the information was retrieved as "Full Name," "Passport Number," "Nationality," "Date of Birth," "Date of Issue," "Place of Issue," "ID Card," "Date of Expiry," and so on. The average recognition rate was high, with the bulk of findings above 92%. Vu, X. S., Bui, Q. A., Nguyen, N. V., Nguyen, T. T. H., Vu, T. presented an overview of the MCOCR challenge "Mobile-captured image document recognition for Vietnamese receipts" [6]. This data challenge provided a new dataset that contains a large number of receipt photos, which are essential for document processing and accounting automation. The annotation method, which used a combination of systematic model-based and human-in-the-loop approaches, helped to create a valuable dataset for future study in automatic document processing in Vietnam. They hoped that the dataset will motivate researchers and machine learning experts to offer their expertise to Vietnam's picture document identification challenge. However, most of these works have just focussed on detection and recognition of text. In our research, we go further by investigating the sentiment of these recognized texts in order to support authorities in their strategic decision making process.

III. Proposed system

The proposed system firstly accepts an image consisting of text on a noise background as input. Then, the image background will be processed with CRAFT-based model to extract the text. Boxes of characters are identified and linked. Theses text containers of words in the image will be the model's output, which will then be rearranged in the same order as the original image, and the text in the image will be clipped out according to the text's bounding box. The output text from this OCR phase are then embedded using Term Frequency - Inverse Document Frequency (TF-IDF) and then fitted into an ensemble model combined of Multinomial Naive Bayes (MNB) and Linear Support Vector Machine (SVM) using Voting Classifier to classify the texts. Figure 1 depicts phases of the proposed model.

A. Text area detection

The text detection model is a Fully Convolutional Network architecture based on VGG-16 (Visual Geometry Group) [16] with batch normalization is adopted as a backbone, composed of two modules: convolutional neural network and upsample. Figure 2 depicts a schematic representation of the network architecture. ConvBlock is a convolution block that consists of a convolution layer, a batch normalization layer, an activation layer, and a pooling layer. UpConvBlock is an upsample convolution, including a convolution layer, a batch normalization layer, an activation layer, and an upsample layer. The CNN section concludes with the layer of convolutional networks and the layer of maximum pooling. Following the convolution kernel operation, the model outputs text image morphological properties such as color and texture. Mathematically, image features are composed of feature vectors calculated by many convolution kernels. These vectors are concatenated further and eventually comprise the output of the convolutional network in the form of a feature map. Prior to upsampling, feature fusion between multiple convolution layers is required to optimize feature expression. The upsample layer is responsible for the convolution feature map's expansion operation. The low-scale eigengraph is expanded to the higher-scale feature map using a linear interpolation approach, which has a negligible effect on the overall distribution. The region and affinity scores are output via distinct branches following the output of four upper sample layers and four subsequent convolution layers.

- **Region Score**: gives the region of the character, localizes the character.
- **Affinity Score**: condenses multiple instances of characters into a single one (a word).

The probability of the central character is then encoded using a Gaussian heat map.

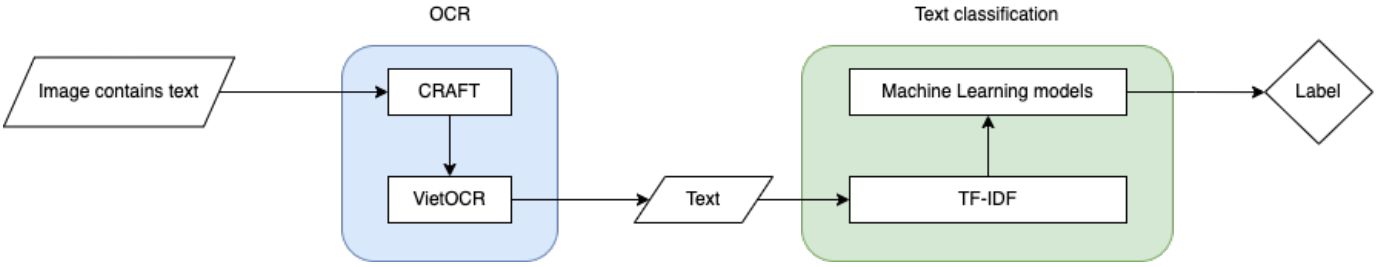
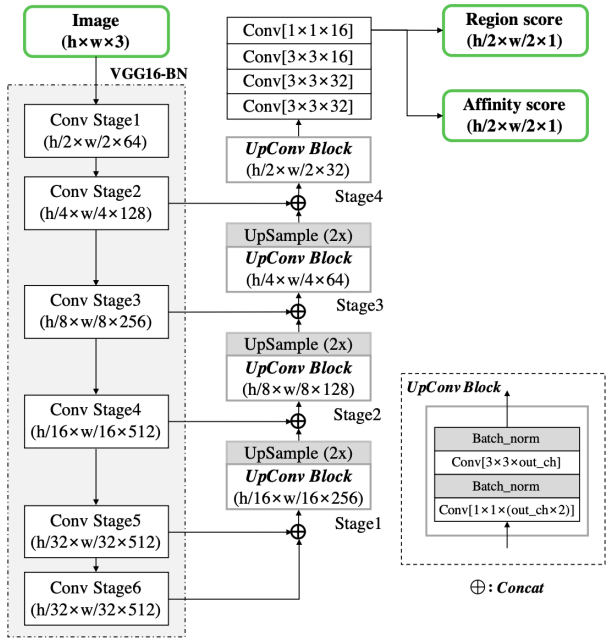


Figure 1: Proposed model



Source: https://arxiv.org/abs/1904.01941

Figure 2: Schematic illustration of network architecture

As shown in Figure 3, the affinity box is defined by adjacent character areas. By connecting the opposing corners of each character area with diagonal lines, it is possible to generate two triangles, the upper and lower character triangles. Then, for each pair of adjacent character regions, a relational region is formed by taking the centroids of the upper and lower character triangles that join to form the corners. The steps to approximate actual labels for area points and relation points are as follows: (i) Prepare a 2-dimensional isotropic Gaussian map. (ii) Calculate the Perspective Transform between the Gaussian map and the character container. (iii) Transform the Gaussian map to match the character container. According to the research findings, the model provides a high level of flexibility in identifying and extracting text images in challenging circumstances such as misaligned, curved, or deformed captured words. Furthermore, because the proposed model is based on the CRAFT concept, a multilingual model, it can be used with languages other than Latin, such as pictographic languages.

The objective function is the mean square error (MSE) loss function specified in formula (1). The variables $\tilde{y}_{\text{region}}$ and $\tilde{y}_{\text{affinity}}$ represent the predicted region score and affinity map, respectively, whereas y_{region} and y_{affinity} represent the ground truth annotation. The model parameters are updated using the Stochastic Gradient Descent (SGD) [17] algorithm, which adjusts the weight of the nodes to minimize the objective function.

$$MSE\ Loss = (\mathrm{y}_{\text{region}} - \tilde{y}_{\text{region}})^2 + (y_{\text{affinity}} - \tilde{y}_{\text{affinity}})^2 \quad (1)$$

B. Text recognition

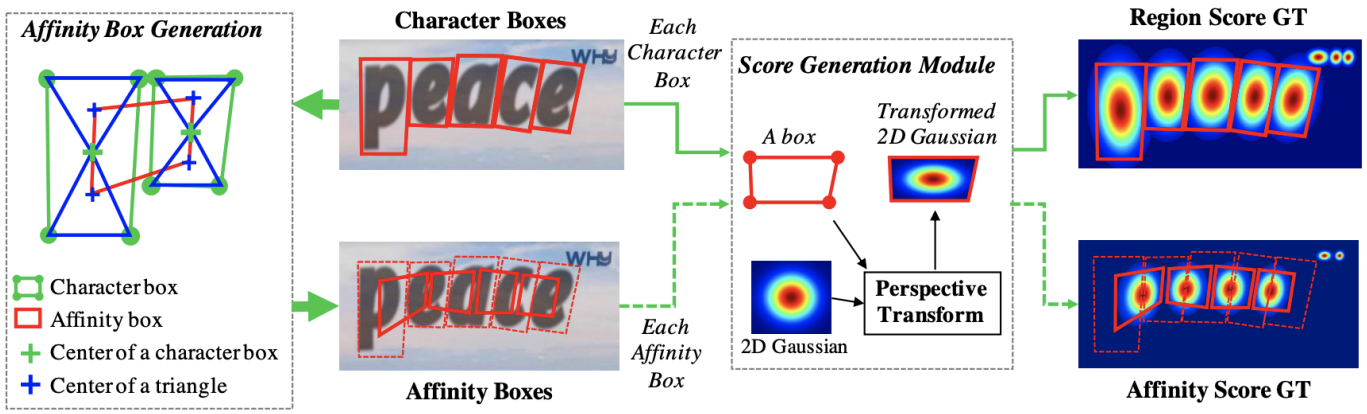
VietOCR is a OCR model that recognizes handwriting and typed letters for Vietnamese. Model architecture is a great combination of Convolutional Neural Network (CNN) and Transformer [18] (which is the foundation of many other models, the most famous being BERT - Bidirectional Encoder Representations from Transformers [19]).

Encoder. The encoder receives an input picture $x_{\text{img}} \in \Re^{3\times H_0\times W_0}$, and resizes it to a fixed size (H, W). Due to the fact that the Transformer encoder cannot interpret raw images that are not a sequence of input tokens, the encoder decomposes the input image into a batch of $N = HW/P2$ foursquare patches with a fixed size of (P, P), whereas the scaled image's width W and height H are guaranteed to be divisible by the patch size P. Following that, the patches are flattened into vectors and linearly projected to D-dimension vectors, which represent the patch embeddings, where D is the Transformer's hidden size over all of its layers.

The attention [20] mechanism is to allocate various levels of attention to the values and output their weighted total, with the weights of the values being determined by the relevant keys and queries. All queries, keys, and values in the self-attention modules originate from the same sequence. The matrix representing the outcome of attention is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

The scaling factor of $\frac{1}{\sqrt{d_k}}$ is applied to avoid the extremely small gradients of the softmax function, where the d_k is the dimension of queries and keys. The multi-head attention is to project the queries, keys and values h times with different learnable weights of projection, which allows the model to



Source: https://arxiv.org/abs/1904.01941

Figure 3: Illustration of ground truth generation procedure

jointly gather the information from different representation subspaces.

$$\begin{aligned} \text{MultiHead}(Q,K,V) &= \text{Concat}\left(\text{head}_1,\ldots,\ \text{head}_{\text{h}}\right)W^O \\ \text{where head} &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{3}$$

Different from the features extracted by the CNN− like network, the Transformer models have no image-specific inductive biases and process the image as a sequence of patches, which enables the model to pay different attention to either the whole image or the independent patches.

Decoder. The original Transformer decoder was used for VietOCR. The standard Transformer decoder also has a stack of identical layers, which have similar structures to the layers in the encoder, except that the decoder inserts the "encoder-decoder attention" between the multi-head self-attention and feedforward network, to distribute different attention on the output of the encoder. In the encoder - decoder attention module, the keys and values come from the encoder output while the queries come from the decoder input. In addition, the decoder leverages the attention masking in the self-attention to prevent from getting more information during training than prediction. Based on the fact that the output of the decoder will right shift one position from the input of the decoder, the attention masking need to ensure the output for the position i can only pay attention to the known output, which is the input on the positions less than i :

$$\begin{aligned} h_i &= \text{Proj}\left(\text{Emb}\left(\text{Token}_i\right)\right) \\ \sigma\left(h_{ij}\right) &= \frac{e^{h_{ij}}}{\sum_{k=1}^{V} e^{h_{ik}}} \quad \text{for } j = 1,2,\ldots,V \end{aligned} \tag{4}$$

The embedding from the decoder is projected from the model dimension to the dimension of the vocabulary size V. The probabilities over the vocabulary are calculated by the softmax function and we use beam search to get the final output. The whole process is depicted in Figure 4.

C. Text classification

Natural language processing (NLP) models that works well with English might not do the same with Vietnamese, since the word formation in each language is different, especially in case of compound words. For example, if Vietnamese words such as ""sinh viên", "giảng viên", "đại học"are separated into single words then they might have no meaning but in case of English, this is possible. Therefore, while training an NLP model for Vietnamese, the fundamental step is to combine single characters into meaningful words. To accomplish this, we use the word_tokenizer of Underthesea (https://github.com/undertheseanlp/underthesea), an Open-Source Vietnamese Natural Language Process Toolkit which provides a lot of pre-trained models for Vietnamese such as Sentence Segmentation, Word Segmentation, POS tagging , Chunking, Named Entity Recognition, etc.. After applying the word_tokenizer function, the characters in the dictionary that are related to each other will be linked together with the symbol "_". For instance, the phrase "Chúng tôi hiện đang là giảng viên đại học"will be replaced with "Chúng_tôi hiện đang là giảng_viên đại_học". These sentences are then tokenized by using TF-IDF and then fitted into an ensemble of machine learning models composed of MNB, Linear SVM and Voting Classifier to perform classification phase.

IV. Result

A. Datasets

Two datasets were built to evaluate the proposed model. One contains 2000 images and the other contains 1756 Vietnamese paragraphs which were crawled from social media platforms such as Facebook and Google. The image dataset is divided into two parts with label is either Normal or Anti-social according to the texts it contains inside. The paragraphs in the second dataset are also labeled similarly. Both labelings are done thanks to expertise knowledge of experts in social

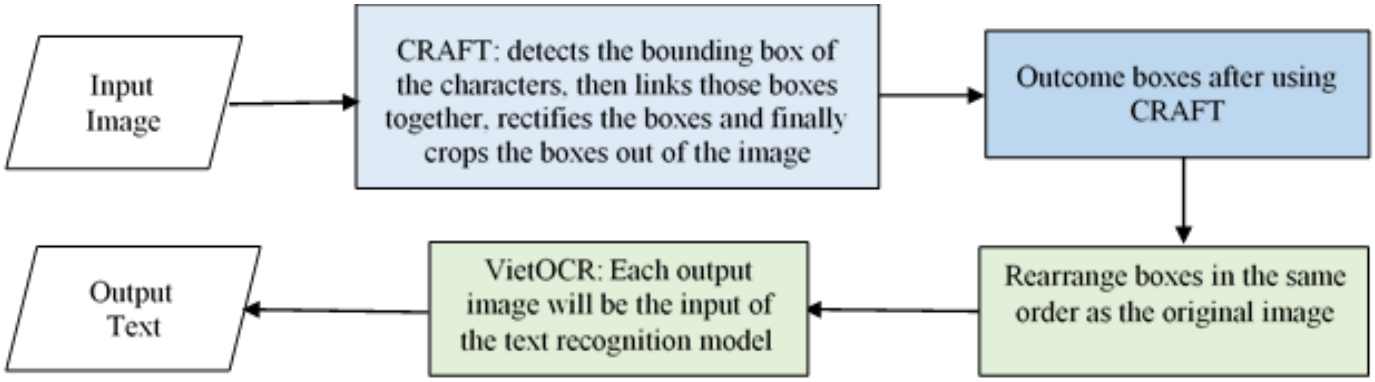


Figure 4: Text detection and recognition phase

security protection field. All the images are subject to brightness, clarity, and reliability requirements while the texts in paragraphs are preprocessed by deleting useless icons and being normalized according to NFC standard and transform all characters to lower case. The text detection and recognition model will be evaluated on the combined dataset of both labels, while the accuracy of the text classification model will be evaluated through the exact division of the two labels.

B. Experimental result

Due to the fact that the proposed system is composed of several models, evaluating the overall performance of the system using error measures is exceedingly challenging. Thus, the authors choose to evaluate the system using actual photos containing text and determining its correctness.

Figure 5: Accurate text detection and recognition

Figure 6: Error due to characters too close to each other

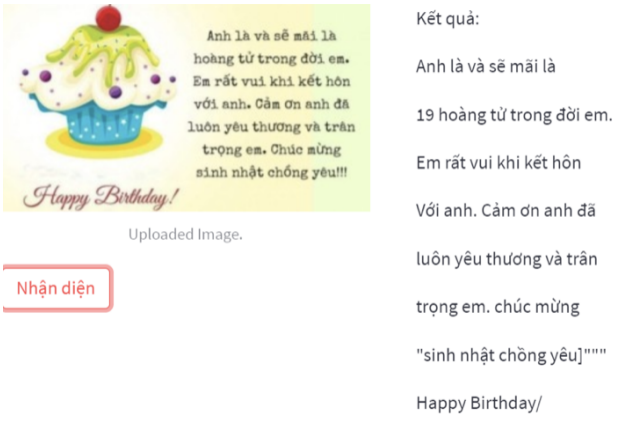


Figure 7: Error due to punctuation marks too close together

Figure 8: Error due to distorted or blurred image

The results of the text detection and recognition phase of the system will be evaluated using the formula

$$Accuracy = \frac{number\_of\_characters\_correctly\_predicted}{total\_number\_of\_recognizable\_characters} \quad (5)$$

The results indicated that, on a dataset of 2000 test images, the proposed system achieves an accuracy rate of up to 88.04 percent. Inaccurate images can be caused by a variety of factors, including insufficient image brightness, distorted images that are blurred or broken, as illustrated in Figure 8, or punctuation marks that are too close together, as depicted in Figure 7, etc. For the text classification model, the results in Figure 9 demonstrate that the model achieves accuracy up to

94%, precision, recall and f1-score on both labels are relatively high (over 91%).

	precision	recall	f1-score	support
0	0.93	0.97	0.95	197
1	0.96	0.91	0.94	171
accuracy			0.94	368
macro avg	0.95	0.94	0.94	368
weighted avg	0.94	0.94	0.94	368

Figure 9: Text classification model's report

V. Conclusion

In this study, we proposed a system that uses the CRAFT technology to locate text in images via bounding boxes, then translates these bounding boxes to text using the VietOCR model. Finally, to identify whether the image contains anti-social content, it employs a text classification model based on a combination of machine learning algorithms which are Multinomial Naive Bayes, Linear Support Vector Machine, and Voting Classifier. The results of numerous tests demonstrated that our system performs well in detecting, recognizing and classifying Vietnamese texts from images. One of the most crucial aspects that determine the performance of the entire system is the dataset used to train the models. To support this study, we had built two datasets as described above. Determining which content corresponds to which label is a complex task that necessitates tremendous effort and domain-specific knowledge. Since labeling the text contained in the dataset has a profound influence on the model's accuracy and the ultimate meaning of the study, we enlisted the assistance of a team of specialists with the domain expertise to assist in the data labeling process. The system does, however, have certain inherent flaws, the most noticeable of which is its sensitivity to the quality of the input data. Specifically, if the original image quality is poor or the text in the image to be classified is written in unusual fonts, the system would be unable to extract the proper text from it, resulting in erroneous predictions. In the future, we are aiming for a solution to automatically improve image quality as a pre-processing step of data before being fitted into the model, thereby boosting the performance of text recognition and extraction from images and, ultimately, making more accurate predictions for the text classification problem.

References

1 Li, H. and Shen, C., 2016. *Reading Car License Plates Using Deep Convolutional Neural Networks and LSTMs.*

2 Liu, X., Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., Bai, X., Shi, B., Karatzas, D., Lu, S. and Jawahar, C., 2019. *ICDAR 2019 Robust Reading Challenge on Reading Chinese Text on Signboard.*

3 Nga, P. T. T., Trang, N. T. H., Phúc, N. V., Quý, T. D., and Bình, V. P., 2017. *Vietnamese text extraction from book covers.* Dalat University Journal of Science, 7(2), 142-152.

4 Van Hoai, D. P., Duong, H. T., and Hoang, V. T., 2021. *Text recognition for Vietnamese identity card based on deep features network.* International Journal on Document Analysis and Recognition (IJDAR), 24(1), 123-131.

5 Hung, P. D., and Loan, B. T., 2020. *AutomaticAutomatic vietnamese passport recognition on android phones.* In International Conference on Future Data and Security Engineering (pp. 476-485). Springer, Singapore.

6 Vu, X. S., Bui, Q. A., Nguyen, N. V., Nguyen, T. T. H., and Vu, T., 2021. *Mc-ocr challenge: Mobile-captured image document recognition for vietnamese receipts.* In 2021 RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-6). IEEE.

7 Baek, Y., Lee, B., Han, D., Yun, S. and Lee, H. (2019). *Character Region Awareness for Text Detection.* arXiv:1904.01941 [cs]. [online]

8 Kibriya, A.M., Frank, E., Pfahringer, B. and Holmes, G. (2004). *Multinomial Naive Bayes for Text Categorization Revisited*. Lecture Notes in Computer Science, pp.488–499.

9 Ghosh, S., Dasgupta, A. and Swetapadma, A. (2019). *A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification*. [online] IEEE Xplore. doi:10.1109/ISS1.2019.8908018.

10 Zhang, Y., Zhang, H., Cai, J. and Yang, B. (2014). *A Weighted Voting Classifier Based on Differential Evolution. Abstract and Applied Analysis*, [online] 2014, pp.1–6.

11 M.S. Akopyan, O.V. Belyaeva, T.P. Plechov and D.Y. Turdakov, *Text recognition on images from social media*, 2019, Ivannikov Memorial Workshop (IVMEM)

12 Matteo Brisinello, Ratko Grbić, Dejan Stefanović and Robert PečkaiKovač, *Optical Character Recognition on images with colorful background*, 2018, IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin).

13 Neha Agrawal, Arashdeep Kaur, *An Algorithmic Approach for Text Recognition from Printed/Typed Text Images*, 2018, 8th International Conference on Cloud Computing, Data Science & Engineering

14 K. Karthick, K.B. Ravindrakumar, R. Francis, S. Ilankannan, *Steps Involved in Text Recognition and Recent Research in OCR; A Study*, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019.

15 Anupriya Shrivastava, Amudha J., Deepa Gupta, Kshitij Sharma, *Deep Learning Model for Text Recognition in Images*, 10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur, Kanpur, India.

16 Simonyan, K. and Zisserman, A. 2014. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [online] arXiv.org

17 Bottou, Léon,. 1998. *Online Algorithms and Stochastic Approximations*. Online Learning and Neural Networks. Cambridge University Press. ISBN 978-0-521-65263-6.

18 Wolf, T., Debut, L., et al. 2020. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv:1910.03771 [cs]. [online]

19 Devlin, J., Chang, M.-W., et al. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv.org

20 Vaswani, A., Brain, G., et al. 2017. *Attention Is All You Need*. [online]